Heterogeneity in Prediction Research:

Methods and applications

David van Klaveren

Heterogeneity in Prediction Research:

Methods and applications

David van Klaveren

FSC
www.fsc.org
MIX
Paper from
responsible sources
FSC® C012854

ISBN: 978-94-6233-493-9

Financial support for the publication of this thesis was generously provided by Cardialysis BV and the Department of Public Health of Erasmus MC.

Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged.

Printed by Gildeprint.

Heterogeneity in Prediction Research:

Methods and applications

Heterogeniteit in predictie-onderzoek:

Methoden en toepassingen

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
vrijdag 13 januari 2017 om 13.30 uur

door

David van Klaveren
geboren te Rotterdam

Erasmus University Rotterdam

PROMOTIECOMMISSIE

Promotor:	Prof.dr. E.W. Steyerberg
Overige leden:	Prof.dr.ir. H. Boersma Prof.dr. S. Sleijfer Prof.dr. T. Stijnen
Copromotor:	Dr. Y. Vergouwe

CONTENTS

PART III APPLICATIONS

1

General introduction

PREDICTION RESEARCH

Prediction models are ubiquitous tools that aim to inform patients, clinicians and health policy makers about the likelihood of uncertain outcomes [1]. Prediction models exploit identifiable heterogeneity of patients by combining characteristics of individual patients to predict the presence of a disease or condition (diagnosis) or a future event (prognosis). Individualized predictions have the potential to support different kinds of medical decisions. For example, the clinical decision to treat individual acute ischemic stroke patients with recombinant tissue plasminogen activator (rt-PA) may be supported by predictions of treatment benefit [2, 3]. Likewise, screening for the presence of Chlamydia is a public health intervention that may be supported by risk stratification, limiting screening and subsequent treatment to high risk individuals [4]. Many other prediction models that can be used to support medical decision making will be presented in this thesis (Table 1.1).

Regression techniques are most widely used in the development of prediction models to estimate the relation between predictors and the outcome of interest in a representative sample of patient data (development data). The value of the prediction model for decision making depends on the model's predictive performance. Model performance is often measured within the development data (internal validation), but model performance in independent data (external validation) is a better indicator for the generalizability of a prediction model [5]. The performance of a prediction model is typically measured in two dimensions: the ability to predict in agreement with observed frequencies across the range of predictions (calibration) and the ability to separate subjects with good outcome from subjects with poor outcome (discrimination) [1].

Prediction research focuses on the development of valid prediction models and on the assessment of their generalizability and applicability. PART I and II of this thesis are about two different forms of heterogeneity in prediction research. One form of heterogeneity occurs when risk differs between clusters of patients. The second form of heterogeneity – treatment effect heterogeneity – manifests itself when treatment effects differ substantially between identifiable patient subgroups. These two forms of heterogeneity in prediction research will be clarified in the next sections.

Table 1.1 Prediction models in this thesis.

Predicted outcome	**Decisions**	**Chapters**
Survival after traumatic brain injury	Treatment intensity (e.g. hyperventilation, barbiturates, or mannitol); admission to intensive care; withdrawal of treatment	2; 5; 7
Survival after hospitalization for congestive heart failure	Patient disposition (discharge, admission to intensive/cardiac care); symptom relief; end-of-life care	3; 6
Survival after treating head and neck cancer patients	Treatment with chemotherapy	4
Survival after revascularization with Coronary Artery Bypass Grafting (CABG) or Percutaneous Intervention (PCI)	Treatment with CABG or PCI	5; 8; 10; 12-14
Cardiovascular disease in the general population	Treatment with statins	9
Functional outcome after treating stroke patients	Treatment with rt-PA	10
Survival after treating acute myocardial infarction patients	Treatment with aggressive thrombolysis	11
Out-of-hospital bleeding during Dual AntiPlatelet Therapy (DAPT) after coronary stent implementation	Duration of DAPT	15
Survival after treating esophageal cancer patients with neoadjuvant chemoradiotherapy (nCRT) and surgery or surgery alone	nCRT before surgery	16
Non-sentinel lymph node metastases of sentinel lymph node positive breast cancer patients	Axillary lymph node dissection	17; 18
Sexually transmitted infection in the general population	Screening	19; 20
Neonatal mortality in the general population of low and middle income countries	Antenatal, intrapartum and postnatal care	21

RISK HETEROGENEITY IN CLUSTERED DATA

Prediction models are often developed and validated in clustered patient data. A typical example is a multicenter study, i.e. where data are available from patients who are treated in different centers. Even when patient inclusion criteria are similar across centers, there often are risk differences between patients from different centers ("unidentifiable heterogeneity") [6]. A comparable type of clustering may occur in patients treated in different countries or in patients treated by different caregivers in the same center. Similarly, in public health research the study population is often clustered in geographical regions such as countries, municipalities or neighborhoods.

Multilevel regression models may reveal that the baseline risk differs between clusters (varying intercepts) and that predictor effects vary across clusters (varying slopes) [7, 8]. Consequently, the performance of a prediction model may vary across clusters. Varying intercepts and slopes are indicative for variation in cluster-specific calibration and discrimination. Furthermore, discrimination is influenced by the level of dissimilarity across individuals within a cluster (case-mix heterogeneity): when individuals within a cluster are more alike it will be harder to separate them [9]. A case-mix corrected discrimination measure has been suggested before as a benchmark value for comparison to the observed discriminative ability in a validation study [10]. Further work to disentangle causes of variation in performance is however necessary, specifically for the context of clustered data.

HETEROGENEITY OF TREATMENT EFFECT

Treatments that demonstrate benefit on average in randomized clinical trials help some patients but not others. A major focus of patient-centered outcomes research and personalized medicine is to identify this heterogeneity of treatment effect (HTE) so that treatment might be targeted to those who benefit, and avoided in those where it is useless or harmful [11-13]. To make optimal decisions it has been suggested to compare absolute treatment benefit – the difference between relevant outcomes in treated and control groups – under different treatment strategies [14, 15]. The absolute treatment benefit for individual patients depends on their risk in the absence of treatment since patients at low risk have little to gain from treatment. Risk prediction models including a constant relative treatment effect are helpful for predicting absolute treatment benefit [16]. In addition, the relative risk reduction from a specific treatment may be different between patients (relative treatment effect heterogeneity) [17, 18]. Incorporating relative treatment effect heterogeneity in risk prediction models – by using predictive factors for differential treatment effect (treatment effect modifiers) – has been recommended, but is sensitive to the pitfall of finding false-positive or false-negative subgroup effects by multiple testing [19-24].

Risk prediction models may well support treatment decision making under the condition that they accurately predict individual treatment benefit, i.e. the difference in potential outcomes under different treatment regimens. But risk prediction models are usually validated for their ability to predict risk, not for their ability to predict treatment benefit. A risk discrimination measure may be informative about a model's ability to separate high risk from low risk patients, but the relevance of such a measure for assessing a model's decision making potential may be questioned. Performance measures for treatment selection should assess how well a model discriminates patients who benefit from those who do not. However, prediction model performance is measured based on predictions and actual outcomes in individual patients. Measuring a model's ability to predict benefit is thus hampered by the fact that the actual benefit for each patient is inherently unobservable, since their potential (counterfactual) outcome under the alternative therapy is not known (Table 1.2) [25, 26].

Table 1.2 Treatment effect at patient level.

Potential outcome of patient with treatment A	Potential outcome of patient with treatment B	Causal effect of treatment B versus treatment A
alive	dead	harm
alive	alive	no effect
dead	dead	no effect
dead	alive	benefit

Cost-effectiveness analysis is a useful tool to weigh the benefits against the harms (including costs) of interventions, ideally over a lifetime horizon [27]. Similar to heterogeneity of treatment effect, the cost-effectiveness of a treatment has increasingly been recognized to be heterogeneous across individual patients [28-30]. Interventions that are cost-effective on average may be of very low value for many (even most) patients [31]. Individualizing cost-effectiveness might thus support more efficient distribution of resources, but requires individualized long-term estimates of treatment benefit, treatment harms, treatment costs, and patient preferences [32-35]. Short-term individualized treatment benefit can be based on outcome data from a clinical trial with detailed information on patient characteristics. However, the individualized long-term treatment benefit also depends on the post-trial life expectancy. Because post-trial survival information is generally lacking, post-trial life expectancy is usually derived from less granular population life tables. However, the underlying assumption of equal post-trial life expectancy for low and high risk patients of the same sex and age is unrealistic.

AIMS OF THIS THESIS

Aim 1 - How to validate prediction models in clustered data?

We will study the generalizability of prediction models across clusters of patients, ultimately to define a framework for assessing model performance in clustered data. We will explore existing and new methods for assessing model performance within clusters and for assessing heterogeneity in model performance across clusters. We will distinguish between the impact of case-mix differences and of predictor effect validity on discriminative ability.

Aim 2 - How to develop and validate prediction models for guiding treatment decisions?

We will compare modeling approaches to estimate the individual survival benefit of treatment with either coronary artery bypass graft surgery (CABG) or percutaneous coronary intervention (PCI) for patients with complex coronary artery disease. We will explore new methods to validate models for their ability to predict treatment benefit rather than risk. We will study the influence of different assumptions for modeling the short-term trial-based risk reduction and the post-trial life expectancy on individualized cost-effectiveness estimates.

Aim 3 – How to apply methods for development and validation of predictions models for guiding treatment decisions?

We will apply methods for development and validation of prediction models to several case studies of guiding clinical and public health decisions. We will use the newly developed methods (Aim 1 and 2) as building blocks for guiding treatment decisions in several of these case studies.

REFERENCES

1. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer: New York, 2009.
2. Kent DM, Selker HP, Ruthazer R, Bluhmki E, Hacke W. The stroke-thrombolytic predictive instrument: a predictive instrument for intravenous thrombolysis in acute ischemic stroke. Stroke 2006; 37: 2957-2962.
3. Kent DM, Ruthazer R, Decker C, Jones PG, Saver JL, Bluhmki E, Spertus JA. Development and validation of a simplified Stroke-Thrombolytic Predictive Instrument. Neurology 2015; 85: 942-949.
4. Gotz HM, van Bergen JE, Veldhuijzen IK, Broer J, Hoebe CJ, Steyerberg EW, Coenen AJ, de Groot F, Verhooren MJ, van Schaik DT, Richardus JH. A prediction rule for selective screening of Chlamydia trachomatis infection. Sex Transm Infect 2005; 81: 24-30.
5. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000; 19: 453-473.
6. Hunink MGM, ., et al. Decision Making in Health and Medicine. Cambridge University Press, 2014.
7. Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press: Cambridge, 2007.
8. Legrand C, Duchateau L, Janssen P, Ducrocq V, Sylvester R. Validation of prognostic indices using the frailty model. Lifetime Data Analysis 2009; 15: 59-78.
9. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Med Res Methodol 2012; 12: 82.
10. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol 2010; 172: 971-980.
11. Rothwell PM. Can overall results of clinical trials be applied to all patients? Lancet 1995; 345: 1616-1619.
12. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004; 82: 661-687.
13. Vickers AJ, Kent DM. The Lake Wobegon Effect: Why Most Patients Are at Below-Average Risk. Ann Intern Med 2015; 162: 866-867.
14. Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. Lancet 2005; 365: 256-265.
15. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA 2007; 298: 1209-1212.
16. Califf RM, Woodlief LH, Harrell FE, Jr., Lee KL, White HD, Guerci A, Barbash GI, Simes RJ, Weaver WD, Simoons ML, Topol EJ. Selection of thrombolytic therapy for individual patients: development of a clinical model. GUSTO-I Investigators. Am Heart J 1997; 133: 630-639.
17. Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 2010; 11: 85.
18. Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, Schroter S, Sauerbrei W, Altman DG, Hemingway H, Group P. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ 2013; 346: e5793.
19. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355: 1064-1069.
20. Stein CR, Kaufman JS, Ford CA, Leone PA, Feldblum PJ, Miller WC. Screening young adults for prevalent chlamydial infection in community settings. Ann Epidemiol 2008; 18: 560-571.

21. Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? Am Heart J 2006; 151: 257-264.
22. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Med Res Methodol 2006; 6: 18.
23. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine--reporting of subgroup analyses in clinical trials. N Engl J Med 2007; 357: 2189-2194.
24. Haukoos JS, Hopkins E, Bender B, Sasson C, Al-Tayyib AA, Thrun MW, Denver Emergency Department HIVTRC. Comparison of enhanced targeted rapid HIV screening using the Denver HIV risk score to nontargeted rapid HIV screening in the emergency department. Ann Emerg Med 2013; 61: 353-361.
25. Holland PW. Statistics and Causal Inference. Journal of the American Statistical Association 1986; 81: 945-960.
26. Rubin DB. Causal Inference Using Potential Outcomes. Journal of the American Statistical Association 2005; 100: 322-331.
27. Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. Cost-Effectiveness in Health and Medicine Oxford University Press: New York, NY, 1996.
28. Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, Luce BR, Studies ITFoGRP--M. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices--Modeling Studies. Value Health 2003; 6: 9-17.
29. Stevens W, Normand C. Optimisation versus certainty: understanding the issue of heterogeneity in economic evaluation. Soc Sci Med 2004; 58: 315-320.
30. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, Augustovski F, Briggs AH, Mauskopf J, Loder E, Force IHEEPG-CGRPT. Consolidated Health Economic Evaluation Reporting Standards (CHEERS)--explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. Value Health 2013; 16: 231-250.
31. Kent DM, Vijan S, Hayward RA, Griffith JL, Beshansky JR, Selker HP. Tissue plasminogen activator was cost-effective compared to streptokinase in only selected patients with acute myocardial infarction. J Clin Epidemiol 2004; 57: 843-852.
32. Basu A, Meltzer D. Value of information on preference heterogeneity and individualized care. Med Decis Making 2007; 27: 112-127.
33. Sculpher M. Subgroups and heterogeneity in cost-effectiveness analysis. PharmacoEconomics 2008; 26: 799-806.
34. Sculpher M. Reflecting heterogeneity in patient benefits: the role of subgroup analysis with comparative effectiveness. Value Health 2010; 13 Suppl 1: S18-21.
35. Ioannidis JPA, Garber AM. Individualized Cost-Effectiveness Analysis. PLoS Med 2011; 8: e1001058.

PART I

RISK HETEROGENEITY IN CLUSTERED DATA

2

Assessing discriminative ability of risk models in clustered data

D van Klaveren
EW Steyerberg
P Perel
Y Vergouwe

BMC Medical Research Methodology 2014

ABSTRACT

Background The discriminative ability of a risk model is often measured by Harrell's concordance-index (c-index). The c-index estimates for two randomly chosen subjects the probability that the model predicts a higher risk for the subject with poorer outcome (concordance probability). When data are clustered, as in multicenter data, two types of concordance are distinguished: concordance in subjects from the same cluster (within-cluster concordance probability) and concordance in subjects from different clusters (between-cluster concordance probability). We argue that the within-cluster concordance probability is most relevant when a risk model supports decisions within clusters (e.g. who should be treated in a particular center). We aimed to explore different approaches to estimate the within-cluster concordance probability in clustered data.

Methods We used data of the CRASH trial (2,081 patients clustered in 35 centers) to develop a risk model for mortality after traumatic brain injury. To assess the discriminative ability of the risk model within centers we first calculated cluster-specific c-indexes. We then pooled the cluster-specific c-indexes into a summary estimate with different meta-analytical techniques. We considered fixed effect meta-analysis with different weights (equal; inverse variance; number of subjects, events or pairs) and random effects meta-analysis. We reflected on pooling the estimates on the log-odds scale rather than the probability scale.

Results The cluster-specific c-index varied substantially across centers (*IQR*=0.70-0.81; I^2=0.76 with 95% confidence interval 0.66 to 0.82). Summary estimates resulting from fixed effect meta-analysis ranged from 0.75 (equal weights) to 0.84 (inverse variance weights). With random effects meta-analysis – accounting for the observed heterogeneity in c-indexes across clusters – we estimated a mean of 0.77, a between-cluster variance of 0.0072 and a 95% prediction interval of 0.60 to 0.95. The normality assumptions for derivation of a prediction interval were better met on the probability than on the log-odds scale.

Conclusion When assessing the discriminative ability of risk models used to support decisions at cluster level we recommend meta-analysis of cluster-specific c-indexes. Particularly, random effects meta-analysis should be considered.

BACKGROUND

Assessing the performance of a risk model is of great practical importance. An essential aspect of model performance is separating subjects with good outcome from subjects with poor outcome (discrimination) [1]. The concordance probability is a commonly used measure of discrimination reflecting the association between model predictions and true outcomes [2, 3]. For binary outcome data it is the probability that a randomly chosen subject from the event group has a higher predicted probability of having an event than a randomly chosen subject from the non-event group. For time-to-event outcome data it is the probability that, for a randomly chosen pair of subjects, the subject who experiences the event of interest earlier in time has a lower predicted value of the time to the occurrence of the event. For both kinds of outcome data the concordance probability is often estimated with Harrell's concordance (c)-index [2].

In risk modelling, clustered data are frequently used. A typical example is multicenter patient data, i.e. data of patients who are treated in different centers with similar inclusion criteria across the centers. Patients treated in the same center are nevertheless more alike than patients from different centers. A comparable type of clustering may occur in patients treated in different countries or in patients treated by different caregivers in the same center. Similarly, in public health research the study population is often clustered in geographical regions like countries, municipalities or neighbourhoods. It has been suggested that clustering should be taken into account in the development of risk models to obtain unbiased estimates of predictor effects [4]. This can be done by using a multilevel logistic regression model for binary outcomes or a frailty model for time-to-event outcomes [5, 6].

It would be natural to take clustering also into account when measuring the performance of a risk model. For multilevel models, it has been proposed to consider the concordance probability of subjects within the same cluster (within-cluster concordance probability) separately from the concordance probability of subjects in different clusters (between-cluster concordance probability) [7, 8]. We propose using the within-cluster concordance probability when risk models are used to support decisions within clusters, e.g. in clinical practice where decisions on interventions are commonly taken within centers. A valuable risk model should then be able to separate subjects within the same cluster into those with good outcome and poor outcome. We consider the within-cluster concordance probability more relevant in this context than the between-cluster or overall concordance probability.

Here, we aimed to estimate the within-cluster concordance probability from clustered data. We explored different meta-analytic methods for pooling cluster-specific concordance probability estimates with an illustration in predicting mortality among patients suffering from traumatic brain injury.

METHODS

Mortality in traumatic brain injury patients

We present a case study of predicting mortality after Traumatic Brain Injury (TBI). Risk models using baseline characteristics provide adequate discrimination between patients with good and poor 6-month outcomes after TBI [9, 10]. We used patients enrolled in the Medical Research Council Corticosteroid Randomisation after Significant Head Injury [11] trial (registration ISRCTN74459797, http://www.controlled-trials.com/), who were recruited between 1999 and 2004. This was a large international double-blind, randomized placebo-controlled trial of the effect of early administration of a 48-h infusion of methylprednisolone on outcome after head injury. The trial included 10,008 adults clustered in 239 centers with Glasgow Coma Scale (GCS) [12] Total Score ≤ 14, who were enrolled within 8 hours after injury. By design the patient inclusion criteria were equal in all 239 centers.

We considered patients with moderate or severe brain injury (GCS Total Score ≤ 12) and observed 6-month Glasgow Outcome Scale (GOS) [13]. Patients who were treated in one of 35 European centers with more than 5 patients experiencing the event (*n* = 2,081), were used to assess the discriminative ability of a prediction model developed with data from 35 centers. Patients who were treated in one of 21 Asian centers with more than 5 patients experiencing the event (*n* = 1,421) were used to assess the discriminative ability at external validation.

We used a Cox proportional hazards model with age, GCS Motor Score and pupil reactivity as covariates similar to previously developed risk models [9, 10]. We modelled center with a Gamma frailty (random effect) to account for heterogeneity in mortality among centers. We estimated parameters on the European selection of patients with the R package survival [14, 15]. As center effect estimates are unavailable when using a risk model in new centers, we calculated individual risk predictions applying the Gamma frailty mean of 1 for each patient.

Cluster-specific concordance probabilities

We estimated the concordance probability within each cluster by Harrell’s c-index [2], i.e. the proportion of all usable pairs of subjects in which the predictions are concordant with the outcomes. A pair of subjects is usable if we can determine the ordering of their outcomes. For binary outcomes, pairs of subjects are usable if one of the subjects had an event and the other did not. For time-to-event outcomes, pairs of subjects are usable if their failure times are not equal and at least the smallest failure time is uncensored. For a usable subject pair the predictions are concordant with the outcomes if the ordering of the predictions is equal to the ordering of the outcomes. Values of the c-index close to 0.5 indicate that the model does not perform much better than a coin-flip in predicting which subject of a randomly chosen pair will have a better outcome. Values of the c-index near 1 indicate that the model is almost perfectly able to predict which subject of a randomly chosen pair will have a favourable outcome. We estimated the variances of the cluster-specific c-indexes with a method proposed by Quade [16]. Formulas are provided in Appendix 2.1.

Pooling cluster-specific concordance probability estimates

The within-cluster concordance C_W can be estimated by pooling the cluster-specific concordance probability estimates into a weighted average. Previously, the cluster-specific concordance probability estimates were pooled with the number of usable subject pairs as weights [7, 8]. Here, we define eight different ways for pooling of cluster-specific estimates – both on the probability scale and on the log-odds scale – based on fixed effect meta-analysis and random effects meta-analysis.

We consider a dataset with subjects in *K* clusters. Let m_k be the number of subjects and e_k be the number of events in cluster *k*. We denote the number of usable subject pairs – pairs of subjects for whom we can determine the ordering of their outcomes – in cluster *k* by n_k. The cluster-specific concordance probability estimate for cluster *k* is denoted by $\hat{C}_k$ with sampling variance estimate $\hat{\sigma}_k^2$.

Fixed effect meta-analysis

Fixed effect meta-analysis assumes that one common within-cluster concordance probability C_W exists that applies to all clusters. The observed cluster-specific estimates vary only because of chance created from sampling subjects. Fixed effect meta-analysis with cluster weights w_k results in:

$$\hat{C}_W = \frac{\sum_k w_k \hat{C}_k}{\sum_k w_k} \text{ with } \hat{\sigma}^2_{\hat{C}_W} = \frac{\sum_k w_k^2 \hat{\sigma}_k^2}{(\sum_k w_k)^2} \tag{1}$$

The simplest approach would be to apply equal weights, $w_k = 1/K$, for each cluster (method 1). This estimator is quite naive when the cluster size varies, because small clusters are given the same weight as large clusters and information about the precision of the cluster-specific estimates is ignored. Heuristic choices of weights taking the cluster size into account are the number of subjects, $w_k = m_k$ (method 2), or the number of events, $w_k = e_k$ (method 3). Analogous to the definition of the c-index a fourth option is the number of usable subject pairs as weights, $w_k = n_k$ (method 4). The pooled estimate is than equal to the proportion of all usable within-cluster subject pairs in which the predictions and outcomes are concordant. Another choice of meta-analysis weights are the inverse variances, $w_k = 1/\hat{\sigma}_k^2$ (method 5). These weights express the precision of the cluster-specific estimates and are commonly used in meta-analysis of study-specific treatment effects.

Random effects meta-analysis

In our context a random effects meta-analysis considers that the cluster-specific estimates vary not only because of sampling variability but also because of differences in true concordance probabilities. This is appropriate for high values of I^2 [17]. I^2 measures the proportion of variability in cluster-specific estimates that is due to between-cluster heterogeneity rather than chance. Random effects meta-analysis assumes that cluster-specific concordance probabilities C_k are distributed about mean μ with between-cluster variance τ^2, with the observed $\hat{C}_k$ normally distributed about C_k with sampling variance σ_k^2. The mean within-cluster concordance probability estimate $\hat{\mu}$ is the average of the cluster-specific estimates with the inverse variances as weights (method 6):

$$\hat{\mu} = \frac{\sum_k w_k \hat{C}_k}{\sum_k w_k} \;,\; \hat{\sigma}_{\hat{\mu}}^2 = \frac{\sum_k w_k^2 \left(\hat{\sigma}_k^2 + \hat{\tau}^2\right)}{(\sum_k w_k)^2} = \frac{1}{\sum_k w_k} \text{ with } w_k = 1/(\hat{\sigma}_k^2 + \hat{\tau}^2) \qquad (2)$$

For estimation of the between-cluster variance τ^2 we used the DerSimonian and Laird [18] method. Alternative estimators for τ^2 can be found in DerSimonian and Kacker [19].

With the additional assumption of normally distributed C_k we can derive a prediction interval for the within-cluster concordance probability C_W in a new or unspecified cluster [20]. If τ^2 were known, then $\hat{\mu} \sim N(\mu, \hat{\sigma}_{\hat{\mu}}^2)$ and $C_W \sim N(\mu, \tau^2)$ imply (assuming independence of C_W and $\hat{\mu}$ given μ) that $C_W - \hat{\mu} \sim N(0, \tau^2 + \hat{\sigma}_{\hat{\mu}}^2)$. Hence the within-cluster concordance probability C_W in a new cluster is normally distributed, with mean $\hat{\mu}$ and variance $\tau^2 + \hat{\sigma}_{\hat{\mu}}^2$ (Figure 2.1). Since τ^2 is estimated, we assume $\frac{C_W - \hat{\mu}}{\sqrt{\hat{\tau}^2 + \hat{\sigma}_{\hat{\mu}}^2}}$ to take a more conservative t-distribution with $K-2$ degrees of freedom instead of the standard normal distribution [20]. Thus, a 95% prediction interval of the within-cluster concordance probability C_W in an unspecified cluster can be approximated by: $\hat{\mu} \pm t_{K-2}^{0.975} \sqrt{\hat{\tau}^2 + \hat{\sigma}_{\hat{\mu}}^2}$ with $t_{K-2}^{0.975}$ denoting the 97.5% percentile of the t-distribution with $K-2$ degrees of freedom.

Meta-analysis scale

When calculating a prediction interval of the within-cluster concordance probability C_W, Riley et al [21] advised to perform a random effects meta-analysis on a scale that helps meet the normality assumption for the random effects. When the normality assumption of the random effects model holds, the C_k are normally distributed with mean μ and variance $\tau^2 + \sigma_k^2$.

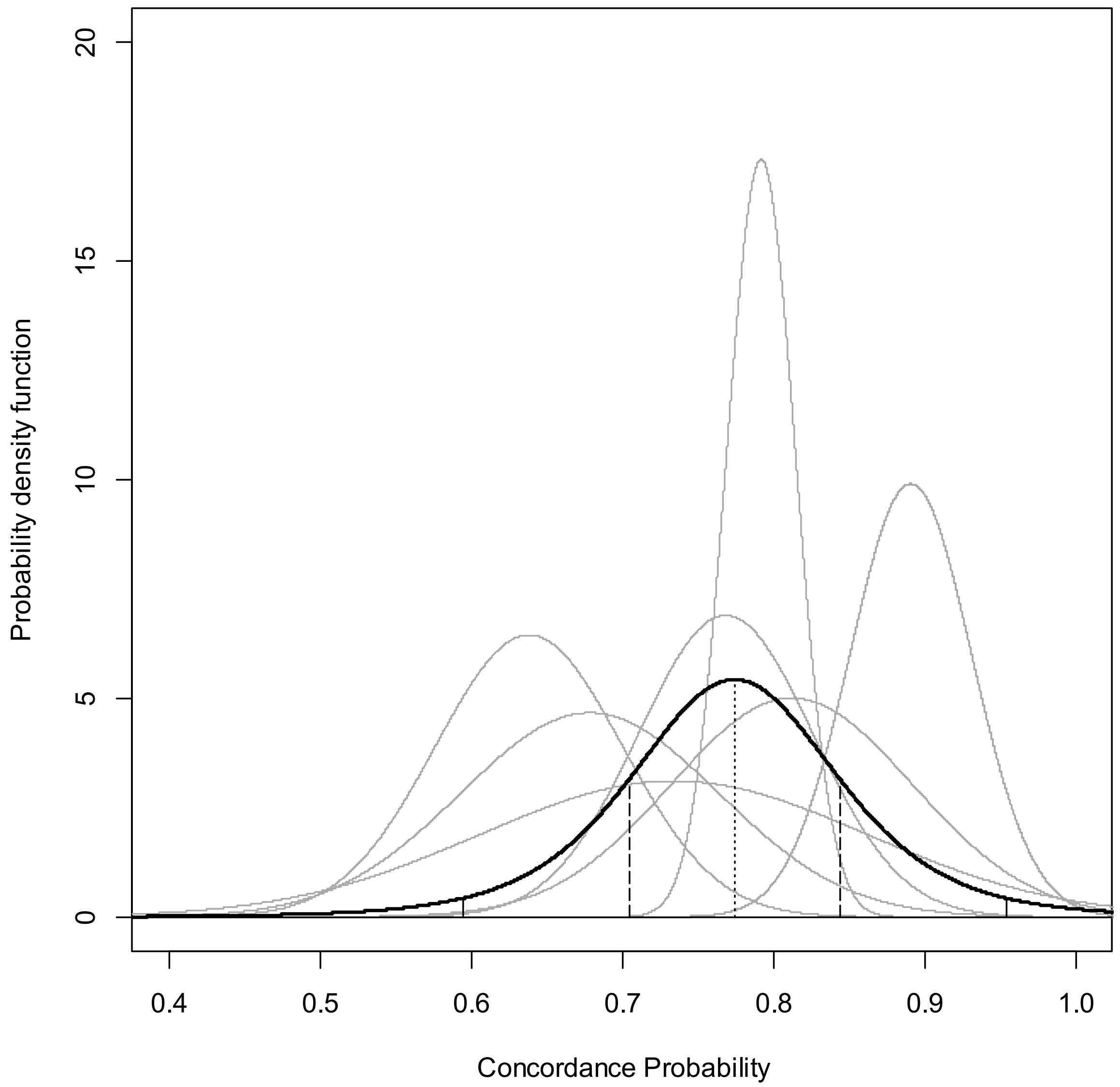


Figure 2.1 Example of random effects meta-analysis of concordance probability estimates in 7 clusters. Cluster-specific estimates are in grey. Under the assumption of normally distributed cluster-specific concordance probabilities, the predictive distribution resulting from a random effects meta-analysis is in black. The mean, the mean ± one standard deviation and the 2.5 and 97.5 percentiles of the predictive distribution are plotted with vertical lines.

As a consequence, the standardized residuals z_k defined below should approximately have a standard normal distribution:

$$z_k = (\hat{C}_k - \hat{\mu}) \Big/ \sqrt{\hat{\tau}^2 + \hat{\sigma}_k^2} \tag{3}$$

To consider if the normality assumption is valid we used a normal probability plot of z_k and applied the Shapiro-Wilk test to z_k [22]. In a normal probability plot z_k is plotted against a theoretical normal distribution in such a way that the points should form an approximate

Table 2.1 Overview of the 8 methods for pooling of cluster-specific concordance probability estimates.

	Fixed effect meta-analysis Assuming the same true (logit) concordance probability within each cluster	**Random effects meta-analysis** Assuming variation in true (logit) concordance probabilities across clusters
Probability scale Meta-analysis of cluster-specific estimates of the concordance probability	1. Equal weight for each cluster 2. Number of subjects in the cluster 3. Number of subjects in the cluster with an event 4. Number of usable subject pairs within the cluster 5. Inverse of the cluster-specific sampling variance estimate	6. Inverse of the sum of the cluster-specific sampling variance estimate and the between-cluster variance estimate
Log-odds scale Meta-analysis of cluster-specific estimates of the logit concordance probability	7. Inverse of the cluster-specific sampling variance estimate on log-odds scale	8. Inverse of the sum of the cluster-specific sampling variance estimate on log-odds scale and the between-cluster variance estimate on log-odds scale

straight line. Departures from this straight line indicate departures from normality. The Shapiro-Wilk test returns the probability of obtaining the test-statistic as least as extreme as the observed one, under the null-hypothesis that z_k are normally distributed (p-value). When the p-value is above significance level α, say 5%, the null hypothesis that z_k is normally distributed is not rejected.

Since the concordance probability is restricted to [0, 1] the normality assumption of random effects meta-analysis may be violated. We considered inverse variance weighted meta-analysis on the log-odds scale as an alternative approach (methods 7 and 8 for fixed effect and random effects meta-analysis respectively). The resulting estimators for the within-cluster concordance probability are defined in Appendix 2.2. The normality assumption on log-odds scale was again assessed by the normal probability plot and the Shapiro-Wilk test.

Table 2.1 contains a summary of the eight pooling methodologies described above. For all the meta-analyses we used the R package rmeta [14, 23].

RESULTS

The European patients were slightly older in comparison with the Asian patients (median age 36 vs. 31 years) and were more likely to have the worst GCS Motor Score of 1, i.e. no motor response (21% versus 4%) compared to the Asian patients (Table 2.2). However, 6 month mortality was lower in the European patients (27%) than in the Asian patients (35%).

Table 2.2 Patient characteristics in selected European and Asian centers.

Characteristic	Measure or Category	Europe		Asia	
Age (years)	Median (25 - 75 percentile)	36	(24 - 53)	31	(22 - 43)
GCS Motor score	No response (1)	445	(21%)	55	(4%)
	Extension (2)	134	(6%)	96	(7%)
	Abnormal flexion (3)	176	(8%)	124	(9%)
	Normal flexion (4)	321	(15%)	261	(18%)
	Localizes / obeys (5/6)	1,005	(48%)	885	(62%)
Pupil reactivity	No pupil reacted	291	(14%)	129	(9%)
	One pupil reacted	123	(6%)	117	(8%)
	Both pupils reacted	1,667	(80%)	1,175	(83%)
Six-month mortality	Dead	553	(27%)	495	(35%)
Patients	Total	2,081		1,421	
Centers	Total	35		21	
Patients per center	Median (25 - 75 percentile)	33	(21 - 64)	34	(20 - 66)

We found that 6-month mortality was clearly associated with higher age, worse GCS Motor Score and less pupil reactivity (Table 2.3). Heterogeneity in mortality among European centers was substantial as indicated by the hazard ratio of 1.7 for the 75 percentile versus the 25 percentile of the random center effect, based on the quartiles of the Gamma frailty distribution with mean 1 and variance estimate 0.146.

Among European centers (overall c-index 0.80) the c-indexes varied substantially with an inter-quartile-range of 0.70 to 0.81 (Figure 2.2). Pooled concordance probability estimates resulting from fixed effect meta-analysis ranged from 0.75 (equal weights) to 0.84 (inverse variance weights). Random effects meta-analysis (method 6) led to a mean concordance probability estimate $\hat{\mu} = 0.77$, a between-cluster variance estimate $\hat{\tau}^2 = 0.0072$ and a wide 95% prediction interval (0.60 to 0.95) reflecting the strong

Table 2.3 Associations between predictors and 6-month mortality in European centers.

Characteristic	Level	HR (95% CI)	
Age (years)	47 versus 23*	2.1	(1.9-2.4)
GCS Motor score	No response (1)	3.1	(2.4-4.0)
	Extension (2)	2.8	(2.0-3.8)
	Abnormal flexion (3)	2.4	(1.7-3.2)
	Normal flexion (4)	1.5	(1.1-2.0)
	Localizes / obeys (5/6)	1.0	(ref)
Pupil reactivity	No pupil reacted	2.8	(2.3-3.5)
	One pupil reacted	1.7	(1.2-2.3)
	Both pupils reacted	1.0	(ref)
Center random effect	75 versus 25 percentile	1.7	

* Interquartile range

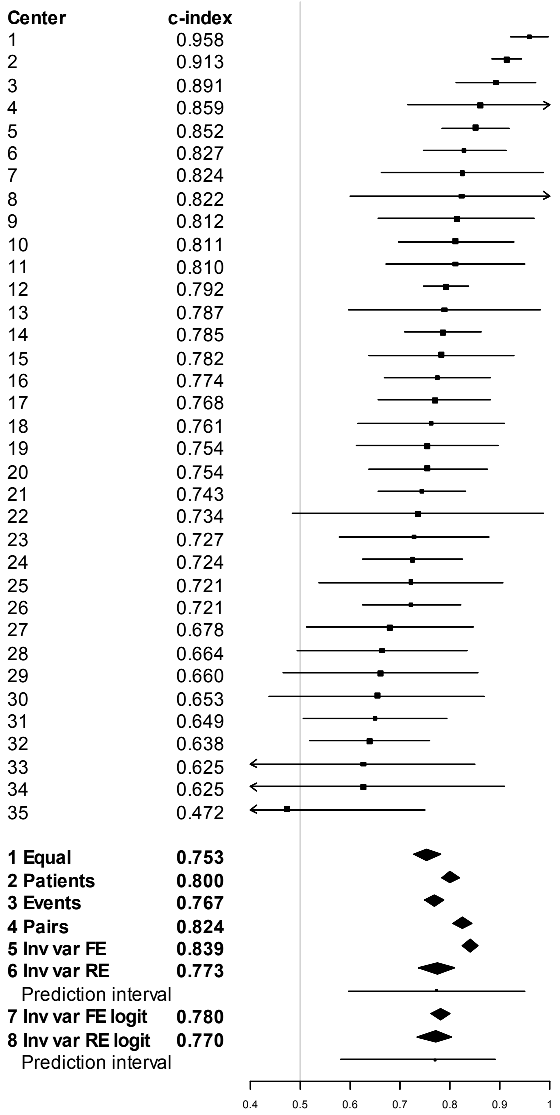
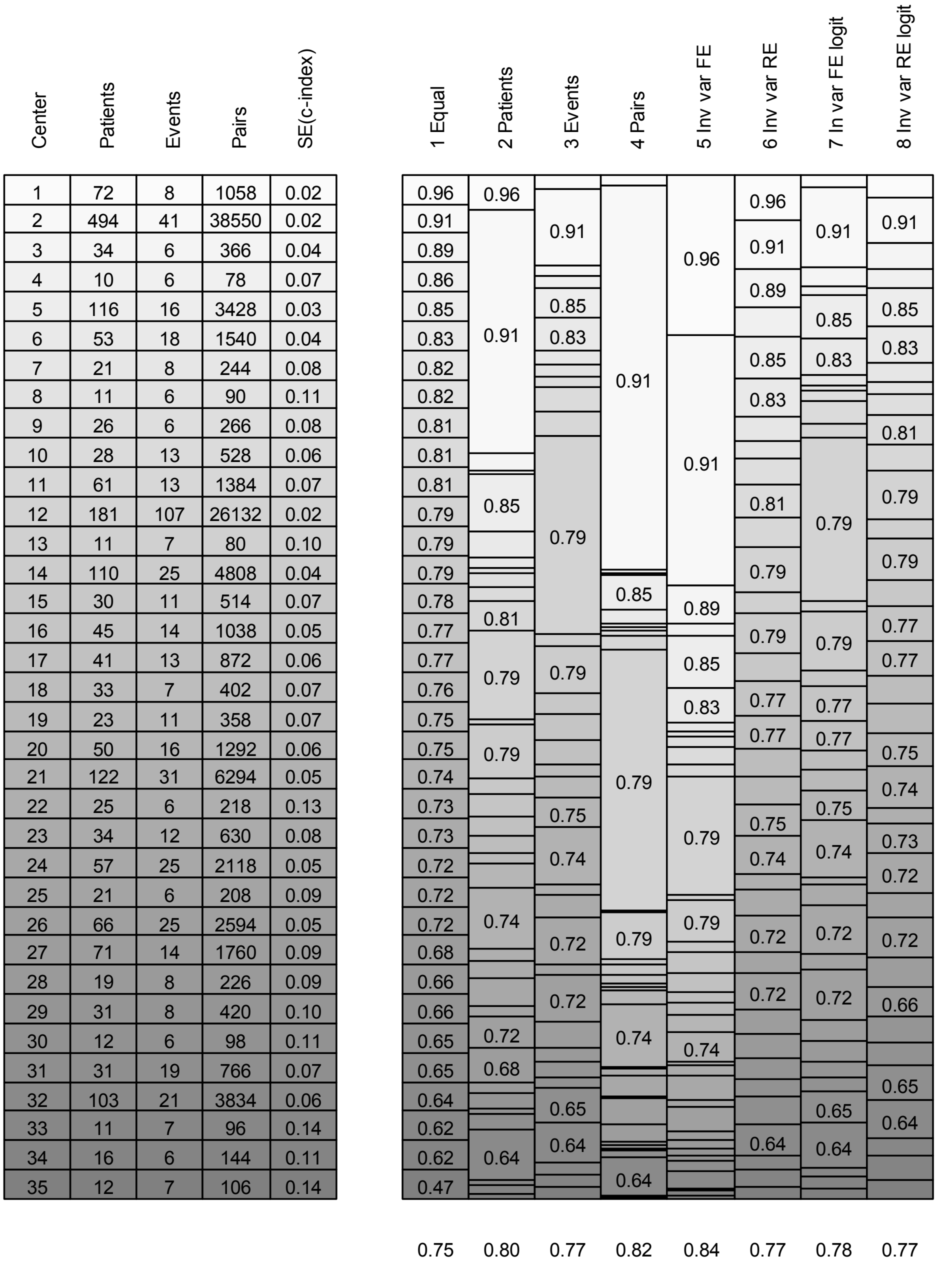


Figure 2.2 Center-specific and pooled concordance probability estimates with 95% confidence intervals for European centers. For pooled estimates based on random effects meta-analysis a 95% prediction interval for the concordance probability is presented by a horizontal line. 1 Equal = Fixed effect meta-analysis with equal weights; 2 Patients = Fixed effect meta-analysis with number of patients as weights; 3 Events = Fixed effect meta-analysis with number of events as weights; 4 Pairs = Fixed effect meta-analysis with number of usable patient pairs as weights; 5 Inv var FE = Fixed effect meta-analysis with inverse variance weights; 6 Inv var RE = Random effects meta-analysis with inverse variance weights; 7 Inv var FE logit = Fixed effect meta-analysis with inverse variance weights on log-odds scale; 8 Inv var RE logit = Random effects meta-analysis with inverse variance weights on log-odds scale.

heterogeneity in the cluster-specific concordance probabilities (I^2 = 0.76 with 95% confidence interval 0.66 to 0.82). Random-effects meta-analysis on log-odds scale (method 8) led to similar results, but with a somewhat smaller asymmetric prediction interval (0.58 to 0.89).

Large differences in pooling weights, together with heterogeneity in the cluster-specific concordance probabilities, led to very different pooled estimates. We analysed the pooling weights to explain the differences in pooled estimates (Figure 2.3). The patient-weighted estimate was dominated by center 2 with 494 of the 2,081 patients. The event-weighted estimate was dominated by center 12 with 107 out of 553 events.



Center	Patients	Events	Pairs	SE(c-index)	1 Equal
1	72	8	1058	0.02	0.96
2	494	41	38550	0.02	0.91
3	34	6	366	0.04	0.89
4	10	6	78	0.07	0.86
5	116	16	3428	0.03	0.85
6	53	18	1540	0.04	0.83
7	21	8	244	0.08	0.82
8	11	6	90	0.11	0.82
9	26	6	266	0.08	0.81
10	28	13	528	0.06	0.81
11	61	13	1384	0.07	0.81
12	181	107	26132	0.02	0.79
13	11	7	80	0.10	0.79
14	110	25	4808	0.04	0.79
15	30	11	514	0.07	0.78
16	45	14	1038	0.05	0.77
17	41	13	872	0.06	0.77
18	33	7	402	0.07	0.76
19	23	11	358	0.07	0.75
20	50	16	1292	0.06	0.75
21	122	31	6294	0.05	0.74
22	25	6	218	0.13	0.73
23	34	12	630	0.08	0.73
24	57	25	2118	0.05	0.72
25	21	6	208	0.09	0.72
26	66	25	2594	0.05	0.72
27	71	14	1760	0.09	0.68
28	19	8	226	0.09	0.66
29	31	8	420	0.10	0.66
30	12	6	98	0.11	0.65
31	31	19	766	0.07	0.65
32	103	21	3834	0.06	0.64
33	11	7	96	0.14	0.62
34	16	6	144	0.11	0.62
35	12	7	106	0.14	0.47

Figure 2.3 Meta-analysis pooling weights for European centers. For methods 1 to 8 the weights are represented by the height of the bars on the right hand side of the figure. C-indexes are printed in the bars if the cluster weight was at least equal to the average weight. 1 Equal = Fixed effect meta-analysis with equal weights; 2 Patients = Fixed effect meta-analysis with number of patients as weights; 3 Events = Fixed effect meta-analysis with number of events as weights; 4 Pairs = Fixed effect meta-analysis with number of usable patient pairs as weights; 5 Inv var FE = Fixed effect meta-analysis with inverse variance weights; 6 Inv var RE = Random effects meta-analysis with inverse variance weights; 7 Inv var FE logit = Fixed effect meta-analysis with inverse variance weights on log-odds scale; 8 Inv var RE logit = Random effects meta-analysis with inverse variance weights on log-odds scale.

The patient-pair-weighted estimate was heavily determined by both center 2 and center 12 as the number of usable patient pairs is related to the number of patients times the number of events. The fixed effect inverse-variance weighted estimate was also strongly influenced by centers with high number of patients or events, because the standard errors of the cluster-specific estimates depend heavily on the number of patients and events. Furthermore, the fixed effect inverse-variance weighted estimate was upwardly influenced by center 1 as a result of the small standard error relative to the small number of patients and events. The random effects inverse-variance weighted estimate was much less dominated by particular centers and close to the equally weighted estimate because of the large amount of heterogeneity. The standard error on the log-odds scale increased with increasing c-index according to equation 10 in Appendix 2.2 and therefore put less weight on the centers with a high concordance probability estimate resulting in lower pooled estimates. The large standard errors for centers with high c-index also decreased the heterogeneity (I^2 = 0.61 with 95% confidence interval 0.44 to 0.73) on the log-odds scale resulting in more similar weights for fixed effect and random effects meta-analysis.

To check the validity of the normality assumption in the random effects meta-analyses, we calculated standardized residuals (equation 3), both on the probability and the log-odds scale. The standardized residuals better fitted to the standard normal distribution on the probability scale than on the log-odds scale (Figure 2.4, p-values for rejection of the normality null hypothesis of 0.666 on probability scale and of 0.030 on log-odds scale).

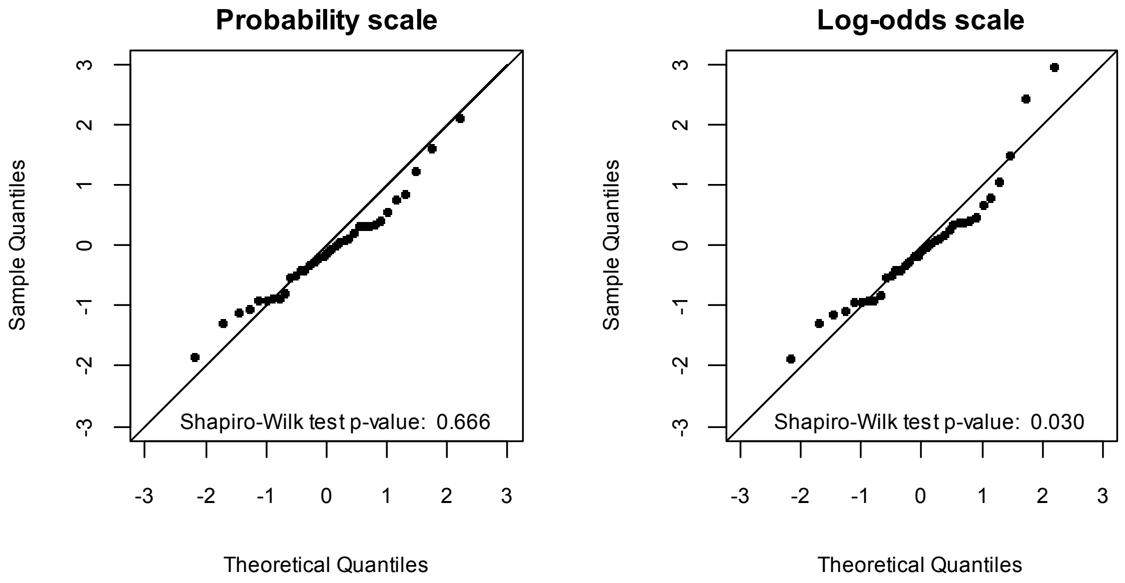


Figure 2.4 Normal probability plot of standardized residuals on probability scale and log-odds scale (European centers). Shapiro-Wilk test results are printed at the bottom of each plot.

To illustrate the comparison in an external validation setting, we repeated the analysis of the within-cluster concordance probability in Asian centers with the same risk model (Figure 2.5). Among Asian clusters (overall c-index 0.74) the c-indexes varied less (*IQR* 0.71-0.78), which was reflected in a lower proportion of variation among clusters that is due to heterogeneity rather than chance (I^2 = 0.32 with 95% confidence interval 0 to 0.60). As a result, different pooling methodologies led to more similar pooled estimates, because differences in cluster weights have less impact when cluster specific estimates are more alike. Based on random effects meta-analysis, estimates of the mean within-cluster concordance probability and the between-cluster variance were $\hat{\mu} = 0.75$ and $\hat{\tau}^2 = 0.0013$ respectively. The resulting prediction interval (0.67 to 0.83) was much smaller than for the European clusters. The heterogeneity disappeared on the log-odds scale (I^2 = 0) leading to equal estimates by fixed effect and random effects meta-analysis.

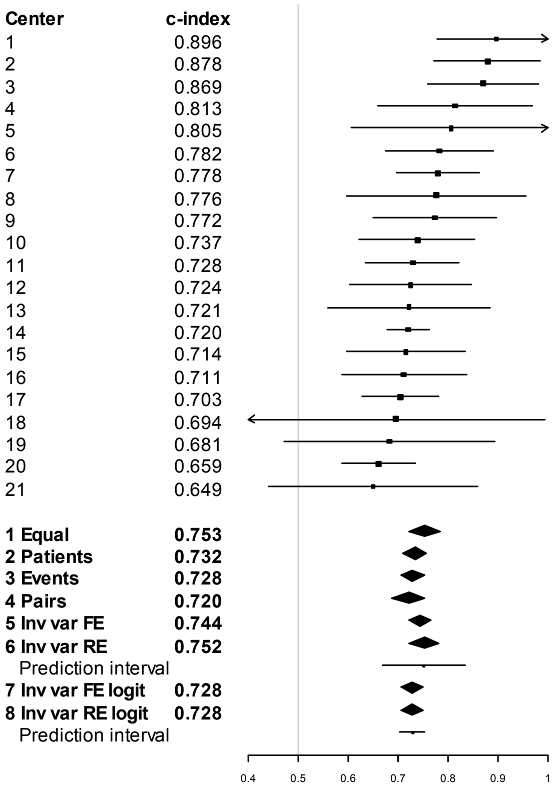


Figure 2.5 Center-specific and pooled concordance probability estimates with 95% confidence intervals for Asian centers (external validation). For pooled estimates based on random effects meta-analysis (methods 6 and 8) a 95% prediction interval for the concordance probability is presented by a horizontal line. 1 Equal = Fixed effect meta-analysis with equal weights; 2 Patients = Fixed effect meta-analysis with number of patients as weights; 3 Events = Fixed effect meta-analysis with number of events as weights; 4 Pairs = Fixed effect meta-analysis with number of usable patient pairs as weights; 5 Inv var FE = Fixed effect meta-analysis with inverse variance weights; 6 Inv var RE = Random effects meta-analysis with inverse variance weights; 7 Inv var FE logit = Fixed effect meta-analysis with inverse variance weights on log-odds scale; 8 Inv var RE logit = Random effects meta-analysis with inverse variance weights on log-odds scale.

DISCUSSION

We studied how to assess the discriminative ability of risk models in clustered data. The within-cluster concordance probability is an important measure for risk models when these models are used to support decisions on interventions within the clusters. The within-cluster concordance probability can be estimated by pooling cluster-specific concordance probability estimates (e.g. c-indexes) with a meta-analysis, similar to pooling of study-specific treatment effect estimates. We considered different pooling strategies (Table 2.1) and recommend random effects meta-analysis in case of substantial variability – beyond chance – of the concordance probability across clusters [20, 21]. To decide if the meta-analysis should be undertaken on the probability scale or the log-odds scale we suggest considering the normality assumptions on both scales by normal probability plots and Shapiro-Wilk tests of the standardized residuals.

The illustration of predicting 6-month mortality after TBI prompted the use of random-effects meta-analysis because of the strong difference – beyond chance – in concordance probability among centers. This was clearly visualized by the forest plot in Figure 2.2. Random effects meta-analysis results can be summarized by the mean concordance probability and a 95% prediction interval for possible values of the concordance probability. By definition, these results give insight into the variation of the discriminative ability among centers as opposed to fixed effect meta-analysis results [20, 21]. By comparing normal probability plots and Shapiro-Wilk test results based on the standardized residuals we concluded the random effects meta-analysis results on probability scale to be the most appropriate (Figure 2.4). Although the methodology is illustrated with time-to-event outcomes of traumatic brain injury patients, it is also applicable to binary outcomes.

Even if a risk model contains regression coefficients that are optimal for the data in each cluster, differences in case mix may lead to different concordance probabilities across clusters [24]. Furthermore, predictor effects may vary because of cluster-specific circumstances, also leading to different cluster-specific concordance probabilities. Given the variability beyond chance in our case study, we consider a random effects meta-analysis of the cluster-specific c-indexes as most appropriate.

The assumption of random effects meta-analysis is that underlying concordance probabilities among clusters are exchangeable, i.e. cluster-specific concordance probabilities are expected to be non-identical, yet identically distributed [20]. If part of the variation can be explained by cluster characteristics, a meta-regression – assuming partial exchangeability – of the concordance probability estimates with cluster characteristics as covariates is preferable.

We chose to analyse the concordance probability as it is the most commonly used measure of discriminative ability of a risk model. However, the same logic of pooling cluster-specific performance measure estimates can be applied to any other performance measure, like the discrimination slope, the explained variation (R^2) or the Brier score [25].

We used Harrell's c-index to estimate cluster-specific concordance probabilities together with Quade's formula for the cluster-specific variances of the c-index [2, 16]. The same methodology of pooling cluster-specific performance measure estimates can be applied to other concordance probability estimators and its variances. Other estimators for the concordance probability in time-to-event data can be found in Gönen and Heller [26] and Uno et al [27]. These estimators are especially favourable when censoring varies by cluster as they are shown to be less sensitive to censoring distributions. Other variance estimators are described by Hanley and McNeil [28], and DeLong et al [29] for binary outcome data and by Nam and D'Agostino [30] and Pencina and D'Agostino [3] for time-to-event outcome data. The variance of the concordance probability estimate can also be estimated with a bootstrap procedure [31].

CONCLUSION

We recommend meta-analysis of cluster-specific c-indexes when assessing discriminative ability of risk models used to support decisions at cluster level. Particularly, random effects meta-analysis should be considered as it allows for and provides insight into the variability of the concordance probability among clusters.

ACKNOWLEDGMENTS

The authors express their gratitude to all of the principal investigators of the CRASH trial for providing the data. We thank Prof. Emmanuel Lesaffre (Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands) for helpful comments. This work was supported by the Netherlands Organisation for Scientific Research (grant 917.11.383).

REFERENCES

1. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW: Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010, 21(1):128-138.
2. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA: Evaluating the yield of medical tests. *JAMA* 1982, 247(18):2543-2546.
3. Pencina MJ, D'Agostino RB: Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* 2004, 23(13):2109-2123.
4. Bouwmeester W, Twisk JW, Kappen TH, van Klei WA, Moons KG, Vergouwe Y: Prediction models for clustered data: comparison of a random intercept and standard regression model. *Medical Research Methodology* 2013 in press, 13.
5. Gelman A, Hill J: *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press; 2007.
6. Duchateau L, Janssen P: *The Frailty Model*: Springer; 2010.
7. Van Oirbeek R, Lesaffre E: An application of Harrell's C-index to PH frailty models. *Stat Med* 2010, 29(30):3160-3171.
8. Van Oirbeek R, Lesaffre E: Assessing the predictive ability of a multilevel binary regression model. *Computational Statistics &amp; Data Analysis* 2012, 56(6):1966-1980.
9. Collaborators MCT, Perel P, Arango M, Clayton T, Edwards P, Komolafe E, Poccock S, Roberts I, Shakur H, Steyerberg E *et al*: Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. *BMJ* 2008, 336(7641):425-429.
10. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JD *et al*: Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Med* 2008, 5(8):e165; discussion e165.
11. Edwards P, Arango M, Balica L, Cottingham R, El-Sayed H, Farrell B, Fernandes J, Gogichaisvili T, Golden N, Hartzenberg B *et al*: Final results of MRC CRASH, a randomised placebo-controlled trial of intravenous corticosteroid in adults with head injury-outcomes at 6 months. *Lancet* 2005, 365(9475):1957-1959.
12. Teasdale G, Jennett B: Assessment of coma and impaired consciousness. A practical scale. *Lancet* 1974, 2(7872):81-84.
13. Jennett B, Bond M: Assessment of outcome after severe brain damage. *Lancet* 1975, 1(7905):480-484.
14. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/. 3-900051-07-0 edition.; 2011.
15. Therneau T. and original Splus->R port by Lumley T.: survival: Survival analysis, including penalised likelihood. R package version 2.36-9. http://CRAN.R-project.org/package=survival. 2011.
16. Quade D: *Nonparametric partial correlation*. *Volume No. 526*. North Carolina: Institute of Statistics Mimeo Series No. 526; 1967.
17. Higgins JP, Thompson SG: Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002, 21(11):1539-1558.
18. DerSimonian R, Laird N: Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986, 7(3):177-188.
19. DerSimonian R, Kacker R: Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials* 2007, 28(2):105-114.
20. Higgins JPT, Thompson SG, Spiegelhalter DJ: A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009, 172(1):137-159.

21. Riley RD, Higgins JPT, Deeks JJ: Interpretation of random effects meta-analyses. *BMJ* 2011, 342.
22. Hardy RJ, Thompson SG: Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* 1998, 17(8):841-856.
23. Lumley T: rmeta: Meta-analysis. R package version 2.16. http://CRAN.R-project.org/package=rmeta. 2009.
24. Vergouwe Y, Moons KG, Steyerberg EW: External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010, 172(8):971-980.
25. Steyerberg EW: *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*: Springer-Verlag New York; 2009.
26. Gönen M, Heller G: Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005, 92(4):965-970.
27. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ: On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011, 30(10):1105-1117.
28. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982, 143(1):29-36.
29. DeLong ER, DeLong DM, Clarke-Pearson DL: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988, 44(3):837-845.
30. Nam BH, D'Agostino RB: Discrimination Index, the Area under the ROC Curve. In *Goodness-of-Fit Tests and Model Validity*. Boston: Birkhauser; 2002.
31. Efron B, Tibshirani R: *An Introduction to the Bootstrap*: Chapman & Hall; 1993.
32. Pepe MS: The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford Statistical Science Series, Oxford University Press; 2003:107.

APPENDIX 2.1

The concordance probability is defined as the probability that a randomly chosen subject pair with different outcomes is concordant. For a randomly chosen subject pair (*i, j*) with outcomes Y_i and Y_j and model predictions $\hat{Y}_i$ and $\hat{Y}_j$ the concordance probability *C* is:

$$C = \Pr\left(\hat{Y}_i < \hat{Y}_j \middle| Y_i < Y_j\right) \tag{4}$$

Harrell's c-index [2] estimates the concordance probability by the proportion of all usable pairs of subjects (n_u) in which the predictions and outcomes are concordant (n_c), with tied predictions (n_t) counted as 1/2:

$$\hat{C} = \frac{n_c + n_t/2}{n_u} \tag{5}$$

For binary outcomes *y*, pairs of subjects are usable if one of the subjects had an event and the other did not. The number of usable subject pairs n_u, the number of concordant subject pairs n_c and the number of tied subject pairs n_t are:

$$\begin{aligned} n_u &= \sum\nolimits_i \sum\nolimits_j I(y_i < y_j) \\ n_c &= \sum\nolimits_i \sum\nolimits_j I(y_i < y_j \ \ and \ \ \hat{y}_i < \hat{y}_j) \\ n_t &= \sum\nolimits_i \sum\nolimits_j I(y_i < y_j \ \ and \ \ \hat{y}_i = \hat{y}_j) \end{aligned} \tag{6}$$

For time-to-event outcomes *y*, pairs of subjects are usable if their survival times are not equal and at least the smallest survival time is uncensored. We have to add the restriction that the smallest observation y_i of each subject pair is uncensored, denoted by $\delta_i = 1$:

$$\begin{aligned} n_u &= \sum\nolimits_i \sum\nolimits_j I(y_i < y_j \ \text{ and } \ \delta_i = 1) \\ n_c &= \sum\nolimits_i \sum\nolimits_j I(y_i < y_j \ \text{ and } \ \delta_i = 1 \ \text{ and } \ \hat{y}_i < \hat{y}_j) \\ n_t &= \sum\nolimits_i \sum\nolimits_j I(y_i < y_j \ \text{ and } \ \delta_i = 1 \ \text{ and } \ \hat{y}_i = \hat{y}_j) \end{aligned} \tag{7}$$

The variance of the c-index can be estimated according to Quade [16]:

$$\hat{\sigma}^2_{\hat{C}} = \frac{\sum n_{u,i}^2 \left(\sum n_{c-d,i}\right)^2 - 2\sum n_{u,i} \sum n_{c-d,i} \sum n_{u,i} n_{c-d,i} + \left(\sum n_{u,i}\right)^2 \sum n_{c-d,i}^2}{\left(\sum n_{u,i}\right)^4} \tag{8}$$

All summations over *i* with $n_{u,i}$ and $n_{,c\text{-}d,i}$ the number of usable and the number of concordant minus discordant subject pairs of which subject *i* is one:

$$\begin{aligned} n_{u,i} &= \sum\nolimits_j I(y_i < y_j \ \text{ and } \ \delta_i = 1) \\ n_{c,i} &= \sum\nolimits_j I(y_i < y_j \ \text{ and } \ \delta_i = 1 \ \text{ and } \ \hat{y}_i < \hat{y}_j) \\ n_{d,i} &= \sum\nolimits_j I(y_i < y_j \ \text{ and } \ \delta_i = 1 \ \text{ and } \ \hat{y}_i > \hat{y}_j) \\ n_{c-d,i} &= n_{c,i} - n_{d,i} \end{aligned} \tag{9}$$

APPENDIX 2.2

Based on the delta method, Pepe [32] gave a variance estimator for the logit of the c-index:

$$\operatorname{var}\left(\operatorname{logit}\left(\hat{\mathrm{C}}\right)\right) = \operatorname{var}\left(\log\left(\frac{\hat{\mathrm{C}}}{1-\hat{\mathrm{C}}}\right)\right) = \frac{\operatorname{var}\left(\hat{\mathrm{C}}\right)}{\left(\hat{\mathrm{C}}\left(1-\hat{\mathrm{C}}\right)\right)^2} \qquad (10)$$

We used this variance estimator to perform a meta-analysis on log-odds scale. The pooling weights (method 7) for a fixed effect inverse variance meta-analysis on log-odds scale are:

$$w_q = \left[\frac{\hat{\sigma}_q^2}{\left(\hat{C}_q\left(1-\hat{C}_q\right)\right)^2}\right]^{-1} \qquad (11)$$

The pooling weights (method 8) for a random effects inverse variance meta-analysis on log-odds scale are:

$$w_q = \left[\frac{\hat{\sigma}_q^2}{\left(\hat{C}_q\left(1-\hat{C}_q\right)\right)^2} + \hat{\tau}^2\right]^{-1} \qquad (12)$$

The resulting pooled estimates together with confidence and prediction intervals are transformed back to probability scale.

3

Geographic and temporal validity of prediction models: different approaches were useful to examine model performance

PC Austin
D van Klaveren
Y Vergouwe
D Nieboer
DS Lee
EW Steyerberg

Journal of Clinical Epidemiology 2016

ABSTRACT

Objective Validation of clinical prediction models traditionally refers to the assessment of model performance in new patients. We studied different approaches to geographic and temporal validation in the setting of multicenter data from two time periods.

Study Design and Setting We illustrated different analytic methods for validation using a sample of 14,857 patients hospitalized with heart failure at 90 hospitals in two distinct time periods. Bootstrap resampling was used to assess internal validity. Meta-analytic methods were used to assess geographic transportability. Each hospital was used once as a validation sample, with the remaining hospitals used for model derivation. Hospital-specific estimates of discrimination (c-statistic) and calibration (calibration intercepts and slopes) were pooled using random effects meta-analysis methods. I^2 statistics and prediction interval width quantified geographic transportability. Temporal transportability was assessed using patients from the earlier period for model derivation and patients from the later period for model validation.

Results Estimates of reproducibility, pooled hospital-specific performance, and temporal transportability were on average very similar, with c-statistics of 0.75. Between-hospital variation was moderate according to I^2 statistics and prediction intervals for c-statistics.

Conclusion This study illustrates how performance of prediction models can be assessed in settings with multicenter data at different time periods.

WHAT IS NEW?

Key findings

- Using data on patients hospitalized with heart failure in the Canadian province of Ontario and a previously-derived clinical prediction model, we found that several strategies to quantify model performance showed similar overall results, with moderate variation in center-specific performance.
- Ninety-five percent prediction intervals for a new hospital-specific c-statistic were moderately wide in each of the two time periods.

What this adds to what is known

- Bootstrap correction for optimism resulted in a similar overall estimate of model performance as a leave-one-hospital-out approach, in which each hospital was used once for model validation.
- Random-effects meta-analysis provided insight into the variability of center-specific performance measures as an indication of geographical transportability of a prediction model, when the focus is on within-center performance of the model.

What is the implication/what should change now

- Appropriate statistical methods should be used to quantify the geographic and temporal portability of clinical prediction models.
- Validation studies of clinical prediction models should carefully describe whether overall validity of a model is reported, or that transportability is addressed by assessment of geographical or temporal variability in performance.

INTRODUCTION

Clinical prediction models permit one to estimate the probability of the presence of disease or of the occurrence of adverse events. These models can inform medical decision making and provide individualized information on patient prognosis. Validation traditionally refers to assessing the performance of a model in subjects other than those in whom it was developed. Validation is an important issue in the scientific development of prediction models towards wide application.

Different frameworks for model validation have been proposed. Internal validation is commonly differentiated from external and temporal validation [1, 2]. Interval validation, also referred to as reproducibility [3, 4], describes how well the model performs in patients who were not included in model development, but who are from the same underlying population. Temporal validation refers to the performance of the model on subsequent patients in settings similar to that in which the model was developed. External validation refers to the process of examining the performance of the model on data from centres different from those which participated in model development. The term transportability refers to a model that maintains its performance in a population that is different from that in which it was developed [3, 4]. Different aspects of transportability have been defined: historical, geographic, methodologic (model performs well when data were collected using different methods), spectrum (model performs well when the distribution of disease severity differs), and follow-up interval (model performs well when the outcome is assessed over a different duration of follow-up time) [3].

We aimed to describe and illustrate methods for assessing the geographic and temporal transportability of clinical prediction models. Accordingly, we analyzed data on patients hospitalized with congestive heart failure (CHF) at a large number of hospitals in two distinct time periods.

METHODS

Data sources

The study used patients from The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study, which was an initiative to improve the quality of care for patients with cardiovascular disease in Ontario [5]. Only patients admitted to those 90 hospitals that participated in both phases of the study were included in the current study. The current study included 7,549 patients hospitalized with CHF during the first phase of the study (April 1999 to March 2001) and 7,308 patients hospitalized during the second phase of the study (April 2004 to March 2005).

There was a notable difference in the inclusion and exclusion criteria between the two phases of the study. Patients were excluded from the first phase if they had had a prior hospitalization for CHF. This exclusion criterion was removed from the second phase of the study. This enabled us to examine both temporal portability and spectrum or methodological portability.

Heart failure mortality prediction model

The EFFECT-HF mortality prediction models estimate the probability of death within 30 days and one year of hospitalization for CHF [6]. The model for predicting one-year mortality uses 11 variables: age, systolic blood pressure on admission, respiratory rate on admission, low sodium serum concentration (< 136 mEq/L), low serum hemoglobin (< 10.0 g/dL), serum urea nitrogren, presence of cerebrovascular disease, presence of dementia, chronic obstructive pulmonary disease, hepatic cirrhosis, and cancer.

Table 3.1 Methods for assessing geographical and temporal model performance.

	Description
Methods that ignore temporal and geographic variation	
Apparent performance	Model performance is assessed in the sample in which it was developed. No adjustment is made for the model being optimized to fit in the sample used for derivation and validation.
Optimism-corrected performance	Model is derived in a bootstrap sample and applied to the overall sample to provide an estimate of model optimism. The average optimism is computed over a large number of bootstrap samples and is subtracted from the estimate of apparent performance.
Geographic transportability	
Internal-external: Leave-one-hospital-out (pooled)	Data from one hospital is withheld and the model is derived using data from the remaining hospitals. The model is then applied to subjects from the withheld hospital to obtain predicted probabilities for each of the withheld subjects. This process is repeated so that each hospital is excluded once from the derivation sample. Model performance is then determined in the pooled sample consisting of the predictions for each subject when that subject's hospital was excluded from the model derivation sample.
Internal-external: Leave-one-hospital-out (meta-analysis)	As for internal-external, but rather than estimating performance on the pooled sample, we combine the hospital-specific estimates of model performance using a random effects meta-analysis.
Temporal transportability (model estimated in Phase 1 and applied in Phase 2)	
Fixed effects regression model	Model contains fixed intercept and fixed effects for all covariates (similar to all the models described above). Model is derived in Phase 1 and validated in Phase 2.
Mixed effects regression model	Model contains hospital-specific random intercepts and fixed effects for all covariates. Model is derived in Phase 1 and validated in Phase 2.
Case-mix adjusted performance	Model is developed in Phase 1 and applied to subjects in Phase 2. Using the predicted probability of the occurrence of the outcome, outcomes are simulated for each subject in Phase 2. Using the simulated outcome and the predicted probability of the occurrence of the outcome, model performance is assessed. This process is repeated 1,000 times to obtain a stable estimate of model performance.
Simultaneous geographic and temporal portability	
Leave-one-hospital-out temporally (meta-analysis)	Data from one hospital is withheld. The model is derived using Phase 1 data from the remaining hospitals. The model is then validated in the excluded hospital using data from Phase 2. Process is repeated so that each hospital is used once for model validation. The hospital-specific estimates of performance are then pooled using a random effects meta-analysis.
Leave-one-hospital-out temporally (pooled)	Data from one hospital is withheld. The model is derived using Phase 1 data from the remaining hospitals. The model is then applied to the excluded hospital using data from Phase 2. Process is repeated so that each hospital is used once for model validation. The estimated probability of the outcome is pooled across all patients at all hospitals and the c-statistic is calculated.

Measures of model performance

Discrimination is a key component of assessing the validity of a clinical prediction model. We quantified discrimination using the c-statistic [7, 8]. We used two methods for assessing model calibration. First, loess smoothers were used to describe graphically the agreement between predicted probabilities and the observed probabilities of the occurrence of the outcome [9]. Second, we used calibration intercepts and slopes as summary measures [10]. The calibration intercept, also known as calibration in the large, is equal to the intercept of a logistic regression model in which the binary outcome is regressed on the estimated linear predictor when the slope is fixed at one [7]. The calibration slope is the slope from a logistic regression model when the binary outcome is regressed on the estimated linear predictor. The predicted probabilities are too low if the calibration intercept is greater than zero and are too high if the calibration intercept is less than zero. A calibration slope smaller than one indicates that the range of observed probabilities is smaller than the range of predicted probabilities [1, 11].

Statistical methods for assessing geographic and temporal validity (Table 3.1)

Model reproducibility: bootstrap estimates of optimism-corrected performance

Apparent performance refers to the performance of the model in the sample in which the model was developed. The apparent estimate of model performance tends to be optimistic, since the model is derived in the same sample in which its performance is being assessed. We may use bootstrapping to adjust for this optimism [7] (Section 1 of the online Appendix). Bootstrap-corrected estimates of performance assess the internal validity of the estimated prediction model (or the reproducibility of the model [3, 4]). This denotes the expected performance of the model if it were to be applied to new patients, from the same population as those used for model derivation. Alternative methods exist to assess model reproducibility. These include split-sample assessment and “leave-one-out” approaches [12, 13]. We did not consider these methods as previous studies that have found them to be inefficient [14, 15] or result in an under-estimation of the c-statistic when the number of events per variable was low [16].

Estimates of temporal transportability

The following model was fit: $\mathrm{logit}(p_{ij}) = \alpha_0 + \mathbf{X}_{ij}\beta$, where $\mathbf{X}_{ij}$ denotes a vector containing the predictor variables, β denotes the vector of regression coefficients and α_0 denotes the intercept, where the subscript “*ij*” denotes the *i*th patient admitted to the *j*th hospital. Using the coefficients estimated in the first phase of the sample, predicted probabilities of the occurrence of the outcome were then obtained for each subject in the second phase of the sample. Both the discrimination and calibration of the model estimated in the first phase of data were assessed using the subjects from the second phase of the study. As noted above, the inclusion and exclusion criteria differed slightly between the two phases of the study. While these methods were primarily intended to assess the temporal portability of the model, they also reflect spectrum transportability.

We further examined whether incorporating hospital-specific random effects in the prediction model estimated in the first phase of the study improved its temporal portability. The prediction model described above was modified to include hospital-specific random effects when fit in the first phase of the study: $\mathrm{logit}(p_{ij}) = \alpha_{0j} + \mathbf{X}_{ij}\beta$, where $\alpha_{0j} \sim N(\alpha_0, \sigma^2)$.

Assessing geographic portability of the model

Within each of the two phases of the study, we examined the degree to which model performance varied across hospitals. One hospital was excluded from the analytic sample. The prediction model was estimated in the remaining hospitals. This process was repeated so that each hospital was excluded once. We considered two different methods for assessing geographic portability. The first, referred to as 'leave-one-hospital-out (pooled)', determined the predicted probability of the occurrence of the outcome for each patient in the excluded hospital using the model fit in the remaining hospitals. The predicted probabilities for all patients at all hospitals were pooled and the performance of the prediction model was assessed. This approach was used for both model discrimination and calibration. This approach can be seen as a form of cross-validation, in which the strata consist of individual centres [17].

The second approach, referred to as 'leave-one-hospital-out (meta-analysis)', is based on work by van Klaveren et al. [18] (Section 2 of the online Appendix). Hospital-specific measures of model performance were obtained at each excluded hospital when the model was fit using the sample of all of the other remaining hospitals. Random effects meta-analyses methods were used to combine the individual hospital-specific estimates of model performance. Pooled estimates of discrimination and calibration were obtained as well as estimates of heterogeneity of the between-hospital variance (τ^2). It has been suggested that I^2 values of 25%, 50% and 75% can be considered to denote low, moderate, and high heterogeneity for treatment effect estimates [19]. We follow this classification in our study. Furthermore, prediction intervals were calculated for the expected performance of the clinical prediction model in centres that did not contribute to their development.

These two models differ only in that the former pools all of the patient-specific predicted probabilities and then computes an overall measure of model performance, while the latter pools hospital-specific estimates of model performance.

Simultaneous geographic and temporal transportability

We examined a 'leave-one-hospital-out' approach to examine geographic and temporal portability. One hospital was selected from the set of 90 hospitals. The model was estimated using patients admitted during Phase 1 to the remaining 89 hospitals. The estimated prediction model was then applied to patients admitted during Phase 2 to the selected hospital. When using the 'leave-one-hospital-out (meta-analysis)' approach, the c-statistic of the model, when applied to patients from this single hospital in Phase 2 was then determined. This process was repeated 90 times, allowing each hospital to serve as the validation sample once. The 90 estimates of the c-statistic were then pooled using a random effects meta-analysis, as described above. In contrast, when using a 'leave-one-hospital-out (pooled)' approach, the predicted probabilities obtained at each of the 90 hospitals (obtained when that hospital was used as the validation sample) were pooled to provide a single c-statistic.

Effects of changes in case-mix on temporal variation in model performance.

We examined whether changes in case-mix between the two phases of the study had an effect on the temporal validity of the prediction model [4]. First, the two phases of the study were pooled and an indicator variable denoting temporal period was regressed on the 11 variables in the clinical prediction model and a binary variable denoting one year mortality. The c-statistic of this model was used as a measure of the degree to which the case-mix of patients differed between the two study periods, also referred to as a membership model [4]. Second, we computed the linear predictor of the original EFFECT-HF model estimated in the first phase of the study and when applied to patients in the second phase. Both the standard deviation and the mean of the linear predictor were determined in each of the two phases. Increased variability of the linear predictor denotes increased heterogeneity of case-mix. As heterogeneity increases, the expected discriminative ability of a model increases [20].

Additionally, we estimated the case-mix corrected c-statistic of the model developed in the first phase of the study, when applied to the second phase of the study [21] (Section 3 of the online Appendix).

RESULTS

Reproducibility

The apparent c-statistic of the EFFECT-HF model was 0.747 in each of the two phases (Table 3.2). Bootstrap validation showed very little optimism in the apparent estimates of performance (decrease by 0.002 to 0.745 in each of the two samples).

Table 3.2 Estimated c-statistics obtained using different approaches.		
	Phase 1	Phase 2
Reproducibility (performance in different patients from the same population)		
Apparent performance	0.747	0.747
Optimism-corrected performance	0.745	0.745
Leave-one-hospital-out (pooled)	0.745	0.745
Leave-one-hospital-out (meta-analysis of model performance)	0.752	0.754
Temporal transportability (estimate in Phase 1 and apply in Phase 2)		
No hospital-specific random effects (model contained a fixed intercept and fixed effects for the predictor variables)	0.745	
With hospital-specific random effects (model contained hospital-specific intercepts and fixed effects for the predictor variables)	0.745	
Case-mix adjusted performance	0.746	
Simultaneous geographic and temporal transportability		
Model estimated in 89 hospitals in Phase 1 and then applied to the excluded hospital in Phase 2 (meta-analytic pooling of performance estimates) ('leave-one-hospital-out (meta-analysis)')	0.753	
Model estimated in 89 hospitals in Phase 1 and then applied to the excluded hospital in Phase 2 ('leave-one-hospital-out (pooled)')	0.745	

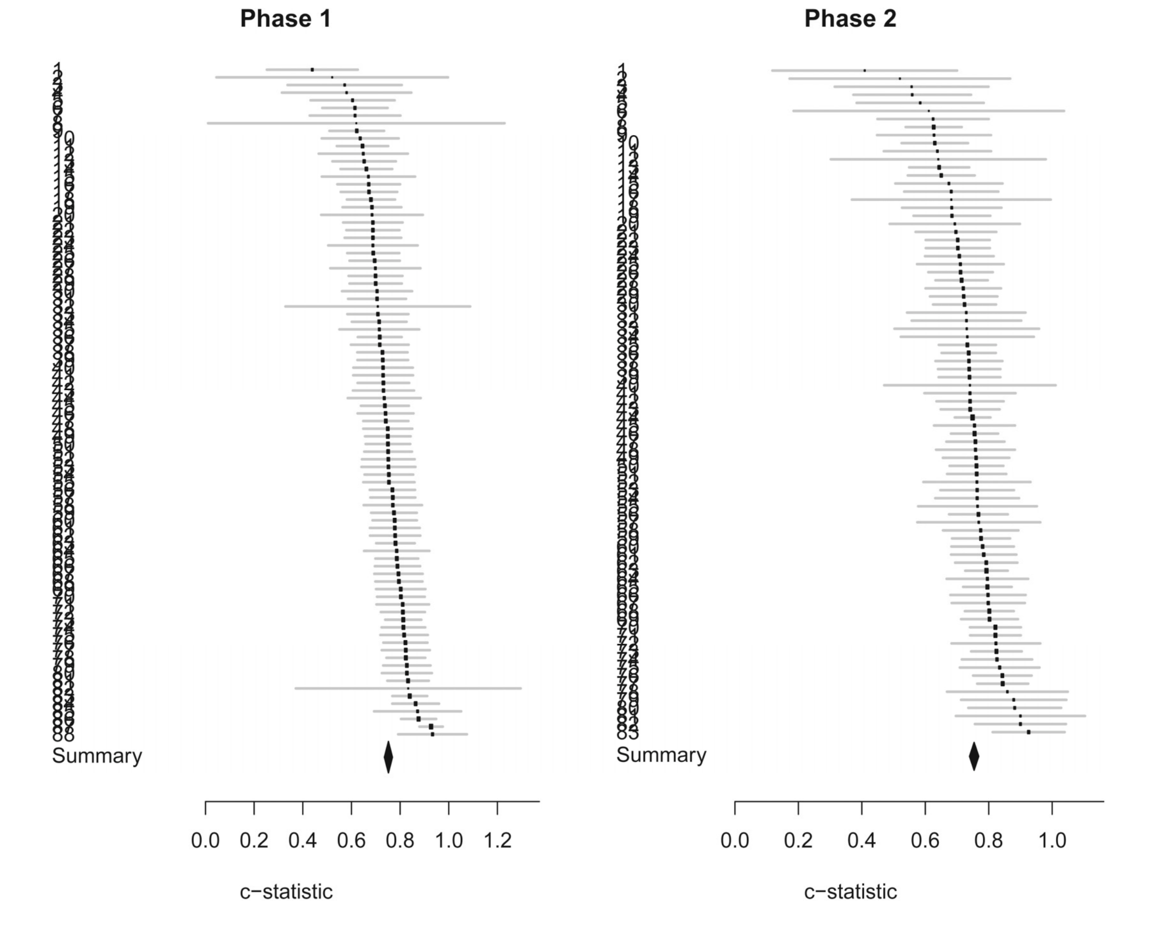


Figure 3.1 Random-effects meta-analyses of hospital-specific c-statistics.

Geographic transportability

When using the 'leave-one-hospital-out (pooled)' approach, the estimate of the c-statistic of the EFFECT-HF model was the same as the bootstrap-corrected estimates of the c-statistics observed above. The random effects meta-analysis estimates of the mean within-hospital c-statistics were slightly higher: 0.752 (95% CI: (0.735, 0.769)) and 0.754 (95% CI: (0.739, 0.769)) in the Phase 1 and 2 samples, respectively. The 95% prediction intervals were wide: (0.644, 0.859) and (0.689, 0.819), respectively. These denote the intervals within which the true hospital-specific c-statistic for a new hospital is likely to lie. The width of these prediction intervals reflects both the degree of between-hospital heterogeneity in the hospital-specific c-statistics (i.e., τ^2) and the standard deviation of the mean (which is influenced by the size of the overall sample). The values of τ (which are estimates of the between-hospital standard deviation of the hospital-specific performance) in the two phases were 0.054 and 0.032, while the values of I^2 (which measures the degree of heterogeneity in the hospital-specific measures of performance) in the two phases were 48.5% and 23.9%. Thus, there was moderately greater heterogeneity in the hospital-specific c-statistics in the earlier time period compared to the later time period (Figure 3.1).

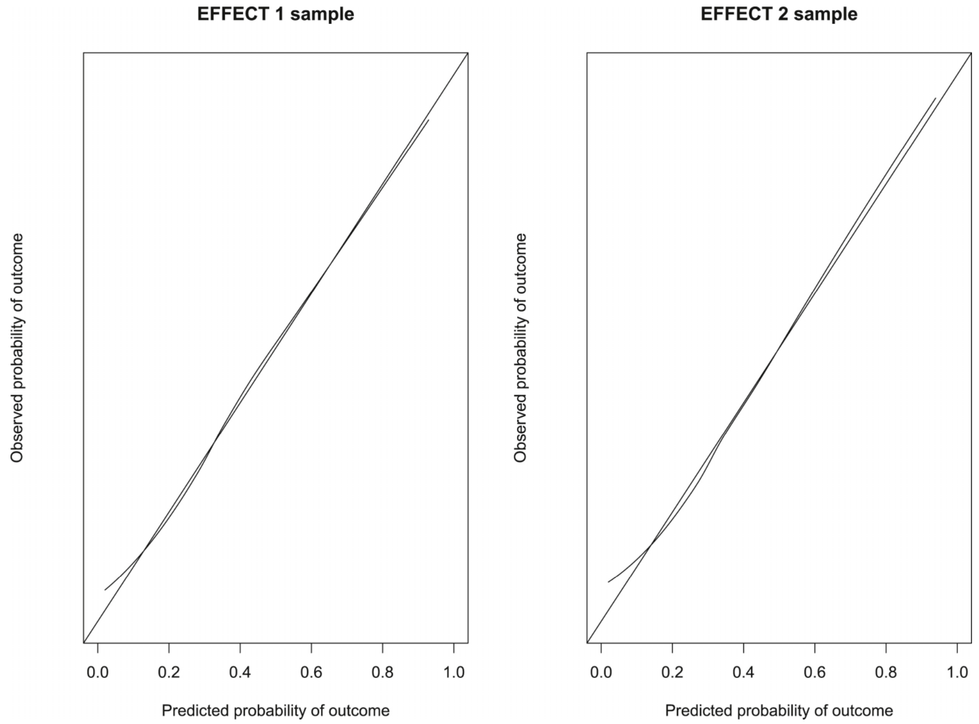


Figure 3.2 Calibration in EFFECT samples (leave-one-hospital-out approach). EFFECT, The Enhanced Feedback for Effective Cardiac Treatment.

The overall calibration was nearly perfect using the 'leave-one-hospital-out (pooled)' approach (Figure 3.2). The model displayed very good calibration in each of the two phases of the study, with some minor suggestion of under-prediction in those patients with the lowest predicted probability of mortality. The random effects meta-analysis estimates of the hospital-specific calibration intercepts for the EFFECT-HF model in the Phase 1 and Phase 2 samples were 0.011 (95% CI: (-0.053, 0.075)) and 0.016 (95% CI: (-0.059, 0.091)), respectively. The 95% prediction intervals were (-0.317, 0.340) and (-0.419, 0.451), respectively. The I^2 statistics in the two phases were 28.5% and 40.1%. Thus, there was low to modest heterogeneity in the hospital-specific mortality. The random effects meta-analysis estimates of the hospital-specific calibration slopes for the EFFECT-HF model in the Phase 1 and Phase 2 samples were 0.968 (95% CI: (0.896, 1.040)) and 0.964 (95% CI: (0.892, 1.036)), respectively. The 95% prediction intervals were (0.643, 1.292) and (0.702, 1.225), with I^2 equal to 21.7% and 14.3% respectively. Thus, there was lower heterogeneity in the hospital-specific calibration slopes (Figure 3.3) than in the hospital-specific c-statistics. Thus, there was no clear evidence of over-fitting or different overall predictor effects when applying the prediction model to patients at different hospital within the same temporal period.

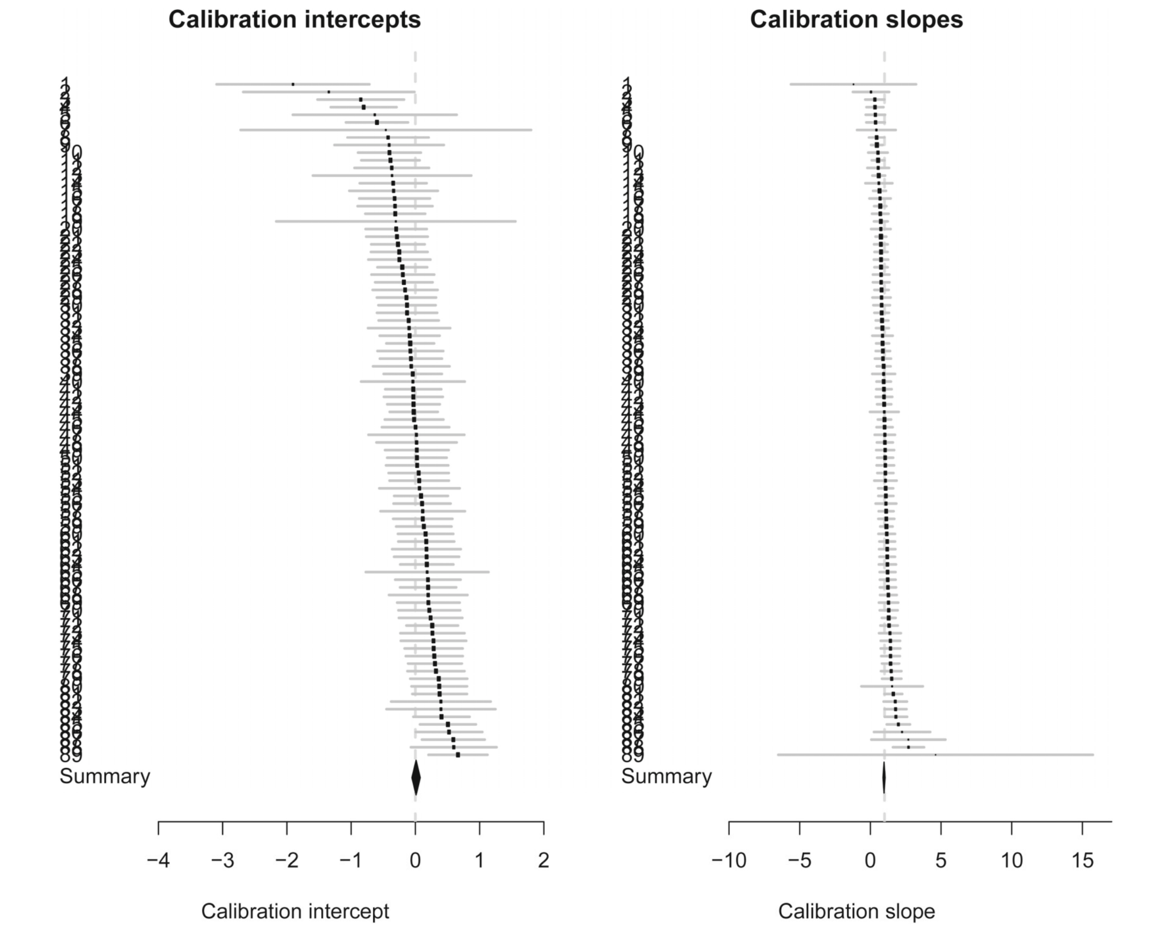


Figure 3.3 Meta-analyses of calibration intercepts and slopes using a leave-one-hospital-out approach.

Temporal transportability

When the EFFECT-HF model was estimated in the first phase and then applied to the second phase, the estimated c-statistic in the second phase was 0.745. When the model was modified to incorporate hospital-specific random effects the variance of the random intercepts was 0.02635 (resulting in a residual intraclass correlation coefficient of 0.008 [22]), and the resultant c-statistic remained unchanged at 0.745 (hospital-specific random effects were incorporated into the linear predictor when making predictions).

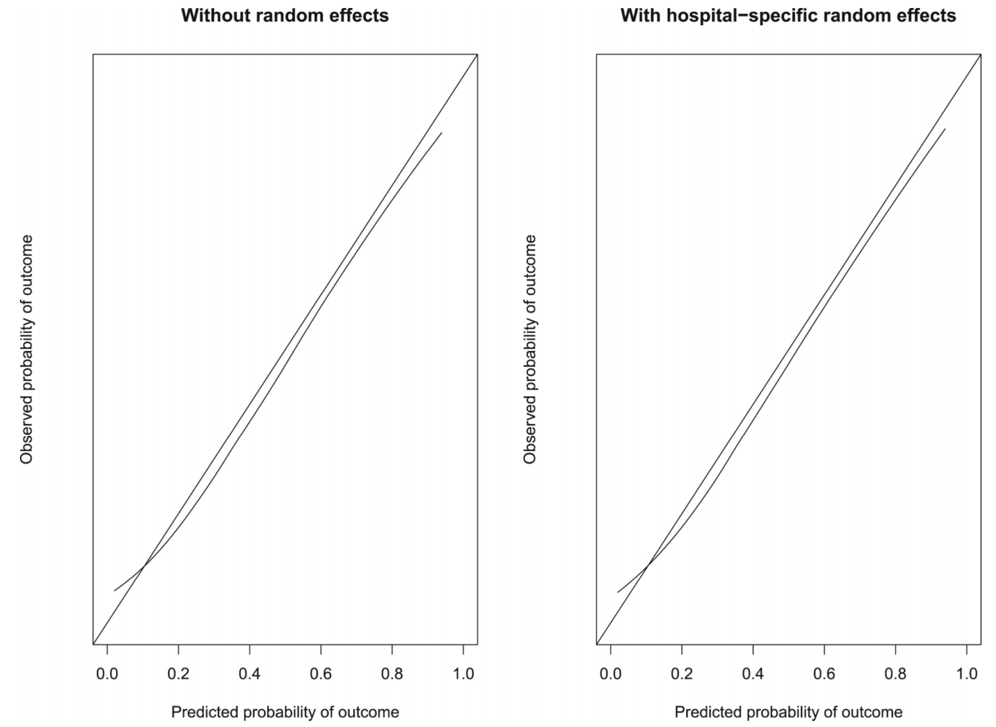


Figure 3.4 Temporal calibration in phase 2 sample with and without random effects.

Calibration for Phase 2 showed a slope close to 1 (0.984, 95% CI: (0.923, 1.046)), and an intercept of -0.121 (95% CI: (-0.175, -0.067)). Results were very similar with random effects (0.979 and -0.115, respectively). So, the probability of mortality was slightly lower in Phase 2 (calibration intercept < 0). Overall calibration plots are described in Figure 3.4.

The c-statistic of the model for predicting study phase was 0.580, suggesting similarity in case-mix between Phase 1 and Phase 2. The means of the linear predictors and the standard deviations of the linear predictors were also very similar. Indeed, the case-mix corrected c-statistic of the model developed in Phase 1 when applied to Phase 2 was 0.746. This differed negligibly from the c-statistic of 0.745 that was obtained when the EFFECT-HF model was developed in the Phase 1 sample and applied to the Phase 2 sample.

Simultaneous geographic and temporal transportability

When using the ‘leave-one-hospital-out (pooled)’ approach, the estimated c-statistic was 0.745. When using the ‘leave-one-hospital-out (meta-analysis)’ approach, the mean hospital-specific c-statistic from the random effects meta-analysis was 0.753, while the estimate of τ was 0.028. The value of the I^2 statistic was 20.6%, with 95% prediction interval (0.693, 0.812).

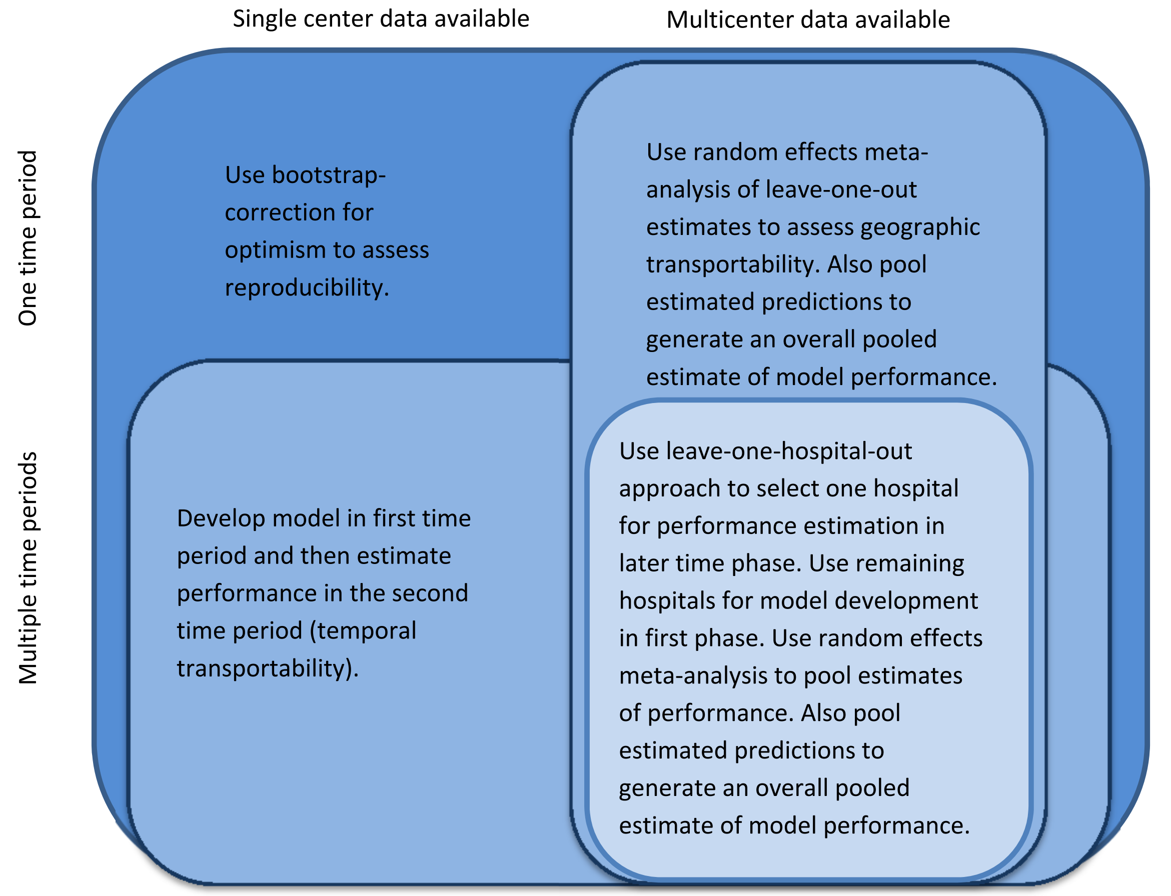
DISCUSSION

We illustrated different strategies for assessing the geographic and temporal performance of a clinical prediction model for mortality in patients with heart failure. We started with conventional strategies such as bootstrapping and leave-one-hospital-out. When using leave-one-hospital-out approaches, we considered a pooled approach in which predicted probabilities were pooled, as well as novel approaches based on random effects pooling of hospital-specific estimates of model performance. All strategies showed similar overall performance, but small to moderate variation in performance by hospital (Table 3.2). In Figure 3.5 we summarize graphically some recommendations for assessing geographic and temporal portability of clinical prediction models based on our reported analyses.

Bootstrap-based methods for optimism correction allow one to assess model reproducibility: how well the model will perform in different patients from the same population in which the model was developed [14, 15]. Frequently, researchers do not have access to subjects from other centres or different time periods with which to externally validate the derived model. Thus, at the first stage of model development and validation, the estimate of model reproducibility often serves as the best *initial* estimate of how well the model will perform in subsequent subjects and in subjects from different centres and regions [3]. The apparent performance was very similar to the bootstrap-corrected for optimism estimate of performance, which is explained by the large sample size available in each of the two phases in the current study. More optimism is to be expected when smaller sample sizes are used for model derivation [14].

A leave-one-hospital-out approach was very useful to examine geographic transportability. The pooled estimates of the model c-statistic were very similar to those obtained using bootstrap-correction for model optimism. This finding may be unsurprising, as both approaches can be seen as different forms of internal validation, with the former being a form of cross-validation. We note that as the number of centres that are included in model development increases, the pooled performance of the model in a different set of centres will likely be comparable to the performance of the model in the full derivation sample. Geographical transportability is more likely to be poor when the model was developed at a single centre than when it was developed using subjects from a large number of centers.

We emphasize that developing a model in a large set of centres does not guarantee that there will be negligible variation in the hospital-specific performance of the model when applied to a new set of centres. This variation can be studied using random effects meta-analytic methods [23]. Such a meta-analytic approach produces an estimate of the pooled hospital-specific c-statistic but also of the variance of the hospital-specific c-statistics. One could argue that geographic transportability is primarily indicated by this variation of performance across the centres, as this denotes the degree to which model performance can be expected to vary across centres (heterogeneity). We found that there was more between-centre heterogeneity in performance in Phase 1 than in Phase 2, and more in c-statistics than in calibration slopes. The latter may reflect that the c-statistic depends both on case-mix differences and differences in model fit to specific centers [4, 20, 21].



Use bootstrap-correction for optimism to assess reproducibility.

Use random effects meta-analysis of leave-one-out estimates to assess geographic transportability. Also pool estimated predictions to generate an overall pooled estimate of model performance.

Develop model in first time period and then estimate performance in the second time period (temporal transportability).

Use leave-one-hospital-out approach to select one hospital for performance estimation in later time phase. Use remaining hospitals for model development in first phase. Use random effects meta-analysis to pool estimates of performance. Also pool estimated predictions to generate an overall pooled estimate of model performance.

Figure 3.5 Recommendations for validating clinical prediction models.

When we simultaneously examined temporal and geographic transportability, the overall c-statistic was identical to the assessment of the temporal transportability. Similarly, this estimate was equal to that obtained in each of the two phases of the study when using a leave-one-hospital-out approach, as described above.

When comparing methods for assessing the temporal transportability of the prediction model, identical estimates of the overall c-statistic were obtained regardless of whether or not one included hospital-specific random effects in the clinical prediction model (with a residual intraclass correlation coefficient of 0.008, the between-centre variation in mortality was low). The ability to omit hospital-specific random effects is advantageous, since these will be of use only when the model is applied to patients admitted to the same hospitals as those in which the model was developed.

In the current study one might argue that we did not conduct a true assessment of external validation. Many of the analyses that we described would constitute “internal-external validation”, while our assessment of model reproducibility would constitute internal validation [24]. The highest standard for external validation would entail validating the derived model in patients from a different temporal period, from a different geographic period and by different investigators from those who developed the original model. Some of our analyses fulfilled the first two criteria. However, the final criterion was not satisfied, as the same study investigators were responsible for the study design and data collection in both phases of the study. The strength of arguments for geographic and temporal transportability in our setting would depend on the differences between the hospitals selected for model derivation and those selected for model validation and the temporal difference between the two time periods.

In the current study, we only considered the inclusion of patient-level characteristics in the clinical prediction model. This reflects the typical development of clinical prediction models, in which hospital or system characteristics are excluded from the model. It is possible that inclusion of hospital characteristics (e.g., hospital volume of the condition in question, academic affiliation, staff training, etc.) can improve the performance of the model. Furthermore, the inclusion of such characteristics may result in models with improved geographic transportability, if the distribution of hospital characteristics differs between the centers that were used for model development and the centers in which the model will ultimately be applied (the variance of the random effects can give some indication of the potential for subsequent improvements). However, the inclusion of such characteristics could result in an unwarranted extrapolation if the hospitals to which the model was applied differed substantially from those used for derivation (i.e., if the model was developed at low-volume centers and then applied at high-volume centers).

In summary, we illustrated the application of a set of analytic methods for assessing the reproducibility, geographic transportability and temporal transportability of clinical prediction models. We focused here on the traditional concept of validity, i.e. assessing performance, specifically calibration and discrimination, in subjects not considered at model development. An alternative perspective is to evaluate geographic and temporal effects within the full data set [24]. We expand on this perspective in a companion article [25]. Understanding the purpose of each validation approach, its strengths and limitations, as well as its interpretation, will permit investigators to better assess the performance of clinical prediction models as well as to assess the quality of validations presented in the literature.

ACKNOWLEDGEMENTS

This study was supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). Dr. Austin is supported in part by a Career Investigator award from the Heart and Stroke Foundation. Dr. Lee is supported by a Clinician-Scientist award from the CIHR. Dr. Steyerberg and Mr. Van Klaveren are supported in part by a U award (U01NS086294, value of personalized risk information). Mr. van Klaveren and Dr. Vergouwe are supported in part by the Netherlands Organisation for Scientific Research (grant 917.11.383). The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) data used in the study was funded by a CIHR Team Grant in Cardiovascular Outcomes Research. These datasets were linked using unique, encoded identifiers and analyzed at the Institute for Clinical Evaluative Sciences (ICES).

SUPPLEMENTARY DATA

An online appendix can be found at http://dx.doi.org/10.1016/j.jclinepi.2016.05.007.

REFERENCES

1. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009; 338:b605
2. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine* 2000; 19(4):453-473
3. Justice AC, Covinsky KE, Berlin JA. Assessing the Generalizability of Prognostic Information. *Annals of Internal Medicine* 1999; 130(6):515-524
4. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology* 2015; 68(3):279-289. DOI: 10.1016/j.jclinepi.2014.06.018
5. Tu JV, Donovan LR, Lee DS, Wang JT, Austin PC, Alter DA, Ko DT. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. *Journal of the American Medical Association* 2009; 302(21):2330-2337
6. Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu JV. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. *Journal of the American Medical Association* 2003; 290(19):2581-2587
7. Steyerberg EW. *Clinical Prediction Models*. Springer-Verlag: New York, 2009.
8. Harrell FE, Jr. *Regression modeling strategies*. Springer-Verlag: New York, NY, 2001.
9. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statisics in Medicine* 2014; 33(3):517-535. DOI: 10.1002/sim.5941
10. Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958; 45(3-4):592-565. DOI: 10.1093/biomet/45.3-4.562
11. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of Probabilistic Predictions. *Medical Decision Making* 1993; 13(1):49-57. DOI: 10.1177/0272989X9301300107
12. Picard RR, Berk KN. Data Splitting. *The American Statistician* 1990; 44(2):140-147. DOI: http://dx.doi.org/10.1080/00031305.1990.10475704
13. Airola A, Pahikkala T, Waegeman W, De Baets B, Salakoski T. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Computational Statistics & Data Analysis* 2011; 55(4):1828-1844. DOI: 10.1016/j.csda.2010.11.018
14. Austin, P. C. and Steyerberg, E. W. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Statistical Methods in Medical Research . 2014.
15. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology* 2001; 54(8):774-781
16. Smith GC, Seaman SR, Wood AM, Royson P, White IR. Correcting for optimistic prediction in small data sets. *American Journal of Epidemiology* 2014; 180(3):318-324. DOI: 10.1093/aje/kwu140
17. Royston P, Parmar MK, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Stat.Med.* 2004; 23(6):907-926. DOI: 10.1002/sim.1691
18. van Klaveren D., Steyerberg EW, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. *BMC Medical Research Methodology* 2014; 14:5. DOI: 10.1186/1471-2288-14-5
19. Chen DG, Peace KE. *Applied Meta-Analysis with R*. CRC Press: Boca Raton, FL, 2013.
20. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC: Medical Research Methodology* 2012; 12:82. DOI: 10.1186/1471-2288-12-82

21. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *American Journal of Epidemiology* 2010; 172(8):971-980
22. Snijders T, Bosker R. *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*. Sage Publications: London, 1999.
23. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011; 342:d549
24. Steyerberg, E. W. and Harrell, F. E., Jr. Prediction models need appropriate internal, internal-external, and external validation. Journal of Clinical Epidemiology . 4-18-2015.
25. Austin, P. C., van Klaveren D., Vergouwe, Y., Nieboer, D., Lee, D. S., and Steyerberg, E. W. Geographic and temporal validity of prediction models: Different approaches were useful to examine heterogeneity. 2016; Unpublished Work

4

Interpretation of concordance measures for clustered data

D van Klaveren
EW Steyerberg
Y Vergouwe

Statistics in Medicine 2014

Mauguen et al [1] extended two censoring-robust estimators of the concordance probability to frailty models. The estimators were proposed by Uno et al [2] and by Gönen and Heller [3] ('Uno' and 'GH' estimators respectively). The authors followed the suggestion of Van Oirbeek and Lesaffre [4] to derive separate concordance probability estimates for patients within the same cluster and for patients in different clusters and to pool them into an overall estimate. Although the proposed techniques add to the assessment of prognostic model performance in clustered survival data, we would like to discuss three issues related to their interpretation and practical use.

First, the model-based GH estimator does not use observed survival times directly in contrast to Harrell's c-index [5] and the Uno estimator. Instead, the effect of observed survival times is mediated through the regression coefficients. As a consequence the concordance probability in a new population is estimated under the assumption that the regression coefficients are correct. The GH estimator should therefore be interpreted with care when applied to new populations. The authors applied the GH estimator in clusters of a validation population, using the regression coefficients of the development population. The resulting GH estimates are similar to benchmark estimates suggested before [6] and differ only from the concordance probability estimates in the development population due to differences in patient heterogeneity ('case-mix'). We undertook a small simulation study with different external validation settings to illustrate the interpretation of the GH estimator (Table 4.1). When both case-mix distribution (X) and coefficient (β) were equal to the development population (validation 1), the concordance probability estimates gave similar results as in the development setting, apart from small differences due to sensitivity for censoring. When we lowered case-mix heterogeneity (validation 2), all concordance measures decreased similarly. When we lowered the coefficient (validation 3) the GH and the Benchmark estimates remained almost the same while the c-index and the Uno estimate decreased further, empirically supporting the above reasoning.

Second, the authors recommended using cluster-specific (conditional) predictions for validation of a prognostic model. They suggested using the validation data to estimate frailties for new clusters. However, using validation data to derive predictions does not correspond to a direct external validation of a prognostic model's performance in new settings. It might better be labelled a form of internal validation [7]. We recommend to use population (marginal) predictions for external validation and to limit the use of cluster-specific predictions to temporal validation, with frailties estimated on development data and validated on more recent data from the same clusters.

Third, the authors did not give guidance when to use within-cluster, between-cluster or overall concordance measures. We propose using the within-cluster concordance probability in clinical practice as decisions on interventions are commonly taken within centers (clusters).

Table 4.1 Simulation study results. Means (empirical standard errors) of concordance probability estimates in a development setting and three validation settings.

	Development	Validation 1	Validation 2	Validation 3
	X ~ Unif [0, 1]	X ~ Unif [0, 1]	X ~ Unif [0.1, 0.9]	X ~ Unif [0.1, 0.9]
	β = 3	β = 3	β = 3	β = 2
Censoring (%)	**0.0 (0.0)**	**0.0 (0.0)**	**0.0 (0.0)**	**0.0 (0.0)**
$\hat{\beta}$	3.015 (0.221)			
c-index	0.710 (0.013)	0.710 (0.013)	0.677 (0.014)	0.625 (0.015)
Uno	0.710 (0.013)	0.710 (0.013)	0.677 (0.014)	0.625 (0.015)
GH	0.710 (0.012)	0.710 (0.011)	0.678 (0.011)	0.678 (0.011)
Benchmark		0.711 (0.017)	0.678 (0.018)	0.678 (0.018)
Censoring (%)	**50.4 (2.4)**	**50.7 (2.6)**	**50.6 (2.5)**	**64.2 (2.5)**
$\hat{\beta}$	3.026 (0.295)			
c-index	0.721 (0.019)	0.720 (0.018)	0.683 (0.020)	0.628 (0.025)
Uno	0.717 (0.018)	0.716 (0.017)	0.681 (0.019)	0.629 (0.025)
GH	0.710 (0.015)	0.711 (0.015)	0.678 (0.014)	0.678 (0.014)
Benchmark		0.717 (0.021)	0.682 (0.021)	0.683 (0.021)

For each setting 1000 replications of 400 patient profiles X were drawn from a uniform distribution (column heading). Survival times were generated by multiplication of exp($\mathrm{X}\beta$) with independent draws from the exponential distribution (β in column headings). Right-censoring times were drawn from a uniform distribution with support [0, c] where c was chosen to target 0% and 50% censoring. Concordance probability estimates were based on predictions $\mathrm{X}\hat{\beta}$ with $\hat{\beta}$ estimated in the development data. The time-dependent Uno estimator was calculated at τ = 0.9c. To obtain the Benchmark estimate we calculated the predicted survival function for each patient in X based on the model fit in the development data. The predicted survival functions were used to sample 400 survival times. The Benchmark estimate was then calculated as the c-index in this new sample.

A valuable prognostic model should be able to separate patients within the same center into those with good outcome and poor outcome. In contrast, we consider the overall concordance measure appropriate when decisions are taken at the population level, where between-center heterogeneity can be used to guide decision making. Following the same line of reasoning when patient data is clustered in clinical trials, we recommend using the within-cluster concordance probability. Our rational is that between-trial heterogeneity is not exploitable in clinical practice.

We dispute the authors' conclusion in a head and neck cancer case study that external validation in a USA population confirmed the performance of a prognostic model developed in a European population. Regardless of the GH overall concordance probability estimate based on cluster-specific predictions (0.625) we consider the Uno within-cluster probability estimates the most appropriate indicators of discriminative ability of the proposed prognostic model. These estimates were significantly lower in the USA validation population (mean 0.488) than in the European development population (mean 0.615). Furthermore, these estimates were similar for the frailty model and the Cox model, but varied substantially across clusters, both in the European and in the USA population. The difference between the GH estimates of the within-cluster concordance (0.570) and the between-cluster concordance (0.612) reflected substantially stronger heterogeneity between patients from different clusters than between patients within the same cluster.

In conclusion, for external validation in clinical practice we recommend using non-parametric within-cluster concordance probability estimates (c-index or Uno), without using cluster-specific (conditional) predictions. The use of GH estimates is valuable for benchmark purposes. Between-cluster concordance probability estimates may be useful when between-cluster heterogeneity in case-mix is exploitable for guidance of decision making.

ACKNOWLEDGMENTS

This work was supported by the Netherlands Organisation for Scientific Research (grant 917.11.383).

REFERENCES

1. Mauguen A, Collette S, Pignon JP, Rondeau V. Concordance measures in shared frailty models: application to clustered data in cancer prognosis. *Stat Med* 2013. DOI: 10.1002/sim.5852.
2. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011; 30: 1105-1117.
3. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005; 92: 965-970.
4. Van Oirbeek R, Lesaffre E. An application of Harrell's C-index to PH frailty models. *Stat Med* 2010; 29: 3160-3171.
5. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982; 247: 2543-2546.
6. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010; 172: 971-980.
7. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000; 19: 453-473.

5

A new concordance measure for risk prediction models in external validation settings

D van Klaveren
M Gönen
EW Steyerberg
Y Vergouwe

Statistics in Medicine 2016

ABSTRACT

Concordance measures are frequently used for assessing the discriminative ability of risk prediction models. The interpretation of estimated concordance at external validation is difficult if the case-mix differs from the model development setting. We aimed to develop a concordance measure that provides insight into the influence of case-mix heterogeneity and is robust to censoring of time-to-event data.

We first derived a model-based concordance measure (*mbc*) that allows for quantification of the influence of case-mix heterogeneity on discriminative ability of proportional hazards and logistic regression models. This *mbc* can also be calculated including a regression slope that calibrates the predictions at external validation (*c-mbc*), hence assessing the influence of overall regression coefficient validity on discriminative ability. We derived variance formulas for both *mbc* and *c-mbc*. We compared the *mbc* and the *c-mbc* with commonly used concordance measures in a simulation study and in two external validation settings.

The *mbc* was asymptotically equivalent to a previously proposed resampling-based case-mix corrected c-index. The *c-mbc* remained stable at the true value with increasing proportions of censoring, while Harrell's c-index and to a lesser extent Uno's concordance measure increased unfavorably. Variance estimates of *mbc* and *c-mbc* were well in agreement with the simulated empirical variances.

We conclude that the *mbc* is an attractive closed-form measure that allows for a straightforward quantification of the expected change in a model's discriminative ability due to case-mix heterogeneity. The *c-mbc* also reflects regression coefficient validity, and is a censoring-robust alternative for the c-index when the proportional hazards assumption holds.

INTRODUCTION

Assessing the performance of a clinical prediction model is of great practical importance to learn about the potential clinical value. An essential aspect of model performance is separating subjects with good outcome from subjects with poor outcome (discrimination) [1]. Concordance measures, also called concordance-statistics or c-statistics, are commonly used to assess the discriminative ability of risk prediction models. A c-statistic estimates for two randomly chosen subjects the probability that the model predicts a higher risk for the subject with poorer outcome (concordance probability) [2, 3]. The observed c-statistic of a risk prediction model in external validation data depends on the validity of the regression coefficients, but also on the heterogeneity of the case-mix [4-6]. Case-mix heterogeneity refers to the variation in subject characteristics and can readily be quantified by the standard deviation of the linear predictor [5].

Harrell's concordance-index (c-index) is the most frequently used c-statistic for binary and for time-to-event outcomes, but is sensitive to censoring of time-to-event outcomes [7, 8]. An inverse probability weighting technique was proposed by Uno et al. to offset the dependence of the c-index on censoring [8]. For validation of proportional hazards regression models within model development data, Gönen and Heller proposed a censoring-robust concordance measure [7]. This model-based concordance measure, which was also suggested by Korn and Simon as a measure of explained variation [9], is a function of the regression coefficients and the covariate distribution and does not use observed event and censoring times. Consequently, in an external validation population it merely assesses the expected discriminative ability of the model, similar to a previously proposed case-mix corrected c-index [10]. This case-mix corrected c-index – based on resampling outcomes under the assumption of correct regression coefficients – was suggested to disentangle the effect of a different case-mix from incorrect regression coefficients on discrimination [4]. Such disentangling is relevant to the interpretation of a difference between the c-statistic at model development versus the observed c-statistic at external validation. We hereto calculate the difference between the c-statistic at model development and the case-mix corrected c-statistic at external validation to indicate the change in discriminative ability attributable to the difference in case-mix heterogeneity. Next, the difference between the observed c-statistic and the case-mix corrected c-statistic in external validation data expresses the change in discriminative ability due to the (in)correctness of the regression coefficients.

We aimed to develop a model-based concordance measure (mbc) to assess the discriminative ability of risk prediction models in external data. Since the most commonly used concordance measures all have their restrictions (Table 5.1), the new measure should be a valuable addition for: (1) assessment of the influence of case-mix heterogeneity on concordance of both logistic regression and proportional hazards regression models; and (2) censoring-robust measurement of a proportional hazards regression model's concordance in external validation data. We studied the behavior of the newly developed concordance measure in external validation settings with simulation and two case studies.

THE MODEL-BASED CONCORDANCE

Notation

We will assess the discriminative ability of previously developed logistic regression models and proportional hazards regression models in new patient populations. Both regression models predict patient outcome Y based on a linear predictor $x^T\beta$, which is a linear combination of the patient's baseline characteristics vector x and regression coefficient vector β. The random outcome variable Y_i and its realization y_i for patient i of n patients in the validation population takes values of 0 or 1 in case of a logistic regression model, and positive time-to-event values in case of a proportional hazards regression model. For a time-to-event realization of patient i we use the indicator δ_i to denote an observed event time $(\delta_i = 1)$ or a censored event time $(\delta_i = 0)$. When the i^{th} row of a population's design matrix X is the baseline characteristics vector x_i, the linear predictor $x_i^T\beta$ of patient i is the i^{th} element of the vector Xβ. Note that an additional first element of x_i is set to 1 for multiplication with a logistic regression model's intercept. A linear predictor $x_i^T\beta$ of a logistic regression model is transformed by the logistic function to obtain prediction p_i. A linear predictor of a proportional hazards regression models is transformed into a prediction p_i by the survival function $S(t|x_i^T\beta) = exp\left\{-\int_0^t \lambda_0(s)e^{x_i^T\beta}ds\right\}$, with $\lambda_0(t)$ the baseline hazard function of the proportional hazards regression model's. Although the baseline hazard function is necessary to obtain absolute risk predictions, we will not need it in the remainder of this paper.

Derivation of the model-based concordance

The concordance probability is defined as the probability that a model predicts for two randomly chosen patients a higher risk for the patient with poorer outcome.

For a given patient population it is the probability that a randomly selected patient pair has concordant predictions and outcomes, divided by the probability that their outcomes are different (not "tied"). The probability that a randomly selected patient pair has concordant predictions and outcomes is [9]:

$$P(\text{concordant}) = \frac{1}{n(n-1)}\sum_i \sum_{j\neq i}\left[I(p_i < p_j)P(Y_i < Y_j) + I(p_i > p_j)P(Y_i > Y_j)\right] \quad (1)$$

Similarly, the probability that a randomly selected patient pair has unequal outcomes is:

$$P(\text{unequal } Y) = \frac{1}{n(n-1)} \sum_{i} \sum_{j \neq i} \left[P(Y_i < Y_j) + P(Y_i > Y_j) \right] \qquad (2)$$

Thus, the concordance probability CP in a patient population is obtained by dividing the probabilities of equation 1 and 2:

$$CP = \frac{\sum_i \sum_{j \neq i} \left[I(p_i < p_j) P(Y_i < Y_j) + I(p_i > p_j) P(Y_i > Y_j) \right]}{\sum_i \sum_{j \neq i} \left[P(Y_i < Y_j) + P(Y_i > Y_j) \right]} \qquad (3)$$

With $P(Y_i < Y_j) = I(y_i < y_j)$ equation 3 returns Harrell's c-index, but to obtain a model-based estimator we derive $P(Y_i < Y_j)$ from a regression model. For a logistic regression model the model-based probabilities $P(Y_i < Y_j)$ are:

$$P(Y_i < Y_j) = P(Y_i = 0) P(Y_j = 1) = \frac{1}{1 + \mathrm{e}^{x_i^T \beta}} \frac{1}{1 + \mathrm{e}^{-x_j^T \beta}} \qquad (4)$$

Combining equations 3 and 4, and replacing $I(p_i < p_j)$ by $I(x_i^T \beta < x_j^T \beta)$ because the predictions are an increasing function of the linear predictor, results in the model-based concordance (mbc) for logistic regression models:

$$mbc(X\beta) = \frac{\sum_i \sum_{j \neq i} \left[\frac{I(x_i^T \beta < x_j^T \beta)}{\left(1 + \mathrm{e}^{x_i^T \beta}\right)\left(1 + \mathrm{e}^{-x_j^T \beta}\right)} + \frac{I(x_i^T \beta > x_j^T \beta)}{\left(1 + \mathrm{e}^{-x_i^T \beta}\right)\left(1 + \mathrm{e}^{x_j^T \beta}\right)} \right]}{\sum_i \sum_{j \neq i} \left[\frac{1}{\left(1 + \mathrm{e}^{x_i^T \beta}\right)\left(1 + \mathrm{e}^{-x_j^T \beta}\right)} + \frac{1}{\left(1 + \mathrm{e}^{-x_i^T \beta}\right)\left(1 + \mathrm{e}^{x_j^T \beta}\right)} \right]} \qquad (5)$$

For a proportional hazards regression model the model-based probabilities $P(Y_i < Y_j)$ are [7]:

$$P(Y_i < Y_j) = -\int_0^\infty S(t | x_j^T \beta)\, dS(t | x_i^T \beta) = \frac{1}{1 + \mathrm{e}^{(x_j - x_i)^T \beta}} \qquad (6)$$

Combining equations 3 and 6, and replacing $I(p_i < p_j)$ by $I(x_i^T\beta > x_j^T\beta)$ because the time-to-event predictions are a decreasing function of the linear predictor, results in the model-based concordance (mbc) for proportional hazards regression models:

$$mbc(X\beta) = \frac{\sum_i \sum_{j\neq i}\left[\frac{I(x_i^T\beta > x_j^T\beta)}{1+\mathrm{e}^{(x_j-x_i)^T\beta}} + \frac{I(x_i^T\beta < x_j^T\beta)}{1+\mathrm{e}^{(x_i-x_j)^T\beta}}\right]}{\sum_i \sum_{j\neq i}\left[\frac{1}{1+\mathrm{e}^{(x_j-x_i)^T\beta}} + \frac{1}{1+\mathrm{e}^{(x_i-x_j)^T\beta}}\right]} \qquad (7)$$

The denominator of equation 7 is equal to $n(n-1)$ since $\left[\frac{1}{1+\mathrm{e}^{(x_j-x_i)^T\beta}} + \frac{1}{1+\mathrm{e}^{(x_i-x_j)^T\beta}}\right] = 1.$

Equation 3 assumes that model predictions p_i and p_j are different for every combination of *i* and *j*. Since model predictions may be equal for some combinations of *i* and *j*, e.g. when x is a binary marker, we can generalize equation 3, and similarly equations 5 and 7, by using $I(p_i \leq p_j)$ instead of $I(p_i < p_j)$. Hence equation 3 can also be written in the familiar c-statistic format:

$$CP = \frac{\sum_i \sum_{j\neq i}\left[I(p_i < p_j)P(Y_i < Y_j) + \frac{1}{2}I(p_i = p_j)P(Y_i < Y_j)\right]}{\sum_i \sum_{j\neq i} P(Y_i < Y_j)} \qquad (3')$$

In an apparent validation of a model with regression coefficient estimates $\hat{\beta}$ the $mbc(X\hat{\beta})$ gives an estimate of the concordance probability. For proportional hazards regression models the $mbc(X\hat{\beta})$ is identical to the censoring-robust estimator proposed before by Gönen and Heller [7]. Gönen and Heller derived their model-based concordance measure from a reversed definition of the concordance probability, i.e. $\frac{\sum_i \sum_{j\neq i}[I(p_i\leq p_j)P(Y_i<Y_j)+I(p_i\geq p_j)P(Y_i>Y_j)]}{\sum_i \sum_{j\neq i}[I(p_i\leq p_j)+I(p_i\geq p_j)]}$, conditioning on weakly ordered predictions. However, for fully continuous predictions and outcomes the two definitions of the concordance probability are equivalent since $p_1 \leq p_2$ implies $p_1 < p_2$ and the summands in the denominator, $[P(Y_i < Y_j) + P(Y_i > Y_j)]$ and $[I(p_i \leq p_j) + I(p_i \geq p_j)]$, both equal 1 [11].

For proportional hazards regression models based on uncensored, continuous time-to-event outcomes, the $mbc(X\hat{\beta})$ is asymptotically equivalent to Harrell's c-index when the proportional hazards assumption holds (Appendix 5.2). The same asymptotic equivalence holds for logistic regression models, with exact equality when the model contains only one categorical predictor (Appendix 5.2). In an external validation setting of a model with regression coefficients β the $mbc(X\beta)$ can be used as a benchmark value to assess the influence of case-mix heterogeneity on the concordance probability – comparable to the case-mix corrected c-index as proposed before [4] – since it assumes correct regression coefficients [10]. Appendix 5.1 contains the derivation of variance estimates of the mbc in model development and external validation settings.

Table 5.1 Use of *mbc*, *c-mbc* and commonly used concordance measures. For proportional hazards regression models and logistic regression models, it is specified: how to measure concordance in an apparent validation setting (in patients whose data was used for model development); how to assess the influence of case-mix heterogeneity on concordance (Concordance assuming correct regression coefficients) in an external validation setting (in new patients); and how to measure concordance in an apparent validation setting.

Apparent validation	**External validation**	
Concordance	**Concordance assuming correct regression coefficients**	**Concordance**
Proportional hazards regression models		
c-index	case-mix corrected c-index	c-index
Uno		Uno
Gönen-Heller	Gönen-Heller	
mbc	*mbc*	*c-mbc*
Logistic regression models		
c-index	case-mix corrected c-index	c-index
mbc	*mbc*	*c-mbc*

Including the calibration slope in the *mbc*

In an external validation setting the $mbc(X\beta)$ does not use observed outcomes and is therefore not influenced by the validity of the regression coefficients in the validation data. Refitting the regression model to the validation data does not give insight into the discriminative ability of the previously developed model. To assess the influence of overall regression coefficient validity on concordance, we first estimate the calibration slope γ_1 in the validation data, i.e. the regression coefficient of a model that regresses the observed outcomes y on the linear predictors $X\beta$ in the validation data [12]. If $\hat{\gamma}_1 = 1$, the regression coefficients are on average correct in the validation data. In contrast, $\hat{\gamma}_1 < 1$ indicates a weaker association between the linear predictor and the outcomes in the validation data. For logistic regression models, an intercept estimate $\hat{\gamma}_0$ is required for estimation of the calibration slope $\hat{\gamma}_1$. With $\hat{\gamma}X\beta$ we will denote $\hat{\gamma}_0 + \hat{\gamma}_1 X\beta$ for logistic regression models and $\hat{\gamma}_1 X\beta$ for proportional hazards regression models. The $mbc(\hat{\gamma}X\beta)$, which we label calibrated model-based concordance (*c-mbc*), assesses both the influence of case-mix heterogeneity and the overall validity of the regression coefficients β. Similar to the original Gönen and Heller estimator, the *c-mbc* does not directly depend on observed survival and censoring times. Instead, it is only based on the regression coefficients $\hat{\gamma}$ and the distribution of the linear predictor $X\beta$. Since the effect of censoring on the bias of $\hat{\gamma}$ is negligible, $mbc(\hat{\gamma}X\beta)$ is expected to be insensitive to censoring as well. Table 5.1 gives an overview of the potential use of the *mbc* in relation to existing concordance measures.

CASE STUDIES

Unfavorable outcome after traumatic brain injury

To illustrate the use of the *mbc* and the *c-mbc* for logistic regression models, we revisit a case study on the prediction of 6-month outcome in patients with traumatic brain injury [4]. A model to predict unfavorable outcome (i.e., death, a vegetative state, or severe disability) was developed with data on 1,118 subjects (456 (41%) had an unfavorable outcome) from the International Tirilazad Trial [13]. The validity of the risk prediction model was studied in 1,041 subjects (395 (38%) had an unfavorable outcome) who were enrolled in the North American Tirilazad Trial [14]. The logistic regression model consisted of three predictors (age, motor score, and pupillary reactivity) for an unfavorable 6-month outcome [15].

The model showed reasonable discrimination in the development sample (c-index 0.749; *mbc* 0.749; Table 5.2). The larger variability of the linear predictor in the external validation sample than in the development sample ($\mathrm{SD}(X\beta)$ = 1.11 and $\mathrm{SD}(X\beta)$ = 1.03, respectively) substantially increased the expected discriminative ability (*mbc* = 0.767; 95% CI 0.759 - 0.775; Table 5.2). Including the validity of the regression coefficients (calibration slope 1.02) indicated a small additional increase in discriminative ability (c-index 0.779; *c-mbc* 0.774; Table 5.2).

Table 5.2 Case study of concordance measures (95% confidence interval) in logistic regression and Cox regression.

	Logistic regression		Cox regression	
	Apparent validation	External validation	Apparent validation	External validation
SD($X\beta$)	1.028	1.112	0.904	0.965
Calibration slope $\hat{\gamma}$	1.000	1.023	1.000	0.785
Harrell's c-index	0.749 (0.719 - 0.778)	0.779 (0.750 - 0.808)	0.744 (0.707 - 0.781)	0.725 (0.700 - 0.750)
Uno (τ = 4)			0.743 0.705 - 0.782)	0.729 (0.687 - 0.771)
$mbc = mbc(X\beta)$	0.749 (0.721 - 0.778)	0.767 (0.759 - 0.775)†	0.707 (0.680 - 0.734)	0.719 (0.715 - 0.722)†
$c\text{-}mbc = mbc(\hat{\gamma}X\beta)$		0.774 (0.746 - 0.803)		0.684 (0.667 - 0.700)

The logistic regression model for unfavorable outcome after traumatic brain injury was developed in the International Tirilazad Trial and validated in the North American Tirilazad Trial. The Cox regression model for survival after revascularization was developed in the SYNTAX trial and validated in the CREDO-KYOTO registry. † 95% confidence interval based on the variance estimate of $mbc(X\beta)$ under the assumption of true β.

Survival after coronary revascularization

To illustrate the use of the *mbc* and the *c-mbc* for proportional hazards regression models, we apply them to a recent validation study of the SYNTAX Score II (SSII) [16]. The SSII has been developed by applying a Cox proportional hazards model to the data of the SYNTAX trial [17, 18]. The SSII uses 2 anatomical variables (anatomical Syntax Score and unprotected left main coronary artery disease) and 6 clinical variables (age, creatinine clearance, left ventricular ejection fraction, sex, chronic obstructive pulmonary disease, and peripheral vascular disease) to predict 4-year mortality after revascularization with CABG or PCI. For validation of SSII we use 3,986 patients of the Coronary REvascularization Demonstrating Outcome Study in Kyoto (CREDO-Kyoto) PCI/CABG registry cohort-2 [19].

There was a substantial difference in the development data between the *mbc* (0.707; Table 5.2) and both the c-index (0.744) and Uno's concordance measure (0.743), probably due to a high proportion of censoring (90.1%). Under the assumption that the proportional hazards assumption holds until the follow-up is complete– the proportional hazards assumption of the Cox regression model was not rejected up till 4 years of follow-up (p=0.63) [20] – we may conclude from the simulations study that the *mbc* gives a better estimate of the concordance probability in this example. The larger variability of the linear predictor in the external validation sample than in the development sample ($\mathrm{SD}(X\beta)$ = 0.97 and $\mathrm{SD}(X\beta)$ = 0.90, respectively) increased the expected discriminative ability (*mbc* = 0.719; 95% CI 0.715 - 0.722; Table 5.2). However, including the validity of the regression coefficients in the external validation sample (calibration slope 0.785) indicated an overall decrease in discriminative ability (c-index = 0.725; Uno = 0.729; *c-mbc* = 0.684; Table 5.2). The difference between the *c-mbc* and both the c-index and Uno's concordance measure was again considerable, likely due to a high proportion of censoring (89.7%). The proportional hazards assumption of the Cox regression model that was refitted to the external validation data was again not rejected (p=0.41).

SIMULATION STUDY

Methods

We simulated validation studies of a logistic regression model that aims to predict a binary endpoint and a proportional hazards regression model for a time-to-event endpoint. Both regression models were characterized by a linear predictor $x^T\beta$, with the baseline characteristic vector x consisting of a continuous predictor $x1$, e.g. age, and a binary predictor $x2$, e.g. sex. To mimic different external validation settings, we generated patient data (10.000 replications of 400 patients per setting) with different case-mix heterogeneity and different true regression coefficients (Table 5.3; Table 5.4). In a base case scenario (A), continuous predictors $x1$ were drawn from a standard normal distribution and binary predictors $x2$ were drawn from a Bernoulli distribution with success probability 0.2. Based on true predictor effects $\beta_1 = \beta_2 = 1$, we generated binary outcomes y from a Bernoulli distribution with success probabilities $[1 + \exp\{-x^T\beta\}]^{-1}$ (true intercept $\beta_0 = -2$) and time-to-event outcomes y from an exponential distribution with mean $\exp\{-x^T\beta\}$.

To study the influence of case-mix heterogeneity on concordance measures we varied the standard deviation of the continuous predictor (0.8 and 1.2 in scenarios B and C respectively) and the success probability of the binary predictor (0.1 and 0.4 in scenarios D and E respectively). We studied the influence of overall regression coefficient validity by varying the true effects of the continuous predictor (0.8 and 1.2 in scenarios F and G respectively), the binary predictor (0.5 and 2 in scenarios H and I respectively), and the true intercept of the logistic regression model (-3 and -1 in scenarios J and K respectively).

Censoring times were generated from an exponential distribution with mean c for different choices of c to analyze the effect of different proportions of censoring. To illustrate the effect of a violation of the proportional hazards assumption, we alternatively generated time-to-event outcomes from an exponential distribution with mean $\exp\{-x^t\beta_s\}$ and time-varying coefficients $\beta_s = \beta\ \exp[-0.5s]$.

In each sample we calculated: the linear predictor $X\beta$ with predictor effects $\beta_1 = \beta_2 = 1$ and intercept $\beta_0 = -2$ in case of binary outcomes; the calibration slope $\hat{\gamma}$ as the regression coefficient of a model with the linear predictor $x^t\beta$ as the sole predictor; the $mbc(X\beta)$ and the $mbc(\hat{\gamma}X\beta)$ with their variance estimates; Harrell's c-index; the case-mix corrected c-index, i.e. the c-index based on either 25, 100 or 400 resampled outcomes for each linear predictor $x^t\beta$; and Uno's concordance measure with the truncation time τ equal to the maximum follow-up time and to 80% of the maximum follow-up time in each replication. We used the rcorr.cens function in R package Hmisc and the Est.Cval function in R package survC1 for calculation of Harrell's c-index and Uno's concordance measure respectively [21-23].

Results

For binary outcomes and for uncensored time-to-event outcomes we found that the means of the mbc and the case-mix corrected c-index were very similar across the different validation settings (Table 5.3; Table 5.4). The empirical standard deviation of the case-mix corrected c-index was slightly higher as a result of resampling 400 binary outcomes and 25 time-to-event outcomes for each patient. However, the case-mix corrected c-index converged to the mbc with increasing numbers of resampled outcomes per patient (Figure 5.1).

Table 5.3 Simulation results for binary outcomes generated from a logistic regression model.

	Simulation settings					Benchmark performance			Overall performance			
	Case-mix		Coefficients									
	sd($x1$)	p(x2)	β_0	β_1	β_2	case-mix corrected c-index	mbc = $mbc(X\beta)$	SE[mbc]	Calibration slope $\hat{\gamma}$	c-index	c-mbc = $mbc(\hat{\gamma}X\beta)$	SE[c-mbc]
A	1	0.2	-2	1	1	0.760 (0.0077)	0.761 (0.0076)	0.0075	1.012 (0.154)	0.761 (0.030)	0.761 (0.030)	0.030
B	**0.8**	0.2	-2	1	1	0.728 (0.0073)	0.728 (0.0071)	0.0071	1.011 (0.175)	0.728 (0.033)	0.728 (0.032)	0.032
C	**1.2**	0.2	-2	1	1	0.790 (0.0077)	0.790 (0.0076)	0.0077	1.015 (0.141)	0.791 (0.028)	0.791 (0.028)	0.027
D	1	**0.1**	-2	1	1	0.755 (0.0077)	0.755 (0.0076)	0.0075	1.015 (0.162)	0.756 (0.032)	0.756 (0.031)	0.031
E	1	**0.4**	-2	1	1	0.765 (0.0075)	0.765 (0.0073)	0.0073	1.012 (0.147)	0.765 (0.029)	0.766 (0.028)	0.028
F	1	0.2	-2	**0.8**	1	0.760 (0.0077)	0.761 (0.0076)	0.0075	0.845 (0.146)	0.728 (0.033)	0.729 (0.032)	0.032
G	1	0.2	-2	**1.2**	1	0.760 (0.0076)	0.761 (0.0074)	0.0075	1.180 (0.162)	0.790 (0.028)	0.790 (0.027)	0.027
H	1	0.2	-2	1	**0.5**	0.760 (0.0076)	0.761 (0.0075)	0.0075	0.923 (0.152)	0.745 (0.032)	0.745 (0.032)	0.032
I	1	0.2	-2	1	**2**	0.760 (0.0076)	0.761 (0.0075)	0.0075	1.159 (0.160)	0.784 (0.028)	0.785 (0.027)	0.026
J	1	0.2	**-3**	1	1	0.760 (0.0078)	0.761 (0.0076)	0.0075	1.020 (0.203)	0.769 (0.042)	0.769 (0.041)	0.040
K	1	0.2	**-1**	1	1	0.760 (0.0076)	0.761 (0.0075)	0.0075	1.012 (0.133)	0.755 (0.025)	0.756 (0.025)	0.025

10,000 replications of 400 patients; casemix-corrected c-index based on 400 resampled outcomes per patient.

Table 5.4 Simulation results for time-to-event outcomes generated from a proportional hazards regression model.

	Simulation settings					Benchmark performance			Overall performance			
	Case-mix		Coefficients									
	sd($x1$)	p($x2$)	β_0	β_1	β_2	case-mix corrected c-index	mbc = $mbc(X\beta)$	SE[mbc]	Calibration slope $\hat{\gamma}$	c-index	c-mbc = $mbc(\hat{\gamma}X\beta)$	SE[c-mbc]
A	1	0.2	-2	1	1	0.736 (0.0062)	0.736 (0.0057)	0.0056	1.003 (0.064)	0.736 (0.013)	0.737 (0.011)	0.011
B	**0.8**	0.2	-2	1	1	0.708 (0.0059)	0.708 (0.0055)	0.0054	1.004 (0.072)	0.708 (0.014)	0.709 (0.012)	0.012
C	**1.2**	0.2	-2	1	1	0.760 (0.0061)	0.761 (0.0056)	0.0057	1.003 (0.059)	0.761 (0.012)	0.761 (0.011)	0.011
D	1	**0.1**	-2	1	1	0.731 (0.0062)	0.731 (0.0057)	0.0056	1.004 (0.065)	0.732 (0.013)	0.732 (0.011)	0.011
E	1	**0.4**	-2	1	1	0.742 (0.0061)	0.742 (0.0056)	0.0056	1.003 (0.063)	0.742 (0.013)	0.742 (0.011)	0.011
F	1	0.2	-2	**0.8**	1	0.736 (0.0062)	0.736 (0.0057)	0.0056	0.826 (0.059)	0.707 (0.014)	0.707 (0.012)	0.012
G	1	0.2	-2	**1.2**	1	0.736 (0.0062)	0.736 (0.0056)	0.0056	1.168 (0.068)	0.760 (0.012)	0.760 (0.011)	0.011
H	1	0.2	-2	1	**0.5**	0.736 (0.0059)	0.736 (0.0056)	0.0056	0.901 (0.063)	0.723 (0.013)	0.720 (0.012)	0.011
I	1	0.2	-2	1	**2**	0.736 (0.0061)	0.736 (0.0056)	0.0056	1.057 (0.067)	0.748 (0.013)	0.745 (0.011)	0.011

10,000 replications of 400 patients; casemix-corrected c-index based on 25 resampled outcomes per patient.

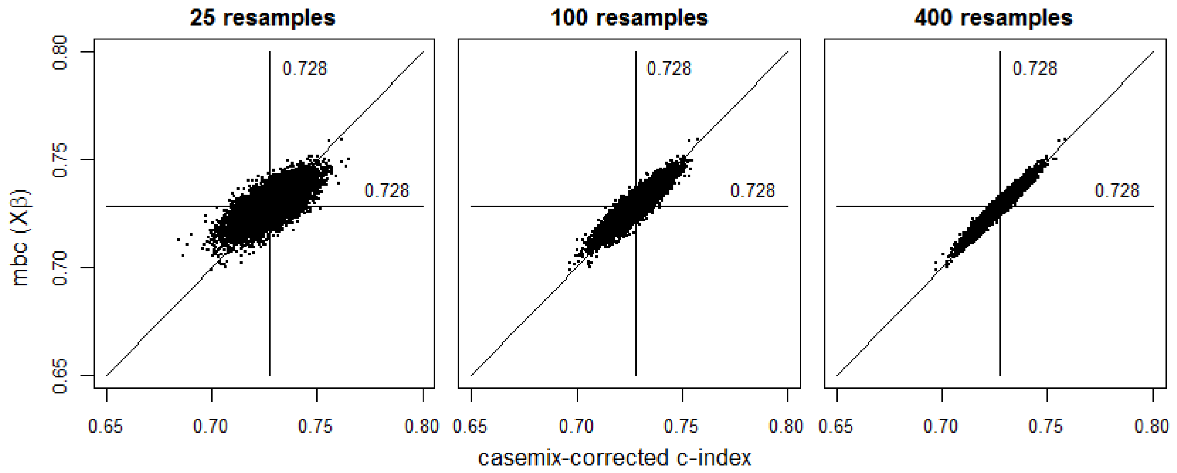


Figure 5.1 *mbc* versus case-mix corrected c-index based on 25, 100 and 400 resampled outcomes per patients respectively. Setting B of the binary outcome simulation; 10,000 replications of 400 patients.

The means of the *c-mbc* and the c-index were very similar as well, although the empirical standard deviation was lower for the *c-mbc* in case of Cox regression (Table 5.3; Table 5.4). Standard deviation estimates of *mbc* and *c-mbc* were well in agreement with the simulated empirical variances (Table 5.3; Table 5.4). Across all validation settings, the *c-mbc* remained stable at the true value with increasing proportions of censoring of time-to-event outcomes, while the c-index and to a lesser extent Uno's concordance measure increased unfavorably (Table 5.5; Supplementary figure 1). The empirical standard deviation of the *c-mbc* – again in good agreement with the standard deviation estimate – was structurally smaller than the standard deviation of the c-index and Uno's measure. When outcomes were sampled from an exponential distribution with time-varying coefficients ($\beta(s) = \beta \exp[-0.5s]$), the proportional hazards assumption of the *c-mbc* was violated leading to an underestimation of the concordance probability, specifically in the absence of time-to-event outcomes (Supplementary table 1). As a result of the decrease of the true regression coefficients in time, all concordance measures increased with increasing proportions of censoring of time-to-event outcomes.

We further analyzed the relation between the c-index and the *mbc*, the calibration slope or the *c-mbc* with scatterplots of validation setting A (Figure 5.2). Variation in the *mbc* was small compared to the c-index, since case-mix heterogeneity ($\mathrm{SD}(X\beta)$) was stable across the samples (left panel). With the limited number of 400 patients in each sample, the calibration slope – representing overall regression coefficient validity – varied substantially across the samples and was strongly related to the c-index (middle panel). Finally, the *c-mbc* – incorporating both case-mix heterogeneity and overall regression coefficient validity – correlated very well with the c-index (right panel).

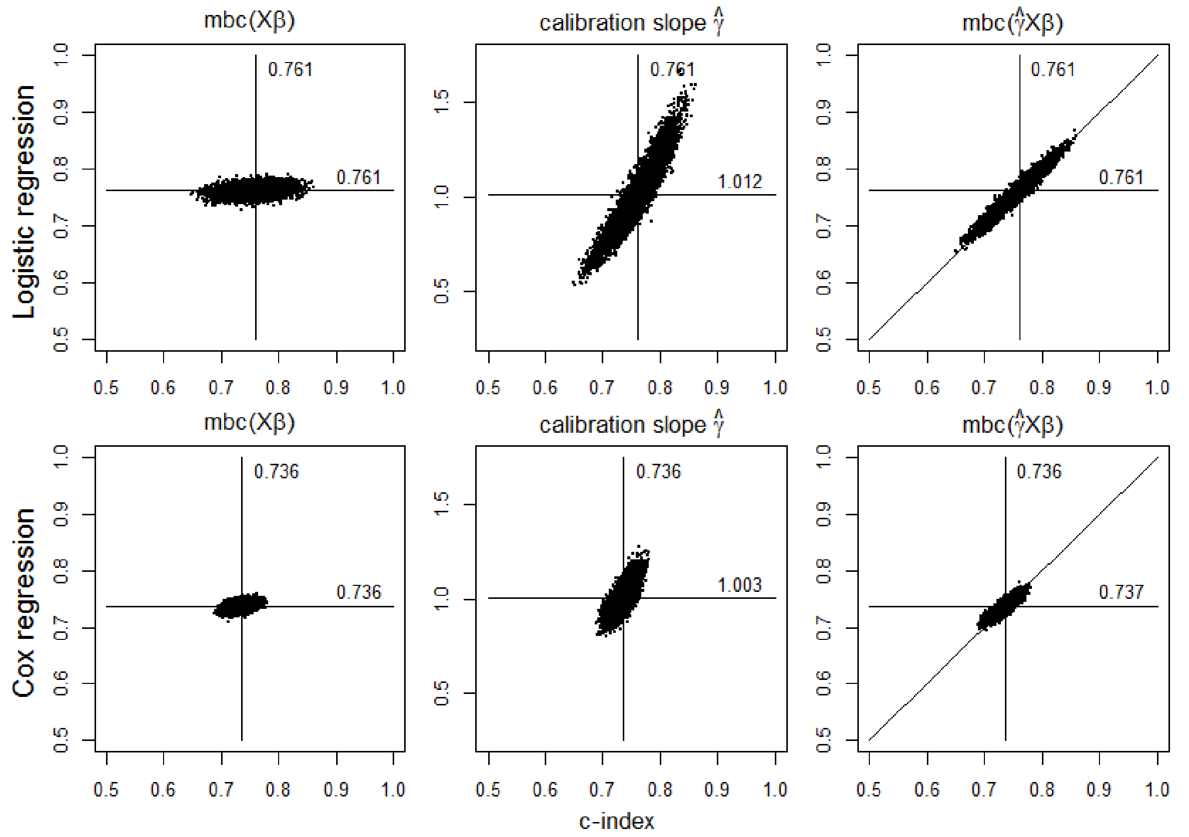


Figure 5.2 Performance measures (y-axis) *mbc*, calibration slope and *c-mbc* versus the c-index (x-axis). Setting A of the binary outcome simulation and the time-to-event outcome simulation; 10,000 replications of 400 patients.

When we changed case-mix heterogeneity (setting B-E), both the mean *mbc* and the mean *c-mbc* changed similarly, since the *mbc*'s assumption of correct regression coefficients held in these validation settings. As expected, when we changed the regression coefficients (settings F-I) the *mbc* remained the same while the *c-mbc* changed in accordance with the calibration slope. Changing the intercept of the logistic regression model (settings J-K) again affected only the *c-mbc*, but with a much smaller impact than a change in the regression coefficients of predictors.

DISCUSSION

We derived the *mbc*, which is a closed-form, censoring-robust alternative for the resampling-based case-mix corrected c-index. We showed that the *mbc* is asymptotically equivalent to a previously proposed, approximate case-mix corrected c-index. The *c-mbc* is comparable to Harrell's c-index in independent data with binary and time-to-event outcomes and furthermore is robust to censoring of time-to-event outcomes, in contrast with the c-index and Uno's concordance measure.

Table 5.5 Simulation results for time-to-event outcomes generated from a proportional hazards regression model with varying amounts of censoring.

							Concordance measures				
	sd(x1)	p(x2)	β_0	β_1	β_2	Cens %	c-index	Uno τ = 0.8 max(FU)	Uno τ = max(FU)	*c-mbc* = $mbc(\hat{\gamma}X\beta)$	SE[*c-mbc*]
A	1	0.2	-2	1	1	0	0.736 (0.013)	0.736 (0.013)	0.736 (0.013)	0.737 (0.011)	
						24	0.743 (0.015)	0.737 (0.014)	0.737 (0.014)	0.737 (0.012)	0.012
						50	0.751 (0.019)	0.738 (0.017)	0.738 (0.018)	0.737 (0.014)	0.014
						73	0.761 (0.025)	0.744 (0.031)	0.743 (0.034)	0.737 (0.017)	0.017
B	**0.8**	0.2	-2	1	1	0	0.708 (0.014)	0.708 (0.014)	0.708 (0.014)	0.709 (0.012)	0.012
						23	0.713 (0.016)	0.709 (0.014)	0.708 (0.014)	0.709 (0.013)	0.013
						50	0.720 (0.020)	0.710 (0.018)	0.709 (0.018)	0.709 (0.015)	0.015
						74	0.729 (0.027)	0.715 (0.033)	0.714 (0.036)	0.709 (0.019)	0.019
C	**1.2**	0.2	-2	1	1	0	0.761 (0.012)	0.761 (0.012)	0.761 (0.012)	0.761 (0.011)	0.011
						25	0.768 (0.014)	0.761 (0.013)	0.761 (0.013)	0.761 (0.011)	0.011
						50	0.778 (0.017)	0.763 (0.016)	0.763 (0.017)	0.761 (0.013)	0.013
						71	0.789 (0.023)	0.769 (0.029)	0.768 (0.032)	0.761 (0.015)	0.015
D	1	**0.1**	-2	1	1	0	0.732 (0.013)	0.732 (0.013)	0.732 (0.013)	0.732 (0.011)	0.011
						26	0.738 (0.015)	0.732 (0.014)	0.732 (0.014)	0.732 (0.012)	0.012
						52	0.746 (0.019)	0.733 (0.018)	0.733 (0.018)	0.732 (0.014)	0.014
						74	0.755 (0.026)	0.740 (0.033)	0.739 (0.036)	0.732 (0.017)	0.018
E	1	**0.4**	-2	1	1	0	0.742 (0.013)	0.742 (0.013)	0.742 (0.013)	0.742 (0.011)	0.011
						21	0.747 (0.014)	0.742 (0.013)	0.742 (0.013)	0.742 (0.012)	0.012
						46	0.755 (0.017)	0.743 (0.016)	0.743 (0.016)	0.742 (0.013)	0.013
						69	0.765 (0.023)	0.748 (0.027)	0.747 (0.029)	0.743 (0.016)	0.016
F	1	0.2	-2	**0.8**	1	0	0.707 (0.014)	0.707 (0.014)	0.707 (0.014)	0.707 (0.012)	0.012
						23	0.712 (0.016)	0.708 (0.014)	0.708 (0.014)	0.708 (0.013)	0.013
						50	0.719 (0.020)	0.709 (0.018)	0.708 (0.018)	0.709 (0.015)	0.015
						74	0.728 (0.027)	0.713 (0.033)	0.713 (0.036)	0.709 (0.019)	0.019
G	1	0.2	-2	**1.2**	1	0	0.760 (0.012)	0.760 (0.012)	0.760 (0.012)	0.760 (0.011)	0.011
						25	0.768 (0.014)	0.761 (0.013)	0.761 (0.013)	0.760 (0.011)	0.011
						50	0.777 (0.017)	0.763 (0.017)	0.762 (0.017)	0.760 (0.013)	0.013
						71	0.788 (0.023)	0.769 (0.030)	0.768 (0.032)	0.759 (0.015)	0.015
H	1	0.2	-2	1	**0.5**	0	0.724 (0.013)	0.724 (0.013)	0.724 (0.013)	0.721 (0.012)	0.011
						25	0.729 (0.015)	0.724 (0.014)	0.724 (0.014)	0.721 (0.013)	0.012
						52	0.737 (0.019)	0.725 (0.018)	0.725 (0.018)	0.721 (0.015)	0.014
						74	0.745 (0.026)	0.731 (0.033)	0.730 (0.037)	0.721 (0.018)	0.018
I	1	0.2	-2	1	**2**	0	0.747 (0.013)	0.747 (0.013)	0.747 (0.013)	0.744 (0.011)	0.011
						23	0.755 (0.014)	0.748 (0.013)	0.748 (0.013)	0.746 (0.012)	0.012
						47	0.767 (0.017)	0.749 (0.016)	0.749 (0.016)	0.749 (0.013)	0.013
						68	0.783 (0.022)	0.756 (0.028)	0.755 (0.030)	0.753 (0.015)	0.015

10,000 replications of 400 patients.

The *mbc* improves the understanding of a difference between the c-statistic at model development versus the observed c-statistic at external validation. The difference between the *mbc* at model development and the *mbc* at external validation indicates the change in discriminative ability attributable to the difference in case-mix heterogeneity. The difference between the *c-mbc* and the *mbc* in external validation data expresses the change in discriminative ability due to the (in)correctness of the regression coefficients. Thanks to their censoring-robustness, the *mbc* and the *c-mbc* facilitate measurements of concordance that are not biased by differences in censoring distributions between the development and the external validation setting.

The *mbc* and the *c-mbc* are model-based, i.e. they are based on the assumption that the true risks fit into the framework of a model. This assumption is necessary to evaluate the probability of the outcomes being ordered, conditional on the risk scores, i.e. $P_{ij} = P\left(Y_i < Y_j \middle| x_i^T\beta, x_j^T\beta\right)$. In this paper we used either a logistic regression model or a proportional hazards regression model to evaluate these probabilities. This may be a limitation compared to Harrell's c-index and Uno's concordance measure, since these pure rank-order statistics are applicable to any risk scoring system. However, since logistic regression and proportional hazards regression are commonly used to model binary outcomes and time-to-event outcomes respectively, the *mbc* and the *c-mbc* may often be valuable.

The *c-mbc* was shown to be very robust to censoring in the simulation study where the proportional hazards assumption held. When the proportional hazards assumption did not hold – as in our sensitivity analysis with time-dependent coefficients – the *c-mbc* gave different estimates than Harrell's c-index and Uno's concordance measure, even without censoring of time-to-event outcomes. In the presence of time-varying coefficients it may be better to assess discriminative ability in a limited follow-up period [8]. This was beyond the scope of our research, but we provided formulas for an *mbc* truncated at a fixed follow-up time in Appendix 5.3. When coefficients are time-dependent, the *c-mbc* could alternatively be based on more sophisticated conditional probabilities $P\left(Y_i < Y_j\right)$. Stare et al. proposed a measure for use with general dynamic event history regression models, including models with time-dependent coefficients, that reduces to the c-index for single-event survival data with neither censoring nor time dependency [24]. However, since all concordance measures for models with time-dependent coefficients will probably be sensitive to censoring, their use in practice needs additional study.

The *c-mbc* assumes a linear relationship (represented by the calibration slope $\hat{\gamma}_1$) between linear predictors and either the log hazard for time-to-event outcomes or the log odds for binary outcomes. In the scenarios F, G, H and I of our simulation study this assumption was clearly violated, since the true effect of only one of two predictors was varied consecutively. Although the *c-mbc* was robust to violation of the linearity assumption in these scenarios – the mean *c-mbc* was very close to the mean c-index (Table 5.3; Table 5.4) – further research is necessary to understand the importance of this assumption. An alternative *c-mbc* that allows for potential non-linear relationships between linear predictors and outcomes could be considered.

We derived variance estimators for the *mbc* – under the assumption of correct regression coefficients – and the *c-mbc* – including regression coefficient uncertainty. Variance estimates were very well in line with the empirical variances of the simulation study. The empirical variances of the *c-mbc* were generally lower than those of the c-index and Uno's concordance measure for proportional hazards regression models, especially in the presence of high proportions of censoring. The higher precision of the *c-mbc* is likely the result of its proportional hazards assumption.

In both case-studies the effect of more case-mix heterogeneity (larger standard deviation of the linear predictor) on discriminative ability was illustrated by an increase in the *mbc* in the validation data compared to the *mbc* in the development data. The influence on discriminative ability of a change in the strength of the association between predictions and outcomes (calibration slope above 1 in the logistic regression case study, calibration slope below 1 in the proportional hazards regression case study), was reflected in the difference between the *c-mbc* and the *mbc* in the validation data. The large difference of the *c-mbc* in comparison with the c-index and Uno's concordance measure, emphasized the importance of using censoring-robust concordance measures when time-to-event outcomes are substantially censored.

In conclusion, the *mbc* is an attractive closed-form measure that allows for straightforward quantification of the expected change in a model's discriminative ability due to case-mix heterogeneity in a validation setting. Moreover, the *c-mbc* also reflects the impact of regression coefficient validity on a model's discriminative ability in external validation data, and is a censoring-robust alternative for the c-index when the proportional hazards assumption holds.

ACKNOWLEDGMENTS

The authors express their gratitude to all of the principal investigators of the International Tirilazad Trial, the SYNTAX trial and the CREDO-Kyoto registry for providing the data. This work was supported by the Netherlands Organisation for Scientific Research (grant 917.11.383).

SUPPLEMENTARY DATA

An online appendix can be found at http://dx.doi.org/ 10.1002/sim.6997.

REFERENCES

1. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; 21: 128-138.
2. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982; 247: 2543-2546.
3. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* 2004; 23: 2109-2123.
4. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010; 172: 971-980.
5. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol* 2012; 12: 82.
6. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol* 2013; 13: 33.
7. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005; 92: 965-970.
8. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011; 30: 1105-1117.
9. Korn EL, Simon R. Measures of explained variation for survival data. *Stat Med* 1990; 9: 487-503.
10. van Klaveren D, Steyerberg EW, Vergouwe Y. Interpretation of concordance measures for clustered data. *Stat Med* 2014; 33: 714-716.
11. Lambert J, Chevret S. Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. *Statistical Methods in Medical Research* 2014.
12. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* Springer: New York, 2009.
13. Hukkelhoven CW, Steyerberg EW, Farace E, Habbema JD, Marshall LF, Maas AI. Regional differences in patient characteristics, case management, and outcomes in traumatic brain injury: experience from the tirilazad trials. *J Neurosurg* 2002; 97: 549-557.
14. Marshall LF, Maas AI, Marshall SB, Bricolo A, Fearnside M, Iannotti F, Klauber MR, Lagarrigue J, Lobato R, Persson L, Pickard JD, Piek J, Servadei F, Wellis GN, Morris GF, Means ED, Musch B. A multicenter trial on the efficacy of using tirilazad mesylate in cases of head injury. *J Neurosurg* 1998; 89: 519-525.
15. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JD, Maas AI. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Med* 2008; 5: e165.
16. Campos CM, van Klaveren D, Iqbal J, Onuma Y, Zhang YJ, Garcia-Garcia HM, Morel MA, Farooq V, Shiomi H, Furukawa Y, Nakagawa Y, Kadota K, Lemos PA, Kimura T, Steyerberg EW, Serruys PW. Predictive Performance of SYNTAX Score II in Patients With Left Main and Multivessel Coronary Artery Disease-analysis of CREDO-Kyoto registry. *Circ J* 2014; 78: 1942-1949.
17. Mohr FW, Morice MC, Kappetein AP, Feldman TE, Stahle E, Colombo A, Mack MJ, Holmes DR, Jr., Morel MA, Van Dyck N, Houle VM, Dawkins KD, Serruys PW. Coronary artery bypass graft surgery versus percutaneous coronary intervention in patients with three-vessel disease and left main coronary disease: 5-year follow-up of the randomised, clinical SYNTAX trial. *Lancet* 2013; 381: 629-638.
18. Farooq V, van Klaveren D, Steyerberg EW, Meliga E, Vergouwe Y, Chieffo A, Kappetein AP, Colombo A, Holmes DR, Jr., Mack M, Feldman T, Morice MC, Stahle E, Onuma Y, Morel MA,

Garcia-Garcia HM, van Es GA, Dawkins KD, Mohr FW, Serruys PW. Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score II. *Lancet* 2013; 381: 639-650.

19. Kimura T, Morimoto T, Furukawa Y, Nakagawa Y, Kadota K, Iwabuchi M, Shizuta S, Shiomi H, Tada T, Tazaki J, Kato Y, Hayano M, Abe M, Tamura T, Shirotani M, Miki S, Matsuda M, Takahashi M, Ishii K, Tanaka M, Aoyama T, Doi O, Hattori R, Tatami R, Suwa S, Takizawa A, Takatsu Y, Takahashi M, Kato H, Takeda T, Lee JD, Nohara R, Ogawa H, Tei C, Horie M, Kambara H, Fujiwara H, Mitsudo K, Nobuyoshi M, Kita T. Long-term safety and efficacy of sirolimus-eluting stents versus bare-metal stents in real world clinical practice in Japan. *Cardiovasc Interv Ther* 2011; 26: 234-245.
20. Grambsch PM, Therneau TM. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika* 1994; 81: 515-526.
21. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/, 2011.
22. Hmisc: Harrell Miscellaneous. R package version 3.9-2. http://CRAN.R-project.org/package=Hmisc, 2012.
23. survC1: C-statistics for risk prediction models with censored survival data. R package version 1.0-2. http://CRAN.R-project.org/package=survC1, 2013.
24. Stare J, Perme MP, Henderson R. A measure of explained variation for event history data. *Biometrics* 2011; 67: 750-759.
25. Quade D. *Nonparametric partial correlation.* Institute of Statistics Mimeo Series No. 526: North Carolina, 1967.

APPENDIX 5.1

Estimating the variance of the *mbc*

The $mbc(X\hat{\beta})$ estimates the concordance probability in an apparent validation setting. To obtain a variance estimate of $mbc(X\hat{\beta})$ we follow the derivation of the variance estimate of Gönen and Heller's concordance probability estimator [7]. It starts with a local linear asymptotic approximation of $mbc(X\hat{\beta})$ in the neighborhood of the point estimate $\beta_0 = E\hat{\beta}$:

$$mbc(X\hat{\beta}) = mbc(X\beta_0) + \left[\frac{\partial E\{mbc(X\beta)\}}{\partial \beta}\right]^T \Bigg|_{\beta=\beta_0} (\hat{\beta} - \beta_0) + o_p(1) \tag{8}$$

With $(\beta_0) = \left[\frac{\partial E\{mbc(X\beta)\}}{\partial \beta}\right]\Big|_{\beta=\beta_0}$, conditional on the covariates, the centered partial likelihood estimator $(\hat{\beta} - \beta_0)$ is asymptotically independent of $mbc(X\beta_0)$. In addition, since $D(\beta_0)$ converges to a constant, the asymptotic variance of $mbc(X\hat{\beta})$ is approximately:

$$\mathrm{var}\{mbc(X\hat{\beta})\} \approx \mathrm{var}\{mbc(X\beta_0)\} + D(\beta_0)^T \mathrm{var}(\hat{\beta}) D(\beta_0) \tag{9}$$

The first part on the right hand side is the variance of the sample statistic $mbc(X\beta_0)$. It is easy to obtain since $mbc(X\beta_0)$ is the ratio of two U-statistics of degree 2:

$$mbc(X\beta_0) = \frac{U_1}{U_2}$$

$$U_1 = \frac{1}{n}\sum\nolimits_i U_1^{(i)} \text{ , } U_1^{(i)} = \frac{1}{n-1}\sum\nolimits_{j\neq i}\left[I(p_i < p_j)P(Y_i < Y_j) + I(p_i > p_j)P(Y_i > Y_j)\right] \tag{10}$$

$$U_2 = \frac{1}{n}\sum\nolimits_i U_2^{(i)} \text{ ,} \qquad U_2^{(i)} = \frac{1}{n-1}\sum\nolimits_{j\neq i}\left[P(Y_i < Y_j) + P(Y_i > Y_j)\right]$$

From U-statistics theory we know that [25]:

$$\frac{\sqrt{n}}{s}\left(\frac{U_1}{U_2}-\frac{\theta_1}{\theta_2}\right)\xrightarrow{d}N(0,1)$$

$$s^2=\frac{4[U_2^2v_{11}-2U_1U_2v_{12}+U_1^2v_{22}]}{U_2^4}$$

$$v_{11}=\frac{1}{n-1}\sum\nolimits_i\left(U_1^{(i)}-U_1\right)^2 \tag{11}$$

$$v_{12}=\frac{1}{n-1}\sum\nolimits_i\left(U_1^{(i)}-U_1\right)\left(U_2^{(i)}-U_2\right)$$

$$v_{22}=\frac{1}{n-1}\sum\nolimits_i\left(U_2^{(i)}-U_2\right)^2$$

Hence

$$\hat{\text{var}}\{mbc(X\beta_0)\}=\frac{s^2}{n} \tag{12}$$

Note that for proportional hazards regression $U_2\equiv 1$ and therefore the variance estimate reduces to $\frac{4v_{11}}{n}$.

The second part on the right hand side of equation 9 represents the variance of $mbc\left(X\hat{\beta}\right)$ as a result of uncertainty in the regression coefficients. We can use the inverse of the likelihood information matrix to estimate $\text{var}\left(\hat{\beta}\right)$. The function $D(\beta_0)$ in equation 9 is estimated through numerical differentiation:

$$\widehat{D}_i(\beta_0)=\frac{mbc(X\beta_0+s_ie_i)-mbc(X\beta_0-s_ie_i)}{2s_i} \tag{13}$$

Where s_i is the standard error of $\hat{\beta}_i$ and e_i is the i^{th} unit vector.

The variance of $mbc\left(X\hat{\beta}\right)$ in an apparent validation setting was decomposed into a part due to sampling patients and a part due to regression coefficient uncertainty (equation 9). When $mbc(X\beta)$ is used in an external validation setting to assess the influence of case-mix heterogeneity on discriminative ability – assuming correct regression coefficients – , its variance is limited to the first part (equation 12). The *c-mbc* ($mbc(\hat{\gamma}X\beta)$) assesses both the influence of case-mix heterogeneity and the validity of the linear predictor $X\beta$ in external data. The derivation of the variance estimate of $mbc(\hat{\gamma}X\beta)$ is similar to equation 8-13. Replacing X by $X\beta$ and β by γ in equation 8 results in:

$$mbc(\hat{\gamma}X\beta) = mbc(\gamma_0 X\beta) + \left[\frac{\partial E\{mbc(\gamma X\beta)\}}{\partial \gamma}\right]^T \Bigg|_{\gamma=\gamma_0} (\hat{\gamma} - \gamma_0) + o_p(1) \qquad (14)$$

Similar to equation 9, the asymptotic variance of $mbc(\hat{\gamma}X\beta)$ is approximately:

$$\mathrm{var}\{mbc(\hat{\gamma}X\beta)\} \approx \mathrm{var}\{mbc(\gamma_0 X\beta)\} + D(\gamma_0)^T \mathrm{var}(\hat{\gamma}) D(\gamma_0) \qquad (15)$$

We can use equations 10-12 to estimate $\mathrm{var}\{mbc(\gamma_0 X\beta)\}$, the inverse of the likelihood information matrix to estimate $\mathrm{var}(\hat{\gamma})$ and numerical differentiation (equation 13) to estimate $D(\gamma_0)$.

APPENDIX 5.2

Comparison of the *mbc* with Harrell's c-index

We will show that within the development data, the *mbc* and the c-index are asymptotically equivalent for logistic regression models and – if the proportional hazards assumption holds – for proportional hazards regression models developed with continuous, uncensored time-to-event outcomes. We will start by showing that the *mbc* and the c-index are exactly equal at internal validation of a logistic regression model with one binary or categorical predictor.

Logistic regression

Harrell's c-index – allowing for ties in predictions – is:

$$c_H(X\hat{\beta}, y) = \frac{\sum_{i,j:\, x_i^T\hat{\beta}<x_j^T\hat{\beta}} I\{y_i < y_j\} + \frac{1}{2}\sum_{i,j:\, x_i^T\hat{\beta}=x_j^T\hat{\beta}} I\{y_i < y_j\}}{\sum_{i,j} I\{y_i < y_j\}} \tag{16}$$

The *mbc* at internal validation can be written similarly as:

$$mbc(X\hat{\beta}) = \frac{\sum_{i,j:\, x_i^T\hat{\beta}<x_j^T\hat{\beta}} P(Y_i < Y_j | x_i^T\hat{\beta}, x_j^T\hat{\beta}) + \frac{1}{2}\sum_{i,j:\, x_i^T\hat{\beta}=x_j^T\hat{\beta}} P(Y_i < Y_j | x_i^T\hat{\beta}, x_j^T\hat{\beta})}{\sum_{i,j} P(Y_i < Y_j | x_i^T\hat{\beta}, x_j^T\hat{\beta})} \tag{17}$$

The denominators of $mbc(X\hat{\beta})$ and $c_H(X\hat{\beta}, y)$ are equal when $\hat{\beta}$ is estimated with logistic regression:

$$\sum_{i,j} P(Y_i < Y_j | x_i^T\hat{\beta}, x_j^T\hat{\beta}) = \sum_i P(Y_i = 0 | x_i^T\hat{\beta}) \sum_j P(Y_j = 1 | x_j^T\hat{\beta}) =$$
$$= \sum_i I\{y_i = 0\} \sum_j I\{y_j = 1\} = \sum_{i,j} I\{y_i < y_j\} \tag{18}$$

With only one binary predictor x, the numerators are equal as well since:

$$\sum_{i,j:\, x_i^T\hat{\beta}<x_j^T\hat{\beta}} P(Y_i < Y_j | x_i^T\hat{\beta}, x_j^T\hat{\beta}) =$$
$$= \sum_{i:\, x_i=0} P(Y_i = 0 | x_i^T\hat{\beta}) \sum_{j:\, x_j=1} P(Y_j = 1 | x_j^T\hat{\beta}) = \tag{19}$$
$$= \sum_{i:\, x_i=0} I\{y_i = 0\} \sum_{j:\, x_j=1} I\{y_j = 1\} = \sum_{i,j:\, x_i^T\hat{\beta}<x_j^T\hat{\beta}} I\{y_i < y_j\}$$

And for ties:

$$T = \sum_{i,j:\, x_i^T\hat{\beta}=x_j^T\hat{\beta}} P\left(Y_i < Y_j \middle| x_i^T\hat{\beta}, x_j^T\hat{\beta}\right) =$$

$$\begin{aligned} &\sum_{i:\, x_i=0} P\left(Y_i = 0 \middle| x_i^T\hat{\beta}\right) \sum_{j:\, x_j=0} P\left(Y_j = 1 \middle| x_j^T\hat{\beta}\right) \\ &\qquad + \sum_{i:\, x_i=1} P\left(Y_i = 0 \middle| x_i^T\hat{\beta}\right) \sum_{j:\, x_j=1} P\left(Y_j = 1 \middle| x_j^T\hat{\beta}\right) = \\ &\sum_{i:\, x_i=0} I\{y_i = 0\} \sum_{j:\, x_j=0} I\{y_j = 1\} + \sum_{i:\, x_i=1} I\{y_i = 0\} \sum_{j:\, x_j=1} I\{y_j = 1\} = \\ &= \sum_{i,j:\, x_i^T\hat{\beta}=x_j^T\hat{\beta}} I\{y_i < y_j\} \end{aligned} \qquad (20)$$

Consequently, $mbc(X\hat{\beta})$ and $c_H(X\hat{\beta}, y)$ are exactly equal for a logistic regression model with one binary predictor. Following the same line of reasoning, it is easy to show that $mbc(X\hat{\beta})$ and $c_H(X\hat{\beta}, y)$ are equal for logistic regression models with one categorical predictor with any number of levels.

For logistic regression models in general, we stratify the subjects in risk groups R_k of increasing linear predictor values. The mbc can be written as:

$$mbc(X\hat{\beta}) = \frac{\sum_{\substack{k,l:\, k<l \\ i\in R_k, j\in R_l}} P\left(Y_i < Y_j \middle| x_i^T\hat{\beta}, x_j^T\hat{\beta}\right) + T' + \frac{1}{2}T}{\sum_{i,j} P\left(Y_i < Y_j \middle| x_i^T\hat{\beta}, x_j^T\hat{\beta}\right)} \qquad (21\ )$$

With T' and T denoting the sum of conditional probabilities $P\left(Y_i < Y_j \middle| x_i^T\hat{\beta}, x_j^T\hat{\beta}\right)$ for respectively the subject pairs with different risk predictions but in the same risk group $\left(i \in R_k, j \in R_k \colon x_i^T\hat{\beta} < x_j^T\hat{\beta}\right)$ and the subject pairs with equal risk predictions $\left(x_i^T\hat{\beta} = x_j^T\hat{\beta}\right)$. The between-risk-group sums in $mbc(X\hat{\beta})$ and $c_H(X\hat{\beta}, y)$ are asymptotically equivalent since:

$$\sum_{i\in R_k, j\in R_l} P\left(Y_i < Y_j \middle| x_i^T\hat{\beta}, x_j^T\hat{\beta}\right) =$$

$$\sum_{i\in R_k} P\left(Y_i = 0 \middle| x_i^T\hat{\beta}\right) \sum_{j\in R_l} P\left(Y_j = 1 \middle| x_j^T\hat{\beta}\right) \sim \tag{22}$$

$$\sum_{i\in R_k} I\{y_i = 0\} \sum_{j\in R_l} I\{y_j = 1\} = \sum\nolimits_{i\in R_k, j\in R_l} I\{y_i < y_j\}$$

For subject pairs with different risk predictions but in the same risk group (T') and for subject pairs with equal risk predictions (T), a similar equivalence between the $mbc(X\hat{\beta})$ and $c_H(X\hat{\beta}, y)$ can be derived. In conclusion, the $mbc(X\hat{\beta})$ and $c_H(X\hat{\beta}, y)$ are asymptotically equivalent for logistic regression models.

Proportional hazards regression

Harrell's c-index – without ties in predictions – is:

$$c_H(X\hat{\beta}, y) = \frac{\sum_i \sum_j \{I\{x_i^T\hat{\beta} > x_j^T\hat{\beta}\} I\{y_i < y_j\}\delta_i\}}{\sum_i \sum_j \{I\{y_i < y_j\}\delta_i\}} \tag{23}$$

For uncensored time-to-event outcomes, it can be written as:

$$c_H(X\hat{\beta}, y) = \frac{\sum_{i,j:\, x_i^T\hat{\beta} > x_j^T\hat{\beta}} I\{y_i < y_j\}}{\sum_{i,j:\, i\neq j} I\{y_i < y_j\}} \tag{24}$$

The mbc at internal validation can be written similarly as:

$$mbc(X\hat{\beta}) \; = \frac{\sum_{i,j:\, x_i^T\hat{\beta} > x_j^T\hat{\beta}} P\left(Y_i < Y_j \middle| x_i^T\hat{\beta}, x_j^T\hat{\beta}\right)}{\sum_{i,j:\, i\neq j} P\left(Y_i < Y_j \middle| x_i^T\hat{\beta}, x_j^T\hat{\beta}\right)} \tag{25}$$

For continuous uncensored time-to-event outcomes, the denominators of $mbc(X\hat{\beta})$ and $c_H(X\hat{\beta}, y)$ are both $\frac{n(n-1)}{2}$ since $P\left(Y_i < Y_j \middle| x_i^T\hat{\beta}, x_j^T\hat{\beta}\right) + P\left(Y_j < Y_i \middle| x_i^T\hat{\beta}, x_j^T\hat{\beta}\right) = 1$.

Under the assumption of proportional hazards, the true conditional probability $P\left(Y_i < Y_j \middle| x_i^T\beta, x_j^T\beta\right)$ given one binary predictor with values $x_i = 1$ and $x_j = 0$ follows from equation 6:

$$P\left(Y_i < Y_j \middle|\, x_i = 1,\, x_j = 0, \beta\right) = \frac{1}{1 + exp\{\beta\}} \tag{26}$$

Since the observed frequency of $I\{y_i < y_j\}$ with $x_i = 1$ and $x_j = 0$ will converge to this true probability when the proportional hazards assumption holds, it follows that:

$$\sum_{i,j:\, x_i^T\hat{\beta} > x_j^T\hat{\beta}} P\left(Y_i < Y_j \middle| x_i^T\hat{\beta}, x_j^T\hat{\beta}\right) = \sum_{\substack{i:\, x_i=1 \\ j:\, x_j=0}} P\left(Y_i < Y_j \middle| x_i^T\hat{\beta}, x_j^T\hat{\beta}\right) = \\ \sum_{\substack{i:\, x_i=1 \\ j:\, x_j=0}} \frac{1}{1 + exp\{\hat{\beta}\}} \sim \sum_{\substack{i:\, x_i=1 \\ j:\, x_j=0}} I\{y_i < y_j\} = \sum_{i,j:\, x_i^T\hat{\beta} > x_j^T\hat{\beta}} I\{y_i < y_j\} \tag{27}$$

Applying the same stratification in risk groups R_k of increasing linear predictor values as for logistic regression models, leads to the conclusion that $mbc\left(X\hat{\beta}\right)$ and $c_H\left(X\hat{\beta}, y\right)$ are asymptotically equivalent for proportional hazards regression models, when the proportional hazards assumption holds.

APPENDIX 5.3

The truncated *mbc*

The truncated concordance probability $CP(\tau)$ in a patient population is:

$$CP(\tau) = \frac{\sum_i \sum_{j \neq i} \left[I(p_i < p_j) P(Y_i < Y_j, Y_i < \tau) + I(p_i > p_j) P(Y_i > Y_j, Y_j < \tau) \right]}{\sum_i \sum_{j \neq i} \left[P(Y_i < Y_j, Y_i < \tau) + P(Y_i > Y_j, Y_j < \tau) \right]} \tag{28}$$

For the truncated model-based concordance $mbc(\tau; X\beta)$ we again derive the probabilities $P(Y_i < Y_j, Y_i < \tau)$ from the proportional hazards regression model:

$$P(Y_i < Y_j, Y_i < \tau) = -\int_0^\tau S(t|x_j^T\beta)\, dS(t|x_i^T\beta) \tag{29}$$

Using integration by parts gives:

$$P(Y_i < Y_j, Y_i < \tau) = -\int_0^\tau \mathrm{d}\left[S(t|x_j^T\beta) S(t|x_i^T\beta)\right] + \int_0^\tau S(t|x_i^T\beta)\, \mathrm{d}S(t|x_j^T\beta) \tag{30}$$

The second integral on the right-hand-side of the equation can be written as:

$$\begin{aligned} \int_0^\tau S(t|x_i^T\beta) \mathrm{d}S(t|x_j^T\beta) &= -\int_0^\tau S(t|x_i^T\beta) S(t|x_j^T\beta) \lambda_0(t) \mathrm{e}^{x_j^T\beta} \mathrm{dt} \\ &= -\,\mathrm{e}^{(x_j - x_i)^T\beta} \int_0^\tau S(t|x_i^T\beta) S(t|x_j^T\beta) \lambda_0(t) \mathrm{e}^{x_i^T\beta} \mathrm{dt} \\ &= \ \mathrm{e}^{(x_j - x_i)^T\beta} \int_0^\tau S(t|x_j^T\beta) \mathrm{d}S(t|x_i^T\beta) \\ &= -\,\mathrm{e}^{(x_j - x_i)^T\beta} P(Y_i < Y_j, Y_i < \tau) \end{aligned} \tag{31}$$

Substituting equation 31 into equation 30 results in:

$$P(Y_i < Y_j, Y_i < \tau) = \frac{1 - S(\tau|x_i^T\beta) S(\tau|x_j^T\beta)}{1 + \mathrm{e}^{(x_j - x_i)^T\beta}} \tag{32}$$

Since $S(\tau|x^T\beta) = S_0(\tau)^{\mathrm{e}^{x^T\beta}}$, equation 32 depends on the linear predictors $x_i^T\beta$ and $x_j^T\beta$, and on the baseline survival function $S_0(\tau)$ at time τ. The truncated model-based concordance results from equations 28 and 32:

$$mbc(\tau; X\beta) = \frac{\sum_i \sum_{j\neq i}\left[\left(1 - S(\tau|x_i^T\beta)S\left(\tau|x_j^T\beta\right)\right)\left(\frac{I\left(x_i^T\beta > x_j^T\beta\right)}{1+\mathrm{e}^{\left(x_j-x_i\right)^T\beta}} + \frac{I\left(x_i^T\beta < x_j^T\beta\right)}{1+\mathrm{e}^{\left(x_i-x_j\right)^T\beta}}\right)\right]}{\sum_i \sum_{j\neq i}\left[\left(1 - S(\tau|x_i^T\beta)S\left(\tau|x_j^T\beta\right)\right)\left(\frac{1}{1+\mathrm{e}^{\left(x_j-x_i\right)^T\beta}} + \frac{1}{1+\mathrm{e}^{\left(x_i-x_j\right)^T\beta}}\right)\right]} \quad (33)$$

In contrast with the original $mbc(X\beta)$ in equation 7, which is defined on the basis of complete follow-up, the truncated $mbc(\tau; X\beta)$ weighs each patient pair by the probability that at least one of the patients encounters the event before follow-up time τ.

6

Geographic and temporal validity of prediction models: different approaches were useful to examine heterogeneity

PC Austin
D van Klaveren
Y Vergouwe
D Nieboer
DS Lee
EW Steyerberg

Submitted

ABSTRACT

Objective To examine methods for assessing temporal and geographic heterogeneity in baseline risk and predictor effects in prediction models.

Study Design and Setting We studied 14,857 patients hospitalized with heart failure at 90 hospitals in Ontario, Canada in two distinct temporal periods. We focussed on geographic and temporal variation in baseline risk (intercept) and predictor effects (regression coefficients) of the EFFECT-HF mortality model for predicting one-year mortality in patients hospitalized for heart failure. We hereto used random effects logistic regression models for the 14,857 patients.

Results The baseline risk of mortality displayed moderate geographic variation, with the hospital-specific probability of one-year mortality for a reference patient lying between 0.168 and 0.290 for 95% of hospitals. Furthermore, the odds of death were 11% lower in the second period than in the first period. However, we found minimal geographic or temporal variation in predictor effects. The one exception was that the effect of hepatic cirrhosis on mortality was weaker in the later period compared to in the earlier period.

Conclusion This study illustrates how temporal and geographic heterogeneity of prediction models can be assessed in settings with a large sample of patients from a large number of centres at different time periods.

WHAT IS NEW?

Key findings

- Using data on patients hospitalized with heart failure in the Canadian province of Ontario and a previously-derived clinical prediction model, a modest temporal improvement in baseline risk was observed.
- Moderate geographic variation in the baseline risk of one-year mortality was observed.
- Predictor effects were stable across geographic regions and in time.

What this adds to what is known

- Assessment of temporal and geographic stability in baseline risk and predictor effects with random effects models complements conventional methods for model validation that are focused on reporting model performance for independent validation samples.

What is the implication/what should change now

- Authors examining the generalizability of clinical prediction models are encouraged to examine the geographic and temporal stability in baseline risk and predictor effects.

INTRODUCTION

Clinical prediction models permit one to estimate the probability of the presence of disease (diagnosis), or the probability of the occurrence of adverse events for patients with specific medical diagnoses or undergoing specific surgical procedures or interventions (prognosis). Classical aspects of model validation include internal validation or reproducibility (how the model performs in patients who were not included in model development, but who are from the same underlying population), temporal validation (how the model performs on subsequent patients at the same centres at which the model was developed), and geographic (how the model performs on patients from centres different from those which participated in model development) [1-5]. The current gold-standard approach to assessing model validity is to report a summary measure of model performance, such as the concordance statistic (*c*) or area under the ROC curve, in a sample different from that in the model was developed [6]. In a companion article, we illustrated the application of different methods for assessing temporal and geographic performance of prediction models in independent samples [7].

Model transportability can also be examined by the temporal or geographic stability of the baseline risk and predictor effects. A desirable property for a prediction model is that the estimated effects are constant across geographic regions and across different temporal periods. Our objective was to describe and illustrate methods for assessing such geographic and temporal stability of baseline risk and predictor effects and to provide guidance on their use. Accordingly, we analyzed data on patients hospitalized with congestive heart failure (CHF) at a large number of hospitals in two distinct time periods.

METHODS

Data source and prediction model

The current study used 7,549 patients hospitalized with CHF during the first phase of the EFFECT study phase (April 1999 to March 2001) and 7,308 patients hospitalized with CHF during the second phase of the study (April 2004 to March 2005) [8]. The EFFECT-HF mortality prediction model for one-year mortality uses 11 variables: age, systolic blood pressure on admission, respiratory rate on admission, low sodium serum concentration (< 136 mEq/L), low serum hemoglobin (< 10.0 g/dL), serum urea nitrogen, presence of cerebrovascular disease, presence of dementia, chronic obstructive pulmonary disease, hepatic cirrhosis, and cancer [9]. For the current analyses, the four continuous variables were standardized to have mean zero. Greater details on the study sample and prediction model are provided elsewhere [7, 9]. All analyses were conducted in the pooled sample consisting of patients from both phases of the study.

Exploring geographic heterogeneity

First, we fit a fixed effects logistic regression model in which the probability of one-year mortality was regressed on the 11 predictors in the EFFECT-HF model (Model 1) (all models are described mathematically in Appendix 6.1). This model ignores both temporal and geographic variability in the probability of one-year mortality. From this model, we extracted from the fitted linear predictor. This is the conventional linear predictor that would be obtained in a study that ignored temporal and geographic effects. This linear predictor will be used in subsequent models where noted.

A series of random effects logistic regression models were fit to explore geographic variation. First, we modified Model 1 by including hospital-specific random intercepts (Model 2). The inclusion of random intercepts allows one to explore geographic variation in the baseline risk of one-year mortality across hospitals, by allowing the log-odds of one-year mortality to vary across hospitals. However, the effect of each predictor variable was constant across hospitals.

We then fit a random intercept model in which the log-odds of one-year mortality was regressed on the marginal linear predictor estimated above (Model 3). This analysis allows one to assess whether the log-odds of death for an arbitrarily-defined reference patient (one whose linear predictor was equal to zero) varies across hospitals. The effect of the linear predictor was uniform across hospitals and no effect of time was considered. This analysis is very similar to random effects meta-analysis of the calibration intercept observed across hospitals, as explored in the companion paper.

We considered an extension of Model 3 in which the effect of the linear predictor was allowed to vary randomly across hospitals (Model 4). This model incorporated both random intercepts and a random slope. Thus, both the baseline log-odds of death for a reference patient and the effect of the linear predictor were allowed to vary across hospitals. This analysis is very similar to the random effect meta-analysis of hospital-specific calibration slopes, as explored in the companion paper.

Finally, we extended Model 4 to allow the effect of each of the 11 predictors to vary across hospitals, after adjusting for the effect of the linear predictor (Model 5). For this particular set of analyses, we centered the estimated linear predictor around its mean for computational reasons. The interpretation of the hospital-specific effect for the given predictor variable (e.g., age) is as a difference in effect compared to the recalibrated effect as estimated by the previous model. A model of this form has been described previously when examining model validation [10]. Eleven versions of this model were fit, in which the effect of one of the 11 predictors was allowed to vary across hospitals, while the overall effects of the predictors were only allowed to vary according to a calibration slope across hospitals. So, there was essentially a random overall factor for the remaining predictors while we focused on the effect of one specific predictor at a time. We also considered a variant of Model 5, where the effect of the remaining predictors was fixed as in Model 2, which showed similar results. For computational reasons, we were unable to fit a full random coefficients model in which the baseline risk (intercept) and all 11 predictive effects (regression coefficients) varied simultaneously across hospitals.

Exploring temporal heterogeneity

We explored heterogeneity in baseline risk across time and between hospitals using a random intercept model that incorporated a fixed effect denoting temporal period and a fixed effect for the linear predictor estimated previously (Model 6). In this model, the intercept was allowed to vary across hospitals. Thus, this model allows the baseline risk of one-year mortality to vary randomly across hospitals as well as systematically between the two time periods.

We then considered temporal variation in the overall predictor effect by extending Model 6 to include an interaction between temporal period and the linear predictor (Model 7). This model allowed the effect of the linear predictor to differ between the two time periods.

In order to examine whether the effect of individual predictors varied temporally, we considered a further extension of the above model, replacing the linear predictor by the 11 covariates in the EFFECT-HF model (Model 8). The resultant model had 12 main effects (one for the temporal period and 11 for the individual predictors) and 11 interactions (interactions between the temporal period and each of the predictors). Thus, the effect of each of the 11 covariates was allowed to differ between the two time periods.

Simultaneous exploration of geographic and temporal heterogeneity of predictor effects

As an extension to Model 7, we fit a random effects logistic regression model to explore simultaneously geographic and temporal variation in estimated overall predictor effects (Model 9). This model included a random intercept that varied across hospitals, an effect due to the linear predictor that varied across hospitals, a temporal effect that varied across hospitals and an interaction between these two effects that varied across hospitals. This model permits: (i) the effect of the linear predictor to vary between hospitals; (ii) the effect of temporal period to vary across hospitals; (iii) the effect of temporal period on the predictor effects to vary across hospitals. For computational reasons, we did not attempt to fit a full random coefficients model with interaction by time, in which the baseline risk (intercept) and all 11 predictive effects (regression coefficients) could vary simultaneously across hospitals and across time.

Table 6.1 Estimated odds ratios (95% Confidence interval) from fixed effects and random intercept model.

Variable	Model 1 fixed effects model		Model 2 random intercept model	
Age (per year increase)	1.042	(1.038, 1.047)	1.043	(1.0388, 1.0470)
Systolic blood pressure (per mmHg)	0.987	(0.985, 0.988)	0.987	(0.9852, 0.9879)
Respiratory rate (per breath)	1.026	(1.019, 1.032)	1.025	(1.0186, 1.0310)
Serum urea nitrogen	1.105	(1.096, 1.114)	1.106	(1.0966, 1.1148)
low sodium serum concentration (< 136 mEq/L)	1.365	(1.249, 1.493)	1.364	(1.2462, 1.4926)
low serum hemoglobin (< 10.0 g/dL)	1.181	(1.057, 1.319)	1.172	(1.0487, 1.3103)
Cancer	1.668	(1.492, 1.864)	1.682	(1.5038, 1.8816)
Chronic obstructive pulmonary disease	1.331	(1.221, 1.450)	1.329	(1.2190, 1.4496)
Cerebrovascular disease	1.328	(1.207, 1.461)	1.326	(1.2043, 1.4599)
Hepatic cirrhosis	1.910	(1.253, 2.910)	1.914	(1.2534, 2.9241)
Dementia	2.124	(1.877, 2.402)	2.136	(1.8872, 2.4185)

RESULTS

Geographic heterogeneity

The regression coefficients were very similar for Model 1 (fixed effects model that ignored geographic and temporal variation) and Model 2 (random intercept model with hospital-specific random intercepts) (Table 6.1). When using Model 2, the hospital-specific random intercepts were estimated to have the following distribution: $N(\mu = -1.25, \sigma = 0.18)$, with the variance being statistically significantly different from zero (P < 0.0001). From the above distribution, the hospital-specific one-year mortality rates for a reference patient (i.e., one whose standardized covariates were all equal to zero) would lie between 0.167 and 0.292 for 95% of hospitals. The median odds ratio (MOR) (computed using the formula $\mathrm{MOR} = \exp\left(\sqrt{2 \times \sigma^2} \times 0.6745\right)$, where σ^2 is the random effects variance estimated above) was equal to 1.19 [11]. Thus, in comparing the odds of death for an individual at a hospital with a higher risk of death with the odds of death for a similar individual at a hospital with a lower risk of death, the median odds ratio over all possible pair-wise comparison of hospitals was 1.19.

The random intercept model in which the intercept varied across hospitals while the effect of the linear predictor was fixed (Model 3) had the following estimated distribution for the random intercepts: $N(\mu = 0, \sigma = 0.18)$, mirroring the same magnitude of between-hospital variation that was observed above. As expected, the estimated regression coefficient for the linear predictor was close to 1 (1.01).

The random coefficients model in which both the intercept and the effect of the linear predictor were allowed to vary across hospitals (Model 4) was found to provide a marginally statistically significant improvement in fit compared to the prior model in which only the intercept varied across hospitals (P = 0.0478). However, there was only modest evidence that the effect of the linear predictor varied across hospitals. Assuming a normal distribution for the random effects, the effect of the linear predictor on mortality lay between 0.76 and 1.26 for 95% of hospitals.

Finally, we considered the set of 11 random coefficients models in which the intercept, the linear predictor and the effect of one of the covariates were allowed to vary across hospitals (Model 5). For each model, we tested whether the three variance-covariance terms associated with the covariate were simultaneously equal to zero. For two of the models (effect of serum urea and the effect of cancer), the test could not be conducted for computational reasons. Of the remaining nine variables, only the presence of low sodium was found to have an effect that varied across hospitals (P = 0.0096). However, for the remaining eight comparisons, the simpler random coefficients model, in which the effect of the covariate was fixed across hospitals, was found to be acceptable (P > 0.24). For the model in which the effect of low sodium was allowed to vary, the hospital-specific regression coefficients for the effect of low sodium were found to come from the following distribution:

$\mathrm{N}(0.01, \sigma = 0.29)$. Thus, the hospital-specific odds ratios for low sodium lay between 0.57 and 1.80 for 95% of hospitals. For all 11 models, the average effect of the covariate, after adjusting for the linear predictor, was not statistically significant (P > 0.69).

We conclude that there was no strong evidence for heterogeneity in predictor effects, while baseline risks substantially varied between hospitals.

Temporal heterogeneity

The regression coefficient for the main effect of temporal period was -0.111 (odds ratio 0.89, P = 0.0050), in the random intercept logistic model in which the outcome was regressed on the linear predictor and an indicator variable denoting temporal period (Model 6). Thus, the odds of death were 11% lower in the second phase than in the first phase of the study, providing evidence of temporal improvement in the risk of one-year mortality.

The interaction between the linear predictor and the temporal period indicator was not statistically significant (P = 0.883, Model 7). Thus, there was no evidence that the effect of the linear predictor differed between the two time periods.

When the above analysis was repeated with the linear predictor replaced by the 11 individual predictor variables (Model 8), comparable results were observed with one exception: while the effects of 10 of the 11 predictor variables did not change over time (P > 0.067), the effect of cirrhosis differed somewhat between the two time periods (odds ratio 2.98 in Phase 1 vs. 1.12 in Phase 2, P-value for interaction = 0.027).

Simultaneous exploration of geographic and temporal stability

In Model 9 we found no evidence that, on average, the effect of the linear predictor differed between the two time periods (P = 0.99). Furthermore, a test of the hypothesis that the four variance-covariance terms associated with the interaction was not statistically significant (P = 0.88). Consequently, we refit Model 9 after eliminating the interaction term (this removed one fixed effect – the interaction term and four variance-covariance terms – those terms involving the correlation between the random effect for the interaction and the random effects for the other three random effects). In this reduced model, a test of the hypothesis that the three variance-covariance terms associated with either the temporal effect were simultaneously equal to zero was not statistically significant (P = 0.10). Consequently, the effect of time did not vary across hospitals. However, a test of the hypothesis that the five variance-covariance terms involving the linear predictor or the temporal effect were all simultaneously equal to zero was statistically significant (P = 0.03). Thus, there was evidence that the effect of the linear predictor varied across hospitals, even after accounting for the temporal effect.

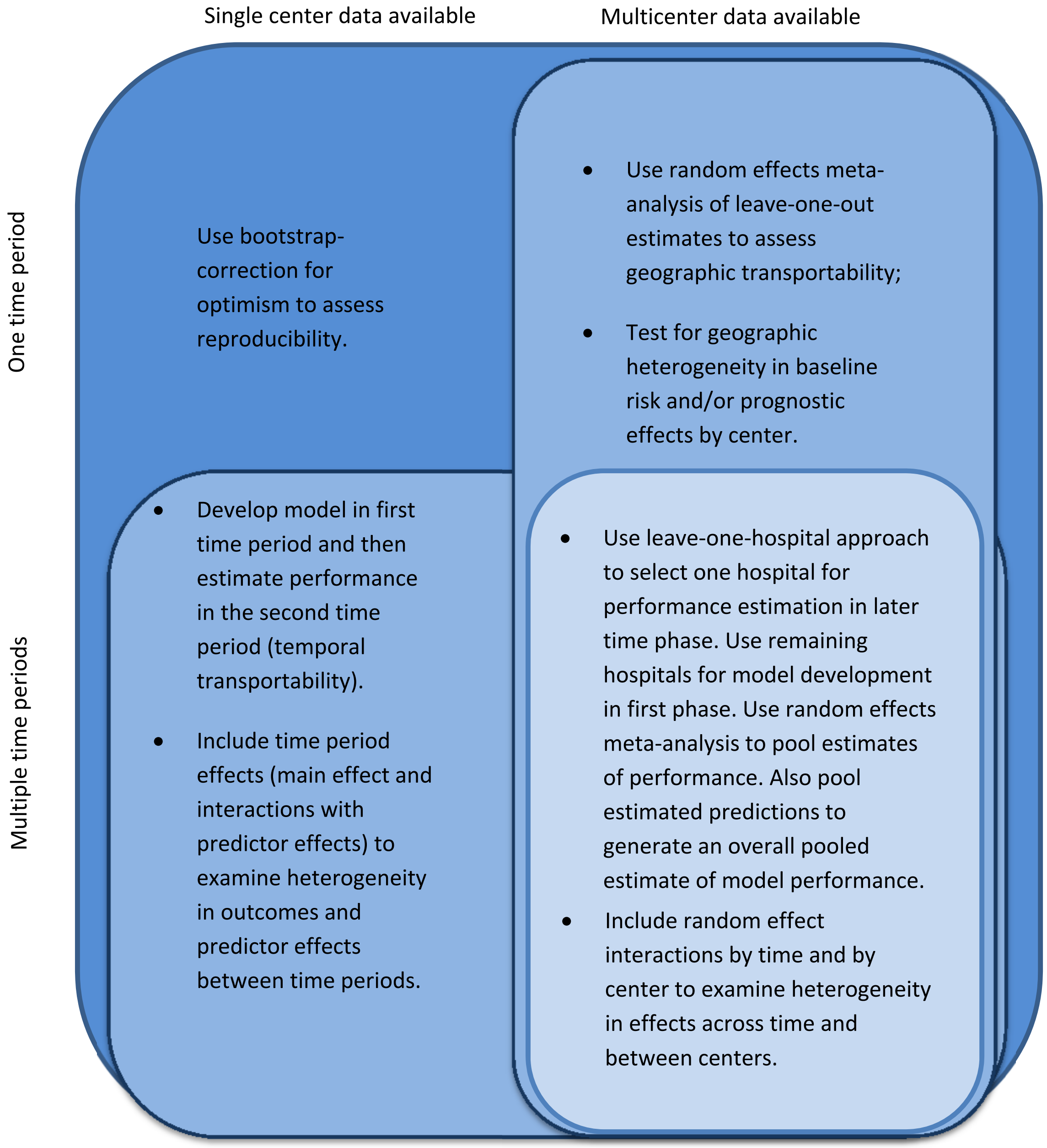
DISCUSSION

Clinical prediction models are intended for widespread application in health care, including use in subjects different from those in whom the model was developed. An emerging aspect of assessing model transportability is assessing the heterogeneity of estimated covariate effects across time and across centres. We illustrated the use of random effects regression models for examining this temporal and geographic heterogeneity in baseline risk and in the estimated predictor effects.

Using data on patients hospitalized with heart failure, we found that temporal and geographic variation in predictor effects was minimal. In contrast, the probability of the occurrence of the outcome ('baseline risk') was found to vary substantially between centres and between time periods. These analyses complement classical methods for assessing model validity reported in the companion article. There, we also found that the EFFECT-HF mortality prediction model displayed good temporal and geographic transportability in terms of discrimination and calibration slope when assessed using an internal-external validation approach. The calibration intercept varied in a similar way to the random effect estimate in the current analyses (Models 3 and 4).

An advantage to the methods illustrated in this paper is that they allow all subjects to be included in model development, without the necessity of withholding some subjects for model validation. This increases model stability, due the larger number of subjects used for model development. In developing prediction models, the desire is for a model that is valid everywhere. The examination of geographic and temporal variation in predictor effects permits an exploration of whether this holds true for a given model. While predictor effects can be anticipated to be fixed geographically or temporally in many settings, this may not be universally true. Certain centres may have more experience and expertise in treating more acutely ill patients, which could diminish the predictive effect of covariates at those hospitals. A more frequent occurrence is that in which the baseline line risk of the outcome varies geographically or temporally. This can result in the developed model displaying lack of calibration when applied in different settings. An example is the validation of the Framingham model to predict cardiovascular disease, in which the baseline risk was found to vary between ethnically-diverse populations [12]. Similar systematic miscalibration was observed for the prediction of indolent prostate cancer in a clinical versus a screening setting [13].

A limitation to relying solely on the methods described in the current paper is the lack of a global measure of model performance such as the c-statistic, the Brier Score, and the generalized R^2 statistic. Such measures can be used for a comparison of the relative performance of competing prediction models. Accordingly, assessing variation in predictor effects can best be seen as complementary, providing important information about the geographic and temporal portability of a particular prediction model.



Use bootstrap-correction for optimism to assess reproducibility.

- Use random effects meta-analysis of leave-one-out estimates to assess geographic transportability;
- Test for geographic heterogeneity in baseline risk and/or prognostic effects by center.

- Develop model in first time period and then estimate performance in the second time period (temporal transportability).
- Include time period effects (main effect and interactions with predictor effects) to examine heterogeneity in outcomes and predictor effects between time periods.

- Use leave-one-hospital approach to select one hospital for performance estimation in later time phase. Use remaining hospitals for model development in first phase. Use random effects meta-analysis to pool estimates of performance. Also pool estimated predictions to generate an overall pooled estimate of model performance.
- Include random effect interactions by time and by center to examine heterogeneity in effects across time and between centers.

Figure 6.1 Recommendations for validating clinical prediction models.

Furthermore, we were unable to fit all of the desired models. We attempted to fit a random coefficients logistic regression model in which the intercept and the effects of all 11 covariates were allowed to vary across hospitals. In Figure 6.1 we summarize graphically some recommendations for assessing geographic and temporal portability of clinical prediction models, based on our analyses in this paper and in the companion study. We provide recommendations for scenarios ranging from the simple, consisting of data from a single centre at a single time period, to the complex, consisting of data from multiple centres or hospitals at multiple time periods. We note that estimates of heterogeneity in baseline risk in the current paper match well with the heterogeneity in calibration intercept in a random effects meta-analysis in the companion paper. Similarly, the limited heterogeneity in effect of the linear predictor was noted here (model 4) and in the meta-analysis of the calibration slope in the companion paper. The extension in the current paper to heterogeneity in effect of individual predictors overall (model 5) or by time (model 8) is not possible in the classical approach to model validation, although this heterogeneity should be reflected in heterogeneity in the c-statistic.

In summary, the estimation-based methods described in the current study complement classical methods for model validation. These methods allow one to directly examine geographic and temporal heterogeneity in baseline risk as well as variation in predictor effects.

ACKNOWLEDGEMENTS

This study was supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). Dr. Austin is supported in part by a Career Investigator award from the Heart and Stroke Foundation. Dr. Lee is supported by a Clinician-Scientist award from the CIHR. Dr. Steyerberg and Mr. Van Klaveren are supported in part by a U award (U01NS086294, value of personalized risk information). Mr. van Klaveren and Dr. Vergouwe are supported in part by the Netherlands Organisation for Scientific Research (grant 917.11.383). The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) data used in the study was funded by a CIHR Team Grant in Cardiovascular Outcomes Research. These datasets were linked using unique, encoded identifiers and analyzed at the Institute for Clinical Evaluative Sciences (ICES).

REFERENCES

1. Justice AC, Covinsky KE, Berlin JA. Assessing the Generalizability of Prognostic Information. *Annals of Internal Medicine* 1999; 130(6):515-524
2. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009; 338:b605
3. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine* 2000; 19(4):453-473
4. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology* 2015; 68(3):279-289. DOI: 10.1016/j.jclinepi.2014.06.018
5. Steyerberg, E. W. and Harrell, F. E., Jr. Prediction models need appropriate internal, internal-external, and external validation. *Journal of Clinical Epidemiology* 2016; 69:245-7. doi: 10.1016/j.jclinepi.2015.04.005
6. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology* 2015; 68(1):25-34. DOI: 10.1016/j.jclinepi.2014.09.007
7. Austin, P. C., van Klaveren D., Vergouwe, Y., Nieboer, D., Lee, D. S., and Steyerberg, E. W. Geographic and temporal validity of prediction models: Different approaches were useful to examine model performance. *Journal of Clinical Epidemiology*. 2016; doi: 10.1016/j.jclinepi.2016.05.007.
8. Tu JV, Donovan LR, Lee DS, Wang JT, Austin PC, Alter DA, Ko DT. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. *Journal of the American Medical Association* 2009; 302(21):2330-2337
9. Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu JV. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. *Journal of the American Medical Association* 2003; 290(19):2581-2587
10. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine* 2004; 23(16):2567-2586. DOI: 10.1002/sim.1844
11. Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, Rastam L, Larsen K. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology and Community Heath* 2006; 60(4):290-297. DOI: 10.1136/jech.2004.029454
12. D'Agostino RB, Sr., Grundy S, Sullivan LM, Wilson P. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA* 2001; 286(2):180-187
13. Steyerberg EW, Roobol MJ, Kattan MW, van der Kwast TH, de Koning HJ, Schroder FH. Prediction of indolent prostate cancer: validation and updating of a prognostic nomogram. *The Journal of Urology* 2007; 177(1):107-112. DOI: 10.1016/j.juro.2006.08.068

APPENDIX 6.1

Mathematical description of statistical models used for studying model variation

Model	Model description	Description
1	$\text{logit}(p_{ij}) = \alpha_0 + \boldsymbol{\beta}\mathbf{X}_{ij}$ where p_{ij} denotes the probability of the outcome for the *i*th patient at the *j*th hospital. From this model, we extracted the linear predictor (LP*ij*).	Fixed effects model, ignoring temporal and geographic heterogeneity
2	$\text{logit}(p_{ij}) = \alpha_{0j} + \boldsymbol{\beta}\mathbf{X}_{ij}$ where the hospital-specific random effects $\alpha_{0j} \sim N(\alpha_0, \sigma^2)$.	Random intercept model, allowing for variation in baseline risk, but assuming common prognostic effects
3	$\text{logit}(p_{ij}) = \alpha_{0j} + \alpha_1 \text{LP}_{ij}$ where the hospital-specific random effects $\alpha_{0j} \sim N(\alpha_0, \sigma^2)$.	Rank 1 model, allowing for common effect of the linear predictor
4	$\text{logit}(p_{ij}) = \alpha_{0j} + \alpha_{1j}\text{LP}_{ij}$ where $\begin{pmatrix} \alpha_{0j} \\ \alpha_{1j} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right)$ The distribution of the random effects was estimated to be $\begin{pmatrix} \alpha_{0j} \\ \alpha_{1j} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} 0.005 \\ 1.008 \end{pmatrix}, \begin{pmatrix} 0.0444 & 0.0139 \\ 0.0139 & 0.0162 \end{pmatrix} \right)$.	Rank 1 model, allowing for heterogeneity in the effect of the linear predictor
5	$\text{logit}(p_{ij}) = \alpha_{0j} + \alpha_{1j}\text{LP}_{ij} + \alpha_{2j} X_{1ij}$ where $\begin{pmatrix} \alpha_{0j} \\ \alpha_{1j} \\ \alpha_{2j} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix} \right)$ and X_{1ij} denotes an individual predictor (e.g., age).	Fully stratified model, allowing for differential prognostic effects (one model per covariate)
6	$\text{logit}(p_{ij}) = \alpha_{0j} + \alpha_1 T_{ij} + \alpha_2 \text{LP}_{ij}$ where the hospital-specific random effects $\alpha_{0j} \sim N(\alpha_0, \sigma^2)$, the fixed effect for LP_{ij} are defined as in Model 3, and T_{ij} denotes the temporal period (T = 0 for Phase 1 vs. T = 1 for Phase 2).	Random intercept model with a fixed main effect for Phase 2 vs Phase 1
7	$\text{logit}(p_i) = \alpha_{0j} + \alpha_1 T_{ij} + \alpha_2 \text{LP}_{ij} + \alpha_3 T_{ij} \times \text{LP}_{ij}$	Random intercept model with a fixed interaction effect for Phase 2 vs Phase 1. The prognostic effect differs between time periods
8	$\text{logit}(p_i) = \alpha_{0j} + \alpha_1 \mathbf{X}_{ij} + \alpha_2 T_{ij} + \alpha_3 T_{ij} \times \mathbf{X}_{ij}$	Random intercept model that allowed effect of each predictor to vary between time periods
9	$\text{logit}(p_{ij}) = \alpha_{0j} + \alpha_{1j} LP_{ij} + \alpha_{2j}\text{T}_{ij} + \alpha_{3j} T_{ij} \times \text{LP}_{ij}$ where $\begin{pmatrix} \alpha_{0j} \\ \alpha_{1j} \\ \alpha_{2j} \\ \alpha_{3j} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{pmatrix} \right)$	The effect of the linear predictor varies between hospitals; the effect of temporal period varies across hospitals; the effect of temporal period on the predictor effects varies across hospitals

7

The calibrated model-based concordance improved assessment of discriminative ability in clustered data of limited size

D van Klaveren
EW Steyerberg
M Gönen
Y Vergouwe

In preparation

ABSTRACT

Objectives Discriminative ability is an important aspect of prediction model performance, but challenging to assess in clustered (e.g. multicenter) data. Concordance (c)-indexes may be too extreme within small clusters. We aimed to define a new approach for the assessment of discriminative ability in clustered data.

Study Design and Setting We assessed discriminative ability of a prediction model for mortality after traumatic brain injury within centers of the CRASH trial. With multilevel regression analysis we estimated cluster-specific calibration slopes which we used to obtain the recently proposed calibrated model-based concordance (*c-mbc*) within each cluster. We compared these *c-mbc*'s with c-indexes in centers of the CRASH trial, and in simulations of clusters with varying slopes.

Results The *c-mbc* was less extreme in distribution than the c-index in 19 European centers (internal validation; *n*=1,716) and 36 non-European centers (external validation; *n*=3,135) of the CRASH trial. In simulations *c-mbc*'s were biased, but less variable than c-indexes, resulting in generally lower root mean squared errors.

Conclusion The *c-mbc*, based on multilevel regression analysis of the calibration slope, is a good alternative to the c-index as a measure of discriminative ability in multicenter studies with patient clusters of limited sample size.

INTRODUCTION

Assessing the performance of a prediction model is of great practical importance. An essential aspect of model performance is separating subjects with good outcome from subjects with poor outcome (discrimination) [1]. Harrell's concordance-index (c-index) is often used to measure discrimination [2]. The c-index estimates the probability that for two randomly chosen subjects with different outcomes the model predicts a higher risk for the subject with poorer outcome (concordance probability). In addition to the c-index, we recently introduced a model-based concordance measure (*mbc*), similar to the concordance probability estimator proposed for proportional hazards regression models by Gönen and Heller [3, 4]. The *mbc* is the expected concordance probability of a regression model under the assumption that the regression coefficients are correct. The *mbc* at external validation is the closed form variant of the previously proposed case-mix corrected c-index [5].The difference between the *mbc* at model development and the *mbc* at external validation indicates the change in discriminative ability attributable to the difference in case-mix heterogeneity between the development and validation data. The calibrated *mbc* (*c-mbc*) – based on predictions recalibrated to the external validation data – also takes (in)correctness of the regression coefficients, including the intercept, into account when measuring the discriminative ability in external data.

In risk modeling, patient data is often clustered. A typical example is multicenter patient data, i.e. data of patients who are treated in different centers. We have suggested summarizing the discriminative ability with random-effects meta-analysis of cluster-specific c-indexes, because the discriminative ability often varies between clusters of patients [6]. However, for small clusters the cluster-specific c-indexes may be too extreme. Extreme estimates are also a problem for cluster-specific calibration intercepts and slopes. Multilevel regression analysis can provide less extreme ("shrunk") random effect estimates, trading off variance with bias [7, 8]. The random effect estimates of calibration intercepts and slopes can also be used for calculation of the *c-mbc*, which is the expected concordance probability under the assumption that the random effect estimates of the calibration intercept and slope are true. Similar to the cluster-specific random intercept and slope estimates, we may expect the cluster-specific *c-mbc*'s to be more stable than the c-indexes.

We aimed to study this new approach for assessment of discriminative ability in clustered data, especially for small clusters. We compare cluster-specific *c-mbc*'s – based on random effect estimates of calibration intercepts and slopes – with cluster-specific c-indexes in a case study with substantial variation in calibration slopes across small clusters. We study the trade-off between variance and bias of cluster-specific c-indexes and *cmbc*'s in a simulation study.

THEORETICAL BACKGROUND

The (calibrated) model-based concordance

The recently proposed *mbc* (equations in Appendix 7.1) estimates a logistic or proportional hazards regression model's concordance probability at apparent validation [4]. The *mbc* is asymptotically equivalent to the c-index, with exact equality when the model contains only one categorical predictor. This *mbc* is a function of the regression coefficients and the covariate distribution and does not use observed outcomes. Consequently, in an external validation population the *mbc* is not influenced by the validity of the regression coefficients and merely assesses the expected discriminative ability of the model, similar to a previously proposed case-mix corrected c-index [10]. To assess the influence of overall regression coefficient validity on the concordance probability, we first estimate the calibration intercept $\gamma_0$and the calibration slope γ_1 in the validation data, i.e. the regression coefficients of a model that regresses the observed outcomes on the linear predictors $X\beta$ in the validation data [9]. If $\hat{\gamma}_1 = 1$, the regression coefficients are on average correct in the validation data. In contrast, $\hat{\gamma}_1 < 1$ indicates a weaker association between the linear predictor and the outcomes in the validation data. The $mbc(\hat{\gamma}_0 + \hat{\gamma}_1 X\beta)$, which we label calibrated model-based concordance (*c-mbc*), incorporates both the influence of case-mix heterogeneity and the overall validity of the regression coefficients β on the discriminative ability of the prediction model. Variance estimates of the *mbc* and the *c-mbc* in model development and external validation settings are easily available as well [4].

The calibrated model-based concordance in clustered data

When data is clustered, we denote with x_{ik} the baseline characteristics vector for patient i in cluster k, and with $z_{ik} = x_{ik}^T\beta$ the corresponding linear predictors of a logistic regression model with regression coefficients β and intercept β_0. We can estimate calibration intercepts γ_{0k} and slopes γ_{1k} for individual clusters with a multilevel logistic regression model [7]:

$$\text{logit}(p_{ik}) = \text{offset}(\beta_0) + \gamma_{0k} + \gamma_{1k} z_{ik}$$

$$\gamma_{0k} \sim N(\gamma_0, \sigma_0^2) \qquad (1)$$

$$\gamma_{1k} \sim N(\gamma_1, \sigma_1^2)$$

Random effects estimates (Best Linear Unbiased Predictors) $\hat{\gamma}_{0k}$ and $\hat{\gamma}_{1k}$ of the intercept and slope in cluster k, can be plugged in equation A5. With $Z_k = X_k\beta$ de linear predictors of patients in cluster k, we obtain the *c-mbc* of a multilevel logistic regression model in cluster *k:*

$$c\text{-}mbc_k = mbc(\beta_0 + \hat{\gamma}_{0k} + \hat{\gamma}_{1k}\, Z_k) \qquad (2)$$

CASE STUDY

We present a case study of predicting mortality after Traumatic Brain Injury (TBI). We used patients enrolled in the Medical Research Council Corticosteroid Randomisation after Significant Head Injury trial (registration ISRCTN74459797, http://www.controlled-trials.com/), who were recruited between 1999 and 2004 [10]. This was a large international double-blind, randomized placebo-controlled trial of the effect of early administration of a 48-h infusion of methylprednisolone on outcome after head injury. We considered patients with moderate or severe brain injury (GCS Total Score ≤ 12) and observed 6-month Glasgow Outcome Scale (GOS) [11, 12]. Patients (*n* = 1,716) who were treated in one of 19 European centers with more than 10 patients experiencing the event were included in the analysis. A logistic regression model was fitted with age, GCS Motor Score and pupil reactivity as covariates, similar to previously developed risk models [13, 14]. To assess model performance within each cluster we estimated cluster-specific calibration intercepts, calibration slopes, and c-indexes. We compared the estimates with random effect estimates of calibration intercepts and slopes (multilevel logistic regression model in equation 1), and *c-mbc*'s (equation 2), respectively. All the analyses were done in R software, multilevel regression analysis was done with the lme4 package [15, 16].

We found substantial heterogeneity in calibration intercepts and slopes ($\sigma_0 = 0.82; \sigma_1 = 0.16$). The cluster-level means of the calibration intercept and slope ($\gamma_0 = 0.24; \gamma_1 = 0.96$) were close to the pooled estimates of the calibration intercept (≡ 0) and the calibration slope (≡ 1). As expected, random effects estimates of calibration intercepts and slopes were less heterogeneous and had narrower 95% confidence intervals than fixed effect estimates (left and middle panels of Figure 7.1; Supplementary table 7.1). Similarly, the *c-mbc*'s based on random effect estimates were less heterogeneous and had narrower 95% confidence intervals than cluster-specific c-indexes (right panel of Figure 7.1). For patients who were treated in one of 36 non-European centers with more than 10 patients experiencing the event (*n* = 3,135), the intercept was poorly calibrated ($\gamma_0 = 1.44$) and the linear predictors slightly overfitted ($\gamma_1 = 0.90$). The heterogeneity in calibration intercepts and slopes was very similar to the European setting ($\sigma_0 = 0.81; \sigma_1 = 0.15$). Differences between fixed effect estimates and random effects estimates, and between c-indexes and *c-mbc*'s were comparable to the European setting (Figure 7.2; Supplementary table 7.2).

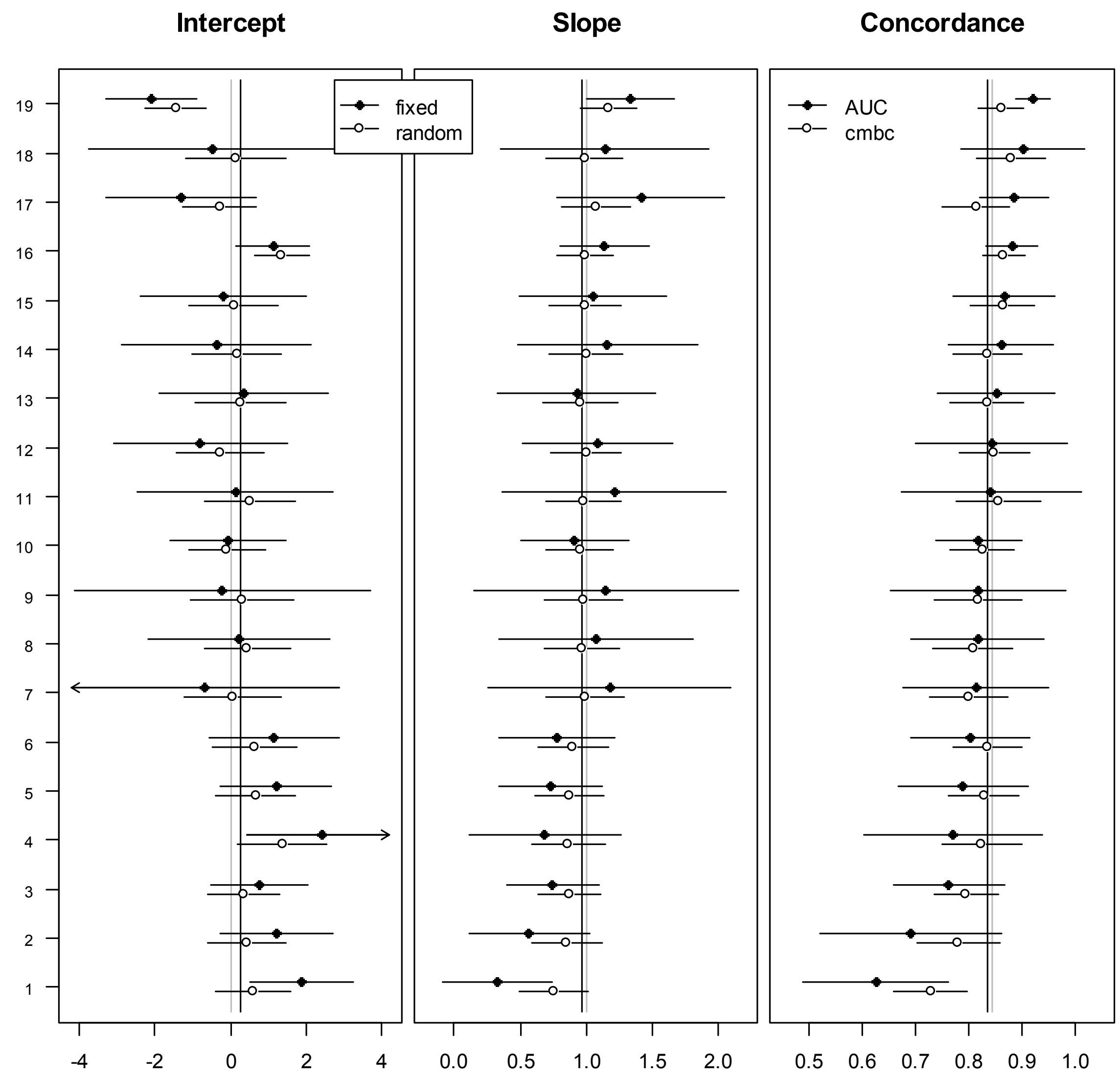


Figure 7.1 Performance measures at internal validation across 19 centers of the CRASH trial. Closed dots represent fixed effect intercept estimates, fixed effect slope estimates and c-indexes in the first, second and third panel respectively. Open dots represent random effects intercept estimates, random effects slope estimates and calibrated model-based concordance estimates in the first, second and third panel respectively. Grey vertical lines represent effect estimates of intercept (0) and slope (1) in the original regression model, together with the expected pooled concordance (0.85). Black vertical lines represent mean effect estimates of intercept (0.24) and slope (0.96) in a multilevel regression model, together with the expected pooled concordance (0.84).

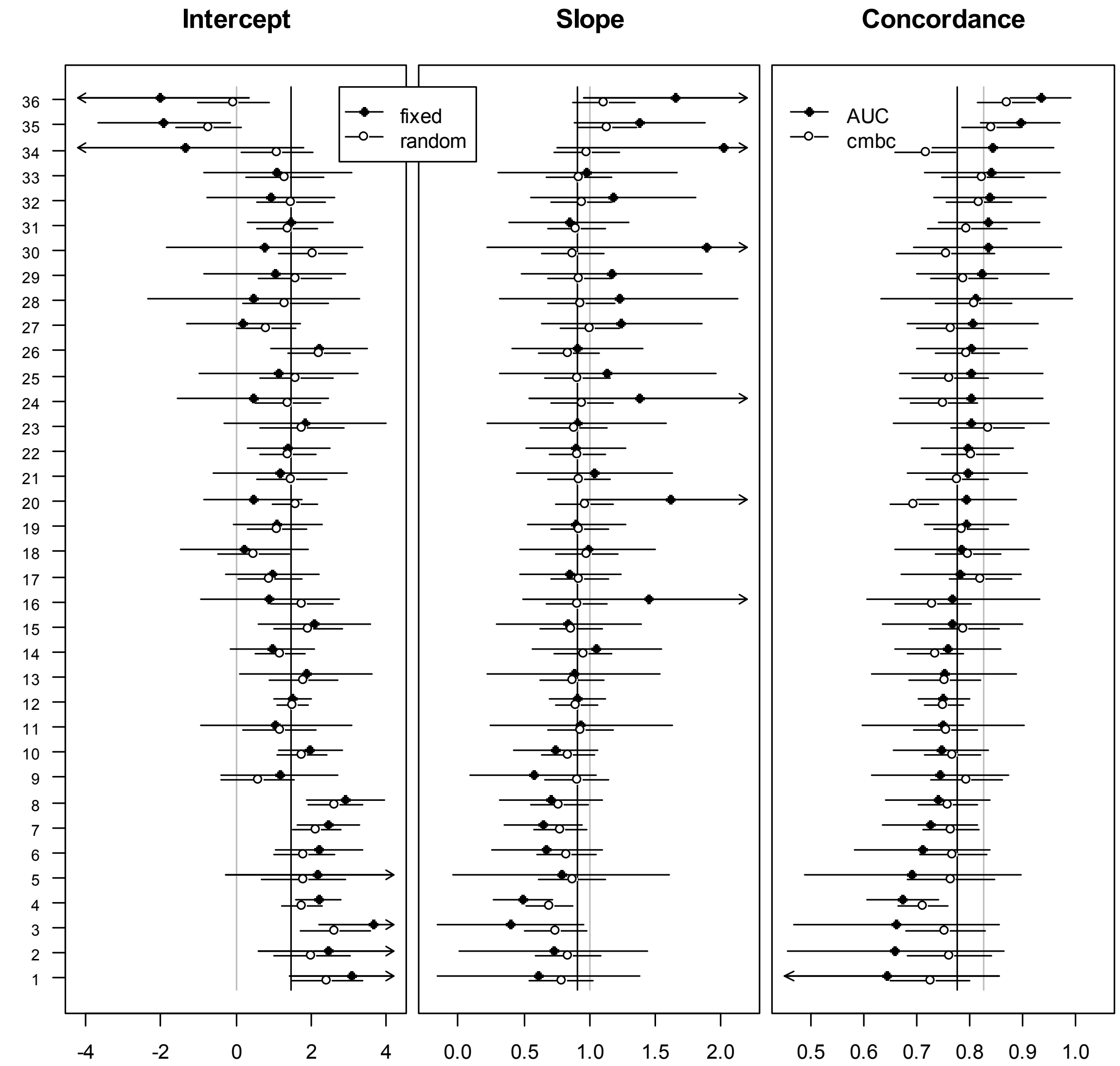


Figure 7.2 Performance measures at external validation across 36 centers of the CRASH trial. Closed dots represent fixed effect intercept estimates, fixed effect slope estimates and c-indexes in the first, second and third panel respectively. Open dots represent random effects intercept estimates, random effects slope estimates and calibrated model-based concordance estimates in the first, second and third panel respectively. Grey vertical lines represent effect estimates of intercept (0) and slope (1) in the original regression model, together with the expected pooled concordance (0.83). Black vertical lines represent mean effect estimates of intercept (1.44) and slope (0.90) in a multilevel regression model, together with the expected pooled concordance (0.78).

SIMULATION STUDY

To study the trade-off between variance and bias of cluster-specific c-indexes and *c-mbc*'s, we simulated validation studies of a logistic regression model in 40 clusters of 200 patients. To incorporate heterogeneity in true intercepts and slopes across clusters, we drew for each cluster k a true calibration intercept γ_{0k} and a true calibration slope γ_{1k} from independent normal distributions with means $\gamma_0 = 0$ and $\gamma_1 = 1$, respectively and standard deviations $\sigma_0 = \sigma_1 = 0.2$.

Table 7.1 Simulation characteristics (2,000 replications) of c-index and *c-mbc* across 40 centers of 200 patients.

	True	Bias		sd		rmse	
Cluster	**concordance**	**c-index**	***c-mbc***	**c-index**	***c-mbc***	**c-index**	***c-mbc***
1	0.664	0.002	0.051	0.055	0.027	0.055	0.057
2	0.666	0.001	0.047	0.059	0.027	0.059	0.054
3	0.695	0.001	0.032	0.053	0.024	0.052	0.040
4	0.699	0.000	0.025	0.056	0.024	0.055	0.034
5	0.702	0.000	0.026	0.054	0.024	0.053	0.034
6	0.705	-0.001	0.030	0.049	0.025	0.048	0.038
7	0.710	-0.002	0.022	0.051	0.023	0.050	0.031
8	0.712	-0.001	0.015	0.056	0.024	0.055	0.027
9	0.712	0.001	0.028	0.048	0.024	0.047	0.035
10	0.718	0.002	0.020	0.051	0.023	0.050	0.029
11	0.726	0.001	0.004	0.056	0.024	0.055	0.022
12	0.727	-0.001	0.001	0.058	0.024	0.057	0.022
13	0.727	0.000	0.017	0.047	0.023	0.046	0.027
14	0.727	0.001	0.010	0.051	0.023	0.050	0.023
15	0.731	0.001	0.008	0.049	0.022	0.048	0.021
16	0.735	-0.001	-0.001	0.053	0.023	0.053	0.021
17	0.735	0.001	0.006	0.050	0.022	0.049	0.021
18	0.738	0.000	0.001	0.052	0.023	0.051	0.020
19	0.741	0.001	0.006	0.047	0.022	0.046	0.020
20	0.744	0.001	0.007	0.045	0.022	0.043	0.020
21	0.750	0.000	0.002	0.046	0.021	0.045	0.019
22	0.751	0.001	-0.004	0.049	0.022	0.047	0.019
23	0.755	0.001	0.006	0.043	0.023	0.041	0.021
24	0.755	0.000	-0.008	0.049	0.022	0.048	0.020
25	0.757	0.002	-0.007	0.046	0.021	0.045	0.020
26	0.757	0.002	-0.013	0.051	0.022	0.049	0.023
27	0.763	-0.002	-0.015	0.050	0.022	0.049	0.024
28	0.763	0.000	-0.014	0.048	0.021	0.047	0.023
29	0.765	0.000	0.000	0.041	0.022	0.040	0.019
30	0.770	0.001	-0.009	0.043	0.022	0.042	0.020
31	0.772	0.002	-0.011	0.043	0.021	0.041	0.021
32	0.774	0.001	-0.015	0.044	0.021	0.042	0.023
33	0.780	0.000	-0.015	0.042	0.021	0.041	0.024
34	0.787	-0.001	-0.029	0.046	0.022	0.044	0.034
35	0.788	0.000	-0.036	0.048	0.023	0.047	0.041
36	0.791	0.001	-0.018	0.039	0.022	0.038	0.026
37	0.795	-0.001	-0.028	0.041	0.022	0.040	0.033
38	0.798	-0.001	-0.031	0.041	0.021	0.040	0.035
39	0.798	0.001	-0.027	0.040	0.022	0.038	0.033
40	0.803	0.000	-0.033	0.040	0.022	0.039	0.038
Mean	0.745	0.000	0.001	0.048	0.023	0.047	0.028

In each of 2,000 replications we generated for patient i in cluster k a continuous baseline linear predictor z_{ik} from a standard normal distribution and a binary outcome from a Bernoulli distribution with success probability $\left[1+\exp\{-(\text{-}2+\gamma_{0k}+\gamma_{1k}z_{ik})\}\right]^{-1}$. We produced cluster-specific (fixed effect) estimates of calibration intercepts and slopes, and c-indexes in each replication. Furthermore, we produced random effect estimates of the calibration intercepts and slopes (multilevel logistic regression model of equation 1), and *c-mbc*'s (equation 2) in each replication.

We summarized the cluster-specific estimates of the calibration intercept, the calibration slope and the concordance probability with the average deviation from the true value (bias), the standard deviation (square root of the variance), and the root of the average squared difference with the true values (root mean squared error [rmse]). We used the $mbc(\text{-}2+\gamma_{0k}+\gamma_{1k}Z_k)$ as the true concordance probability within each cluster, because it is equal to the mean c-index in infinitely many replications of cluster k assuming that γ_{0k} and γ_{1k} are true [4].

Cluster-specific c-indexes were unbiased (Table 7.1). The bias of *c-mbc*'s increased with the deviation of the true cluster-specific concordance probability from the overall average. Due to a positive trade-off with variance (lower standard deviation) the rmse of the *c-mbc* was generally lower than the rmse of the c-index. Similar plots as for the case study (Figures 1 and 2) could be drawn for each replication of the simulation study. We plotted the estimates from the first replication, including true cluster-specific values (Figure 7.3). Again, random effects estimates of calibration intercept and slope, and *c-mbc*'s were less heterogeneous and had narrower 95% confidence intervals than fixed effect estimates and c-indexes, respectively.

We varied simulation settings to visualize the impact on our proposed approach. Without between-cluster heterogeneity of true intercepts and slopes, the random effects estimates, and the *c-mbc*'s were much closer to the true value than the fixed effect estimates and the c-indexes (Supplementary figure 7.1). As a consequence of the unbiasedness of the *c-mbc*, the rmse of *c-mbc* was substantially lower compared to the c-index (Supplementary table 7.3). When we doubled the number of patients in each cluster to 400, the standard deviation of the c-index, the bias of the *c-mbc* and the average difference between the rmse of the *c-mbc* and the rmse of the c-index all were lower than in the simulations with 200 patients in each cluster (Supplementary table 7.4). Finally, we varied the case-mix heterogeneity across clusters by generating a cluster-specific standard deviation of the predictor from a uniform distribution between 0.75 and 1.25, and we reduced overall predictiveness by a true slope of 0.75. Both scenarios were well presented in cluster-specific estimates, by more variation in *c-mbc* (Supplementary figure 7.2) and lower mean *c-mbc* (Supplementary figure 7.3), respectively.

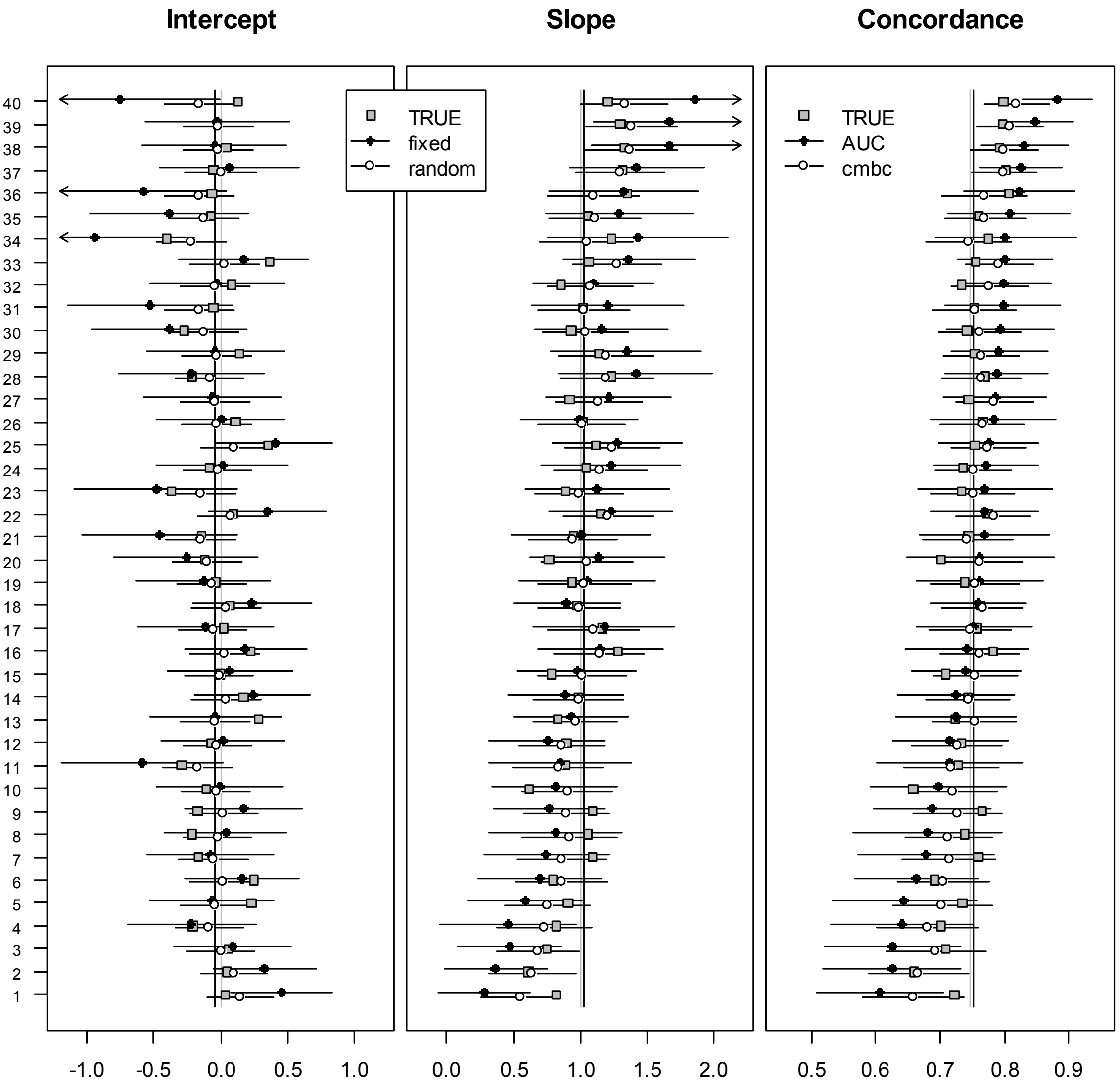


Figure 7.3 Performance measures across 40 centers of 200 simulated patients. Grey squares represent true values of intercept, slope and concordance probability. Closed dots represent fixed effect intercept estimates, fixed effect slope estimates and c-indexes in the first, second and third panel respectively. Open dots represent random effects intercept estimates, random effects slope estimates and calibrated model-based concordance estimates in the first, second and third panel respectively. Grey vertical lines represent effect estimates of intercept and slope (1) in the original regression model, together with the expected pooled concordance. Black vertical lines represent fixed effect estimates of intercept and slope in a multilevel regression model, together with the expected pooled concordance.

DISCUSSION

We proposed a new approach for assessing discriminative ability of prediction models in clustered data. The measure is a modification of the previously proposed calibrated model-based concordance (*c-mbc*), that is the expected concordance under the assumption that the estimates of calibration intercept and slope of the prediction model are true. The *c-mbc* for clustered data uses the random effect estimates of the calibration intercept and slope per cluster provided by a multilevel logistic regression model with the linear predictor as only covariate. The *cmbc* was less extreme in distribution than the c-index in a case study. In simulations with heterogeneous calibration slopes, the random effect estimates of calibration intercept and slope, and thus the *c-mbc*'s were biased, but less variable than the unbiased fixed effect estimates and c-indexes. The trade-off between bias and variance resulted in a generally lower root mean squared error of the *c-mbc* compared to the c-index.

We compared the *c-mbc* based on random effect estimates of the calibration intercept and slope with the c-index. The comparison is basically between a random effect concordance probability estimator and a fixed effect concordance probability estimator, because the c-index is asymptotically equivalent to the *c-mbc* based on fixed effect estimates of the calibration intercept and slope [4]. This explains the observed variance bias trade-off which is typical for the choice between fixed effect and random effect estimates. It is well recognized that unbiasedness is not the only property of an estimator that is important, and that much could be gained by compromising unbiasedness to improve the precision of an estimator [8, 17].

We have recently suggested summarizing the discriminative ability with random-effects meta-analysis of cluster-specific c-indexes, because the discriminative ability often varies between clusters of patients [6, 18]. Random effects meta-analytic techniques inform about the mean and the variation in cluster-specific concordance probabilities, ideally with a prediction interval [19]. However, meta-analytic techniques do not add information about the concordance probability in individual clusters. The techniques proposed in this paper enhance the assessment of discriminative ability in individual clusters of patients.

The patients in our case study were clustered in hospitals. A comparable type of clustering may occur in patients treated in different countries or in patients treated by different caregivers in the same center. Similarly, in public health research the study population is often clustered in geographical regions like countries, municipalities or neighborhoods.

We focused on measuring the performance of logistic regression models in clustered data, using multilevel logistic regression and calibration intercepts, calibration slopes, c-indexes and *c-mbc*'s. This methodology could easily be extended to proportional hazards regression models, based on mixed effects Cox models or shared frailty models, and similar definitions of calibration slopes, c-indexes and *c-mbc*'s in survival data [2, 4, 20].

We initially simulated validation studies of a logistic regression with moderate heterogeneity in true intercepts and slopes across 40 rather small clusters of 200 patients. Obviously, the difference in the rmse of the *c-mbc* compared to the c-index depends on the characteristics of the setting. With negligible heterogeneity in true intercepts and slopes the difference in rsme was higher. With growing numbers of patients per cluster the difference in rsme was lower. Ultimately, the *c-mbc* converges to the c-index with increasing numbers of patients per cluster, because the random effect estimates converge to the fixed effect estimates [4].

The proposed approach depends on the ability of a multilevel regression model to estimate the between-cluster variances of the intercept and the slope. The minimum number of clusters needed to estimate these variances is in the order of 10, but depends on the specific setting [7].

CONCLUSION

The *c-mbc*, based on random effect estimates of the calibration intercept and slope, is a good alternative to the c-index as measure of discriminative ability in clustered data when clusters are of limited size.

ACKNOWLEDGMENTS

The authors express their gratitude to all of the principal investigators of the CRASH trial for providing the data. This work was supported by the Netherlands Organisation for Scientific Research (grant 917.11.383). This work was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA008748.

REFERENCES

1. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; 21: 128-138.
2. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982; 247: 2543-2546.
3. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005; 92: 965-970.
4. van Klaveren D, Gonen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings. *Stat Med* 2016.
5. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010; 172: 971-980.
6. van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. *BMC Med Res Methodol* 2014; 14: 5.
7. Gelman A, Hill J. *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press: Cambridge, 2007.
8. Greenland S. Principles of multilevel modelling. *Int J Epidemiol* 2000; 29: 158-167.
9. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* Springer: New York, 2009.
10. Edwards P, Arango M, Balica L, Cottingham R, El-Sayed H, Farrell B, Fernandes J, Gogichaisvili T, Golden N, Hartzenberg B, Husain M, Ulloa MI, Jerbi Z, Khamis H, Komolafe E, Laloe V, Lomas G, Ludwig S, Mazairac G, Munoz Sanchez Mde L, Nasi L, Olldashi F, Plunkett P, Roberts I, Sandercock P, Shakur H, Soler C, Stocker R, Svoboda P, Trenkler S, Venkataramana NK, Wasserberg J, Yates D, Yutthakasemsunt S, collaborators Ct. Final results of MRC CRASH, a randomised placebo-controlled trial of intravenous corticosteroid in adults with head injury-outcomes at 6 months. *Lancet* 2005; 365: 1957-1959.
11. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. *Lancet* 1974; 2: 81-84.
12. Jennett B, Bond M. Assessment of outcome after severe brain damage. *Lancet* 1975; 1: 480-484.
13. Collaborators MCT, Perel P, Arango M, Clayton T, Edwards P, Komolafe E, Poccock S, Roberts I, Shakur H, Steyerberg E, Yutthakasemsunt S. Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. *BMJ* 2008; 336: 425-429.
14. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JD, Maas AI. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Med* 2008; 5: e165.
15. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/, 2013.
16. Robinson GK. That BLUP is a Good Thing: The Estimation of Random Effects. 1991: 15-32.
17. Efron B. Biased versus unbiased estimation. *Advances in Mathematics* 1975; 16: 259-277.
18. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: Different approaches were useful to examine model performance. *J Clin Epidemiol* 2016.
19. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011; 342: d549.
20. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model.* Springer, 2000.
21. Korn EL, Simon R. Measures of explained variation for survival data. *Stat Med* 1990; 9: 487-503.

APPENDIX 7.1

The model-based concordance (mbc) is a model-based estimator of the concordance probability [4]. The concordance probability is defined as the probability that a model predicts for two randomly chosen patients with different outcomes a higher risk for the patient with poorer outcome. For a given patient population (or cluster of patients) it is the probability that a randomly selected patient pair has concordant predictions and outcomes, divided by the probability that their outcomes are different (not “tied”). Patient i has binary outcome Y_i, baseline characteristics vector x_i, linear predictor $x_i^T\beta$ of a logistic regression model, and prediction $p_i = \text{logit}^{-1}(\beta_0 + x_i^T\beta)$. The probability that a randomly selected patient pair has concordant predictions and outcomes is [21]:

$$P(\text{concordant}) = \frac{1}{n(n-1)}\sum_i\sum_{j\neq i}\left[I(p_i < p_j)P(Y_i < Y_j) + I(p_i > p_j)P(Y_i > Y_j)\right] \quad \text{(A1)}$$

Similarly, the probability that a randomly selected patient pair has unequal outcomes is:

$$P(\text{unequal } Y) = \frac{1}{n(n-1)}\sum_i\sum_{j\neq i}\left[P(Y_i < Y_j) + P(Y_i > Y_j)\right] \quad \text{(A2)}$$

Thus, the concordance probability CP in a patient population is obtained by dividing the probabilities of equation A1 and A2:

$$CP = \frac{\sum_i\sum_{j\neq i}\left[I(p_i < p_j)P(Y_i < Y_j) + I(p_i > p_j)P(Y_i > Y_j)\right]}{\sum_i\sum_{j\neq i}\left[P(Y_i < Y_j) + P(Y_i > Y_j)\right]} \quad \text{(A3)}$$

For a logistic regression model the model-based probabilities $P(Y_i < Y_j)$ are:

$$P(Y_i < Y_j) = P(Y_i = 0)P(Y_j = 1) = \frac{1}{1 + \mathrm{e}^{\beta_0 + x_i^T\beta}}\,\frac{1}{1 + \mathrm{e}^{-(\beta_0 + x_j^T\beta)}} \quad \text{(A4)}$$

Combining equations A3 and A4, and replacing $I(p_i < p_j)$ by $I(x_i^T\beta < x_j^T\beta)$ because the predictions are an increasing function of the linear predictor, results in the model-based concordance (mbc) for logistic regression models:

$$mbc(\beta_0 + X\beta) =$$

$$= \frac{\sum_i \sum_{j\neq i}\left[\frac{I\left(x_i^T\beta < x_j^T\beta\right)}{\left(1+\mathrm{e}^{\beta_0+x_i^T\beta}\right)\left(1+\mathrm{e}^{-\left(\beta_0+x_j^T\beta\right)}\right)} + \frac{I\left(x_i^T\beta > x_j^T\beta\right)}{\left(1+\mathrm{e}^{-\left(\beta_0+x_i^T\beta\right)}\right)\left(1+\mathrm{e}^{\beta_0+x_j^T\beta}\right)}\right]}{\sum_i \sum_{j\neq i}\left[\frac{1}{\left(1+\mathrm{e}^{\beta_0+x_i^T\beta}\right)\left(1+\mathrm{e}^{-\left(\beta_0+x_j^T\beta\right)}\right)} + \frac{1}{\left(1+\mathrm{e}^{-\left(\beta_0+x_i^T\beta\right)}\right)\left(1+\mathrm{e}^{\beta_0+x_j^T\beta}\right)}\right]} \quad (A5)$$

When model predictions may be equal for some combinations of *i* and *j*, e.g. when *x* is a binary marker, we can generalize A5 by using $I\left(p_i \leq p_j\right)$ instead of $I\left(p_i < p_j\right)$.

Supplementary table 7.1 Cluster sizes and performance measures across 19 European centers of the CRASH trial. Fixed effect intercept estimates, fixed effect slope estimates and c-indexes are in the fourth, sixth and eighth column respectively. Random effects intercept estimates, random effects slope estimates and calibrated model-based concordance estimates are in the fifth, seventh and ninth panel respectively.

Cluster	Patients	Events	Intercept		Slope		Concordance	
			fixed	random	fixed	random	c-index	*c-mbc*
1	103	21	1.9 (0.70)	0.6 (0.50)	0.33 (0.21)	0.75 (0.13)	0.63 (0.07)	0.73 (0.04)
2	71	14	1.2 (0.76)	0.4 (0.53)	0.57 (0.23)	0.85 (0.14)	0.69 (0.09)	0.78 (0.04)
3	122	31	0.7 (0.65)	0.3 (0.48)	0.75 (0.18)	0.87 (0.12)	0.76 (0.05)	0.80 (0.03)
4	31	19	2.4 (1.00)	1.4 (0.61)	0.69 (0.29)	0.87 (0.14)	0.77 (0.09)	0.83 (0.04)
5	66	25	1.2 (0.75)	0.7 (0.54)	0.73 (0.20)	0.87 (0.13)	0.79 (0.06)	0.83 (0.03)
6	57	25	1.1 (0.87)	0.6 (0.57)	0.77 (0.22)	0.90 (0.13)	0.80 (0.06)	0.83 (0.03)
7	34	12	-0.7 (1.82)	0.1 (0.66)	1.17 (0.47)	0.99 (0.15)	0.81 (0.07)	0.80 (0.04)
8	41	13	0.2 (1.23)	0.4 (0.59)	1.08 (0.37)	0.97 (0.15)	0.82 (0.06)	0.81 (0.04)
9	23	11	-0.2 (1.99)	0.3 (0.69)	1.15 (0.51)	0.98 (0.15)	0.82 (0.08)	0.82 (0.04)
10	110	25	-0.1 (0.79)	-0.1 (0.52)	0.91 (0.21)	0.95 (0.13)	0.82 (0.04)	0.83 (0.03)
11	30	11	0.1 (1.31)	0.5 (0.62)	1.21 (0.43)	0.97 (0.14)	0.84 (0.09)	0.86 (0.04)
12	61	13	-0.8 (1.17)	-0.3 (0.60)	1.09 (0.29)	1.00 (0.14)	0.84 (0.07)	0.85 (0.03)
13	45	14	0.3 (1.14)	0.3 (0.61)	0.93 (0.30)	0.95 (0.14)	0.85 (0.06)	0.84 (0.04)
14	50	16	-0.4 (1.28)	0.2 (0.60)	1.16 (0.35)	1.00 (0.14)	0.86 (0.05)	0.84 (0.03)
15	53	18	-0.2 (1.11)	0.1 (0.60)	1.05 (0.28)	0.98 (0.14)	0.87 (0.05)	0.86 (0.03)
16	181	107	1.1 (0.49)	1.4 (0.36)	1.14 (0.17)	0.99 (0.11)	0.88 (0.02)	0.87 (0.02)
17	116	16	-1.3 (1.02)	-0.3 (0.49)	1.41 (0.32)	1.07 (0.13)	0.89 (0.03)	0.81 (0.03)
18	28	13	-0.5 (1.67)	0.1 (0.68)	1.14 (0.40)	0.99 (0.15)	0.90 (0.06)	0.88 (0.03)
19	494	41	-2.1 (0.61)	-1.5 (0.41)	1.33 (0.17)	1.17 (0.11)	0.92 (0.02)	0.86 (0.02)

Supplementary table 7.2 Cluster sizes and performance measures at external validation across 36 non-European centers of the CRASH trial. Fixed effect intercept estimates, fixed effect slope estimates and c-indexes are in the fourth, sixth and eighth column respectively. Random effects intercept estimates, random effects slope estimates and calibrated model-based concordance estimates are in the fifth, seventh and ninth panel respectively.

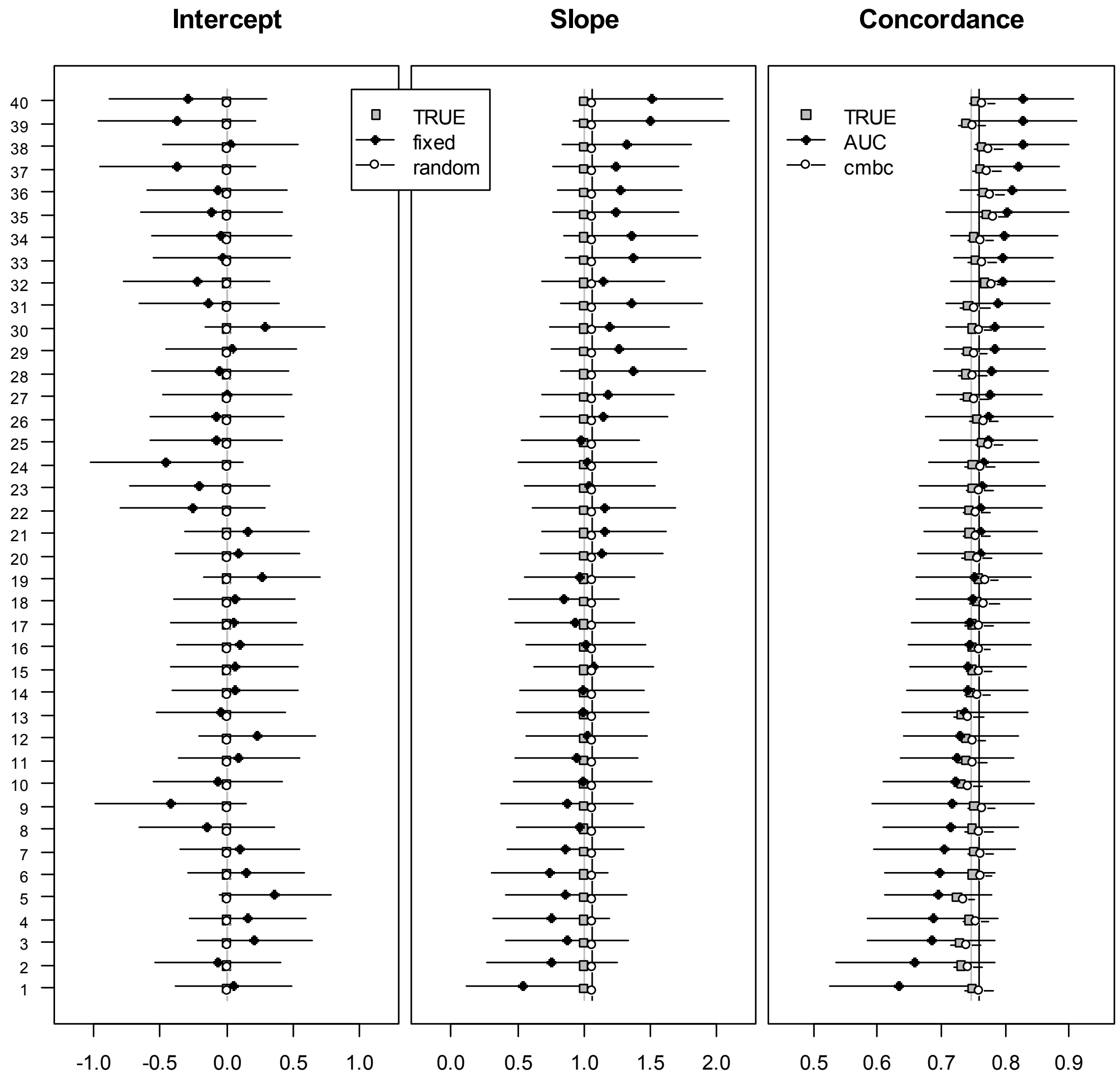
Cluster	Patients	Events	Intercept		Slope		Concordance	
			fixed	random	fixed	random	c-index	*c-mbc*
1	27	14	3.1 (0.85)	2.4 (0.49)	0.61 (0.39)	0.78 (0.12)	0.64 (0.11)	0.73 (0.04)
2	29	15	2.5 (0.95)	2.0 (0.52)	0.73 (0.37)	0.84 (0.13)	0.66 (0.10)	0.76 (0.04)
3	32	19	3.6 (0.73)	2.6 (0.48)	0.40 (0.28)	0.74 (0.12)	0.66 (0.10)	0.75 (0.04)
4	276	84	2.2 (0.31)	1.7 (0.27)	0.50 (0.11)	0.70 (0.09)	0.67 (0.03)	0.71 (0.02)
5	25	14	2.2 (1.23)	1.8 (0.57)	0.78 (0.42)	0.87 (0.13)	0.69 (0.10)	0.77 (0.04)
6	71	30	2.2 (0.59)	1.8 (0.41)	0.67 (0.21)	0.83 (0.11)	0.71 (0.07)	0.77 (0.03)
7	131	63	2.5 (0.42)	2.1 (0.33)	0.65 (0.15)	0.77 (0.10)	0.73 (0.05)	0.76 (0.03)
8	96	61	2.9 (0.53)	2.6 (0.37)	0.71 (0.20)	0.77 (0.11)	0.74 (0.05)	0.76 (0.03)
9	66	12	1.2 (0.80)	0.6 (0.50)	0.57 (0.25)	0.90 (0.12)	0.74 (0.07)	0.79 (0.04)
10	133	50	2.0 (0.43)	1.7 (0.33)	0.74 (0.16)	0.83 (0.10)	0.75 (0.05)	0.77 (0.03)
11	46	15	1.1 (1.02)	1.2 (0.50)	0.93 (0.35)	0.93 (0.13)	0.75 (0.08)	0.76 (0.03)
12	466	146	1.5 (0.26)	1.5 (0.21)	0.90 (0.11)	0.90 (0.08)	0.75 (0.02)	0.75 (0.02)
13	48	24	1.9 (0.90)	1.8 (0.47)	0.88 (0.33)	0.87 (0.12)	0.75 (0.07)	0.75 (0.04)
14	131	33	1.0 (0.58)	1.2 (0.34)	1.05 (0.25)	0.95 (0.11)	0.76 (0.05)	0.74 (0.03)
15	47	24	2.1 (0.77)	1.9 (0.47)	0.84 (0.28)	0.86 (0.12)	0.77 (0.07)	0.79 (0.03)
16	50	17	0.9 (0.94)	1.7 (0.42)	1.45 (0.49)	0.90 (0.12)	0.77 (0.08)	0.73 (0.04)
17	94	28	1.0 (0.63)	0.9 (0.43)	0.85 (0.19)	0.92 (0.11)	0.78 (0.06)	0.82 (0.03)
18	78	19	0.2 (0.86)	0.5 (0.48)	0.99 (0.26)	0.98 (0.12)	0.79 (0.06)	0.80 (0.03)
19	125	49	1.1 (0.61)	1.1 (0.40)	0.90 (0.19)	0.92 (0.11)	0.79 (0.04)	0.78 (0.03)
20	124	36	0.4 (0.67)	1.6 (0.31)	1.62 (0.35)	0.96 (0.11)	0.79 (0.05)	0.70 (0.02)
21	60	30	1.2 (0.91)	1.5 (0.48)	1.04 (0.30)	0.92 (0.12)	0.80 (0.06)	0.78 (0.03)
22	107	43	1.4 (0.56)	1.4 (0.39)	0.89 (0.19)	0.90 (0.11)	0.80 (0.04)	0.80 (0.03)
23	32	20	1.8 (1.09)	1.7 (0.57)	0.91 (0.35)	0.88 (0.13)	0.80 (0.08)	0.84 (0.04)
24	52	17	0.5 (1.03)	1.4 (0.45)	1.38 (0.43)	0.95 (0.12)	0.80 (0.07)	0.75 (0.03)
25	37	16	1.1 (1.07)	1.6 (0.50)	1.14 (0.42)	0.91 (0.13)	0.80 (0.07)	0.76 (0.04)
26	65	38	2.2 (0.66)	2.2 (0.42)	0.90 (0.25)	0.84 (0.11)	0.80 (0.05)	0.80 (0.03)
27	96	20	0.2 (0.76)	0.8 (0.40)	1.24 (0.31)	1.00 (0.12)	0.81 (0.06)	0.76 (0.03)
28	27	13	0.5 (1.43)	1.3 (0.58)	1.22 (0.46)	0.94 (0.13)	0.81 (0.09)	0.81 (0.04)
29	44	20	1.0 (0.96)	1.6 (0.49)	1.17 (0.35)	0.92 (0.12)	0.83 (0.06)	0.79 (0.03)
30	34	15	0.7 (1.33)	2.0 (0.46)	1.88 (0.85)	0.87 (0.12)	0.84 (0.07)	0.75 (0.05)
31	71	20	1.4 (0.58)	1.4 (0.41)	0.84 (0.23)	0.90 (0.11)	0.84 (0.05)	0.80 (0.04)
32	56	26	0.9 (0.87)	1.5 (0.46)	1.18 (0.32)	0.94 (0.12)	0.84 (0.05)	0.82 (0.03)
33	34	13	1.1 (1.00)	1.3 (0.53)	0.98 (0.35)	0.92 (0.13)	0.84 (0.07)	0.83 (0.04)
34	44	12	-1.4 (1.62)	1.1 (0.48)	2.02 (0.65)	0.98 (0.13)	0.84 (0.06)	0.72 (0.03)
35	185	19	-1.9 (0.90)	-0.7 (0.45)	1.38 (0.25)	1.13 (0.12)	0.90 (0.04)	0.84 (0.03)
36	96	19	-2.0 (1.20)	-0.1 (0.48)	1.65 (0.36)	1.11 (0.12)	0.94 (0.03)	0.87 (0.03)

Supplementary table 7.3 Simulation characteristics (2,000 replications) of c-index and *c-mbc* across 40 centers of 200 patients without between center heterogeneity in intercept and slope.

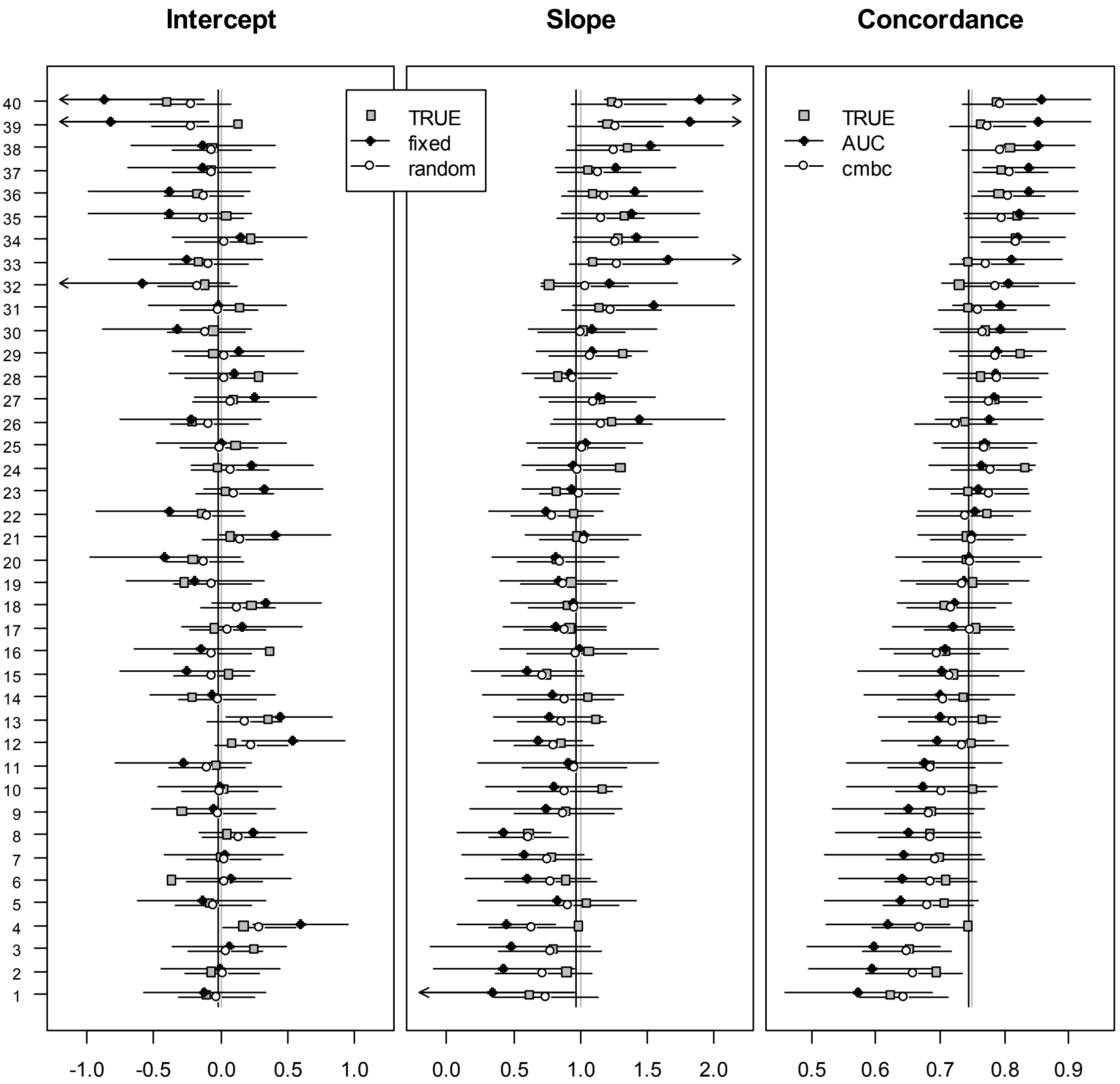
	True	Bias		sd		rmse	
Cluster	concordance	c-index	*c-mbc*	c-index	*c-mbc*	c-index	*c-mbc*
1	0.746	0.001	0.001	0.047	0.014	0.046	0.009
2	0.746	-0.002	0.000	0.049	0.014	0.048	0.009
3	0.746	0.001	0.001	0.049	0.014	0.048	0.009
4	0.746	0.000	0.000	0.046	0.014	0.045	0.009
5	0.746	0.002	0.001	0.048	0.014	0.047	0.009
6	0.746	0.000	0.000	0.048	0.014	0.047	0.009
7	0.746	0.000	0.000	0.049	0.014	0.047	0.009
8	0.746	0.002	0.001	0.049	0.014	0.047	0.009
9	0.746	-0.001	0.000	0.049	0.014	0.047	0.009
10	0.746	0.001	0.000	0.048	0.014	0.047	0.009
11	0.746	-0.001	0.000	0.047	0.014	0.046	0.009
12	0.746	0.001	0.000	0.049	0.014	0.047	0.009
13	0.746	0.001	0.001	0.047	0.014	0.046	0.009
14	0.746	0.001	0.000	0.048	0.014	0.046	0.009
15	0.746	0.000	0.000	0.048	0.014	0.047	0.009
16	0.746	-0.002	0.000	0.046	0.013	0.045	0.009
17	0.746	0.002	0.001	0.046	0.014	0.045	0.009
18	0.746	0.000	0.000	0.048	0.014	0.047	0.009
19	0.746	0.001	0.000	0.049	0.014	0.047	0.009
20	0.746	0.000	0.000	0.047	0.014	0.046	0.009
21	0.746	0.002	0.000	0.048	0.014	0.047	0.009
22	0.746	0.001	0.000	0.047	0.014	0.046	0.009
23	0.746	0.000	0.001	0.047	0.014	0.046	0.009
24	0.746	0.001	0.000	0.047	0.013	0.046	0.009
25	0.746	0.001	0.001	0.048	0.014	0.046	0.009
26	0.746	0.000	0.000	0.049	0.014	0.048	0.009
27	0.746	0.000	0.000	0.048	0.014	0.047	0.009
28	0.746	-0.002	0.000	0.047	0.013	0.046	0.009
29	0.746	0.000	0.001	0.048	0.014	0.047	0.009
30	0.747	0.000	0.000	0.048	0.014	0.047	0.009
31	0.747	0.002	0.001	0.048	0.014	0.047	0.009
32	0.747	0.001	0.000	0.047	0.014	0.046	0.009
33	0.747	0.001	0.000	0.048	0.014	0.046	0.009
34	0.747	0.002	0.001	0.047	0.014	0.046	0.009
35	0.747	-0.001	0.000	0.048	0.014	0.047	0.009
36	0.747	0.000	0.000	0.048	0.014	0.047	0.009
37	0.747	0.000	0.000	0.048	0.014	0.046	0.009
38	0.747	0.001	0.001	0.047	0.014	0.046	0.009
39	0.747	-0.002	0.000	0.047	0.014	0.046	0.009
40	0.747	0.001	0.001	0.047	0.014	0.046	0.009
Mean	0.746	0.000	0.000	0.048	0.014	0.047	0.009

Supplementary table 7.4 Simulation characteristics (2,000 replications) of c-index and *c-mbc* across 40 centers of 400 patients.

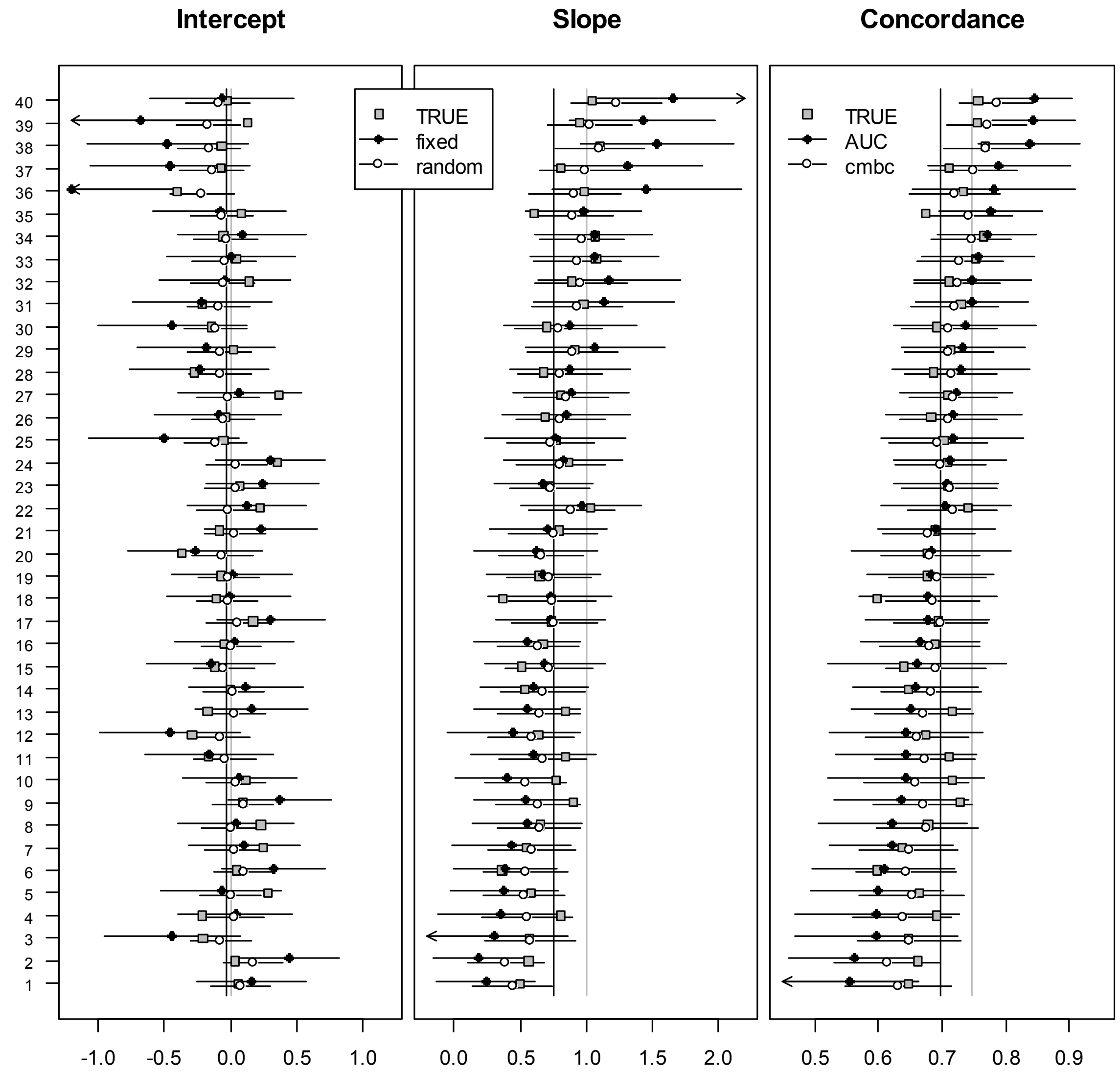
	True	Bias		sd		rmse	
Cluster	concordance	c-index	*c-mbc*	c-index	*c-mbc*	c-index	*c-mbc*
1	0.664	0.000	0.038	0.039	0.025	0.039	0.045
2	0.666	-0.001	0.033	0.043	0.025	0.042	0.041
3	0.695	0.000	0.024	0.038	0.022	0.038	0.032
4	0.699	-0.001	0.017	0.039	0.022	0.039	0.027
5	0.703	0.000	0.019	0.037	0.021	0.037	0.028
6	0.705	0.001	0.025	0.035	0.022	0.034	0.032
7	0.710	-0.001	0.016	0.037	0.021	0.036	0.025
8	0.712	-0.001	0.021	0.033	0.021	0.033	0.029
9	0.712	0.000	0.009	0.038	0.021	0.038	0.022
10	0.718	0.001	0.015	0.034	0.020	0.034	0.024
11	0.726	-0.001	0.000	0.039	0.020	0.038	0.019
12	0.727	0.000	0.006	0.036	0.020	0.035	0.020
13	0.727	0.001	-0.003	0.039	0.021	0.038	0.020
14	0.727	0.001	0.015	0.033	0.020	0.032	0.024
15	0.732	0.002	0.006	0.035	0.019	0.034	0.019
16	0.735	0.001	-0.003	0.038	0.020	0.038	0.019
17	0.735	0.000	0.004	0.035	0.019	0.035	0.018
18	0.738	0.000	-0.001	0.037	0.020	0.036	0.018
19	0.741	0.000	0.005	0.032	0.018	0.032	0.018
20	0.743	0.000	0.006	0.033	0.019	0.032	0.018
21	0.750	0.000	0.002	0.032	0.018	0.031	0.017
22	0.751	0.001	-0.003	0.033	0.019	0.033	0.017
23	0.755	0.001	0.007	0.029	0.019	0.028	0.019
24	0.755	0.000	-0.006	0.033	0.019	0.032	0.018
25	0.757	-0.002	-0.007	0.034	0.018	0.033	0.018
26	0.757	0.000	-0.012	0.034	0.018	0.034	0.021
27	0.763	0.000	-0.013	0.034	0.019	0.033	0.021
28	0.764	0.000	-0.013	0.034	0.018	0.033	0.021
29	0.764	0.000	0.002	0.029	0.019	0.028	0.017
30	0.770	0.000	-0.005	0.030	0.018	0.029	0.017
31	0.772	0.000	-0.008	0.030	0.017	0.029	0.017
32	0.775	0.000	-0.011	0.030	0.018	0.029	0.019
33	0.780	0.001	-0.009	0.029	0.018	0.028	0.018
34	0.786	0.000	-0.023	0.032	0.018	0.031	0.028
35	0.788	0.000	-0.032	0.033	0.019	0.032	0.036
36	0.791	0.000	-0.012	0.027	0.018	0.026	0.019
37	0.795	0.000	-0.020	0.030	0.018	0.029	0.026
38	0.798	0.000	-0.020	0.028	0.018	0.027	0.025
39	0.798	0.000	-0.023	0.029	0.018	0.028	0.028
40	0.803	0.000	-0.025	0.029	0.018	0.028	0.030
Mean	0.745	0.000	0.001	0.034	0.019	0.033	0.023



Supplementary figure 7.1 Performance measures across 40 centers of simulated 200 patients without between center heterogeneity in intercept and slope. Grey squares represent true values of intercept, slope and concordance probability. Closed dots represent fixed effect intercept estimates, fixed effect slope estimates and c-indexes in the first, second and third panel respectively. Open dots represent random effects intercept estimates, random effects slope estimates and calibrated model-based concordance estimates in the first, second and third panel respectively. Grey vertical lines represent effect estimates of intercept (0) and slope (1) in the original regression model, together with the expected pooled concordance (0.76). Black vertical lines represent mean effect estimates of intercept (0.00) and slope (1.06) in a multilevel regression model, together with the expected pooled concordance (0.75).



Supplementary figure 7.2 Performance measures across 40 centers of 200 simulated patients with varying case-mix heterogeneity. Grey squares represent true values of intercept, slope and concordance probability. Closed dots represent fixed effect intercept estimates, fixed effect slope estimates and c-indexes in the first, second and third panel respectively. Open dots represent random effects intercept estimates, random effects slope estimates and calibrated model-based concordance estimates in the first, second and third panel respectively. Grey vertical lines represent effect estimates of intercept and slope (1) in the original regression model, together with the expected pooled concordance. Black vertical lines represent fixed effect estimates of intercept and slope in a multilevel regression model, together with the expected pooled concordance.



Supplementary figure 7.3 Performance measures across 40 centers of 200 simulated patients with reduced overall predictiveness. Grey squares represent true values of intercept, slope and concordance probability. Closed dots represent fixed effect intercept estimates, fixed effect slope estimates and c-indexes in the first, second and third panel respectively. Open dots represent random effects intercept estimates, random effects slope estimates and calibrated model-based concordance estimates in the first, second and third panel respectively. Grey vertical lines represent effect estimates of intercept and slope (1) in the original regression model, together with the expected pooled concordance. Black vertical lines represent fixed effect estimates of intercept and slope in a multilevel regression model, together with the expected pooled concordance.

PART II

HETEROGENEITY OF TREATMENT EFFECT

8

Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions

D van Klaveren
Y Vergouwe
V Farooq
PW Serruys
EW Steyerberg

Journal of Clinical Epidemiology 2015
Awarded with the Journal of Clinical Epidemiology Young Investigator Award 2015

ABSTRACT

Objectives We aimed to compare modelling approaches to estimate the individual survival benefit of treatment with either coronary artery bypass graft surgery (CABG) or percutaneous coronary intervention (PCI) for patients with complex coronary artery disease.

Study Design and Setting We estimated survival with Cox regression models that included the treatment variable (CABG/PCI) interacting with either an internally developed overall prognostic index or with individual prognostic factors. We analyzed data of patients who were randomized in the SYNTAX trial (1800 patients, 178 deaths).

Results A negligible interaction with the prognostic index (p=0.51) led to 4-year survival estimates in favor of CABG for all patients. In contrast, individual interactions indicated substantial relative treatment effect heterogeneity (overall interaction p=0.004), and estimates of 4-year survival were numerically in favor of CABG for 1275 of 1800 patients (71%; 519 with 95% confidence). To test the more complex model with individual interactions we first employed penalized regression, resulting in smaller but largely consistent individual estimates of the survival difference between CABG and PCI. Second, strong treatment interactions were confirmed at external validation in 2891 patients from a multinational registry.

Conclusion Modelling strategies that omit interactions may result in misleading estimates of absolute treatment benefit for individual patients with the potential hazard of suboptimal decision making.

WHAT IS NEW?

Key findings

- Modelling treatment interactions with prognostic factors, rather than a constant relative treatment effect, caused a major shift in the predicted most favorable treatment among the SYNTAX trial patients.
- The model with treatment interactions was supported by a better model fit, robustness in penalized regression analyses, and external validation.

What this adds to what was known

- Although relative treatment effect is often considered to be constant in clinical trials, it may differ substantially across patients and influence the optimal choice of treatment for individual patients.

What is the implication, and what should change now

- We recommend careful analysis of treatment interactions in clinical trial data to reveal possible relative treatment effect heterogeneity and to optimize individual treatment decision making.

INTRODUCTION

Randomized clinical trials provide strong evidence of the benefits and harms of treatments. The estimated overall treatment effect is an important summary result of a clinical trial, but is insufficient to decide which treatment is best suited for an individual patient [1, 2]. Stratified medicine aims to make optimal treatment decisions for individual patients by predicting their response to treatment (treatment benefit) from baseline information. To make optimal decisions it has been suggested to compare absolute treatment benefit – the difference between relevant outcomes in treated and control groups (e.g. mortality reduction) – under different treatment strategies [3]. The absolute treatment benefit for individual patients depends on their risk, e.g. 1-year mortality in the absence of treatment ("baseline risk") since patients at low risk have little to gain from treatment. The absolute benefit is often well estimated by assuming a constant relative risk reduction from a specific treatment. For example, when a treatment has a constant relative 1-year mortality reduction of 20% across patients, the absolute treatment benefit of a patient with 10% baseline mortality will be 2% (20%*10%), twice the absolute treatment benefit of 1% for a patient with 5% baseline mortality (20%*5%). In contrast heterogeneity in the relative risk reduction from a specific treatment (relative treatment effect heterogeneity) would make that absolute treatment effects differ for patients with equal baseline risk. For example, two patients with baseline risk of 10%, but different relative risk reductions of 10% and 20% show absolute risk reductions of 1% (10%*10%) or 2% (10%*20%), respectively [4].

Individual baseline risk can well be assessed with prognostic factors summarized in a prognostic index [5]. Relative treatment effect heterogeneity can be assessed following various approaches. One attractive option is to model a treatment interaction with the prognostic index that represents baseline risk [6-8]. This approach is more parsimonious than considering statistical interactions with each of the prognostic factors. The latter more flexible modelling approach might be reasonable if we expect that treatment response depends on one or more prognostic factors, e.g. because of different underlying biological mechanisms [4]. Such a factor may be referred to as a predictive factor for differential treatment effect [4] or treatment effect modifier [9] . Although modelling of treatment interactions has been recommended [10], it is sensitive to the pitfall of finding false-positive or false-negative subgroup effects [11-15].

We aimed to compare different modelling approaches to estimation of the absolute treatment effect for complex coronary artery disease (CAD) patients, who are treated with either coronary artery bypass graft surgery (CABG) or percutaneous coronary intervention (PCI). We consider relative treatment effect heterogeneity by modelling treatment interactions with a prognostic index and with individual prognostic factors. We specifically aimed to study how to assess the validity of using treatment interactions for guiding treatment decisions.

METHODS

Patient data

We analyzed data of 1800 patients with unprotected left main coronary artery (ULMCA) disease or de-novo three-vessel disease. Patients were randomized on a 1:1 basis to either CABG or PCI with first generation drug eluting stents in the Synergy between Percutaneous Coronary Intervention with Taxus and Cardiac Surgery (SYNTAX) trial (ClinicalTrials.gov, number NCT00114972) [16, 17]. We used eight prognostic factors for mortality (Table 8.1; 178 deaths during 4 years of follow up) which were associated with mortality in either or in both treatment arms [18, 19]: SYNTAX score (www.syntaxscore.com [20, 21]), ULMCA disease, age, female gender, creatinine clearance, left ventricular ejection fraction (LVEF), peripheral vascular disease (PVD) and chronic obstructive pulmonary disease (COPD). To complete a small number of missing prognostic factor values (SYNTAX score 0.6%; creatinine clearance 9.9%; LVEF 1.6%) we employed a multiple imputation strategy (5 imputations; aregImpute function in R package Hmisc [22, 23]). This strategy takes all aspects of uncertainty in the imputations into account by using the bootstrap for drawing predicted values from a full Bayesian predictive distribution.

Table 8.1 Characteristics of patients enrolled in the SYNTAX trial.

Factor	Level	Patients	All-cause Death	Kaplan Meier 4-year survival (95% CI)	Univariable Hazard ratio (95% CI)
SYNTAX score	<23	573	47	91.4 (89.1,93.8)	1.0
	23-32	610	60	89.7 (87.3,92.2)	1.2 (0.8,1.8)
	>=33	606	71	87.9 (85.2,90.5)	1.5 (1.0,2.1)
Age	<60	566	23	95.8 (94.1,97.5)	1.0
	61-69	595	53	90.8 (88.4,93.2)	2.2 (1.4,3.7)
	>=70	639	102	83.2 (80.2,86.2)	4.3 (2.7,6.7)
Creatinine Clearance (mL/min)	<70	546	86	83.8 (80.7,87.0)	1.0
	71-94	546	38	92.8 (90.6,95.0)	0.4 (0.3,0.6)
	>=94	546	35	93.4 (91.3,95.5)	0.4 (0.3,0.6)
LVEF (%)	<50	347	53	84.1 (80.2,88.1)	1.0
	>=50	1425	123	91.0 (89.5,92.5)	0.5 (0.4,0.7)
ULMCA disease	no	1095	101	90.3 (88.6,92.1)	1.0
	yes	705	77	88.7 (86.3,91.1)	1.2 (0.9,1.6)
Gender	female	402	51	86.6 (83.2,90.1)	1.0
	male	1398	127	90.6 (89.0,92.1)	0.7 (0.5,1.0)
COPD	no	1646	148	90.6 (89.2,92.1)	1.0
	yes	154	30	79.8 (73.6,86.6)	2.4 (1.6,3.5)
Peripheral vascular disease	no	1623	135	91.3 (89.9,92.7)	1.0
	yes	177	43	74.7 (68.5,81.6)	3.3 (2.3,4.7)
Treatment	CABG	897	74	91.2 (89.3,93.1)	1.0
	PCI	903	104	88.3 (86.2,90.4)	1.3 (1.0,1.8)

Abbreviations: SYNTAX, Synergy between Percutaneous Coronary Intervention with Taxus and Cardiac Surgery; CI, confidence interval; LVEF, left ventricular ejection fraction; ULMCA, unprotected left main coronary artery; COPD, chronic obstructive pulmonary disease; CABG, coronary artery bypass graft surgery; PCI, percutaneous coronary intervention.

Modelling of relative treatment effects

We employed Cox proportional hazards regression models (R package rms [22, 24]) to analyze the effects of the treatment and prognostic factors on all-cause mortality. First we fitted a model ignoring the treatment to define a prognostic index (PI) for mortality, regardless of the treatment [25]. The hazard function $h(t)$ for a patient with prognostic factors $\mathbf{x}_{\mathbf{PF}}$(SYNTAX score; ULMCA disease; age; female gender; creatinine clearance; LVEF; PVD; COPD) was modeled by the baseline hazard function $h_0(t)$ times the exponent of the linear predictor $\mathbf{x'}_{\mathbf{PF}}\boldsymbol{\pi}_{\mathbf{PF}}$. The effect of each prognostic factor on the log hazard is modeled by the parameters in $\boldsymbol{\pi}_{\mathbf{PF}}$:

$$h(t) = h_0(t)\exp[\mathbf{x'}_{\mathbf{PF}}\boldsymbol{\pi}_{\mathbf{PF}}]$$

The prognostic performance of the PI was quantified by Harrell's c-index [26].

Then we fitted 4 prognostic models. Model 1 contained only an overall treatment effect β_T:

$$h(t) = h_0(t)\exp[x_T\beta_T]$$

The parameter β_T represents the relative increase in log hazard for treatment with PCI ($x_T = 1$) versus treatment with CABG ($x_T = 0$). In model 2 we adjusted for the same prognostic factors $\mathbf{x_{PF}}$ as in the prognostic index:

$$h(t) = h_0(t)\exp[x_T\beta_T + \mathbf{x'_{PF}}\,\boldsymbol{\beta}_{\mathbf{PF}}]$$

The resulting $\hat{\beta}_T$ has been recommended as an efficient estimate of a constant relative treatment effect [27-29]. Model 2 allows for calculation of absolute risk predictions for individual patients, assuming a constant relative treatment effect across all patients. Model 3 included the treatment ($x_T = 1$ for PCI), the internally developed prognostic index ($PI = \mathbf{x'_{PF}}\,\hat{\boldsymbol{\pi}}_{\mathbf{PF}}$) and their interaction $x_T\,PI$:

$$h(t) = h_0(t)\exp[x_T\,\beta_T + PI\,\beta_{PI} + x_T\,PI\,\beta_{T*PI}]$$

The interaction effect estimate $\hat{\beta}_{T*PI}$ expresses the heterogeneity in relative treatment effect for patients with different baseline risk [7]. Model 4 comprised the treatment ($x_T = 1$ for PCI), the prognostic factors ($\mathbf{x_{PF}}$) and a treatment interaction $x_T\mathbf{x_{PF}}$ for each individual prognostic factor:

$$h(t) = h_0(t)\exp[x_T\beta_T + \mathbf{x'_{PF}}\,\boldsymbol{\beta}_{\mathbf{PF}} + x_T\mathbf{x'_{PF}}\,\boldsymbol{\beta}_{\mathbf{T*PF}}]$$

The interaction effect estimates in $\hat{\boldsymbol{\beta}}_{\mathbf{T*PF}}$ express the difference – when comparing treatment with PCI ($x_T = 1$) versus treatment with CABG ($x_T = 0$) – in effect on the log hazard of each prognostic factor. The prognostic factors are each considered here as predictive of the relative treatment effect.

Confirmation of statistical interactions

Statistical significance of interactions in models 3 and 4 was quantified by the p-value of the overall likelihood ratio test statistic with 1 and 8 degrees of freedom respectively [30]. We used Akaike's Information Criterion (AIC) to balance the goodness-of-fit of models 1-4 with their complexity [30].

To examine the sensitivity of the predicted most favorable treatment for overfitting of interaction effects, we employed Cox regression with an L2 (ridge) penalty [31, 32] in model 4 (R-package penalized [22, 33]). We compared penalized with unpenalized interaction effect estimates. Moreover, for each patient in the SYNTAX trial, we compared the most favorable treatment resulting from penalized regression versus that resulting from unpenalized regression.

To further assess the validity of the interaction effect estimates, an external validation study was done. We analyzed patients from the Drug Eluting stent for Left main coronary Artery disease (DELTA) Registry, a multinational, non-randomized, all-comers registry including 2891 patients with ULMCA disease [34]. We compared estimates of interaction effects in the DELTA registry with those in the SYNTAX trial.

Comparison of absolute treatment effects

Based on each of the 4 Cox regression models, we estimated individual patient absolute 4-year risk of all-cause mortality, both for CABG and PCI. We determined a confidence interval for the difference in individual PCI and CABG risk predictions by using the covariance matrix of the parameter estimates.

We assumed that the optimal treatment would be chosen based on the highest absolute treatment effect, expressed as difference in 4-year mortality. To assess the survival benefit of using one model over the other for such treatment decision making we employed a reclassification analysis. To estimate the survival benefit of one model over the other, we could use – for patients with a different treatment recommendation – either the predicted or the observed difference in 4-year mortality. We chose to use the observed mortality difference since it is less sensitive for overfitting of the models: individual risk estimates are used for determining the treatment recommendations, but not for estimating the mortality difference. We estimated the gain in survival by multiplying the proportion of patients with different treatment recommendations with their observed mortality difference in the randomized treatment (CABG and PCI) arms. These gains were also graphically shown in a "benefit graph", as a visual representation of the reclassification analysis [35].

RESULTS

The prognostic index (hazard ratios in Table 8.2) well discriminated high-risk from low-risk patients (c-index 0.73). The variation in individual 4-year mortality predictions (IQR 4.4%-12.7%) implied substantial differences in absolute treatment effect, when the relative treatment effect was assumed constant across patients.

We visualized the results of models 1-4 with scatterplots of the predicted log hazards for CABG versus PCI (Figure 8.1). The overall treatment effect in model 1 was in favor of CABG (unadjusted HR [95% CI] for PCI vs. CABG 1.35 [1.00-1.82]; p-value 0.049; Table 8.2). Adjusting for prognostic factors in model 2 also led to an overall treatment effect in favor of CABG (adjusted HR [95% CI] for PCI vs. CABG 1.47 [1.08-1.98]; p-value 0.013; Table 8.2). The interaction of treatment with the prognostic index in model 3 was far from statistically significant (p-value 0.51), i.e. the relative treatment effect was hardly dependent on baseline risk (Table 8.2). Based on the predictions of model 3, CABG was favored for all 1800 patients.

Table 8.2 Hazard ratios (95% CI) for: the prognostic index (combining the effect of all prognostic factors); models 1-4; model 4 with penalized regression; and external validation of model 4.

	Prognostic index	Model 1	Model 2	Model 3	Model 4	Penalized regression	External validation
PCI vs CABG		1.35 (1.00, 1.82)	1.47 (1.08, 1.98)	1.35 (0.93, 1.97)	1.46 (0.99, 2.16)	1.26	1.31 (0.97, 1.76)
	HR_{CABG}						
SYNTAX sc (10)	1.13 (1.00, 1.29)		1.14 (1.01, 1.29)		0.97 (0.79, 1.18)	1.02	1.12 (0.95, 1.32)
Age (10 yr)	1.53 (1.23, 1.89)		1.52 (1.23, 1.88)		1.88 (1.34, 2.64)	1.52	1.46 (1.15, 1.85)
CrCl (10 mL/min)	0.86 (0.78, 0.96)		0.86 (0.78, 0.95)		0.91 (0.77, 1.07)	0.89	0.91 (0.78, 1.06)
LVEF (10%)	0.66 (0.54, 0.81)		0.65 (0.53, 0.80)		0.84 (0.61, 1.16)	0.83	0.59 (0.47, 0.75)
ULMCA disease	1.06 (0.79, 1.43)		1.06 (0.78, 1.43)		1.47 (0.93, 2.34)	1.33	
Women	1.13 (0.80, 1.58)		1.12 (0.79, 1.57)		0.59 (0.32, 1.10)	0.75	0.52 (0.31, 0.87)
COPD	1.87 (1.25, 2.79)		1.87 (1.25, 2.79)		2.84 (1.64, 4.90)	2.46	3.63 (1.31, 10.04)
PVD	2.55 (1.79, 3.63)		2.64 (1.85, 3.76)		2.79 (1.66, 4.71)	2.53	1.37 (0.68, 2.79)
Prognostic index				2.60 (2.05, 3.31)			
	HR_{PCI}						
SYNTAX sc (10)	1.13 (1.00, 1.29)		1.14 (1.01, 1.29)		1.27 (1.08, 1.50)	1.24	1.32 (1.20, 1.46)
Age (10 yr)	1.53 (1.23, 1.89)		1.52 (1.23, 1.88)		1.29 (0.97, 1.71)	1.34	1.34 (1.19, 1.52)
CrCl (10 mL/min)	0.86 (0.78, 0.96)		0.86 (0.78, 0.95)		0.82 (0.72, 0.93)	0.83	0.93 (0.86, 1.00)
LVEF (10%)	0.66 (0.54, 0.81)		0.65 (0.53, 0.80)		0.56 (0.43, 0.73)	0.56	0.57 (0.50, 0.65)
ULMCA disease	1.06 (0.79, 1.43)		1.06 (0.78, 1.43)		0.82 (0.54, 1.23)	0.86	
Women	1.13 (0.80, 1.58)		1.12 (0.79, 1.57)		1.70 (1.11, 2.60)	1.59	1.09 (0.82, 1.46)
COPD	1.87 (1.25, 2.79)		1.87 (1.25, 2.79)		1.35 (0.74, 2.47)	1.40	1.97 (0.88, 4.42)
PVD	2.55 (1.79, 3.63)		2.64 (1.85, 3.76)		2.79 (1.72, 4.53)	2.78	1.77 (1.01, 3.09)
Prognostic index				2.91 (2.33, 3.63)			
	HR_{PCI}/HR_{CABG}						
SYNTAX sc (10)	1		1		1.32 (1.01, 1.71)	1.22	1.18 (0.98, 1.42)
Age (10 yrr)	1		1		0.69 (0.44, 1.07)	0.88	0.92 (0.70, 1.21)
CrCl (10 mL/min)	1		1		0.89 (0.73, 1.10)	0.93	1.02 (0.86, 1.21)
LVEF (10%)	1		1		0.67 (0.44, 1.00)	0.68	0.96 (0.72, 1.27)
ULMCA disease	1		1		0.56 (0.30, 1.03)	0.65	
Women	1		1		2.87 (1.35, 6.07)	2.12	2.09 (1.16, 3.76)
COPD	1		1		0.48 (0.21, 1.08)	0.57	0.54 (0.20, 1.47)
PVD	1		1		1.00 (0.49, 2.04)	1.10	1.29 (0.51, 3.22)
Prognostic index				1.12 (0.81, 1.55)			
Df	8	1	9	10	17		
AIC	2516	2633	2511	2513	2504		

Hazard ratios are presented for CABG (HR_{CABG}) and PCI (HR_{PCI}) separately since models 3 and 4 assume different prognostic effects for CABG and PCI. The interaction effects of the prognostic factors with the treatment are illustrated by HR_{PCI}/HR_{CABG}. AIC and degrees of freedom are listed for the prognostic index and model 1-4 for comparison of model adequacy.

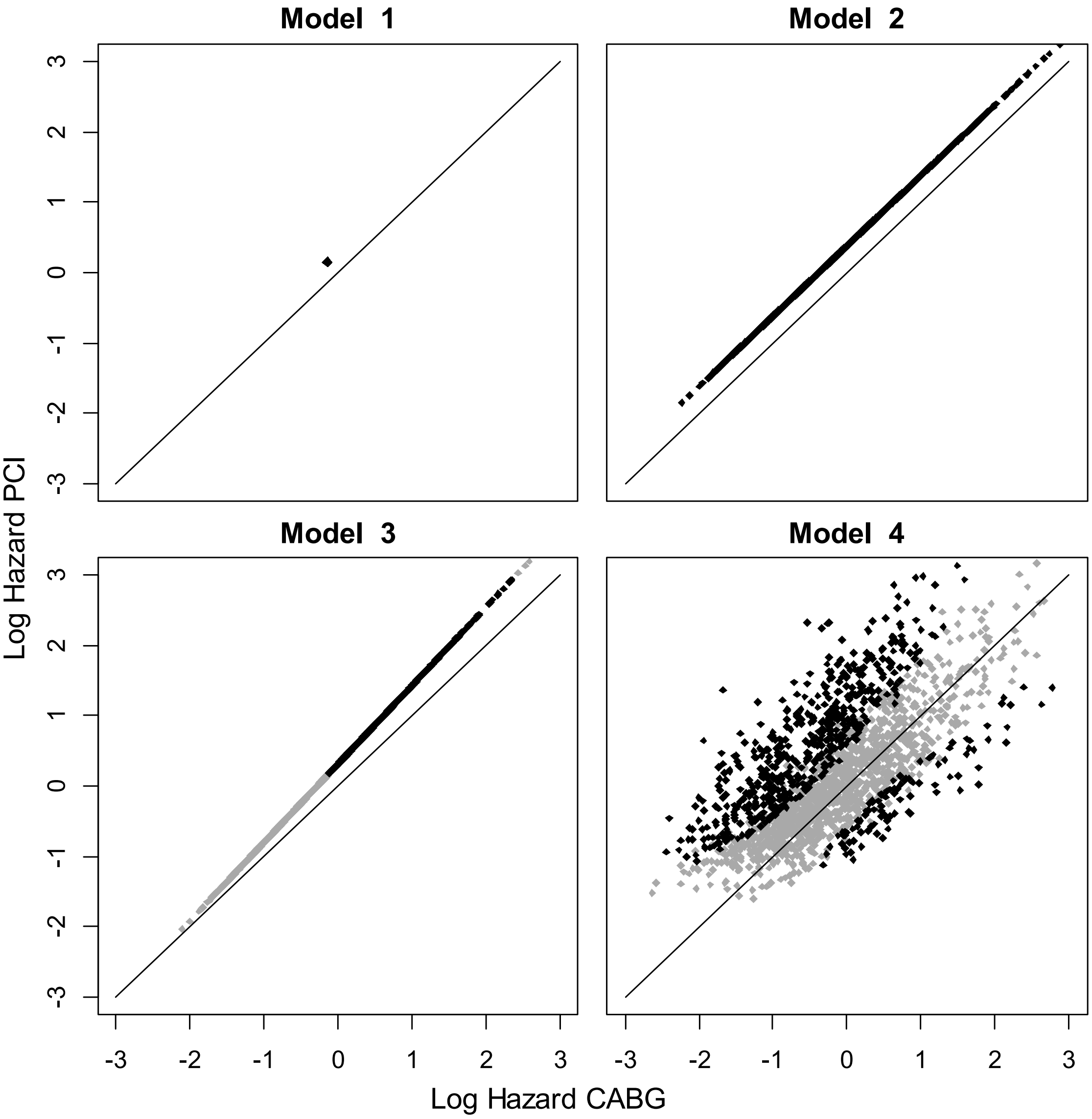


Figure 8.1 CABG vs. PCI log hazard predictions for each patient by model. When 0 is outside the 95% confidence interval for the difference between CABG and PCI predictions, the patient dots are colored black, otherwise they are colored grey.

However, for 1003 patients the predictions were in favor of CABG with less than 95% confidence. Adding the flexibility of treatment interactions with each prognostic factor in model 4 showed substantial heterogeneity of relative treatment effect with a p-value of 0.004 for the overall interaction test based on 8 degrees of freedom. When balancing for model complexity, model 4 still showed the optimal adequacy with an AIC of 2504 against 2511 and 2513 for model 2 and 3 respectively (Table 8.2). Moderate to strong interactions were observed with SYNTAX score, age, LVEF, ULMCA disease, COPD, and female gender (interaction p<0.10). This more flexible model caused a major shift in the predicted most favorable treatment among the 1800 SYNTAX patients. Estimates of 4-year survival were in favor of PCI (PCI 91.4%; CABG 87.2%) for 525 patients (98 with 95% confidence) and in favor of CABG (CABG 93.0%; PCI 86.5%) for 1275 patients (519 with 95% confidence).

When applying penalized regression to model 4, the interaction effect estimates were by definition shrunken, but the strongest interactions effects (SYNTAX score, age, LVEF, ULMCA disease, gender and COPD) remained substantial (Table 8.2). We visualized which treatment was favorable, i.e. had the lowest log hazard prediction according to the penalized regression model, split by patients for whom model 4 predicted favorable outcome with PCI (left panel of Figure 8.2) and for whom model 4 predicted favorable outcome with CABG (right panel of Figure 8.2). With penalized regression, 91.0% of the treatment recommendations for the 1800 patients were equal to model 4; All treatment recommendations were equal for the 617 patients for whom model 4 gave a recommendation with 95% confidence (black dots in Figure 8.2). A penalized regression analysis without penalties on the main effects, i.e. penalties on the interactions effects only, led to very similar treatment recommendations (results not shown). The effect estimates, particularly the strongest interaction effect estimates (SYNTAX score, gender and COPD), were similar in the DELTA registry (Table 8.2).

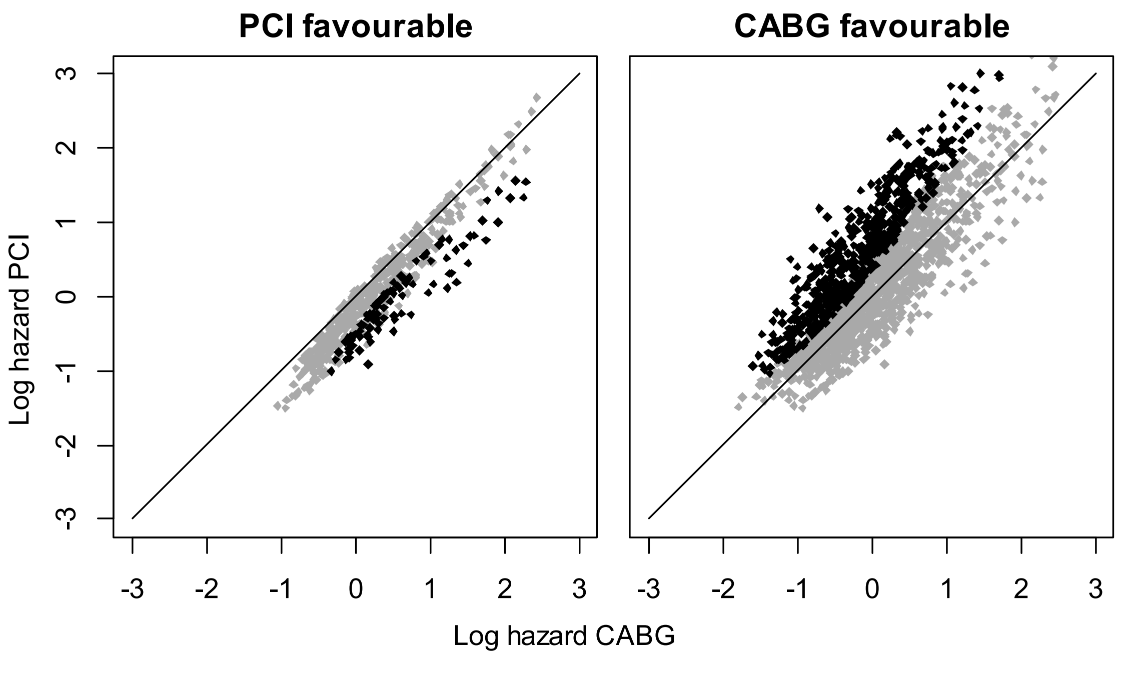


Figure 8.2 CABG vs. PCI log hazard predictions based on penalized regression. Left panel shows patients for whom model 4 predicted favorable outcome with PCI. Right panel shows patients for whom model 4 predicted favorable outcome with CABG. Black dots: 95% confidence interval for the difference between CABG and PCI predictions of model 4 does not include 0; grey dots: 95% confidence interval does include 0.

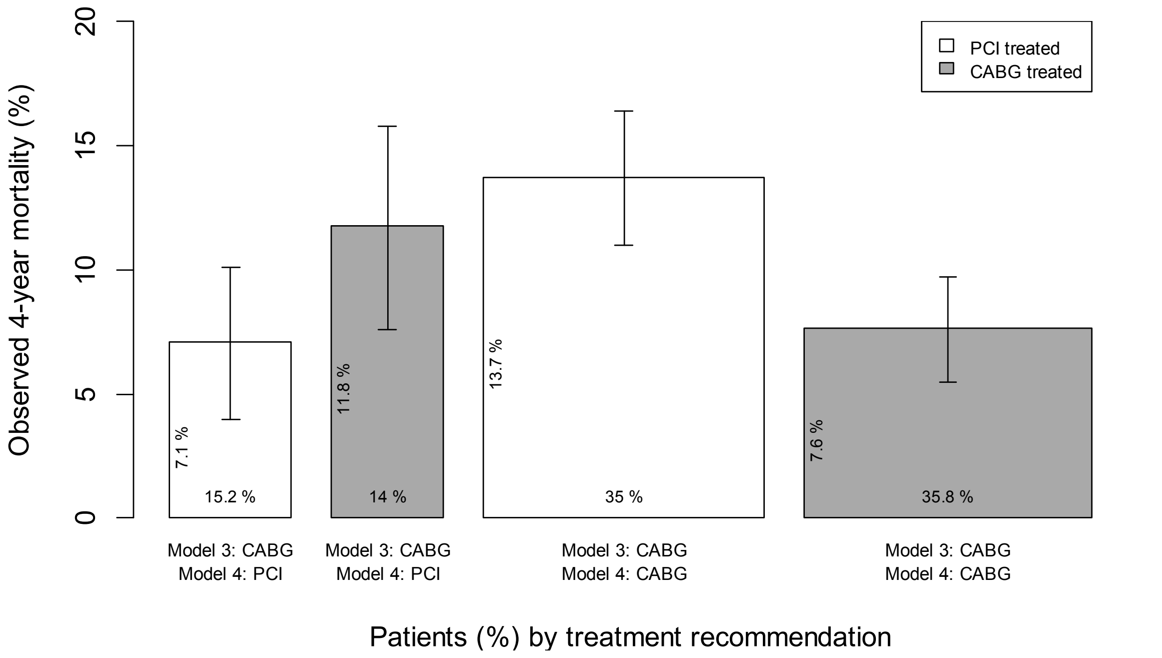


Figure 8.3 Benefit graph of using absolute risk predictions for decision making between PCI and CABG. The width of the bars represents the number of patients in the SYNTAX trial; the height of the bars represents the observed 4-year mortality in the SYNTAX trial (whiskers indicate 95% confidence intervals). Reclassification from PCI according to model 3 to either PCI or CABG according to model 4 does not appear in the graph, since model 3 only recommends CABG.

The difference in survival benefit of using model 4 instead of model 3 for making treatment decisions can be derived from the benefit graph in Figure 8.3. The benefit graph visualizes for the combinations of treatment recommendation according to model 3 and model 4 (CABG/PCI; CABG/CABG), the proportion of patients (width of the bars) and the observed mortality (height of the bars) in both randomized treatment arms (PCI and CABG). For patients for whom both models recommended CABG, CABG (7.6% 4-year mortality) was more effective than PCI (13.7%), but for patients for whom model 4 recommended PCI, PCI (7.1%) was more effective than CABG (11.8%). We estimated a 1.4% survival benefit of using model 4 instead of model 3 by multiplication of the 29% of patients with a different treatment recommendation (total width of the first two bars in Figure 8.3) with their 4.7% observed difference in 4-year mortality between the randomized treatment arms (difference in height of the first two bars in Figure 8.3).

DISCUSSION

We explored different modelling approaches for estimating heterogeneity in treatment effect across individual patients with complex CAD where CABG or PCI could be performed. Treatment interactions with each of the prognostic factors (model 4) fitted much better to the data compared to the treatment interaction with predicted prognosis as a single prognostic index (model 3). Penalized regression, specifically shrinking treatment interactions to the average treatment effect [31, 32] led to largely similar decisions for the individual patients, although relative risk differences – and consequently most absolute risk differences – between CABG and PCI predictions were smaller. These interactions were largely confirmed at external validation in the DELTA registry [19]. The major differences in expected treatment benefit for individual patients between models 3 and 4 (Figure 8.1**),** together with the survival benefit of using model 4 for making treatment decisions, indicated the importance of allowing for treatment interaction with each of the prognostic factors.

Our study confirms that prognostic effects may be very important for estimation of the individual benefit of treatment, since absolute treatment benefit will be higher for those at high risk when the relative risk reduction of a treatment is constant across patients [7, 25]. Furthermore, covariate adjustment may lead to an individualized and efficient treatment effect estimate [27-29]. We therefore recommend to always include prognostic factors when estimating treatment effects that should support decision making for individual patients.

The causal effect of treatment for an individual patient in our study is the difference between the outcome when the patient would have been treated with CABG and the outcome when the same patient would have been treated with PCI [36]. To estimate the difference in outcome when treatment choices are changed, we used randomized data from the all comers SYNTAX trial. In contrast, the use of observational data may produce biased treatment effect estimates when the variation between differently treated patients is not completely controlled for. Specifically, a recent study concluded that documented surgical ineligibility is common and associated with significantly increased long-term mortality among CAD patients undergoing PCI, even after adjustment for known risk factors [37].

We used an internally developed prognostic index for modelling the interaction of the treatment with baseline risk, because it will be relatively easy to obtain in future studies. Although an externally developed baseline risk score is attractive [25], it has been shown with simulations of randomized clinical trials that internally developed baseline risk scores – blinded to the treatment – produce relatively unbiased estimates of treatment effects across the spectrum of risk and are preferred to risk scores developed on the control population [38].

Although sub-group analysis based on interactions may be considered superior to classical sub-group analysis of single factors separately [10], it has similar pitfalls, such as a risk of false-positive findings if large numbers of interactions are assessed, and lack of power to detect interaction effects [11-15]. Similar to classical subgroup effect testing, our approach requires a clear biological motivation for differential mechanisms of treatment effects. In our study, more complex anatomy of the vessel makes PCI treatment a relatively less attractive treatment option. In other studies, when treatment modalities are less different or sample size is small, there may be less potential for predicting differential treatment effect. Ideally, the analysis of differential treatment effect focuses on confirming pre-specified interactions, but exploratory analyses of differential treatment effects could be considered if sample size is large. Exploratory analyses require even more emphasis on careful modelling strategies, model interpretation and model validation [5, 25, 32]. We focused on the overall significance of all interactions considered, similar to overall tests in prediction modelling [32]. We did not select interactions based on statistical significance of individual terms in the multivariable analysis, which might be considered as an alternative modelling approach.

We assumed that treatment decisions would be made on the basis of 4-year survival predictions. In clinical practice, a multidisciplinary heart team will also consider patient preferences, economic costs [39, 40] and other clinical outcomes – namely myocardial infarction, stroke and all-cause revascularization – compared to mortality alone.

CONCLUSION

This study illustrates that different modelling strategies may result in very different estimates of absolute treatment benefit for individual patients. Modelling treatment interactions with individual prognostic factors may be superior to a single interaction with a prognostic index to guide individualized decision making. Further validation and prospective evaluation of this approach across different settings is required.

ACKNOWLEDGMENTS

The authors express their gratitude to all of the principal investigators of the SYNTAX trial and the DELTA trial for providing the data. This work was supported by the Netherlands Organisation for Scientific Research (grant 917.11.383) and the National Institutes of Health (grant U01NS086294).

REFERENCES

1. Rothwell PM. Can overall results of clinical trials be applied to all patients? Lancet. 1995;345:1616-9.
2. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q. 2004;82:661-87.
3. Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. Lancet. 2005;365:256-65.
4. Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ. 2013;346:e5793.
5. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2009.
6. Hayward AC, Goldsmith K, Johnson AM, Surveillance Subgroup of S. Report of the Specialist Advisory Committee on Antimicrobial Resistance (SACAR) Surveillance Subgroup. J Antimicrob Chemother. 2007;60 Suppl 1:i33-42.
7. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA. 2007;298:1209-12.
8. Pocock SJ, Lubsen J. More on subgroup analyses in clinical trials. N Engl J Med. 2008;358:2076; author reply -7.
9. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
10. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Med Res Methodol. 2006;6:18.
11. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet. 2000;355:1064-9.
12. Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? Am Heart J. 2006;151:257-64.
13. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine--reporting of subgroup analyses in clinical trials. N Engl J Med. 2007;357:2189-94.
14. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol. 2004;57:229-36.
15. Schmidt AF, Groenwold RH, Knol MJ, Hoes AW, Nielen M, Roes KC, et al. Exploring interaction effects in small samples increases rates of false-positive and false-negative findings: results from a systematic review and simulation study. J Clin Epidemiol. 2014;67:821-9.
16. Ong AT, Serruys PW, Mohr FW, Morice MC, Kappetein AP, Holmes DR, Jr., et al. The SYNergy between percutaneous coronary intervention with TAXus and cardiac surgery (SYNTAX) study: design, rationale, and run-in phase. Am Heart J. 2006;151:1194-204.
17. Serruys PW, Morice MC, Kappetein AP, Colombo A, Holmes DR, Mack MJ, et al. Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. N Engl J Med. 2009;360:961-72.
18. Farooq V, Vergouwe Y, Raber L, Vranckx P, Garcia-Garcia H, Diletti R, et al. Combined anatomical and clinical factors for the long-term risk stratification of patients undergoing percutaneous coronary intervention: the Logistic Clinical SYNTAX score. European heart journal. 2012;33:3098-104.
19. Farooq V, van Klaveren D, Steyerberg EW, Meliga E, Vergouwe Y, Chieffo A, et al. Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score II. Lancet. 2013;381:639-50.

20. SYNTAX score calculator: www.syntaxscore.com. SYNTAX working-group. Launched 19th May 2009.
21. Sianos G, Morel MA, A.P. K, Morice MC, Colombo A, Dawkins K, et al. The SYNTAX Score: an angiographic tool grading the complexity of coronary artery disease. EuroIntervention2005 p. 219-27.
22. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/. 3-900051-07-0 ed2011.
23. Harrell FE, Jr. Hmisc: Harrell Miscellaneous. R package version 3.9-2. http://CRAN.R-project.org/package=Hmisc. 2012.
24. Harrell FE, Jr. rms: Regression Modeling Strategies. R package version 3.4-0. http://CRAN.R-project.org/package=rms. 2012.
25. Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials. 2010;11:85.
26. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982;247:2543-6.
27. Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? Control Clin Trials. 1998;19:249-56.
28. Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics? Am Heart J. 2000;139:745-51.
29. Hernandez AV, Eijkemans MJ, Steyerberg EW. Randomized controlled trials with time-to-event outcomes: how much does prespecified covariate adjustment increase power? Ann Epidemiol. 2006;16:41-8.
30. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis: Springer-Verlag New York; 2001.
31. Verweij PJ, Van Houwelingen HC. Penalized likelihood in Cox regression. Stat Med. 1994;13:2427-36.
32. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87.
33. Goeman JJ. penalized: L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model. R package, version 0.9-42. http://CRAN.R-project.org/package=penalized. 2012.
34. Chieffo A, Meliga E, Latib A, Park SJ, Onuma Y, Capranzano P, et al. Drug-eluting stent for left main coronary artery disease. The DELTA registry: a multicenter registry evaluating percutaneous coronary intervention versus coronary artery bypass grafting for left main treatment. JACC Cardiovasc Interv. 2012;5:718-27.
35. Steyerberg EW, Vedder MM, Leening MJ, Postmus D, D'Agostino RB, Sr., Van Calster B, et al. Graphical assessment of incremental value of novel markers in prediction models: From statistical to decision analytical perspectives. Biom J. 2014.
36. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974;66:688-701.
37. Waldo SW, Secemsky EA, O'Brien C, Kennedy KF, Pomerantsev E, Sundt TM, et al. Surgical Ineligibility and Mortality Among Patients with Unprotected Left Main or Multivessel Coronary Artery Disease Undergoing Percutaneous Coronary Intervention. Circulation. 2014.
38. Burke JF, Hayward RA, Nelson JP, Kent DM. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. Circ Cardiovasc Qual Outcomes. 2014;7:163-9.
39. Cohen DJ, Lavelle TA, Van Hout B, Li H, Lei Y, Robertus K, et al. Economic outcomes of percutaneous coronary intervention with drug-eluting stents versus bypass surgery for patients

with left main or three-vessel coronary artery disease: one-year results from the SYNTAX trial. Catheter Cardiovasc Interv. 2012;79:198-209.

40. Magnuson EA, Farkouh ME, Fuster V, Wang K, Vilain K, Li H, et al. Cost-effectiveness of percutaneous coronary intervention with drug eluting stents versus bypass surgery for patients with diabetes mellitus and multivessel coronary artery disease: results from the FREEDOM trial. Circulation. 2013;127:820-31.

9

Refining the American guidelines for prevention of cardiovascular disease

D van Klaveren
Y Vergouwe
EW Steyerberg

Lancet 2014

We would like to propose a compromise in the debate on the clinical application of the recently proposed American guidelines for prevention of cardiovascular disease [1].

Paul Ridker and Nancy Cook (Nov 30, p 1762 [2]) criticize the current guidelines' assumption of a constant relative risk reduction and its subsequent focus on absolute (baseline) risk predictions. They plea for statins prescription based on treatment effects observed in specific trial populations. In in our view, combining absolute risk predictions with individualized estimates of relative risk reduction is required to quantify the absolute treatment benefit of statins. Relative risk reduction across subgroups with different levels of baseline risk can be based on the results of the individual patient data meta-analysis of statin trials [3]. Ideally, the individualized relative risk reductions are estimated in a re-analysis of these trial data, by adding a statistical treatment interaction with the risk predictions according to 2013 guidelines [4].

Ridker and Cook recommend recalibration of the guidelines' new prediction model in additional external validation cohorts. Rather, we should use already available contemporary validation cohorts to adjust poorly calibrated risk predictions for time trends. To account for well-recognized cardiovascular disease risk differences across the ethnic groups of Hispanics, Asians and native-Americans, recalibration might be based on available external data as well [5].

In conclusion, we recommend building guidelines on adequate estimates of absolute treatment benefit, requiring a recalibrated absolute risk prediction model in conjunction with individualized estimates of the relative risk reduction.

REFERENCES

1. Stone NJ, Robinson J, Lichtenstein AH, Merz CN, Blum CB, Eckel RH, Goldberg AC, Gordon D, Levy D, Lloyd-Jones DM, McBride P, Schwartz JS, Shero ST, Smith SC, Jr., Watson K, Wilson PW. 2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2013.
2. Ridker PM, Cook NR. Statins: new American guidelines for prevention of cardiovascular disease. *The Lancet* 2013.
3. Cholesterol Treatment Trialists C, Mihaylova B, Emberson J, Blackwell L, Keech A, Simes J, Barnes EH, Voysey M, Gray A, Collins R, Baigent C. The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. *Lancet* 2012;380:581-90.
4. Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials* 2010;11:85.
5. D'Agostino RB, Sr., Grundy S, Sullivan LM, Wilson P, Group CHDRP. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA* 2001;286:180-7.

10

Measuring discrimination and calibration of models that predict treatment benefit: two examples guiding cardiovascular interventions

D van Klaveren
EW Steyerberg
PW Serruys
DM Kent

Submitted

ABSTRACT

Background It is impossible to directly observe treatment benefit in individual patients since their (counterfactual) outcome under the alternative therapy is unknown. Thus, clinical prediction models that are used to support treatment decisions are usually evaluated for their ability to predict the risk of an outcome rather than treatment benefit.

Objective To define performance metrics to describe a model's ability to predict treatment benefit.

Design Predictive modeling study.

Patients Non-acute coronary artery disease patients in the SYNTAX trial (n=1800); acute ischemic stroke patients in 3 recombinant tissue plasminogen activator (rt-PA) trials (n=1205).

Measurements We assessed predictions in alternative models with a conventional concordance (c)-statistic for outcome risk and a novel treatment benefit c-statistic. We defined observed treatment benefit by the outcomes in *pairs* of patients matched on predicted benefit but discordant for treatment assignment (1 for benefit, 0 for no effect, and -1 for harm).

Results In the SYNTAX trial, compared to a model without treatment interactions, the SYNTAX Score II had improved ability to discriminate by treatment benefit (treatment benefit c-statistic 0.590 versus 0.552), despite having similar discrimination for outcome risk (conventional c-statistic 0.725 versus 0.719). However, for the simplified Stroke TPI versus the original Stroke TPI, the conventional c-statistic and the treatment benefit c-statistic were both similar (0.790 versus 0.811; and 0.584 versus 0.578, respectively), indicating no loss of discrimination for treatment benefit prediction with the simplified model.

Limitations More experience is needed to understand benchmarks for treatment benefit c-statistics.

Conclusions The proposed methodology has the potential to measure and communicate information about a prediction model's ability to predict treatment benefit not captured with conventional performance metrics.

INTRODUCTION

Treatments that demonstrate benefit on average in clinical trials help some patients but not others. A major focus of patient-centered outcomes research and personalized medicine is to better understand this heterogeneity of treatment effect (HTE) so that treatment might be targeted to those who benefit, and avoided in those where it is useless or harmful [1-3]. A risk modeling approach to clinical trial analysis has been proposed [4]; examples of this approach to trial analysis have been described for selecting patients for surgical versus medical therapy in carotid endarterectomy [5], for thrombolytic choice in acute myocardial infarction [6], for percutaneous coronary intervention versus coronary artery bypass in non-acute coronary artery disease patients [7], for continuation of dual antiplatelet therapy following PCI [8] and for better targeting prevention programs such as diabetes prevention [9] or lung cancer screening [10].

Risk prediction models can well support treatment decision making when they accurately predict individual treatment benefit, i.e. the difference in potential outcomes under different treatment regimens [11, 12]. However, risk prediction models are usually validated for their ability to predict risk, not for their ability to predict treatment benefit – the difference between outcome risk with *versus* without therapy (or with two alternative strategies). Optimizing commonly used performance measures, such as the concordance-statistic (c-statistic) used to assess a risk prediction model's discriminative ability (i.e. its ability to predict higher risks for those patients with the outcome compared to those without the outcome [13-15]), does not necessarily optimize a model's decision making potential, because performance measures for treatment selection should assess how well a model discriminates patients who benefit from those who do not [16-18]. However, discrimination measures are based on comparing predictions to actual outcomes in individual patients; measuring discriminative ability of benefit predictions is thus hampered by the fact that the actual benefit for each patient is inherently unobservable, since their potential (counterfactual) outcome under the alternative therapy is not known [19, 20].

Thus, we aimed to develop methods for validation of models that are used for predicting treatment benefit. We hereto adapt popular measures of predictive performance, and apply these measures to two previously developed prediction models intended to support decision making on reperfusion therapy: the SYNTAX Score II – developed to stratify patients according to their benefit from coronary artery bypass graft (CABG) compared to percutaneous coronary intervention (PCI) – and the Stroke-Thrombolytic Predictive Instrument (Stroke-TPI), which predicts benefit from thrombolysis in acute ischemic stroke.

METHODS

Theoretical background

The performance of a prediction model is typically measured in two dimensions: discrimination and calibration [21]. Discrimination is a model's ability to separate low risk subjects from high risk subjects. For binary or time-to-event outcomes, it is usually measured by a c-statistic, which is the proportion of all possible pairs of observations discordant on the outcome (i.e. one with the outcome and one without), in which the subject with the outcome had a higher predicted probability than the one without the outcome [13, 14]. Calibration refers to the agreement of a model's predicted probabilities with observed frequencies across the range of predictions [22]. Calibration may be measured by the difference between predicted and observed outcomes in groups defined by quantiles of predicted risk [23]. When we similarly aim to measure a prediction model's performance in predicting treatment benefit this requires predictions and observations of benefit instead of risk.

With clinical trial data, we can predict treatment benefit based on a multivariable model which regresses the outcome of interest with both the treatment and baseline characteristics. The individual predicted (absolute) treatment benefit can be defined as the predicted risk with one treatment minus the predicted risk with an alternative treatment. For example, when an individual's predicted mortality is 10% under one treatment and 8% under the alternative treatment, then the predicted absolute treatment benefit of the alternative treatment is 2%. Often a constant relative treatment effect is assumed, implying no statistical interactions between the treatment and the prognostic factors. In contrast, relative treatment effect heterogeneity – i.e. variation of relative treatment effect across individuals – can be modeled by including treatment interactions in the prediction model [4, 24, 25].

While calculating the predicted benefit in an individual is straightforward, we cannot directly observe the actual treatment benefit for an individual patient in a clinical trial, because the counterfactual outcome is missing. Since we are interested in observed treatment benefit conditional on predicted treatment benefit, we propose to define observed treatment benefit as the difference in outcomes between 2 patients with the same predicted benefit but different treatment assignments. With a binary outcome (say, alive or dead), there are only 4 possible outcome combinations for a pair of patients of whom the first is in the intervention and the second is in the control arm: the first alive and the second dead indicates treatment benefit, both alive and both dead indicate no treatment effect, and the first dead while the second is alive indicates treatment harm (Table 10.1) [26].

Having matched each patient from one trial arm with a patient from the other trial arm with similar (ideally identical) predicted treatment benefit, the agreement between predictions and observations of treatment benefit can be assessed with standard validation techniques, using these pairs of patients (with a trinary response variable) instead of individuals (with a binary response variable). The treatment benefit c-statistic for benefit is defined as the proportion of all possible *pairs* of patient *pairs* discordant on observed benefit in which the patient pair receiving greater treatment benefit was predicted to do so. Calibration for benefit assesses whether the absolute observed benefit matches the predicted benefit. This can be examined by ordering observations according to predicted benefit, and grouping into subgroups (e.g. quintiles of predicted benefit).

Table 10.1 Definition of observed treatment benefit for a pair of matched patients.

Observed outcome of patient in treatment arm A	Observed outcome of patient in treatment arm B	Observed treatment benefit of treatment B versus treatment A
0 (alive)	1 (dead)	-1 (harm)
0 (alive)	0 (alive)	0 (no effect)
1 (dead)	1 (dead)	0 (no effect)
1 (dead)	0 (alive)	1 (benefit)

Case studies

To understand how informative the evaluation of benefit might be in comparison with a more conventional risk-focused approach, we examined alternative versions of two published predictive models that were developed to predict treatment benefit.

The SYNTAX II score was derived in the Synergy between Percutaneous Coronary Intervention with Taxus and Cardiac Surgery (SYNTAX) trial (ClinicalTrials.gov, number NCT00114972) [27, 28]. In the SYNTAX trial 1,800 patients (178 deaths during 4 years of follow-up) with unprotected left main coronary artery (ULMCA) disease or de novo three vessel disease that were randomized to either CABG or PCI with first generation drug eluting stents (baseline characteristics in Table 10.2). The SYNTAX Score II includes eight prognostic factors for mortality and treatment interactions with each of the prognostic factors [7]. For comparison we also considered a prediction model with the eight prognostic factors and an overall treatment effect, i.e. without treatment interactions [25]. The models are described in Table 10.3.

For each patient the predicted benefit (of treatment with CABG) was calculated as the absolute risk prediction when treating with PCI minus the absolute risk prediction when treating with CABG. To realize an equal number of patients in both trial arms we randomly selected 897 out of 903 patients in the PCI arm. Each of the 897 patients in the CABG arm was matched to one of the 897 selected patients in the PCI arm, based on their rank of predicted benefit within the treatment arm. For each pair of patients, the predicted benefit of treating with CABG was set to the average of their individual benefit predictions. The observed benefit for each patient pair was calculated as the mortality outcome (0 = alive; 1 = dead) of the patient in the PCI arm minus the mortality outcome of the patient in the CABG arm (Table 10.1). For the set of 897 predicted and observed treatment benefits, we calculated the treatment benefit c-statistic, and we plotted average observed benefit versus predicted benefit in quintiles of predicted benefit (calibration plot). When a model is validated in the same data that was used for model derivation, estimates of model performance will be too optimistic. We therefore corrected the treatment benefit c-statistic and the calibration plot for optimism with a bootstrap procedure with 1000 resamples [29]. The models were refitted in each bootstrap sample. The optimism for each bootstrap resample was calculated as the difference in performance between validation in the

Table 10.2 Baseline characteristics of SYNTAX and Stroke TPI case studies.

Characteristic	Metric	SYNTAX	ATLANTIS[c]	NINDS
Treatment arm[a]	% (ratio)	50.2 (903/1800)	49.3 (303/614)	49.1 (290/591)
Age (years)	median [IQR] (n)	66 [58,72] (1800)	67.9 [58.8,74.7] (614)	68.4 [59.7,75.4] (591)
Male sex	% (ratio)	77.7 (1398/1800)	60.3 (370/614)	42.8 (253/591)
SYNTAX score	median [IQR] (n)	28 [20,36] (1789)		
Creatinine clearance (mL/min)	median [IQR] (n)	81.3 [64.5,103.3] (1638)		
LVEF (%)	median [IQR] (n)	60 [50,66] (1126)		
poor LVEF (<30%)	% (ratio)	1.9 (34/1772)		
moderate LVEF (30-49%)	% (ratio)	17.7 (313/1772)		
good LVEF (≥50%)	% (ratio)	80.4 (1425/1772)		
ULMCA disease	% (ratio)	39.2 (705/1799)		
COPD	% (ratio)	8.6 (154/1800)		
Peripheral vascular disease	% (ratio)	9.8 (177/1800)		
Diabetes	% (ratio)		21.0 (129/614)	21.2 (125/591)
Prior stroke	% (ratio)		15.1 (93/614)	13.0 (77/591)
SBP (mm Hg)	median [IQR] (n)		153 [137,168] (614)	152 [140,170] (591)
Minutes from stroke to treatment	median [IQR] (n)		273 [240,293] (614)	110 [89,157] (591)
NIHSS	median [IQR] (n)		10 [7,15] (614)	15 [9,20] (591)
3-variable stroke severity scale	median [IQR] (n)		2 [0,3] (614)	
Outcome[b]	% (ratio)	9.9 (178/1800)	40.4 (248/614)	34.5 (204/591)

a. Treatment arm: PCI in SYNTAX; rt-PA in ATLANTIS and NINDS

b. Outcome: All cause death in SYNTAX; Functionally normal/near-normal (mRS score 0 or 1) in ATLANTIS and NINDS

c. Patients of the ATLANTIS A trial and the ATLANTIS B trial were merged

Table 10.3 Model descriptions of SYNTAX and Stroke TPI case studies.

A. Multivariable hazard ratios (95% confidence interval) for mortality of non-acute coronary artery disease patients. For continuous factors the hazard ratio of the interquartile range is presented. When a factor interacts with treatment (all factors in the SYNTAX Score II) the treatment-specific hazard ratio is calculated based on the combination of a factor's main effect and the factor's treatment interaction effect. Hazard ratios above 1 indicate an increase in mortality risk.

	Constant relative treatment effect		SYNTAX Score II	
Factor	CABG	PCI	CABG	PCI
PCI vs. CABG[a]		1.47 (1.08,1.98)		1.44 (0.97,2.13)
SYNTAX score 36 vs 20	1.23 (1.01,1.51)	1.23 (1.01,1.51)	0.95 (0.69,1.31)	1.47 (1.13,1.91)
Age (years) 72 vs 58	1.80 (1.33,2.43)	1.80 (1.33,2.43)	2.68 (1.57,4.55)	1.49 (0.96,2.31)
CrCl (mL/min) 90 vs 65	0.68 (0.53,0.89)	0.68 (0.53,0.89)	0.79 (0.52,1.20)	0.60 (0.43,0.83)
LVEF (%) 50 vs 10	0.18 (0.08,0.40)	0.18 (0.08,0.40)	0.50 (0.14,1.81)	0.10 (0.04,0.28)
ULMCA disease	1.06 (0.78,1.43)	1.06 (0.78,1.43)	1.47 (0.93,2.34)	0.82 (0.54,1.23)
Male sex	0.90 (0.64,1.26)	0.90 (0.64,1.26)	1.68 (0.91,3.12)	0.59 (0.38,0.90)
COPD	1.87 (1.25,2.79)	1.87 (1.25,2.79)	2.84 (1.64,4.90)	1.35 (0.74,2.47)
PVD	2.64 (1.85,3.76)	2.64 (1.85,3.76)	2.79 (1.66,4.71)	2.79 (1.72,4.53)

B. Multivariable odds ratios (95% confidence interval) for good outcome after acute ischemic stroke. For continuous factors the odds ratio of the interquartile range is presented. When a factor interacts with treatment (male sex, prior stroke, SBP and minutes to treatment in the original Stroke TPI; SBP and minutes to treatment in the simplified Stroke TPI) the treatment-specific odds ratio is calculated based on the combination of a factor's main effect and the factor's treatment interaction effect. Odds ratios above 1 indicate an increase in the probability of good outcome.

	Original Stroke TPI		Simplified Stroke TPI	
Factor	Control	rt-PA	Control	rt-PA
rt-PA[a]		1.51 (1.23,1.86)		1.45 (1.18,1.78)
Age (years) 75 vs 59[b]	0.73 (0.63,0.85)	0.73 (0.63,0.85)	0.75 (0.65,0.87)	0.75 (0.65,0.87)
Male sex	1.45 (1.07,1.96)	0.95 (0.71,1.28)	1.03 (0.83,1.26)	1.03 (0.83,1.26)
Diabetes	0.48 (0.37,0.62)	0.48 (0.37,0.62)	0.50 (0.38,0.65)	0.50 (0.38,0.65)
Prior stroke	1.46 (0.99,2.15)	0.67 (0.45,1.00)	0.93 (0.70,1.23)	0.93 (0.70,1.23)
SBP (mm Hg) 169 vs 139	0.86 (0.69,1.07)	0.61 (0.49,0.76)	0.91 (0.73,1.13)	0.62 (0.50,0.78)
Minutes from stroke to treatment 288 vs 155	1.04 (0.81,1.33)	0.63 (0.50,0.81)	0.88 (0.64,1.20)	0.49 (0.37,0.66)
NIHSS 17 vs 8[c]	0.17 (0.14,0.21)	0.17 (0.14,0.21)		
3-variable stroke severity scale 3 vs 1[c]			0.24 (0.20,0.29)	0.24 (0.20,0.29)

a. At average prognostic factor levels

b. At average NIHSS and average 3-variable stroke severity scale

c. At average age

bootstrap sample and validation in the original data. The optimism was subtracted from the apparent performance estimate.

The Stroke TPI was developed on a pooled database of 5 clinical trials testing recombinant tissue plasminogen activator (rt-PA) against placebo in the treatment of acute ischemic stroke (n = 2184 patients). The Stroke TPI predicts the probability of a good functional outcome (modified Rankin Score mRS <=1) and the probability of severe disability or death (mRS >=5) with and without rt-PA. The model for prediction of good functional outcome contained 7 prognostic factors, the treatment and 4 interactions between treatment and prognostic factors [30]. Recently, a modified Stroke TPI was developed, primarily to improve ease-of-use by non-specialists, by replacing a full NIH stroke severity score with a simplified 3-item version. Importantly, 2 of the 4 treatment interactions were eliminated [31]. Both models are described in Table 10.3. While discriminatory performance was shown to be relatively well maintained in the simplified compared to the original model, a remaining concern is that the simplified Stroke TPI may have a reduced ability to segregate patients by their probability of benefiting. Herein, we test these models on the NINDS trial and the ATLANTIS A and B trial (baseline characteristics in Table 10.2) which comprise a subset of the development dataset for these tools [32-34]. The NINDS Trial (n=591) included patients treated from 0 to 180 minutes from symptom onset; the ATLANTIS A Trial (n=48) from 0 to 360 minutes; and the ATLANTIS B Trial (n=566) from 180 to 300 minutes. The ECASS 2 Trial (n=778) was excluded from validation in our study [35].

For each patient in this 1205 patients sample the predicted benefit (of treatment with rt-PA) was calculated as the predicted probability of a good functional outcome when treated with rt-PA minus the predicted probability of a good functional outcome when treated with placebo, for both the original and the simplified Stroke TPI. The methods used for matching of patients, calculation of observed treatment benefit and assessment of discrimination and calibration of treatment benefit predictions (with calibration plots and with c-statistics, respectively) were identical to the methods described above for the SYNTAX trial.

Table 10.4 C-statistics (95% confidence intervals) for risk and for benefit.

Measure	Constant relative treatment effect	SYNTAX Score II	Original Stroke TPI	Simplified Stroke TPI
Risk c-statistic				
Apparent	0.729 (0.691,0.766)	0.744 (0.707,0.781)	0.811 (0.786,0.835)	0.790 (0.764,0.815)
Optimism	0.009	0.018		
Corrected	0.719 (0.682,0.757)	0.725 (0.689,0.762)		
Treatment benefit c-statistic				
Apparent	0.555 (0.501,0.608)	0.620 (0.566,0.674)	0.578 (0.538,0.618)	0.584 (0.544,0.625)
Optimism	0.003	0.030		
Corrected	0.552 (0.498,0.605)	0.590 (0.536,0.644)		

Correction for optimism in the SYNTAX trial without treatment interactions (“Constant relative treatment effect”) and the SYNTAX Score II was based on internal validation with 1,000 bootstrap samples.

RESULTS

In non-acute coronary artery disease in the SYNTAX trial, the model assuming a constant relative treatment effect (i.e. without treatment interactions) discriminated high-risk from low-risk patients only slightly worse than the SYNTAX Score II (risk c-statistics corrected for optimism 0.719 vs 0.725; Table 10.4). The SYNTAX trial arms were well balanced for predicted benefit regardless of the two models that were used to predict benefit (Figure 10.1). Due to the treatment interactions in the SYNTAX Score II, the range of predicted treatment benefit was much wider and, notably, contained negative benefit (treatment with PCI favorable). The treatment benefit c-statistic was substantially higher for the SYNTAX Score II than for the model assuming a constant relative treatment effect (treatment benefit c-statistics corrected for optimism 0.590 vs 0.552; Table 10.4).

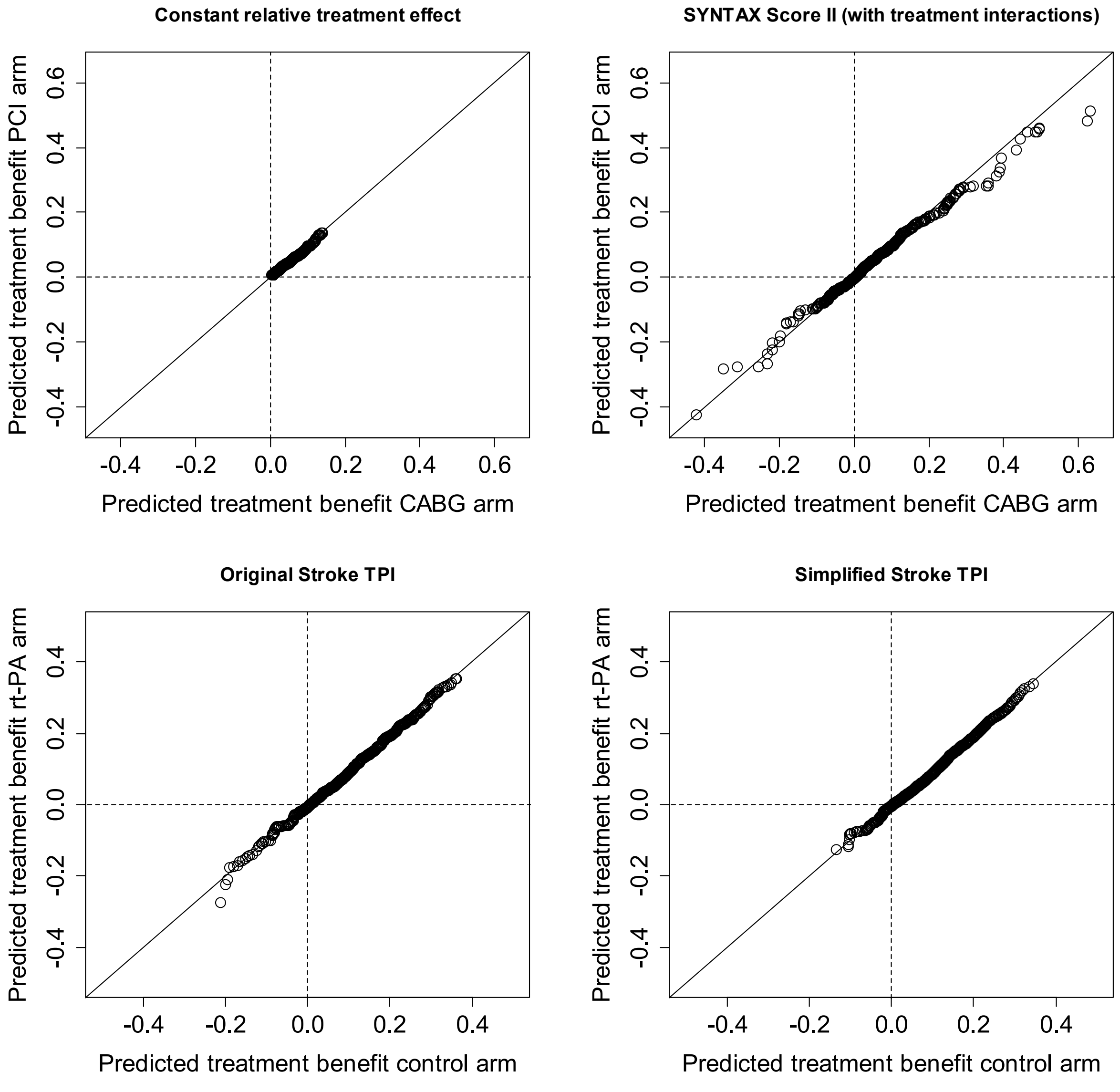


Figure 10.1 Matching patient pairs on predicted treatment benefit. For each matched patient pair the predicted treatment benefit in one trial arm is plotted versus the predicted treatment benefit in the other trial arm. Perfect matching would result in dots located exactly on the diagonal. Upper left: SYNTAX Score II without treatment interactions (“Constant relative treatment effect”) ; Upper right: SYNTAX Score II; Lower left: Original Stroke TPI; Lower right: Simplified Stroke TPI.

Due to the uncertainty in the interaction effect estimates the performance measures of the SYNTAX Score II required a 10 times larger correction for optimism (treatment benefit c-statistic corrections 0.030 vs. 0.003; Table 10.4). Calibration in quintiles of predicted benefit was fairly good for the SYNTAX Score II (Figure 10.2). Again, the SYNTAX Score II required a larger correction for optimism than the model without interactions.

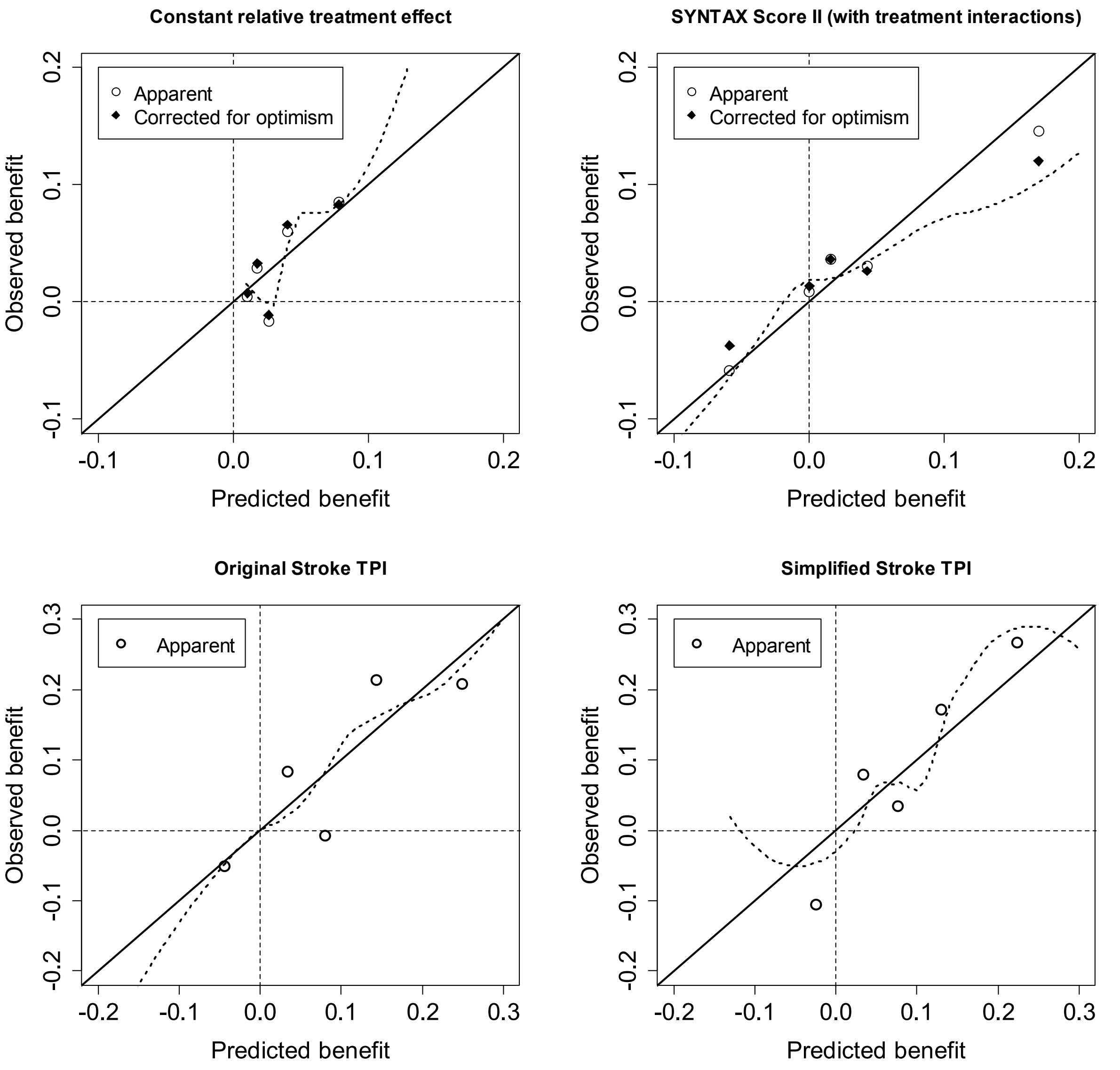


Figure 10.2 Calibration plots: observed treatment benefit versus predicted treatment benefit in 5 equally sized patient pair groups of increasing predicted treatment benefit (quintiles). Upper left: Benefit of treatment with CABG according to the SYNTAX Score II without treatment interactions ("Constant relative treatment effect") ; Upper right: Benefit of treatment with CABG according to the SYNTAX Score II; Lower left: Benefit of treatment with rt-PA according to the Original Stroke TPI; Lower right: Benefit of treatment with rt-PA according to the Simplified Stroke TPI.

As expected, the discriminative ability of the simplified Stroke TPI was somewhat less than the original Stroke TPI (risk c-statistics 0.790 vs 0.811, respectively; Table 10.4). Again, trial arms were well balanced for predicted benefit for both models (Figure 10.1). Due to less treatment interactions in the simplified Stroke TPI, the range of predicted treatment benefit was smaller as compared to the original Stroke TPI, especially for patients with predicted treatment harm (negative benefit). Surprisingly, the treatment benefit c-statistic was not worse for the simplified Stroke TPI (0.584 vs 0.578 for the original Stroke TPI; Table 10.4), and neither was calibration (Figure 10.2). These results indicate that the additional complexity of the original Stroke TPI does not increase the ability to predict treatment benefit for the patients in this validation subpopulation.

DISCUSSION

We proposed measures to validate predictions of treatment benefit, extending widely used approaches to validate predictions of risk. Hereto we defined *observed* treatment benefit as the difference in outcomes between 2 patients with the same predicted benefit but discordant on treatment assignment. The proposed methodology gave interpretable measures of discrimination and calibration, both for the SYNTAX score II and a simplified alternative [25], and for the original and the simplified Stroke TPI [30, 31]. The metrics proposed may help refocus the goals of prediction modeling from discrimination on the basis of outcome risk to discrimination on the basis of potential for treatment benefit, which can theoretically better support optimizing outcomes when considering two alternative treatment strategies [18].

The increase in the treatment benefit c-statistic when using treatment interaction terms in the SYNTAX Score II (from 0.552 to 0.590) indicated a major improvement in discriminative ability. The absolute difference with the c-statistic of a coin toss (0.5; random treatment assignment) almost doubles from 5.2% to 9.0%. The superiority of the SYNTAX Score II is likewise reflected in the calibration curves. While all patient subgroups appear to benefit from bypass surgery compared to PCI when the simplified score is used, there are clearly two “low benefit” quintiles identified with the SYNTAX Score II, for whom bypass surgery may not to be justified – particularly among the lowest benefit quintile for whom PCI appears clearly superior.

In contrast, the original Stroke TPI had a similar treatment benefit c-statistic compared to the simplified Stroke TPI, indicating that the additional treatment interactions included in the original equation – while increasing the range of predicted benefit (Figure 10.1) – did not improve the discriminative ability for treatment benefit in the trial population included in this study. This contrasted with the improvement in discrimination in outcome risk seen with the full (original) compared to the simplified model reflected in the conventional c-statistic. The lack of improvement in predicting benefit was also reflected in the calibration curve (Figure 10.2), which shows predicted harm in a single quintile in either model.

In addition to highlighting the novel statistical metrics, these examples also emphasize the importance of a risk modeling approach to clinical trial analysis when there are alternative treatments with distinct trade-offs. For the acute ischemic stroke example, there appear to be patients who are more likely to be harmed from thrombolysis despite the overall benefit; presumably because the risks of thrombolytic-related intracranial hemorrhage in these selected patients (those with especially high blood pressure arriving late in the treatment window) exceeds any benefits anticipated from reperfusion. Similarly, bypass surgery provides substantial mortality benefit for most patients, but for some the far less invasive PCI strategy would appear to be superior. Carefully developed and validated models that optimize performance metrics for treatment benefit can better segregate patients to the appropriate treatment strategy, in contrast to more conventional evidence based medicine approaches, which emphasize broad application of the best treatment on average.

We note that the treatment benefit c-statistics seen in these examples is in a range that would typically be considered only weakly predictive for conventional risk c-statistics. We would caution that new benchmarks are needed to interpret treatment benefit c-statistics. Because improvements in discrimination for benefit are generally more relevant than improvements in performance measures for risk, these relatively small improvements might be of great clinical importance. Although more experience with this approach is needed, we anticipate that treatment benefit c-statistics above 0.6 might be very unusual, except in the presence of highly deterministic markers that indicate a treatment mechanism. This reflects the greater difficulty in predicting benefit compared to predicting outcome risk, the main challenge of personalized medicine.

We and others have previously emphasized the importance of both prognostic factors and treatment effect interactions for determining benefit [6, 24]. This is reflected in the improvement of the treatment benefit c-statistic in the SYNTAX example. The treatment benefit c-statistic might increase the appreciation of the importance of these interaction terms for predicting benefit. However, while treatment effect interactions are particularly helpful, we note that these interactions can be highly unreliable. Because less is known about effect modifiers than risk predictors and because power is much poorer to detect treatment effect interactions, statistically significant interactions often turn out "false positive" findings, particularly when multiple interaction effects are explored [36, 37]. Interactions should be motivated by external information, and including interaction effects makes external validation for *benefit* (i.e. on a sample including both treatment arms) even more critical than validation for risk. This is reflected in both examples: in SYNTAX, the treatment benefit c-statistic showed a higher optimism correction than the risk c-statistic; in the Stroke-TPI statistically significant interactions failed to improve the treatment benefit c-statistic. These considerations suggest that trials aimed at exploring HTE will need to be substantially larger than conventional trials to be adequately powered for interaction effects in addition to main effects.

Analogous to the methods we proposed here, a ROC measure was proposed for treatment-selection markers using a potential outcomes framework [38]. This is a model-based measure, calculating the expected ROC curve for predicted benefit under the assumption that the benefit predictions are correct. Instead of assuming a correct benefit prediction model, we matched patients with discordant treatment on predicted treatment benefit to obtain observed values of treatment benefit.

Various other methods have been proposed to measure the ability of prediction models to support clinical decisions. Benefit graphs were proposed to measure the difference in outcome when two different models are used for treatment decision making, by multiplying the proportion of patients with different treatment recommendations with their observed mortality difference in the randomized treatment arms [25, 39]. Net benefit of treatment can also be evaluated across a range of treatment benefit thresholds, where benefit is the difference between the probabilities of an event under treatment and under control [40]. Net benefit is then determined for those patients for whom the treatment recommendation was congruent with the treatment allocation in the trial, and compared to the net benefit of all patients in the treatment arm ("treat all"). These evaluations can be shown in decision curves. A key element of decision curves is that the risk threshold at which a patient would opt for treatment is used both to determine the sensitivity (true-positive-rate) and specificity (true-negative-rate) of the prediction model and to weigh the relative harms of false-positive and false-negative predictions [41]. This dual role of the risk threshold makes the curves useful for evaluating the quality of decision making based a risk model, but (anecdotally) also make it cognitively demanding and confusing for non-experts. The advantage of the methods we propose here is that they leverage simple and widely used metrics of model performance, i.e. measures for discrimination and calibration and might therefore be more easily understood by non-experts.

In case of time-to-event outcomes instead of binary outcomes, the proposed methodology is still applicable, although the definition of observed benefit is somewhat different. Treatment benefit (+1) and treatment harm (-1) are defined as a shorter observed time-to-event for the patient in the intervention arm and in the control arm, respectively. When the shorter time-to-event outcome is censored there is no observed treatment effect (0) for a patient pair.

Nonetheless, there are limitations to our approach. The treatment benefit c-statistic inherits the main limitation of the c-statistic in that it measures statistical discrimination, which does not take into account the decisional context. Thus, a given c-statistic may be adequate in one context and not in another. Additionally, more experience using the proposed measures in developing and (especially) validating models promoted to predict benefit is needed to evaluate their usefulness, and to better understand benchmarks. While extension to trials with unbalanced treatment groups is easily accomplished through weighted matching, more work is needed to understand how these procedures might be adapted in an observational setting, where the balance between treatment groups might vary over different levels of predicted benefit.

These caveats aside, our proposed measures are closely aligned with the goal of prediction for personalized medicine – that is, to segregate the population by their likelihood of benefiting from one therapy versus another, thus supporting treatment decisions in individual patients that optimize outcomes taking as much information about each patient into account as possible [42]. To deliver on this promise of more patient-centered evidence, new methods are necessary to explore HTE; the proposed methodology has the potential to measure and communicate a prediction model's ability to predict treatment benefit, in contrast to conventional metrics which measure a prediction model's ability to predict outcome risk.

ACKNOWLEDGMENTS

The authors express their gratitude to all of the principal investigators and patients participating in the SYNTAX trial and in the rt-PA trials NINDS (trial 1 and trial 2) and ATLANTIS (trial A and trial B). The SYNTAX trial was sponsored by Boston Scientific Corporation. The ATLANTIS trial was sponsored by Genentech. This work was supported by the National Institutes of Health (grant U01NS086294).

REFERENCES

1. Rothwell PM. Can overall results of clinical trials be applied to all patients? Lancet 1995; 345: 1616-1619.
2. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004; 82: 661-687.
3. Vickers AJ, Kent DM. The Lake Wobegon Effect: Why Most Patients Are at Below-Average Risk. Ann Intern Med 2015.
4. Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 2010; 11: 85.
5. Rothwell PM, Warlow CP. Prediction of benefit from carotid endarterectomy in individual patients: A risk-modelling study. Lancet 1999; 353: 2105-2110.
6. Califf RM, Woodlief LH, Harrell FE, Jr., Lee KL, White HD, Guerci A, Barbash GI, Simes RJ, Weaver WD, Simoons ML, Topol EJ. Selection of thrombolytic therapy for individual patients: development of a clinical model. GUSTO-I Investigators. Am Heart J 1997; 133: 630-639.
7. Farooq V, van Klaveren D, Steyerberg EW, Meliga E, Vergouwe Y, Chieffo A, Kappetein AP, Colombo A, Holmes DR, Jr., Mack M, Feldman T, Morice MC, Stahle E, Onuma Y, Morel MA, Garcia-Garcia HM, van Es GA, Dawkins KD, Mohr FW, Serruys PW. Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score II. Lancet 2013; 381: 639-650.
8. Yeh RW, Secemsky EA, Kereiakes DJ, Normand SL, Gershlick AH, Cohen DJ, Spertus JA, Steg PG, Cutlip DE, Rinaldi MJ, Camenzind E, Wijns W, Apruzzese PK, Song Y, Massaro JM, Mauri L, Investigators DS. Development and Validation of a Prediction Rule for Benefit and Harm of Dual Antiplatelet Therapy Beyond 1 Year After Percutaneous Coronary Intervention. JAMA 2016; 315: 1735-1749.
9. Sussman JB, Kent DM, Nelson JP, Hayward RA. Improving diabetes prevention with benefit based tailored treatment: risk based reanalysis of Diabetes Prevention Program. BMJ 2015; 350: h454.
10. Kovalchik SA, Tammemagi M, Berg CD, Caporaso NE, Riley TL, Korch M, Silvestri GA, Chaturvedi AK, Katki HA. Targeting of low-dose CT screening according to the risk of lung-cancer death. N Engl J Med 2013; 369: 245-254.
11. Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. Lancet 2005; 365: 256-265.
12. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA 2007; 298: 1209-1212.
13. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143: 29-36.
14. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 1982; 247: 2543-2546.
15. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010; 21: 128-138.
16. Tajik P, Oude Rengerink K, Mol BW, Bossuyt PM. SYNTAX score II. Lancet 2013; 381: 1899.
17. Farooq V, van Klaveren D, Steyerberg EW, Serruys PW. SYNTAX score II - Authors' reply. Lancet 2013; 381: 1899-1900.
18. Kent DM, Hayward RA, Dahabreh IJ. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centered evidence. Int J Epidemiol 2016; In press.

19. Holland PW. Statistics and Causal Inference. Journal of the American Statistical Association 1986; 81: 945-960.
20. Rubin DB. Causal Inference Using Potential Outcomes. Journal of the American Statistical Association 2005; 100: 322-331.
21. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer: New York, 2009.
22. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016.
23. Harrell FE, Jr., Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984; 3: 143-152.
24. Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, Schroter S, Sauerbrei W, Altman DG, Hemingway H, Group P. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ 2013; 346: e5793.
25. van Klaveren D, Vergouwe Y, Farooq V, Serruys PW, Steyerberg EW. Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions. J Clin Epidemiol 2015.
26. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 1986; 15: 413-419.
27. Ong AT, Serruys PW, Mohr FW, Morice MC, Kappetein AP, Holmes DR, Jr., Mack MJ, van den Brand M, Morel MA, van Es GA, Kleijne J, Koglin J, Russell ME. The SYNergy between percutaneous coronary intervention with TAXus and cardiac surgery (SYNTAX) study: design, rationale, and run-in phase. Am Heart J 2006; 151: 1194-1204.
28. Serruys PW, Morice MC, Kappetein AP, Colombo A, Holmes DR, Mack MJ, Stahle E, Feldman TE, van den Brand M, Bass EJ, Van Dyck N, Leadley K, Dawkins KD, Mohr FW, Investigators S. Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. N Engl J Med 2009; 360: 961-972.
29. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996; 15: 361-387.
30. Kent DM, Selker HP, Ruthazer R, Bluhmki E, Hacke W. The stroke-thrombolytic predictive instrument: a predictive instrument for intravenous thrombolysis in acute ischemic stroke. Stroke 2006; 37: 2957-2962.
31. Kent DM, Ruthazer R, Decker C, Jones PG, Saver JL, Bluhmki E, Spertus JA. Development and validation of a simplified Stroke-Thrombolytic Predictive Instrument. Neurology 2015; 85: 942-949.
32. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. N Engl J Med 1995; 333: 1581-1587.
33. Clark WM, Albers GW, Madden KP, Hamilton S. The rtPA (alteplase) 0- to 6-hour acute stroke trial, part A (A0276g) : results of a double-blind, placebo-controlled, multicenter study. Thromblytic therapy in acute ischemic stroke study investigators. Stroke 2000; 31: 811-816.
34. Clark WM, Wissman S, Albers GW, Jhamandas JH, Madden KP, Hamilton S. Recombinant tissue-type plasminogen activator (Alteplase) for ischemic stroke 3 to 5 hours after symptom onset. The ATLANTIS Study: a randomized controlled trial. Alteplase Thrombolysis for Acute Noninterventional Therapy in Ischemic Stroke. JAMA 1999; 282: 2019-2026.
35. Hacke W, Kaste M, Fieschi C, von Kummer R, Davalos A, Meier D, Larrue V, Bluhmki E, Davis S, Donnan G, Schneider D, Diez-Tejedor E, Trouillas P. Randomised double-blind placebo-controlled trial of thrombolytic therapy with intravenous alteplase in acute ischaemic stroke (ECASS II). Second European-Australasian Acute Stroke Study Investigators. Lancet 1998; 352: 1245-1251.

36. Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? Am Heart J 2006; 151: 257-264.
37. Burke JF, Sussman JB, Kent DM, Hayward RA. Three simple rules to ensure reasonably credible subgroup analyses. BMJ 2015; 351: h5651.
38. Huang Y, Gilbert PB, Janes H. Assessing treatment-selection markers using a potential outcomes framework. Biometrics 2012; 68: 687-696.
39. Steyerberg EW, Vedder MM, Leening MJG, Postmus D, D'Agostino RB, Van Calster B, Pencina MJ. Graphical assessment of incremental value of novel markers in prediction models: From statistical to decision analytical perspectives. Biometrical Journal 2014: n/a-n/a.
40. Vickers AJ, Kattan MW, Daniel S. Method for evaluating prediction models that apply the results of randomized trials to individual patients. Trials 2007; 8: 14.
41. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006; 26: 565-574.
42. Pauker SG, Kassirer JP. Decision Analysis. New England Journal of Medicine 1987; 316: 250-258.

11

Biases in individualized cost-effectiveness analysis: influence of choices in modeling short-term, trial-based, mortality risk reduction and post-trial life expectancy

D van Klaveren
JB Wong
DM Kent
EW Steyerberg

Medical Decision Making (in press)

ABSTRACT

Background The benefit and costs of a treatment are typically heterogeneous across individual patients. Randomized clinical trials permit examination of individualized treatment benefit over the trial horizon, but extrapolation to lifetime horizon usually involves combining trial-based individualized estimates of short-term risk reduction with less detailed (less granular) population life tables. However, the underlying assumption of equal post-trial life expectancy for low and high risk patients of the same sex and age is unrealistic. We aimed to study the influence of unequal granularity between models of short-term risk reduction and life expectancy on individualized estimates of cost-effectiveness of aggressive thrombolysis for patients with an acute myocardial infarction.

Methods To estimate life years gained, we multiplied individualized estimates of short-term risk reduction either with less granular and with equally granular post-trial life expectancy estimates. Estimates of short-term risk reduction were obtained from GUSTO trial data (30,510 patients) using logistic regression analysis with treatment, sex and age as predictor variables. Life expectancy estimates were derived from sex and age-specific US life tables.

Results Based on sex and age-specific short-term risk reductions but average population life expectancy (less granularity), aggressive thrombolysis was cost-effective (incremental cost-effectiveness ratio below $50,000) for women above age 49 and men above age 53 (92% and 69% of the population respectively). Considering sex- and age-specific short-term mortality risk reduction and correspondingly sex- and age-specific life expectancy (equal granularity), aggressive thrombolysis was cost-effective for men above age 45 and women above age 50 (95% and 76% of the population respectively).

Conclusions Failure to model short-term risk reduction and life expectancy at an equal level of granularity may bias our estimates of individualized cost-effectiveness and misallocate resources.

INTRODUCTION

The cost-effectiveness of a treatment has increasingly been recognized to be heterogeneous across individual patients [1-3]. Interventions that are cost-effective on average may be of very low value on many (even most) patients. Estimation of individualized (or stratified) cost-effectiveness can thus support more efficient distribution of resources, but requires individualized estimation of treatment benefit, treatment harms, treatment costs, and patient preferences [4-8].

Individualized treatment benefit estimation is often focused on heterogeneity in short-term mortality risk differences under different treatment regimens. These differences can commonly be based on outcome data from a clinical trial with excellent information on patient characteristics. However, the individualized long-term treatment benefit also depends on the post-trial life expectancy. Because post-trial survival information is generally lacking, post-trial life expectancy is usually derived from less granular (less detailed) population life tables. However, the underlying assumption of equal post-trial life expectancy for low and high risk patients of the same sex and age is unrealistic.

For example, in economic analyses of different thrombolytic reperfusion therapies to restore blood flow after acute myocardial infraction (MI), the individualized cost-effectiveness estimates of treatment with accelerated tissue plasminogen activator (t-PA) rather than streptokinase were based on predictions of the 30 day mortality reduction using 11 predictors, while post-trial life expectancies were derived from sex and age specific life tables [9]. Measures of heart attack severity (such as electrocardiographic infarction-size and location, and vital signs) were incorporated into the prediction of short-term mortality (and benefit), but no such measures were included in the estimation of long term survival, despite the fact that infarct size and severity undoubtedly contribute also to longer term mortality [10-12]. Consequently, the long-term treatment benefit in terms of life years may have been overestimated for high-risk patients, and underestimated for low-risk patients.

In general, age, severity of disease and presence of comorbidities are typically strong determinants of both short-term and long-term survival, i.e. patients at high short-term mortality risk likely have a low life expectancy after short-term survival. Therefore, to estimate the long-term treatment benefit for an individual patient, the short-term risk reduction and the post-trial life expectancy should be individualized consistently, i.e. with equal granularity for all determinants of mortality risk.

Although short-term and long-term risk models are ideally based on the same level of detail in information, we do not know the impact of combining detailed short-term risk models with less granular long-term risk models. . It is therefore essential to explore and understand the impact of the imbalance between the level of granularity in models of short-term risk and life expectancy. We aimed to study the influence of unequal granularity between models of short-term risk reduction and life expectancy on individualized cost-effectiveness estimates.

METHODS

We compared individualized cost-effectiveness estimates of t-PA treatment versus streptokinase treatment after an acute myocardial infarction, for different models of short-term risk and post-trial life expectancy. To obtain life years gained at the individual level, we multiplied estimates of short-term risk reductions either with less granular and with equally granular post-trial life expectancy estimates. We adopted a relatively simple methodology to analyze the cost-effectiveness analysis of t-PA treatment versus streptokinase treatment [13]. We analyzed 30,510 patients with an acute myocardial infarction who were randomized to treatment with different forms of thrombolysis in the Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries (GUSTO) trial.

First, we estimated the short-term mortality risk reduction of t-PA treatment using a logistic regression model with treatment, sex and age as predictor variables for 30-day mortality. Thus, the absolute risk reduction of t-PA treatment was estimated by the predicted 30-day mortality without t-PA treatment minus the predicted 30-day mortality with t-PA treatment, assuming a constant relative treatment effect [14]. We estimated post-trial life expectancy from sex and age-specific US life tables in combination with an additional 2% yearly excess hazard to capture the increased long-term mortality risk of cardiovascular patients [9, 15]. To obtain estimates of life years gained for individual patients, sex and age-specific estimates of short-term risk reductions were either multiplied with sex and age-specific post-trial life expectancy estimates (equal level of granularity) or with the average of the post-trial life expectancy estimates in our patient population (less granularity). To interpret the difference between using the population average life expectancy versus a sex- and age-specific life expectancy we first compared estimates of life years gained and of incremental cost-effectiveness ratios (ICERs) for quintiles of short-term mortality risk. We further extended the comparison of ICERS to percentiles of short-term mortality risk for individual female and male patients separately.

Second, to explore the use of additional risk information about the severity of heart failure, we included the Killip classification for estimation of the short-term mortality reduction of t-PA treatment [16]. Since patients with higher Killip class are known to also have worse long-term prognosis [11, 12], we analyzed the sensitivity of cost-effectiveness estimates to also using the Killip class for estimation of the post-trial life expectancy. Hereto we multiplied post-trial yearly mortality hazards with short-term trial-based hazard ratios of Killip classes II, III and IV versus Killip class I and calibrated the excess hazard to produce an equal average life expectancy of the overall patient population. Because the assumption of equal associations of both short-term and long-term mortality with the Killip class may be too rigorous, we also multiplied with the square root of the short-term hazard ratios (i.e. half the effect size) to represent a milder association of long-term risk with the Killip class. To obtain estimates of life years gained for individual patients, sex, age and Killip class-specific estimates of short-term risk reductions were either multiplied with sex, age and Killip class-specific post-trial life expectancy estimates (equal level of granularity) or with sex and age-specific post-trial life expectancy estimates (less granularity).

Incremental costs of t-PA treatment were assumed to be constant across patients at $2,700, and not different for the t-PA group after the first year [9]. To estimate life-expectancy with and without treatment over a life-time time horizon, estimates of 30-day mortality were combined with annual rates of mortality, with discounting at 3% per year. Treatment with t-PA was considered cost-effective for incremental costs per life year gained falling below $50,000. All analyses were done with R statistical software [17].

Table 11.1 Multivariable associations between predictors and 30-day mortality in the GUSTO trial data.

Variable	Level	Excluding Killip class	Including Killip class
Therapy	Streptokinase	1	1
	t-PA	0.83 (0.75, 0.91)	0.81 (0.74, 0.90)
Age	10 years	2.2 (2.1, 2.3)	2.1 (2.0, 2.2)
Sex	Male	1	1
	Female	1.4 (1.2, 1.5)	1.3 (1.2, 1.5)
Killip class	I		1
	II		2.5 (2.4, 2.7)
	III		6.5 (5.7, 7.4)
	IV		16 (13, 20)

Odds ratios (95% confidence interval) for logistic regression models excluding and including the Killip classification

RESULTS

Short-term mortality was significantly associated with the treatment (t-PA versus Streptokinase), and with the sex, age and Killip class of the patient (Table 11.1). Treatment with t-PA reduced overall short-term mortality by 1.0% (6.3% with t-PA versus 7.3% with Streptokinase). With an average discounted post-trial life expectancy of 12.3 years (undiscounted 16.9 years) and fixed incremental costs of $2,700, t-PA treatment was on average cost-effective at the $50,000 threshold (ICER $21,895).

Based on sex and age-specific short-term risk reductions but average population life expectancy, there was a substantial heterogeneity in expected life years gained between the lowest and highest risk quintiles of short-term mortality risk (0.03 life years in 1st quintile vs 0.32 life years in 5th quintile; upper panel of Figure 11.1). The ICER in the lowest risk quintile exceeded the $50,000 threshold (lower panel of Figure 11.1). When both short-term mortality risk reduction and life expectancy were sex and age-specific, i.e. equal in level of granularity, the difference in life years gained between the lowest and highest risk quintile was attenuated (0.04 life years in 1st quintile vs 0.20 life years in 5th quintile; first panel of Figure 11.1). The difference between the ICERs in the first and fifth quintiles decreased accordingly (lower panel of Figure 11.1).

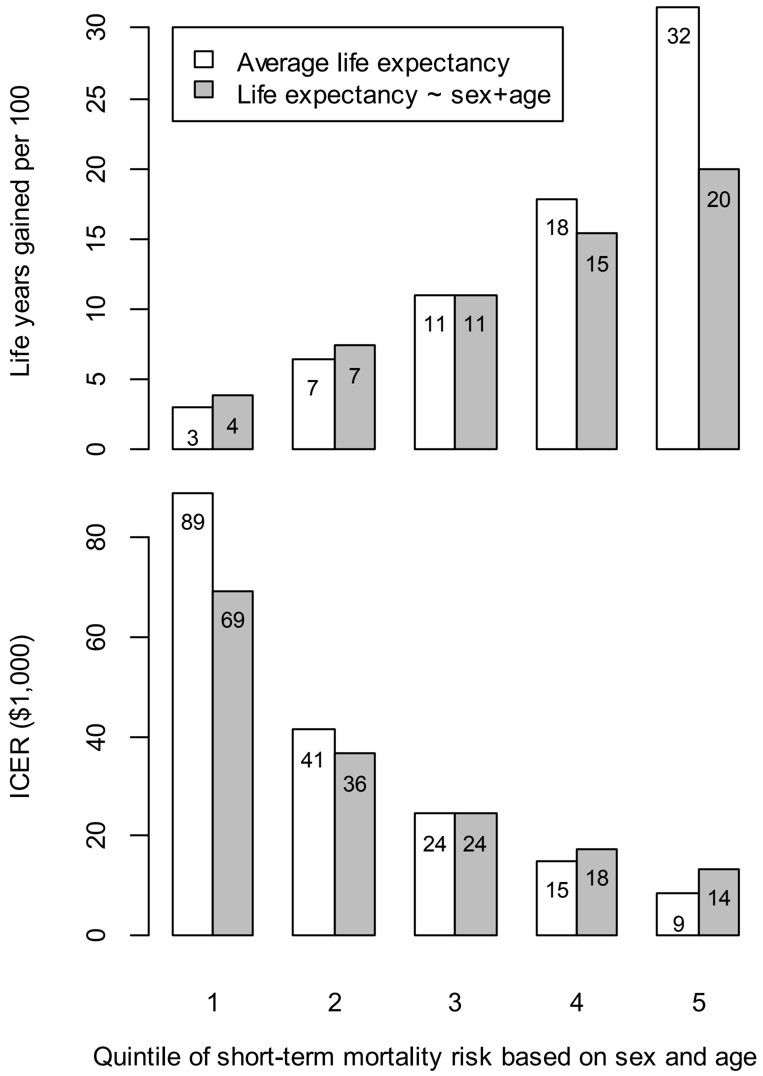


Figure 11.1 Life years gained per 100 patients with an acute MI and incremental cost effectiveness ratio (ICER) by quintile of sex- and age-specific short-term mortality risk. Patients are ordered from low short-term mortality risk to high short-term mortality risk and are divided into 5 equally sized groups (Quintile 1 to Quantile 5). Based on average population life expectancy (white bars) and on sex- and age-specific life expectancy (grey bars).

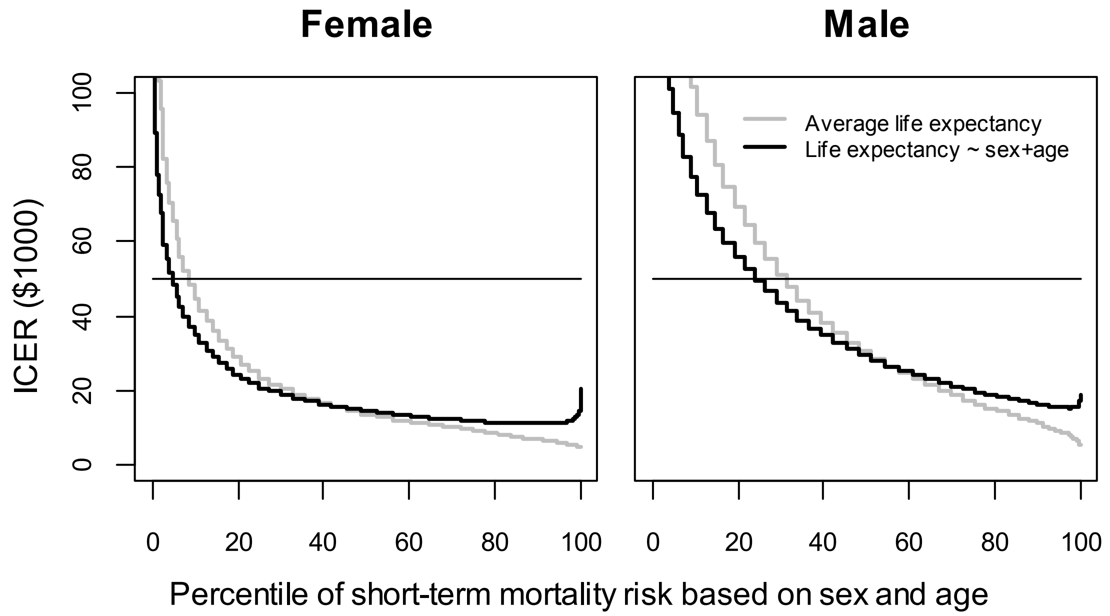


Figure 11.2 Incremental cost-effectiveness ratio (ICER) by percentile of sex- and age-specific short-term mortality risk for female and male patients. Patients are ordered from low short-term mortality risk (Percentile 0) to high short-term mortality risk (Percentile 100). Based on average population life expectancy (grey lines) and on sex- and age-specific life expectancy (black lines). The grey line is below $50,000 for women above the age of 49 and for men above the age of 53. The black line is below $50.000 for women above 45 and for men above 50 years of age.

The attenuation of ICER heterogeneity is further visualized by percentiles of short-term mortality risks for female and male patients separately (Figure 11.2). When only the short-term mortality risk reduction was modelled sex- and age-specific, t-PA treatment was cost-effective for women above the age of 49 (92% of the female population; left panel of Figure 11.2) and men above the age of 53 (69% of the male population; right panel of Figure 11.2). ICER estimates in female patients were lower because of a higher short-term mortality reduction. When both short-term mortality risk reduction and life expectancy were modelled sex- and age-specific, the heterogeneity in cost-effectiveness estimates decreased. As a result cost-effectiveness of t-PA treatment was extended to women above 45 (95% of the female population; left panel of Figure 11.2) and men above 50 years of age (76% of the male population; right panel of Figure 11.2). When future life years were not discounted, ICERs decreased, but the attenuation of ICER heterogeneity was stronger, because the impact of life expectancy assumptions was not moderated by discounting (Figure 11.3).

The effect of negative correlation between mortality risk reduction and life expectancy for individual patients is further illustrated by adding Killip classification as a risk factor. As expected, the effect of applying short-term Killip class hazard ratios to post-trial mortality was stronger with increasing Killip class and with stronger assumptions about the long-term mortality effects of Killip class (Figure 11.4). When short-term and long-term mortality effects of Killip class were assumed equal, t-PA treatment was not cost-effective for older patients with Killip class IV (female above 79; male above 80; lowest right panel of Figure 11.4) because of a low post-trial life expectancy. t-PA treatment was cost-effective for most other patients with worse heart failure than Killip class I.

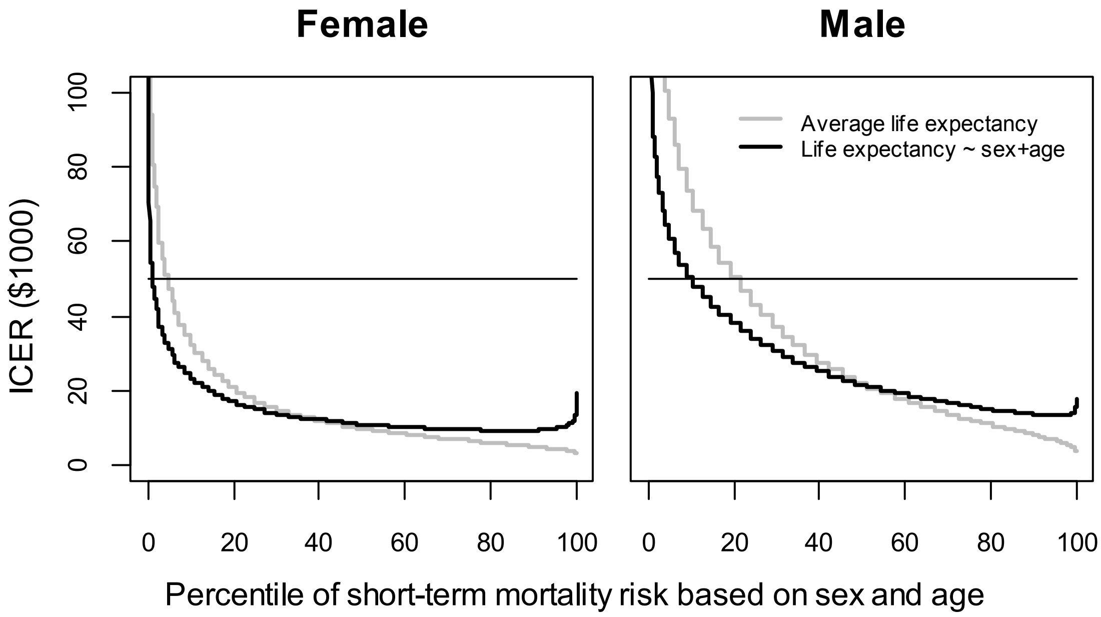


Figure 11.3 Undiscounted incremental cost-effectiveness ratio (ICER) by percentile of sex- and age-specific short-term mortality risk for female and male patients. Patients are ordered from low short-term mortality risk (Percentile 0) to high short-term mortality risk (Percentile 100). Based on average population life expectancy (grey lines) and on sex- and age-specific life expectancy (black lines). The grey line is below $50,000 for women above the age of 45 and for men above the age of 49. The black line is below $50.000 for women above 38 and for men above 44 years of age.

DISCUSSION

We presented two examples of the effect of a negative correlation between the reduction in short-term mortality and post-trial life expectancy at the individual level, i.e. patients at high age had a high short-term mortality risk reduction but a low life expectancy, and similarly, patients in severe Killip class IV had a high risk reduction but a low life expectancy. These examples illustrated the substantial impact on individualized cost-effectiveness estimates of modeling individualized instead of average long-term life expectancy, e.g. the estimate of life years gained for the 20% of patients with the highest sex and age-specific short-term risk decreased from 0.32 to 0.20 when using sex and age-specific life expectancy instead of average population life expectancy. Consequently, when short-term risk reduction and long-term life expectancy were modelled at unequal levels of granularity, individualized estimates

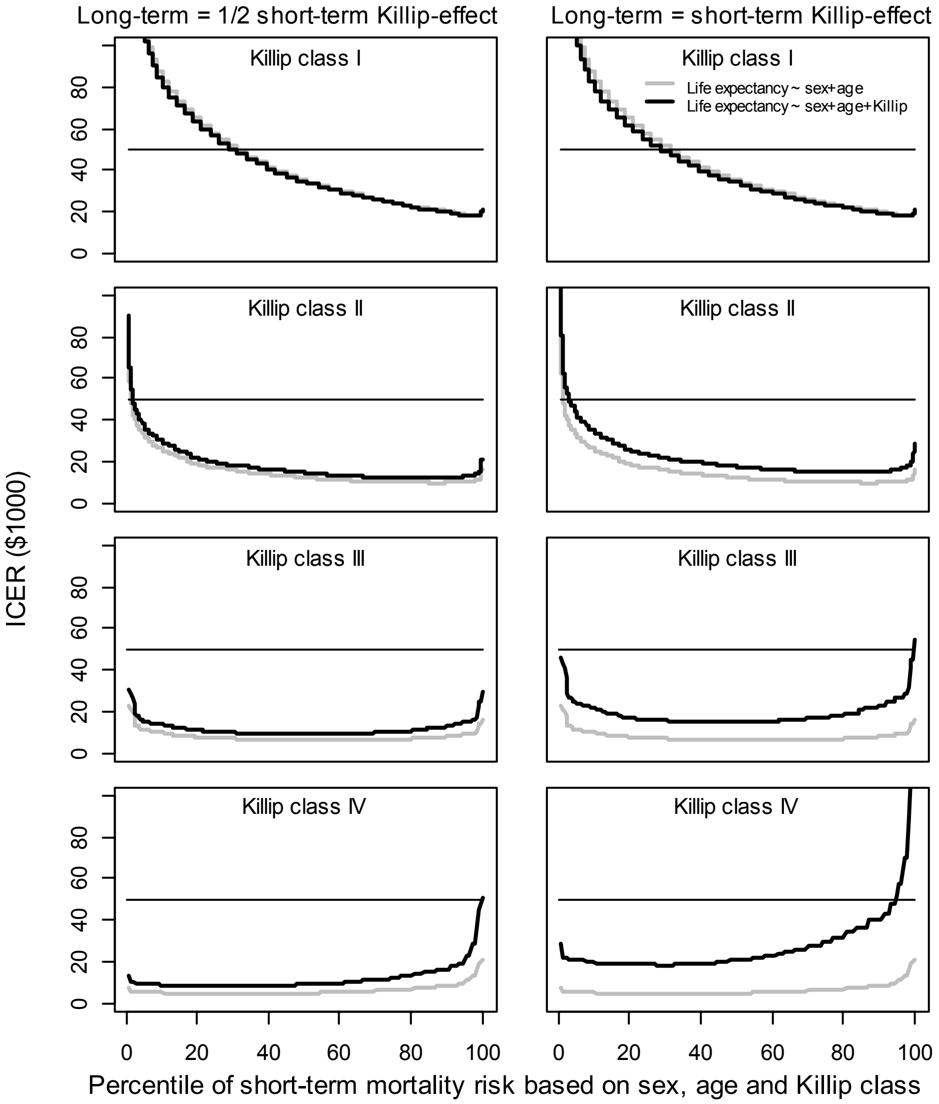


Figure 11.4 Incremental cost-effectiveness ratio (ICER) by percentile of sex- , age- and Killip class-specific short-term mortality risk for male patients. Patients are ordered from low short-term mortality risk (Percentile 0) to high short-term mortality risk (Percentile 100). Based on sex- and age-specific life expectancy (grey lines) and on sex-, age- and Killip class-specific life expectancy (black lines). Short-term Killip hazard ratios (1/2 the effect size, i.e. the square root of the short-term hazard ratios on the left-hand-side) were applied to long-term mortality hazards for individual patients.

of cost-effectiveness were biased, and heterogeneity in cost-effectiveness between low risk and high risk patients (or quintiles of patients) was overestimated.

Because short-term risk models can frequently be developed on detailed clinical data and life expectancy functions are typically obtained from life tables, there is often imbalance between the level of granularity that leads to an exaggeration of the true heterogeneity of individualized cost-effectiveness (and the expected value of individualized care [4]). One example of combining individualized short-term risk estimates with less granular group-level life expectancy estimates can be found in the cost-effectiveness analysis of CT Screening in the NLST [18] which considered risk-based subgroups (quintiles), following the increased attention for individualized screening [19-21]. Risk quintiles were based on the risk factors age, race/ethnicity, education, BMI, COPD, personal history of cancer, family history of lung cancer, and smoking status, intensity, duration, and quit time [20]. Within-trial life-years were observed for approximately 6 years of follow-up. Post-trial life-years were based on 2009 U.S. life-tables, adjusted for smoking status and stage-specific lung cancer mortality, but not for the other lung cancer risk factors. This inconsistency may cause overestimation of the post-trial life-years in the highest risk quintiles, since some risk factors are ignored for their impact on reducing life-expectancy, e.g. smoking intensity [22]. Estimates of long-term benefit may have been too favorable in the two highest lung cancer risk quintiles. The impact on stratified cost-effectiveness estimates of CT screening will probably be less than in our case study of t-PA cost-effectiveness, because the post-trial life expectancy estimates are less influential with a longer trial follow-up (6 years versus 30 days).

We raise awareness that using simple models for life expectancy estimates may lead to overestimation of individualized cost-effectiveness heterogeneity. On the other hand, using simpler models for short-term risk reduction estimates may lead to underestimation of the individualized cost-effectiveness heterogeneity. Instead of limiting the individualization of short-term risk we would like to promote a better understanding of the consequences of unequally granular risk models by means of sensitivity analysis. In our case study, we accounted for correlation between short-term and long-term mortality due to severity of heart failure by applying short-term risk ratios of Killip classes to long-term post-trial mortality. This may be a practical approach to analyze the sensitivity of individualized cost-effectiveness estimates when information about the effects of short-term risk factors on long-term mortality is missing. It prevents overenthusiasm on the value of individualization of risk estimates, which would occur by considering short-term impact only. Ideally the effects of short-term risk factors on long-term mortality are derived from a comprehensive data set. Using published models may also be sensible, and better than ignoring impact of predictors on long-term life expectancy. Regardless of the approach that is chosen for harmonizing the granularity of short-term and long-term mortality, the uncertainty in individualized cost-effectiveness estimates depends on the uncertainty in estimates of both short-term and long-term mortality. Since this paper is about biases rather than uncertainty, quantification of uncertainty in individualized cost-effectiveness estimates requires further research.

We assumed a constant relative treatment effect in our analyses, which implies that the absolute risk reduction of t-PA treatment is an increasing function of baseline risk [9, 13, 14]. Relative treatment effect heterogeneity – i.e. variation of relative treatment effect across individuals – could be incorporated in an analysis of individual cost-effectiveness, but may be sensitive to the pitfalls of subgroup analysis [23].

Individualizing cost-effectiveness requires valid prediction models, both in terms of discrimination and calibration. Overestimation of the average risk (miscalibration) leads to overoptimistic cost-effectiveness estimates. When regression model coefficients are overestimated, the discriminative ability is overoptimistic [24]. Hence, differences in individual cost-effectiveness between individuals are overstated. Further research is necessary to study the relation between classical measures for model performance such as the Area Under the receiver operating characteristic Curve (AUC) and the value of using individualized cost-effectiveness estimates for guiding treatment decisions.

Heterogeneity of patient preferences may also be an important aspect when individualizing cost-effectiveness assessments. For example, the decision whether to screen for prostate cancer or treat patients with localized prostate cancer depends heavily on patients' individual valuations of the outcomes of treatment because screening and treatment result in life expectancy gain but quality-adjusted life expectancy loss [25, 26]. We focused on the relationship between baseline risk and life expectancy after MI. Since quality of life may be correlated with baseline risk, individualized estimates of utility are desirable, again at an equal level of granularity [27].

CONCLUSION

Failure to model short-term risk reduction and life expectancy at an equal level of granularity may bias individualized estimates of cost-effectiveness, exaggerate the true heterogeneity of individualized cost-effectiveness and therefore misallocate resources when risk models are used for targeting care. Individualized cost-effectiveness analysis should be based on estimates of long-term life expectancy at the same level of granularity as the estimates of the short-term risk reduction.

ACKNOWLEDGMENTS

The authors express their gratitude to all of the principal investigators and patients participating in the GUSTO trial. This work was supported by the National Institutes of Health (grant U01NS086294).

REFERENCES

1. Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, Luce BR, Studies ITFoGRP--M. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices--Modeling Studies. *Value Health* 2003; 6: 9-17.
2. Stevens W, Normand C. Optimisation versus certainty: understanding the issue of heterogeneity in economic evaluation. *Soc Sci Med* 2004; 58: 315-320.
3. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, Augustovski F, Briggs AH, Mauskopf J, Loder E, Force IHEEPG-CGRPT. Consolidated Health Economic Evaluation Reporting Standards (CHEERS)--explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. *Value Health* 2013; 16: 231-250.
4. Basu A, Meltzer D. Value of information on preference heterogeneity and individualized care. *Med Decis Making* 2007; 27: 112-127.
5. Sculpher M. Subgroups and heterogeneity in cost-effectiveness analysis. *PharmacoEconomics* 2008; 26: 799-806.
6. Sculpher M. Reflecting heterogeneity in patient benefits: the role of subgroup analysis with comparative effectiveness. *Value Health* 2010; 13 Suppl 1: S18-21.
7. Ioannidis JPA, Garber AM. Individualized Cost-Effectiveness Analysis. *PLoS Med* 2011; 8: e1001058.
8. Vickers AJ, Kent DM. The Lake Wobegon Effect: Why Most Patients Are at Below-Average Risk. *Ann Intern Med* 2015; 162: 866-867.
9. Kent DM, Vijan S, Hayward RA, Griffith JL, Beshansky JR, Selker HP. Tissue plasminogen activator was cost-effective compared to streptokinase in only selected patients with acute myocardial infarction. *J Clin Epidemiol* 2004; 57: 843-852.
10. Lee KL, Woodlief LH, Topol EJ, Weaver WD, Betriu A, Col J, Simoons M, Aylward P, Van de Werf F, Califf RM. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction. Results from an international trial of 41,021 patients. GUSTO-I Investigators. *Circulation* 1995; 91: 1659-1668.
11. Parakh K, Thombs BD, Bhat U, Fauerbach JA, Bush DE, Ziegelstein RC. Long-term Significance of Killip Class and Left Ventricular Systolic Dysfunction. *The American Journal of Medicine* 2008; 121: 1015-1018.
12. de Carvalho LP, Gao F, Chen Q, Sim L-L, Koh T-H, Foo D, Ong H-Y, Tong K-L, Tan H-C, Yeo T-C, Chow K-Y, Richards AM, Peterson ED, Chua T, Chan MY. Long-Term Prognosis and Risk Heterogeneity of Heart Failure Complicating Acute Myocardial Infarction. *The American Journal of Cardiology* 2015; 115: 872-878.
13. Mark DB, Hlatky MA, Califf RM, Naylor CD, Lee KL, Armstrong PW, Barbash G, White H, Simoons ML, Nelson CL. Cost effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction. *N Engl J Med* 1995; 332: 1418-1424.
14. Califf RM, Woodlief LH, Harrell FE, Jr., Lee KL, White HD, Guerci A, Barbash GI, Simes RJ, Weaver WD, Simoons ML, Topol EJ. Selection of thrombolytic therapy for individual patients: development of a clinical model. GUSTO-I Investigators. *Am Heart J* 1997; 133: 630-639.
15. Arias E. United States Life Tables, 2009. *National Vital Statistics Reports* 2014; 62.
16. Killip T, 3rd, Kimball JT. Treatment of myocardial infarction in a coronary care unit. A two year experience with 250 patients. *Am J Cardiol* 1967; 20: 457-464.
17. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/, 2013.

18. Black WC, Gareen IF, Soneji SS, Sicks JD, Keeler EB, Aberle DR, Naeim A, Church TR, Silvestri GA, Gorelick J, Gatsonis C, National Lung Screening Trial Research T. Cost-effectiveness of CT screening in the National Lung Screening Trial. *N Engl J Med* 2014; 371: 1793-1802.
19. Raji OY, Duffy SW, Agbaje OF, Baker SG, Christiani DC, Cassidy A, Field JK. Predictive accuracy of the Liverpool Lung Project risk model for stratifying patients for computed tomography screening for lung cancer: a case-control and cohort validation study. *Ann Intern Med* 2012; 157: 242-250.
20. Tammemagi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, Chaturvedi AK, Silvestri GA, Riley TL, Commins J, Berg CD. Selection criteria for lung-cancer screening. *N Engl J Med* 2013; 368: 728-736.
21. Kovalchik SA, Tammemagi M, Berg CD, Caporaso NE, Riley TL, Korch M, Silvestri GA, Chaturvedi AK, Katki HA. Targeting of low-dose CT screening according to the risk of lung-cancer death. *N Engl J Med* 2013; 369: 245-254.
22. Rosenberg MA, Feuer EJ, Yu B, Sun J, Henley SJ, Shanks TG, Anderson CM, McMahon PM, Thun MJ, Burns DM. Chapter 3: Cohort life tables by smoking status, removing lung cancer as a cause of death. *Risk Anal* 2012; 32 Suppl 1: S25-38.
23. Ramsey S, Willke R, Briggs A, Brown R, Buxton M, Chawla A, Cook J, Glick H, Liljas B, Petitti D, Reed S. Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA Task Force report. *Value Health* 2005; 8: 521-533.
24. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* Springer: New York, 2009.
25. Krahn MD, Mahoney JE, Eckman MH, Trachtenberg J, Pauker SG, Detsky AS. Screening for prostate cancer. A decision analytic view. *JAMA* 1994; 272: 773-780.
26. Cowen ME, Miles BJ, Cahill DF, Giesler RB, Beck JR, Kattan MW. The danger of applying group-level utilities in decision analyses of the treatment of localized prostate cancer in individual patients. *Med Decis Making* 1998; 18: 376-380.
27. Lewis EF, Li Y, Pfeffer MA, Solomon SD, Weinfurt KP, Velazquez EJ, Califf RM, Rouleau JL, Kober L, White HD, Schulman KA, Reed SD. Impact of cardiovascular events on change in quality of life and utilities in patients after myocardial infarction: a VALIANT study (valsartan in acute myocardial infarction). *JACC Heart Fail* 2014; 2: 159-165.

PART III

APPLICATIONS

12

Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score II

V Farooq*
D van Klaveren*
EW Steyerberg
E Meliga
Y Vergouwe
A Chieffo
AP Kappetein
A Colombo
DR Holmes
M Mack
T Feldman
MC Morice
E Stahle
Y Onuma
MA Morel
HM Garcia-Garcia
GA van Es
KD Dawkins
FW Mohr
PW Serruys

Lancet 2013

* Equal contribution

ABSTRACT

Background The anatomical SYNTAX score is advocated in European and US guidelines as an instrument to help clinicians decide the optimum revascularisation method in patients with complex coronary artery disease. The absence of an individualised approach and of clinical variables to guide decision making between coronary artery bypass graft surgery (CABG) and percutaneous coronary intervention (PCI) are limitations of the SYNTAX score. SYNTAX score II aimed to overcome these limitations.

Methods SYNTAX score II was developed by applying a Cox proportional hazards model to results of the randomised all comers SYNTAX trial (n=1800). Baseline features with strong associations to 4-year mortality in either the CABG or the PCI settings (interactions), or in both (predictive accuracy), were added to the anatomical SYNTAX score. Comparisons of 4-year mortality predictions between CABG and PCI were made for each patient. Discriminatory performance was quantified by concordance statistics and internally validated with bootstrap resampling. External validation was done in the multinational all comers DELTA registry (n=2891), a heterogeneous population that included patients with three-vessel disease (26%) or complex coronary artery disease (anatomical SYNTAX score ≥33, 30%) who underwent CABG or PCI. The SYNTAX trial is registered with ClinicalTrials.gov, number NCT00114972.

Findings SYNTAX score II contained eight predictors: anatomical SYNTAX score, age, creatinine clearance, left ventricular ejection fraction (LVEF), presence of unprotected left main coronary artery (ULMCA) disease, peripheral vascular disease, female sex, and chronic obstructive pulmonary disease (COPD). SYNTAX score II significantly predicted a difference in 4-year mortality between patients undergoing CABG and those undergoing PCI ($p_{interaction}$ 0.0037). To achieve similar 4-year mortality after CABG or PCI, younger patients, women, and patients with reduced LVEF required lower anatomical SYNTAX scores, whereas older patients, patients with ULMCA disease, and those with COPD, required higher anatomical SYNTAX scores. Presence of diabetes was not important for decision making between CABG and PCI ($p_{interaction}$ 0.67). SYNTAX score II discriminated well in all patients who underwent CABG or PCI, with concordance indices for internal (SYNTAX trial) validation of 0.725 and for external (DELTA registry) validation of 0.716, which were substantially higher than for the anatomical SYNTAX score alone (concordance indices of 0.567 and 0.612, respectively). A nomogram was constructed that allowed for an accurate individualised prediction of 4-year mortality in patients proposing to undergo CABG or PCI.

Interpretation Long-term (4-year) mortality in patients with complex coronary artery disease can be well predicted by a combination of anatomical and clinical factors in SYNTAX score II. SYNTAX score II can better guide decision making between CABG and PCI than the original anatomical SYNTAX score.

INTRODUCTION

The anatomical SYNTAX score is an important instrument that can help clinicians to establish the optimum revascularisation approach in patients with complex coronary artery disease (with or without unprotected left main coronary artery [ULMCA] involvement) [1–5]. It is advocated in both European and US revascularisation guidelines [6,7]. These guidelines also state that clinical variables should be taken into account during discussion between multidisciplinary teams consisting of a clinical cardiologist, cardiac surgeon, and interventional cardiologist (the so-called heart team approach) when deciding the best treatment method; absence of clinical variables is a limitation of the SYNTAX score.

In patients with ULMCA disease, a low-intermediate SYNTAX score (<33) was shown to have much the same long-term clinical outcomes—including all-cause mortality and major cardiovascular and cerebrovascular events—with coronary artery bypass graft surgery (CABG) and percutaneous coronary intervention (PCI) in the SYNTAX trial [1,4,8]. This finding formed the basis for the US Food and Drug Administration's decision to accept a SYNTAX score of less than 33 as an entry criterion to the ongoing international multicentre EXCEL (Evaluation of XIENCE prime versus Coronary artery bypass surgery for Effectiveness of Left main revascularisation) trial, aiming to recruit 2600 patients with ULMCA disease (NCT01205776) [9].

In patients with three-vessel disease, a low SYNTAX score (<23) was shown to have much the same long-term clinical outcomes between CABG and PCI in the SYNTAX trial [1,8]. A substudy of the SYNTAX trial [10] has, however, suggested that patients with high clinical comorbidity (i.e. additive EuroSCORE ≥ 6 [11]) with three- vessel disease, irrespective of the anatomical complexity (SYNTAX score), might potentially derive a prognostic benefit from undergoing CABG compared with PCI, provided an acceptable threshold of operative risk is not exceeded. Some researchers have therefore suggested that the SYNTAX score, and other category-based scores [10,12–15] that labelled patients as low risk, intermediate risk, or high risk might be concealing higher risk patients in lower risk groups, or vice versa.

The purpose of our study was to augment the SYNTAX score with prognostically important clinical variables to form SYNTAX score II, to better guide decision making between CABG and PCI. Additionally, SYNTAX score II should provide the basis for individualised decision making between CABG and PCI, by contrast with the present strategy of grouping (low, intermediate, or high) risk to patients.

METHODS

SYNTAX trial

The SYNTAX trial was a randomised, prospective, multicentre trial (85 centres in 18 countries) with an allcomers design [1,5,8]. Exclusion criteria were minimal and consisted of previous coronary revascularisation, concomitant cardiac surgery (valve or resection of aortic or left ventricular aneurysm) or acute myocardial infarction, and cardiac enzymes more than twice as high as the normal limit. Patients with ULMCA disease (isolated or associated with one-vessel, two-vessel, or three-vessel disease) or de-novo three-vessel disease were randomised on a 1:1 basis to CABG or PCI with TAXUS Express paclitaxel-eluting stent (Boston Scientific Corporation, Natick, MA, USA; n=1800). Randomisation of patients was stratified by clinical site, absence or presence of ULMCA disease, and medically treated diabetes (requiring oral medications or insulin). Patients deemed unsuitable for randomisation by the cardiologist were nested in registries. An independent clinical events committee reviewed all primary clinical endpoints [1].

The anatomical SYNTAX score [1–3] combines the importance of a diseased coronary artery segment by vessel-segment weighting (Leaman score), adverse lesion characteristics (American College of Cardiology/American Heart Association lesion classification, and total occlusion characteristics from the European TOTAL Surveillance Study), and the Medina classification system for bifurcation lesions [14,16–18]. Calculation of anatomical SYNTAX score was done by the heart team before randomisation, and corroborated by an independent core laboratory (Cardialysis BV, Rotterdam, Netherlands), blinded to treatment assignment. Clinical variables were also prospectively collected as part of the original SYNTAX trial [1]. Chronic obstructive pulmonary disease (COPD) was defined as the long-term use of bronchodilators or steroids for lung disease (EuroSCORE definition [11]). Peripheral vascular disease was defined as aorta and arteries other than coronaries, with exercise-related claudication, or revascularisation surgery, or reduced or absent pulsation, or angiographic stenosis of more than 50%, or combinations of these characteristics (Arterial Revascularisation Therapies Study Part I [ARTS I] definition [19]). Preprocedural left ventricular ejection fraction (LVEF) was taken by transthoracic echocardiography or diagnostic left ventriculography, and categorised as good (≥50%), moderate (30–49%), or poor (<30%). Creatinine clearance, a measure of estimated glomerular filtration rate, was defined by the Cockcroft and Gault formula [20].

Within the SYNTAX trial, most predictor values were more than 98% complete. Creatinine clearance was 91% complete, LVEF was 98.4% complete when recorded categorically (good, moderate, and poor) and 62.6% complete when recorded continuously (numerical value). An advanced multiple imputation strategy, which takes the correlation between all potential predictors (method of chained equations [Hmisc package version 3.8-3, in R software version 2.13.2]), and sensitivity analyses were done to account for missing values [21,22]. For the multiple imputation of missing values of the continuous variable LVEF, categories of LVEF were additionally considered. All analyses were done for the imputed datasets, and repeated with only complete data, which gave much the same results. The SYNTAX trial is registered with ClinicalTrials.gov, number NCT00114972.

SYNTAX score II

Combination of anatomical SYNTAX score with three simple clinical variables (age, creatinine or creatinine clearance, LVEF) has been shown to contain most of the prognostic information in predicting mortality after PCI (including the SYNTAX score—logistic clinical SYNTAX score [23]) or CABG (excluding the SYNTAX score—ACEF score [24,25]). Consequently, we developed SYNTAX score II on the basis of a core model consisting of anatomical SYNTAX score, age, creatinine clearance, and LVEF. Other common independent predictors of mortality, using the baseline characteristics of the CABG and PCI cohorts of the SYNTAX trial, were screened and identified with a multivariable Cox proportional hazards model (appendix), the findings of which have been reported [24]. As a result, we added peripheral vascular disease to the core model. Notably, medically treated diabetes—despite being stratified at randomisation in the SYNTAX trial [5] and reported in 26% of patients—was not added to the core model, because it was shown not to be an independent predictor of mortality in the CABG and PCI groups of the SYNTAX trial [24].

The SYNTAX score aids decision making between CABG and PCI; it is more predictive of clinical outcomes in patients undergoing PCI than in those undergoing CABG, for whom it is not predictive [2,10]. This discrepancy means that the SYNTAX score has a significant interaction effect [25] with CABG and PCI in establishing long-term mortality. Interaction for a particular baseline characteristic was defined as the hazard ratio (HR) of mortality associated with that characteristic among patients undergoing PCI (HR_{PCI}), divided by the HR for the same characteristic among patients undergoing CABG (HR_{CABG})—i.e., HR_{PCI}/HR_{CABG}. Consistent with this principle, we screened baseline characteristics (appendix), and added variables to the core model when they showed an interaction between CABG and PCI ($p<0.10$) in affecting 4-year mortality, using a multivariable Cox proportional hazards model. To aid visualisation of the interaction effects, we constructed graphs by plotting the log HR for 4-year mortality (CABG and PCI) on the y axis, and the predictors on the x axis.

Since the SYNTAX trial data originated from 18 countries, we assessed whether between-country heterogeneity existed using gamma frailty in the Cox proportional hazards model [26]. Between-country heterogeneity had a negligible effect on parameter estimates (data not shown). Consequently we present pooled country results.

We constructed scatter plots to visualise how the SYNTAX score II predicted long-term mortality in each individual patient from the CABG and PCI cohorts of the SYNTAX trial (n=1800). On the basis of the Cox proportional hazards model parameter estimates, we made individual risk predictions for each patient as if they had undergone CABG or PCI. The predicted hazards (relative to the average predicted hazard) for CABG versus PCI were plotted. We used a log scale to allow for good separation of individual predicted risks. Bootstrap analyses (1000 resamples with replacement) were done to assess whether individual predictions for CABG were either higher or lower compared with the predictions for PCI with 95% confidence (p<0.05), and are highlighted in the scatter plots [22]. Similar scatter plots were produced for the SYNTAX score tertiles [2] and clinical variables in SYNTAX score II to allow judgment of their collective effect on long-term mortality predictions.

We did reclassification analyses (i.e. what proportion of patients moved from high to low risk, and from low to high risk) following the principles of Net Reclassification Improvement [27]. The proportion of patients reclassified was calculated with bootstrap analyses (1000 resamples), and are represented in the scatter plots.

External validation of SYNTAX score II

External validation was done in the Drug Eluting stent for LefT main coronary Artery disease (DELTA) registry [28]. Briefly, this is a multinational (14 centres in Europe, the USA, and South Korea), non-randomised, allcomers registry (n=2891) of ULMCA disease (isolated, or associated with single or multivessel disease) treated with CABG (n=902, 31.2%) or first generation sirolimus-eluting or paclitaxel-eluting stents (n=1989, 68.8%). SYNTAX scores of 33 or more were reported in 30% (n=871) of the study population and three-vessel disease in 26% (n=744). Previous CABG or PCI was permitted. Follow-up was similar to the SYNTAX trial, with a median period of 3.5 years (IQR 2.5–4.6 years). Data were reported by the study sites in 2039 of 2891 (70.5%) patients for the SYNTAX score, 2785 (96.3%) for age, 2891 (100%) for creatinine (categorical variable [above 150 µmol/L]) and sex, 2747 (95.0%) for LVEF (continuous variable), 1363 (47.1%) for peripheral vascular disease, and 1061 (36.7%) for COPD. Because renal function was collected categorically in the DELTA registry, median renal function from the SYNTAX trial data, with or without a creatinine concentration above 150 µmol/L, was used to replace creatinine with a continuous variable. Since renal function had only a weak inter- action effect in affecting mortality between CABG and PCI, this approach was judged to be acceptable. Multiple imputation of missing values and sensitivity analyses, identical to that described earlier, yielded much the same results for the complete dataset. Specifically, sensitivity analyses [21,22,29] testing the robustness of the interaction effects for COPD and peripheral vascular disease gave similar results.

SYNTAX score II assessments

The most important aspects of the SYNTAX score II are the interaction effects, since they drive decision making between CABG and PCI. Other measures are needed to ensure the accuracy of individual mortality predictions for CABG and PCI, and the magnitude of their differences (e.g. 10% vs 5% or 2% vs 1% predicted risks for CABG and PCI), which will further support decision making. These measures include discrimination with the concordance index, and calibration (agreement between predicted and observed risks) with validation plots. The concordance index estimates the probability that, of two randomly chosen patients, the patient with the higher prognostic score will outlive the one with the lower prognostic score [25]. Values of the concordance index range from 0.5 (no discrimination) to a theoretical maximum of 1. The concordance index was internally validated with a bootstrap procedure (100 resamples with replacement) to correct for optimism in parameter estimates [22].

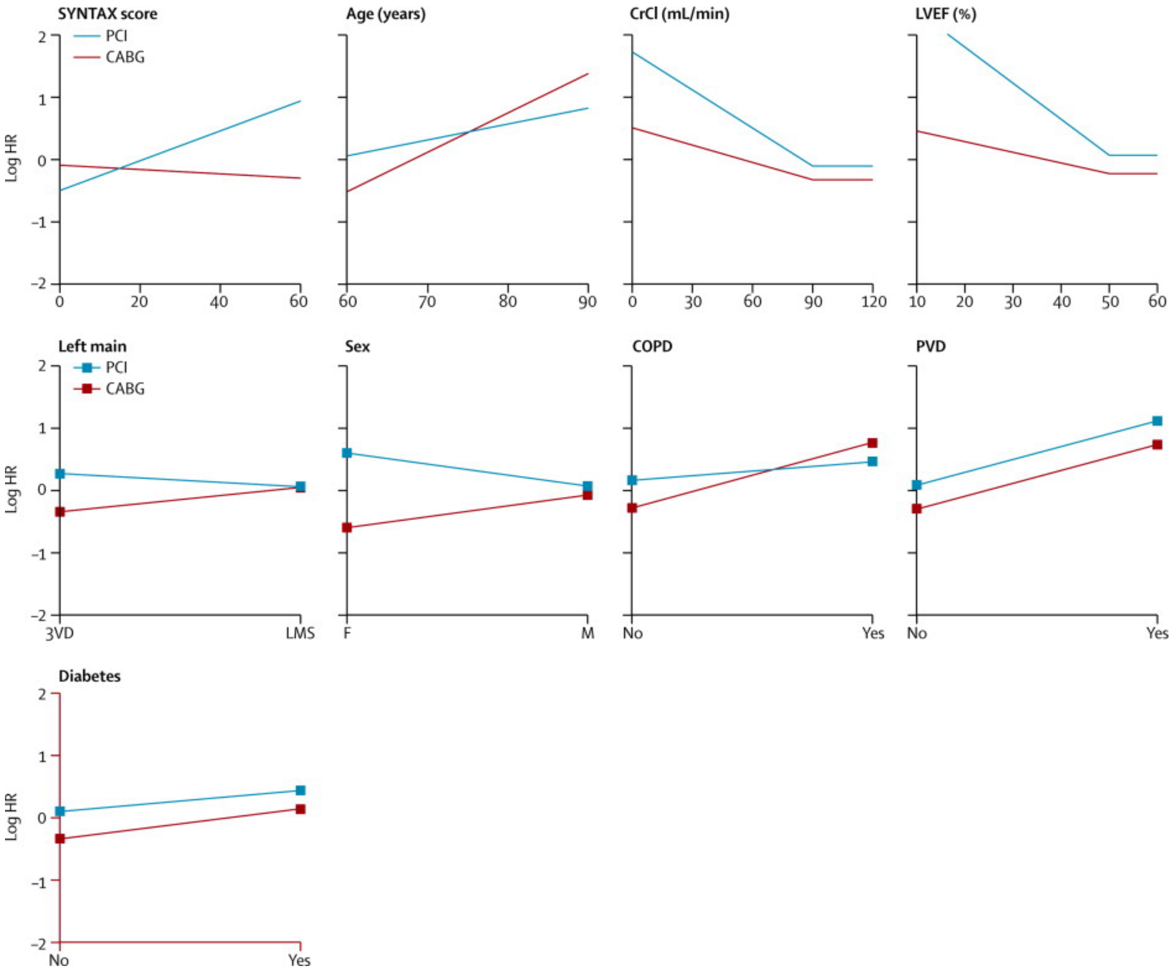


Figure 12.1 Predictor effects for CABG and PCI in SYNTAX score II. Predictor effects are represented visually as a log HR for CABG and PCI on the y-axis for each predictor. Each predictor is expressed on the x-axis continuously (upper panels) or categorically (lower panels), for a person of mean baseline characteristics. Diabetes is included to show its absence of interaction when included in the analyses. Note the differing gradients of the hazards for PCI and CABG, leading to the hazards crossing at an anatomical SYNTAX score of 15. At this cross-over point of hazards, the mortality risk is much the same between CABG and PCI. This threshold of cross-over of hazards will vary according to the level of other variables, namely being lower for female sex, reduced LVEF, and younger age, and higher for COPD, left main disease, and older age. Because both peripheral vascular disease (p=1.00) and diabetes (p=0.67) lacked an interaction effect, as shown by almost parallel HRs (ie, similar increase in mortality risk), their presence would have no effect on decision making between CABG and PCI. Since diabetes was also not an independent predictor of mortality (appendix) [24], it was excluded from SYNTAX score II. CABG=coronary artery bypass surgery. PCI=percutaneous coronary intervention. HR=hazard ratio. CrCl=creatinine clearance. LVEF=left ventricular ejection fraction. Left main=unprotected left main coronary artery disease. 3VD=three-vessel disease. LMS=left main stem. COPD=chronic obstructive pulmonary disease. PVD=peripheral vascular disease.

SYNTAX score II presentation

We present SYNTAX score II as a nomogram, with scores assigned for the presence and magnitude of each predictor directly based on the Cox proportional hazards model coefficients [22]. All statistical analyses were done with Harrell's Regression Modelling Strategies (rms version 3.4-0) package in R software (version 2.13.2) [21,30,31].

Role of the funding source

SYNTAX trial design and conduct was overseen by the SYNTAX steering committee, on which representatives of the study sponsor served. Data analysis and interpretation and writing of the report were independent from the study sponsor. The authors had unrestricted access to the full study database. PWS and EWS took the final responsibility for the decision to submit for publication.

Table 12.1 Development (SYNTAX Trial) and validation (DELTA Registry) data for SYNTAX score II.

	Multivariable adjusted HR (95% CI)		Interaction effect (HR_{PCI}/HR_{CABG}); HR (95% CI; p value)
	CABG 4-year mortality	PCI 4-year mortality	
Development population , SYNTAX trial (n=1800)			
Anatomical SYNTAX score (per 10 point increase)	0.97 (0.79–1.18)	1.27 (1.08–1.50)	1.32 (1.01–1.71; p=0.039)
Age (per 10 year increase)	1.88 (1.34–2.64)	1.29 (0.97–1.71)	0.69 (0.44–1.07; p=0.095)
Creatinine clearance† (per 10 mL/min increase)	0.91 (0.77–1.07)	0.82 (0.72–0.93)	0.89 (0.73–1.10; p=0.30)
LVEF (per 10% increase)	0.84 (0.61–1.16)	0.56 (0.43–0.73)	0.67 (0.44–1.00; p=0.053)
Peripheral vascular disease*	2.79 (1.66–4.71)	2.79 (1.72–4.53)	1.00 (0.49–2.04; p=1.00)
ULMCA disease	1.47 (0.93–2.34)	0.82 (0.54–1.23)	0.56 (0.30–1.03; p=0.062)
Women	0.59 (0.32–1.10)	1.70 (1.11–2.60)	2.87 (1.35–6.07; p=0.0059)
COPD	2.84 (1.64–4.90)	1.35 (0.74–2.47)	0.48 (0.21–1.08; p=0.074)
External validation population, DELTA registry (n=2891)			
Anatomical SYNTAX score (per 10 point increase)	1.12 (0.95–1.32)	1.32 (1.20–1.46)	1.18 (0.98–1.42; p=0.083)
Age (per 10 year increase)	1.46 (1.15–1.85)	1.34 (1.19–1.52)	0.92 (0.70–1.21; p=0.56)
Creatinine clearance (per 10 mL/min increase)	0.91 (0.78–1.06)	0.93 (0.86–1.00)	1.02 (0.86–1.21; p=0.82)
LVEF (per 10% increase)	0.59 (0.47–0.75)	0.57 (0.50–0.65)	0.96 (0.72–1.27; p=0.75)
Peripheral vascular disease	1.37 (0.68–2.79)	1.77 (1.01–3.09)	1.29 (0.51–3.22; p=0.59)
Women	0.52 (0.31–0.87)	1.09 (0.82–1.46)	2.09 (1.16–3.76; p=0.014)
COPD	3.63 (1.31–10.04)	1.97 (0.88–4.42)	0.54 (0.20–1.47; p=0.23)

Hazard ratios (HR) in a multivariable Cox proportional hazards model for SYNTAX score II are shown for the CABG and PCI cohorts, followed by the interaction effects (HR_{PCI}/HR_{CABG}) in affecting long-term mortality between CABG and PCI. CABG=coronary artery bypass graft. PCI=percutaneous coronary intervention. HR=hazard ratio. ULMCA=unprotected left main coronary artery. LVEF=left ventricular ejection fraction. COPD=chronic obstructive pulmonary disease. *Retained in SYNTAX score II to improve the predictive accuracy (discrimination) of the 4-year mortality predictions in the CABG and PCI cohorts.

RESULTS

In the randomised SYNTAX population (n=1800), base- line demographics and clinical characteristics for the CABG (n=897) and PCI (n=903) groups were well balanced and have been described previously (appendix) [1]. At 4-years follow-up, clinical data were available in 819 of 897 patients in the CABG group and 879 of 903 patients in the PCI group. 178 all-cause deaths were recorded (CABG 9.0%, 74 all-cause deaths; PCI 11.8%, 104 all- cause deaths, log rank p value=0.063).

The final developed SYNTAX score II consisted of two anatomical (SYNTAX score and ULMCA disease) and six clinical variables (age, creatinine clearance, LVEF, sex, COPD, and peripheral vascular disease). The interaction effect of the SYNTAX score II in collectively affecting long-term mortality predictions between CABG and PCI was significant ($p_{interaction}$=0.0037). Figure 12.1 shows the interaction effects of the eight SYNTAX score II variables. Six of the eight SYNTAX score II variables—anatomical SYNTAX score, age, LVEF, ULMCA disease, COPD, and female sex—showed a moderate to strong interaction effect in affecting long-term mortality predictions with CABG and PCI ($p_{interaction}$<0.10; Table 12.1). Creatinine clearance ($p_{interaction}$=0.30) and peripheral vascular disease ($p_{interaction}$=1.00) showed weak or negligible interaction effects. Diabetes was not included in the SYNTAX score II because it was not an independent predictor of mortality [24] and did not have an interaction effect (p=0.67) with CABG and PCI for long-term mortality (Figure 12.1).

Figure 12.2 shows scatter plots for individual patients in the left main cohort and three-vessel disease cohort of the SYNTAX trial, and by tertiles of the anatomical SYNTAX score. Individual predictions plotted to the left of the diagonal line favoured CABG, and to the right favoured PCI. Individual predictions for CABG and PCI that could not be separated with 95% confidence (i.e. p>0.05) are highlighted in grey, and had a similar 4-year mortality in the SYNTAX trial (i.e. could not be statistically separated).

For the left main cohort, on the basis of the numerical values of the mortality predictions, CABG was favoured in 50.1% (353) and PCI in 49.9% (352) of the SYNTAX population (figure 2). 62.8% (140) of patients in the low (0–22) SYNTAX score tertile, 61.7% (121) in the intermediate (23–32) tertile, and 31.8% (91) in the high (>32) tertile had numerically lower 4-year mortality predictions for PCI compared with CABG. In 79.7% (562) of patients, 4-year mortality predictions between CABG and PCI could not be significantly separated (p>0.05). 18.8% (42) of patients in the low SYNTAX score tertile had mortality predictions separated with statistical significance (p<0.05) in favour of PCI, and 19.2% (55) of patients in the high SYNTAX score tertile, had statistically significant mortality predictions in favour of CABG.

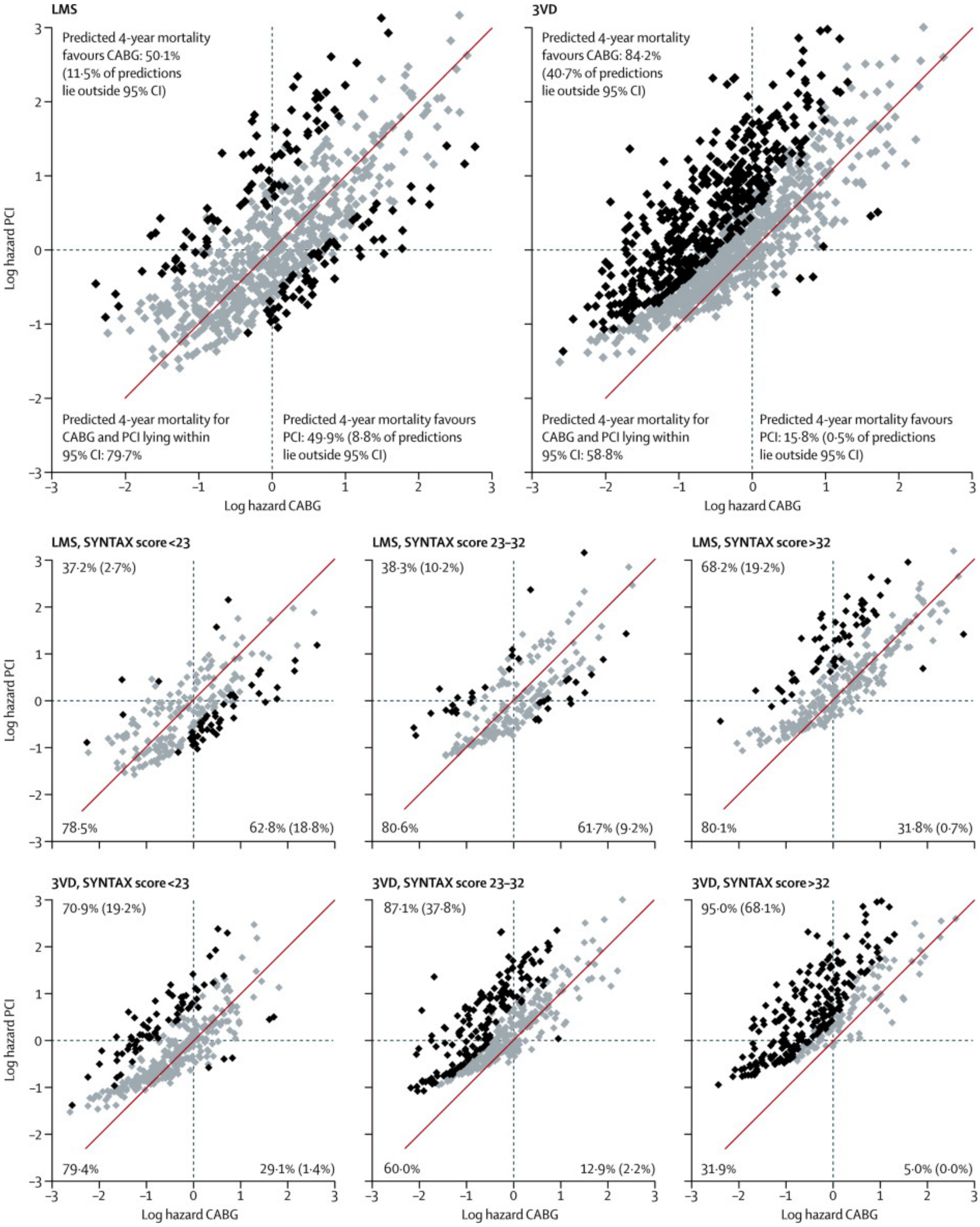
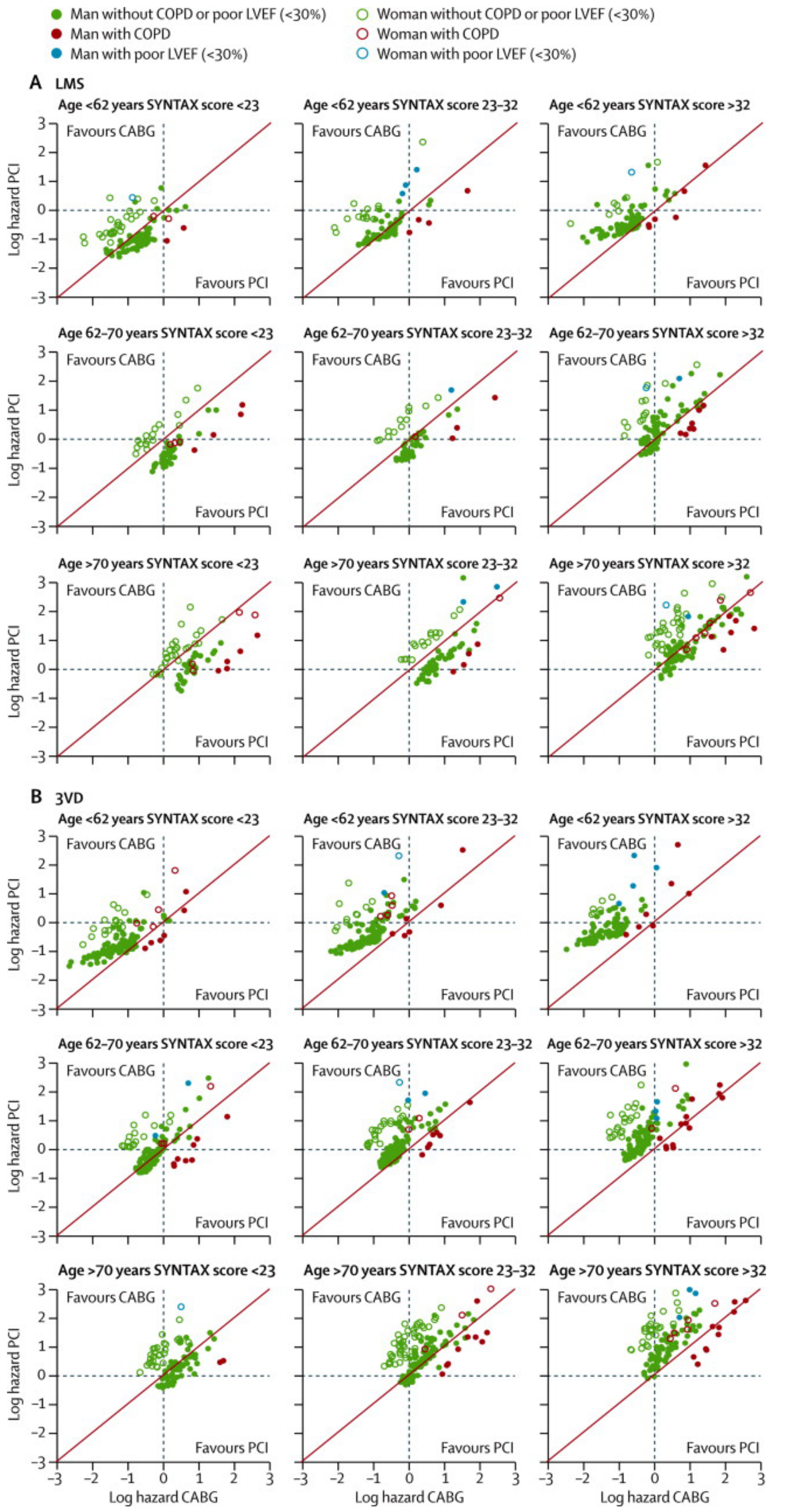


Figure 12.2 Mortality predictions for CABG versus PCI for each individual patient in the randomised SYNTAX trial. The SYNTAX trial included 1800 participants, separated into LMS cohort and 3VD cohort (upper panels), and by tertiles of the anatomical SYNTAX score (lower panels). The diagonal line represents identical mortality predictions for CABG and PCI. Individual predictions plotted to the left of the diagonal line favour CABG (actual percentages shown in top left corner), and to the right favour PCI (actual percentages shown in bottom right corner). Individual mortality predictions for CABG or PCI that could be separated with 95% confidence (p<0.05) are coloured black (actual percentage shown in parentheses in respective corners). Mortality predictions that could not be separated with 95% confidence (p>0.05) are highlighted in grey, and identify patients with similar 4-year mortality. Percentages of patients in each category are shown. CABG=coronary artery bypass surgery. PCI=percutaneous coronary intervention. LMS=left main stem. 3VD=three-vessel disease.



Figures 12.3 Collective effect of SYNTAX score and other anatomical and clinical variables on mortality predictions. LMS cohort (A) 3VD cohort (B). Scatter plots are for illustrative purposes only. CABG= coronary artery bypass surgery. PCI=percutaneous coronary intervention. LMS=left main stem. 3VD=three-vessel disease.

For the three-vessel disease cohort, on the basis of the numerical values of the mortality predictions, CABG was favoured in 84.2% (922) and PCI in 15.8% (173) of the SYNTAX population (figure 2). 29.1% (103) of patients in the low SYNTAX score tertile (0–22), 12.9% (54) in the intermediate tertile (23–32), and 5.0% (16) in the high tertile (>32) had numerically lower 4-year mortality predictions for PCI compared with CABG. In 58.8% (643) of patients, 4-year mortality predictions between CABG and PCI could not be significantly separated (p>0.05); an effect that was more prevalent in the low to indeterminate SYNTAX score tertiles (0–32). In the high SYNTAX score tertile (>32), 68.1% (220) of patients had mortality predictions separated with statistical significance (p<0.05) in favour of CABG.

Figure 12.3 displays scatter plots showing how the anatomical SYNTAX score, presence of ULMCA disease, and clinical variables (tertiles of age, sex, COPD, poor LVEF [<30%]) collectively affect 4-year mortality predictions. On the basis of the crossing of interaction effects between CABG and PCI for the anatomical SYNTAX score shown in figure 1, the presence of specific characteristics of patients altered the threshold value of the anatomical SYNTAX score at which CABG and PCI had much the same 4-year mortality. Younger age, female sex, and reduced LVEF favoured CABG compared with PCI. Thus, in patients with these characteristics, a lower anatomical SYNTAX score (compared with the rest of the population) would be required for the long-term mortality risk to be similar between CABG and PCI. By contrast, older age, COPD, or ULMCA disease favoured PCI compared with CABG and thus, in patients with these characteristics, a higher anatomical SYNTAX score (compared with the rest of the population) would be needed for the long-term mortality risks to be similar. These effects were more prominent in the low-intermediate SYNTAX score tertiles (0–32) than in the high tertile (>32).

On the basis of comparisons of interaction effects (HR_{PCI}/HR_{CABG}), and therefore decision making between CABG and PCI, all variables in the SYNTAX score II interacted in much the same way in the SYNTAX trial and DELTA registry, with the exception of age and LVEF, which had minimal interactions in the DELTA registry. SYNTAX score II discriminated well in all patients who underwent either CABG or PCI, with an internally (SYNTAX trial) validated concordance-index of 0.725 and an externally (DELTA registry) validated concordance- index of 0.716, which were substantially higher than for SYNTAX score alone (internal concordance index 0.567, external concordance index 0.612). The SYNTAX score II was well calibrated (i.e. a good agreement between predicted and actual risks) on validation plots in the SYNTAX trial and DELTA registry (appendix). Analyses in the stratum of patients with three-vessel disease (26% of the DELTA registry) yielded much the same results, with a concordance index for the SYNTAX score II of 0.763, and good calibration (expected 4-year survival 88.2%, actual 4-year survival 86.2%). A nomogram for the bedside application of the SYNTAX score II is detailed in figure 12.4. The nomogram can be used to obtain long-term mortality predictions for individual patients proposing to undergo CABG or PCI.

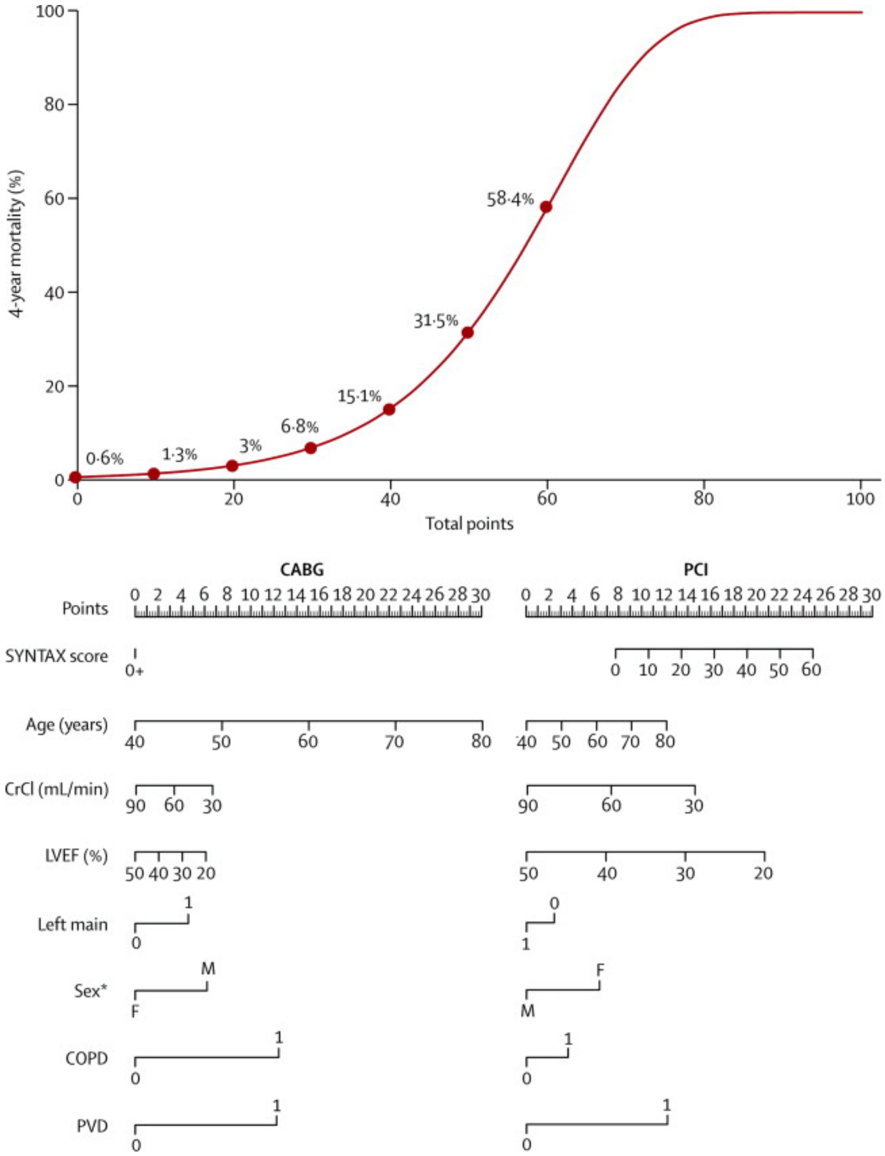


Figure 12.4 SYNTAX Score II nomogram for bedside application. Total number of points for 8 factors can be used to accurately predict 4-year mortality for the individual patient proposing to undergo for CABG or PCI. For example, a 60 year old man with an anatomical SYNTAX score of 30, unprotected left main coronary artery disease, creatinine clearance of 60 mL/min, an LVEF of 50%, and COPD, would have 41 points (predicted 4-year mortality 16.3%) to undergo CABG and 33 points (predicted 4-year mortality 8.7%) to undergo PCI respectively. The same example without COPD included would lead to identical points (29 points) and 4-year mortality predictions (6.3%) for CABG and PCI. COPD defined with EuroSCORE [11] definition, long-term use of bronchodilators or steroids for lung disease. PVD defined according to ARTS I [19] definition, aorta and arteries other than coronaries, with exercise-related claudication, or revascularisation surgery, or reduced or absent pulsation, or angiographic stenosis of more than 50%, or combinations of these characteristics. CABG=coronary artery bypass surgery. PCI=percutaneous coronary intervention. CrCl=creatinine clearance. LVEF=left ventricular ejection fraction. Left main=unprotected left main coronary artery disease. 3VD=three-vessel disease. COPD=chronic obstructive pulmonary disease. PVD=peripheral vascular disease. *Because of the rarity of complex coronary artery disease in premenopausal women, mortality predictions in younger women are predominantly based on the linear relation of age with mortality. The differences in mortality predictions in younger women between CABG and PCI will therefore be affected by larger 95% CIs than those in older women.

DISCUSSION

The main findings of this study are: first, that a personalised, individual assessment of long-term mortality was achievable for patients with complex coronary artery disease (ULMCA or de-novo three-vessel disease) proposing to undergo CABG or PCI; second, that in addition to the anatomical SYNTAX score, other factors had a direct effect on decision making between CABG and PCI, requiring lower (younger age, female sex, lower LVEF) and higher (older age, COPD, ULMCA disease) SYNTAX scores to achieve similar 4-year mortality, findings that were validated in the DELTA registry, with the exception of age and LVEF; third, the presence of diabetes in itself was shown not to be important for decision making between CABG and PCI; fourth, that the SYNTAX score II clearly identified patients for whom either CABG or PCI had a more favourable long-term outlook, and patients for whom long-term outlooks between CABG and PCI were much the same; and fifth, that the individualised approach of the SYNTAX score II, using anatomical and clinical variables that directly improved decision making between CABG and PCI, was more useful than the anatomical SYNTAX score (panel).

During development of SYNTAX score II, we suggested and then showed that the low, intermediate, and high categories of anatomical complexity in the SYNTAX score were concealing lower risk patients in the higher SYNTAX score groups, and vice versa. This principle is well established in epidemiological literature, and necessitates careful reclassification analyses to ensure that patients with high or low risk are appropriately recategorised [22,27,37]. With the individualised approach of SYNTAX score II, a subset of patients with low (<23), intermediate (23–32), or high (>32) anatomical SYNTAX scores were objectively identified, that would have lower, similar, or higher 4-year mortality predictions for CABG or PCI. Importantly, these findings were validated in the DELTA registry [28]. Additionally, the present study shows the important principle of combination of anatomical and clinical variables, which interact with CABG and PCI to affect 4-year mortality (i.e. are more predictive of mortality in one or the other revascularisation methods), and therefore drive decision making between CABG and PCI. The presence of ULMCA disease drove mortality predictions in favour of PCI, requiring higher anatomical SYNTAX scores among PCI patients to achieve similarity in long-term prognosis between CABG and PCI. The main explanation for this finding is that a sizeable proportion of the SYNTAX score can be attributed to the presence of the left main disease. Conversely, in patients with three-vessel disease and no left main involvement, the SYNTAX score would represent more complex downstream coronary anatomical disease, compared with a left main patient with an identical anatomical SYNTAX score, and therefore patients with three-vessel disease would derive a greater prognostic benefit in undergoing CABG [9,10].

Notably, diabetes was not a useful variable in the SYNTAX score II, despite medically treated diabetes being stratified at randomisation in the SYNTAX trial and reported in more than a quarter of patients. Several reasons might explain this apparent paradox. First, diabetes in itself did not produce an interaction effect in affecting long-term mortality between CABG and PCI (Figure 12.1). Second, diabetes is a metabolic, systemic disorder, the severity and duration of which has a specific effect on organs such as the heart, detected by complex coronary anatomy (anatomical SYNTAX score) and LVEF; the brain, detected by the presence of peripheral vascular disease, a sign of systemic atherosclerosis; kidney function, detected by the creatinine clearance; and age, older patients are representative of a longer duration of diabetes and its consequent multiorgan effect. The risk factors in the SYNTAX score II (predominantly the core model: SYNTAX score, age, creatinine clearance, and LVEF) are the dimensions that are relevant for the outlook for the patient, and are why diabetes falls out of the multivariable model. Although diabetes is the common denominator, it cannot be regarded as the one, direct, causative factor, in view of the many other risk factors associated with coronary artery disease, such as hyper-cholesterolaemia and hypertension. These findings are exemplified by diabetes previously being shown not to be an independent predictor of mortality in the CABG and PCI groups of the SYNTAX trial [24], and in a pooled analysis of seven contemporary stent trials (n>6000), after SYNTAX score, age, creatinine clearance, and LVEF were accounted for [23]. Additionally, a large population-based cohort study and meta-analysis involving 128 505 individuals with diabetes showed that individuals without diabetes but with chronic kidney disease and proteinuria had a stronger association with the risk of myocardial infarction, and a higher rate of mortality, compared with those with diabetes [38], and that the relative risk of long-term mortality associated with chronic kidney disease was "much the same irrespective of the presence or absence of diabetes" [39]. Third, two meta-analyses of randomised controlled trials [40,41] comparing CABG against PCI before drug-eluting stents were available (balloon angioplasty or bare metal stents) have shown a survival advantage for individuals with diabetes undergoing CABG (compared with PCI) at 4 years (but not at 6.5 years) in one study [40], and at a median follow up of 5.9 years in the other study [41]. Importantly, significant selection bias in recruiting patients before randomisation occurred in most of these studies (e.g. 2–12% of screened patients were randomised in most studies), whereas in the SYNTAX trial, selection of patients was mandated to be allcomers to overcome these issues. Furthermore, in these two meta-analyses [40,41], most patients had substantially less complex coronary artery disease compared with those in the SYNTAX trial, with single or double vessel disease occurring in almost two-thirds [41] of patients (without left main involvement). Consequently, most of these patients probably would have had anatomical SYNTAX scores that lay below the cross- over point of 15 for the hazards for CABG and PCI (Figure 12.1) in the present study.

Since drug-eluting stents became available, the FREEDOM Trial (n=1900) [42] has shown a mortality benefit for CABG compared with drug-eluting stents in individuals with diabetes with predominantly three- vessel disease (without left main involvement) at a median follow-up of 3.8 years (minimum 2 years). Notably, FREEDOM showed an apparent absence of association of SYNTAX score with 5-year mortality. However, these analyses were underpowered to assess anatomical SYNTAX score, since the numbers at 5 year follow-up were less than a quarter (n=440) of the study population. Therefore, a significant subset of lower risk patients might exist in FREEDOM, and in other reported studies examining multivessel disease (ASCERT registry [43]), in whom CABG and PCI would have much the same long-term clinical outcomes.

Female sex was shown to have a significant interaction effect in the development (SYNTAX trial) and validation (DELTA registry) populations, requiring lower anatomical SYNTAX scores for women to achieve similar long- term mortality after CABG or PCI. Notably, female sex was recently reported to be an independent predictor of long-term mortality in the PCI group of the SYNTAX trial, despite adjustment for risk factors [24], and contrary to reported scientific literature [44,45]. The main hypotheses put forward were that women in the SYNTAX trial had substantially higher anatomical SYNTAX scores (mean 26.5, SD 11.9), and therefore plaque burden, compared with contemporary stent trials (12.9, 8.4) [45], and that the plaque burden might be associated with more unfavourable plaque composition [24,46]. Another perspective is that, because women with complex coronary artery disease are more likely to have greater clinical comorbidity [44], then a lower anatomical SYNTAX score is needed for women to achieve similar long-term mortality between CABG and PCI, as shown by the interaction analyses (Figure 12.1).

SYNTAX score II provides an impartial, evidence-based assessment [47] of the decision making process for clinicians weighing anatomical and clinical factors to establish the optimum revascularisation technique for individual patients with complex coronary artery disease. Such an instrument might help to more clearly and objectively define the often uncertain line that separates patients for whom PCI or CABG should be considered, as reported in appropriate-use criteria for coronary revascularisation [48]. SYNTAX score II should be used by multidisciplinary teams consisting of a clinical cardiologist, cardiac surgeon, and interventionalist to comply with international revascularisation guidelines (class 1 indication) [6,7], and to remove any possibility of individual bias in interpretation.

Importantly, SYNTAX score II was externally validated in the DELTA registry (n=2891), a heterogeneous population that included patients with complex coronary artery disease (anatomical SYNTAX score ≥33 existed in 30% of the DELTA registry) who would have not been suitable for enrolment in the EXCEL trial [9], and three- vessel disease (26% of the DELTA registry). Additionally, all the variables in SYNTAX score II, with the exception of age and LVEF, were externally validated in the DELTA registry. One of the major limitations, inherent to all observational registry studies, is that decision making between CABG and PCI had already been done by clinicians. These decisions, although not objective, are based on the clinical judgment of cardiologists and cardiac surgeons, and thus lead to (often appropriate) selection bias. Evidence to support this notion comes from Hannan and colleagues [49] who examined the New York State registries and showed much the same mortality outcomes between CABG and PCI with drug- eluting stents in patients with multivessel disease when the data were not adjusted for baseline characteristics [49,50]. Notably, after adjustment of the data for baseline characteristics, which included all clinical variables recorded in the SYNTAX score II, a mortality benefit was shown for CABG. The inherent, unavoidable (and often appropriate) selection bias in registries might therefore be the pre- dominant reason for not identifying an interaction effect for age and LVEF in the DELTA registry.

Future validation studies of the SYNTAX score II should ideally be done in sufficiently powered, randomised, allcomers studies comparing CABG against PCI with drug-eluting stents, in which selection bias would be minimised. Such so-called mega-trials are at present rare, and include the recently reported FREEDOM Trial (n=1900) [42] and the ongoing EXCEL trial investigating ULMCA disease (n=2600) [9]. Prospective trials are being planned that will use SYNTAX score II to recruit patients, and include its further validation as one of the endpoints.

The SYNTAX score II nomogram (Figure 12.4) provides individual mortality predictions for CABG and PCI, and a measure of the magnitude of their differences, with clinically applicable accuracy. It is, however, currently limited by being unable to provide an indicator as to whether the mortality predictions can be statistically separated. An online version of the SYNTAX score II is under development, which will provide this additional information. Some of the incomplete data in the DELTA registry might have affected validation of certain variables. This effect would have been minimised since multiple imputation and sensitivity analyses were done [22,29], as evidenced by COPD being shown to have similar interaction coefficients in the SYNTAX trial and DELTA registry, and peripheral vascular disease not to have had an interaction effect in either study. Despite the SYNTAX trial being the only randomised, sufficiently powered, allcomers trial comparing CABG with PCI, with long-term follow-up, we cannot exclude the possibility that a larger sample size might have had an effect on interaction effects between CABG and PCI for certain factors. This uncertainty includes diabetes, although medically treated diabetes was prestratified at randomisation as a powered subgroup in the SYNTAX trial, and was present in more than a quarter of the study patients (26%). Because of the rarity of complex coronary artery disease in premenopausal women, mortality predictions in younger women with the SYNTAX score II are predominantly based on the linear association between age with mortality. The differences in mortality predictions in younger women between CABG and PCI will therefore be affected by larger 95% CIs. The interobserver variability of the SYNTAX score can potentially affect the mortality predictions. Appropriate training has been shown to limit this occurrence [51]. Work is underway to develop a non-invasive calculated anatomical SYNTAX score, using multislice CT that will incorporate non-invasive functional assessment of lesions [14,52]. The possibility of improved mortality cannot be excluded with newer generation drug-eluting stents, particularly since improvements in drug-eluting stent design have shown reductions in the incidence of stent thrombosis and composite clinical outcomes [53,54]. These findings should be balanced by a study of a pooled analysis of more than 6000 patients in seven con- temporary drug-eluting stent trials [23], showing stent generation (newer generation vs first generation) not to be a predictor of mortality. Additionally, when deaths associated with the Academic Research Consortium [55] definition of definite or probable stent thrombosis were removed from the SYNTAX trial (on the basis of the assumption that these patients would be alive), there was a 0.45% (definite stent thrombosis) or 1.5% (definite and probable stent thrombosis) reduction in mortality (appendix). In view of the relative infrequency of these deaths, it is unlikely that they would have had a significant effect on the coefficients for the SYNTAX score II and its ability to affect decision making. The allcomers concept of the SYNTAX trial, although representative of contemporary clinical practice, might be unavoidably limited by the inability to gain appropriate informed consent or refusal to participate from consecutive patients [56].

ACKNOWLEDGMENTS

We thank all the study centres and participants whose work made this study possible. We also thank Jian Huang and Peggy Pereda (Boston Scientific Corporation, Natick, MA, USA) for their invaluable technical support in accessing the study database. The SYNTAX trial was funded by Boston Scientific Corporation.

SUPPLEMENTARY DATA

An online appendix can be found at http://dx.doi.org/10.1016/S0140-6736(13)60108-7.

RESEARCH IN CONTEXT

Systematic review

We searched PubMed using the terms “drug-eluting stents” and “coronary artery bypass graft surgery” and “randomised controlled trials” or “registries,” with the last search done in January, 2013. Search results were filtered by hand and targeted studies which validated a score that stratified the long-term risk for patients undergoing PCI or CABG, or that could potentially help with decision making between these procedures in patients with complex coronary artery disease. Most studies validated the anatomical SYNTAX score as an instrument to guide decision making between PCI and CABG in the context of unprotected left main or multivessel coronary artery disease [14,32]. Two studies did not [33,42], but seemed to be underpowered to allow for such assessment [9,32,34]. Other studies amalgamated the anatomical SYNTAX score with cardiac-surgery-based risk scores to guide decision making between PCI and CABG [10,12,14,15]. These studies were, however, limited by using risk scores previously developed for predicting in-hospital mortality after cardiac surgery. Additionally, they did not use an individualised approach to decision making between PCI and CABG, and instead were reliant on categorising risk into low, intermediate, or high-risk groups, which has previously been shown to be potentially misleading [10,15]. Other studies focused on improvement of longer term clinical predictions in patients undergoing either PCI or CABG, without being specifically developed to directly improve decision making between these procedures [13,23,35,36].

Interpretation

Decision making between PCI and CABG with drug-eluting stents has traditionally been an area free from a compelling evidence base. The SYNTAX trial established that anatomical complexity—as assessed by the SYNTAX score—helped to guide decision making between PCI and CABG. The SYNTAX score II further improved decision making between PCI and CABG by augmenting the anatomical SYNTAX score with anatomical and clinical variables that would change the threshold value of the anatomical SYNTAX score in which PCI and CABG would offer comparable long-term mortality. Such a score would provide more objective, evidence-based decision making between PCI and CABG, when undertaken during multidisciplinary discussions between clinical cardiologists, cardiac surgeons, and interventional cardiologists, in working out the best possible revascularisation method in patients with complex coronary artery disease.

REFERENCES

1. Serruys PW, Morice MC, Kappetein AP, et al. Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. N Engl J Med 2009; 360: 961–72.
2. Serruys PW, Onuma Y, Garg S, et al. Assessment of the SYNTAX score in the Syntax study. EuroIntervention 2009; 5: 50–56.

3. Sianos G, Morel MA, Kappetein AP, et al. The SYNTAX Score: an angiographic tool grading the complexity of coronary artery disease. EuroIntervention 2005; 1: 219–27.
4. Morice MC, Serruys PW, Kappetein AP, et al. Outcomes in patients with de novo left main disease treated with either percutaneous coronary intervention using paclitaxel-eluting stents or coronary artery bypass graft treatment in the Synergy Between Percutaneous Coronary Intervention with TAXUS and Cardiac Surgery (SYNTAX) trial. Circulation 2010; 121: 2645–53.
5. Ong AT, Serruys PW, Mohr FW, et al. The SYNergy between percutaneous coronary intervention with TAXus and cardiac surgery (SYNTAX) study: design, rationale, and run-in phase. Am Heart J 2006; 151: 1194–204.
6. Wijns W, Kolh P, Danchin N, et al. Guidelines on myocardial revascularization: the Task Force on Myocardial Revascularization of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS). Eur Heart J 2010; 31: 2501–55.
7. Levine GN, Bates ER, Blankenship JC, et al. 2011 ACCF/AHA/SCAI Guideline for Percutaneous Coronary Intervention. A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines and the Society for Cardiovascular Angiography and Interventions. J Am Coll Cardiol 2011; 58: e44–122.
8. Mohr FW, Morice MC, Kappetein AP, et al. Coronary artery bypass graft surgery versus percutaneous coronary intervention in patients with three-vessel disease and left main coronary disease: 5-year follow-up of the randomised, clinical SYNTAX trial. Lancet 2013; 381: 629–38.
9. Farooq V, Serruys PW, Stone GW, Virmani R, Chieffo A, Fajadet J. Left main coronary artery disease. Percutaneous interventional cardiovascular medicine. In: The PCR-EAPCI Textbook, EUROPA edn. Toulouse, France: PCR Publishing, 2012: 329–405.
10. Serruys PW, Farooq V, Vranckx P, et al. A global risk approach to identify patients with left main or 3-vessel disease who could safely and efficaciously be treated with percutaneous coronary intervention: the SYNTAX trial at 3 years. JACC Cardiovasc Interv 2012; 5: 606–17.
11. Roques F, Michel P, Goldstone AR, Nashef SA. The logistic EuroSCORE. Eur Heart J 2003; 24: 881–82.
12. Garg S, Sarno G, Garcia-Garcia HM, et al. A new tool for the risk stratification of patients with complex coronary artery disease: the clinical SYNTAX score. Circ Cardiovasc Interv 2010; 3: 317–26.
13. Chen S-L, Chen JP, Mintz G, et al. Comparison between the NERS (New Risk Stratification) score and the SYNTAX (Synergy Between Percutaneous Coronary Intervention With Taxus and Cardiac Surgery) score in outcome prediction for unprotected left main stenting. J Am Coll Cardiol Interv 2010; 3: 632–41.
14. Farooq V, Brugaletta S, Serruys PW. Contemporary and evolving risk scoring algorithms for percutaneous coronary intervention. Heart 2011; 97: 1902–13.
15. Capodanno D, Miano M, Cincotta G, et al. EuroSCORE refines the predictive ability of SYNTAX score in patients undergoing left main percutaneous coronary intervention. Am Heart J 2010; 159: 103–09.
16. Leaman DM, Brower RW, Meester GT, Serruys P, van den Brand M. Coronary artery atherosclerosis: severity of the disease, severity of angina pectoris and compromised left ventricular function. Circulation 1981; 63: 285–99.
17. Medina A, Suarez de Lezo J, Pan M. A new classification of coronary bifurcation lesions. Rev Esp Cardiol 2006; 59: 183.
18. Hamburger JN, Serruys PW, Scabra-Gomes R, et al. Recanalization of total coronary occlusions using a laser guidewire (the European TOTAL Surveillance Study). Am J Cardiol 1997; 80: 1419–23.
19. Serruys PW, Unger F, Sousa JE, et al. Comparison of coronary-artery bypass surgery and stenting for the treatment of multivessel disease. N Engl J Med 2001; 344: 1117–24.

20. Cockcroft DW, Gault MH. Prediction of creatinine clearance from serum creatinine. Nephron 1976; 16: 31–41.
21. Harrell FE Jr, et al. Hmisc: Harrell Miscellaneous. R package version 3.8-3. 2012. http://CRAN.R-project.org/package=Hmisc.
22. Steyerberg EW. Clinical prediction models. a practical approach to development, validation, and updating. New York: Springer, 2009.
23. Farooq V, Vergouwe Y, Raber L, et al. Combined anatomical and clinical factors for the long-term risk stratification of patients undergoing percutaneous coronary intervention: the logistic clinical SYNTAX score. Eur Heart J 2012; 33: 3098–104.
24. Farooq V, Serruys PW, Bourantas C, et al. Incidence and multivariable correlates of long-term mortality in patients treated with surgical or percutaneous revascularization in the synergy between percutaneous coronary intervention with taxus and cardiac surgery (SYNTAX) trial. Eur Heart J 2012; 33: 3105–13.
25. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag, 2001.
26. Hougaard P. Frailty models for survival data. Lifetime Data Anal 1995; 1: 255–73.
27. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008; 27: 157–72.
28. Chieffo A, Meliga E, Latib A, et al. Drug-eluting stent for left main coronary artery disease: the DELTA registry: a multicenter registry evaluating percutaneous coronary intervention versus coronary artery bypass grafting for left main treatment. JACC Cardiovasc Intv 2012; 5: 718–27.
29. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med 2012; 367: 1355–60.
30. R Development Core Team. R: a language and environment for statistical computing. 2011. http://www.R-project.org.
31. Harrell FE Jr. rms: Regression Modeling Strategies. R package version 3.4-0. 2012. http://CRAN.R-project.org/package=rms.
32. Head S, Farooq V, Serruys PW, Kappetein AP. The SYNTAX score and its clinical implications. Heart (in press).
33. Park SJ, Kim YH, Park DW, et al. Randomized trial of stents versus bypass surgery for left main coronary artery disease. N Engl J Med 2011; 364: 1718–27.
34. Serruys PW, Farooq V. Is FREEDOM casting doubt on the SYNTAX score? N Engl J Med (in press).
35. Singh M, Holmes DR, Lennon RJ, Rihal CS. Development and validation of risk adjustment models for long-term mortality and myocardial infarction following percutaneous coronary interventions —clinical perspective. Circ Cardiovasc Interv 2010; 3: 423–30.
36. Shahian DM, O'Brien SM, Sheng S, et al. Predictors of long-term survival after coronary artery bypass grafting surgery: results from the Society of Thoracic Surgeons Adult Cardiac Surgery Database (the ASCERT study). Circulation 2012; 125: 1491–500.
37. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010; 21: 128–38.
38. Tonelli M, Muntner P, Lloyd A, et al. Risk of coronary events in people with chronic kidney disease compared with those with diabetes: a population-level cohort study. Lancet 2012; 380: 807–14.
39. Fox CS, Matsushita K, Woodward M, et al. Associations of kidney disease measures with mortality and end-stage renal disease in individuals with and without diabetes: a meta-analysis. Lancet 2012; 380: 1662–73.
40. Hoffman SN, TenBrook JA, Wolf MP, Pauker SG, Salem DN, Wong JB. A meta-analysis of randomized controlled trials comparing coronary artery bypass graft with percutaneous

transluminal coronary angioplasty: one- to eight-year outcomes. J Am Coll Cardiol 2003; 41: 1293–304.

41. Hlatky MA, Boothroyd DB, Bravata DM, et al. Coronary artery bypass surgery compared with percutaneous coronary interventions for multivessel disease: a collaborative analysis of individual patient data from ten randomised trials. Lancet 2009; 373: 1190–97.
42. Farkouh ME, Domanski M, Sleeper LA, et al. Strategies for multivessel revascularization in patients with diabetes. N Engl J Med 2012; 367: 2375–84.
43. Weintraub WS, Grau-Sepulveda MV, Weiss JM, et al. Comparative effectiveness of revascularization strategies. N Engl J Med 2012; 366: 1467–76.
44. Chieffo A, Hoye A, Mauri F, et al. Gender-based issues in interventional cardiology: a consensus statement from the Women in Innovations (WIN) initiative. EuroIntervention 2010; 5: 773–79.
45. Stefanini GG, Kalesan B, Pilgrim T, et al. Impact of sex on clinical and angiographic outcomes among patients undergoing revascularization with drug-eluting stents. JACC Cardiovasc Interv 2012; 5: 301–10.
46. Lansky AJ, Ng VG, Maehara A, et al. Gender and the extent of coronary atherosclerosis, plaque composition, and clinical outcomes in acute coronary syndromes. JACC Cardiovasc Imaging 2012; 5 (suppl 3): S62–72.
47. Ellis J, Mulligan I, Rowe J, Sackett DL. Inpatient general medicine is evidence based. A-Team, Nuffield Department of Clinical Medicine. Lancet 1995; 346: 407–10.
48. Patel MR, Dehmer GJ, Hirshfeld JW, Smith PK, Spertus JA. ACCF/SCAI/STS/AATS/AHA/ASNC/HFSA/SCCT 2012 appropriate use criteria for coronary revascularization focused update: a report of the American College of Cardiology Foundation Appropriate Use Criteria Task Force, Society for Cardiovascular Angiography and Interventions, Society of Thoracic Surgeons, American Association for Thoracic Surgery, American Heart Association, American Society of Nuclear Cardiology, and the Society of Cardiovascular Computed Tomography. J Am Coll Cardiol 2012; 59: 857–81.
49. Hannan EL, Wu C, Walford G, et al. Drug-eluting stents vs coronary-artery bypass grafting in multivessel coronary disease. N Engl J Med 2008; 358: 331–41.
50. Daemen J, Kukreja N, Serruys PW. Drug-eluting stents vs coronary-artery bypass grafting. N Engl J Med 2008; 358: 2641–42.
51. Genereux P, Palmerini T, Caixeta A, et al. SYNTAX score reproducibility and variability between interventional cardiologists, core laboratory technicians, and quantitative coronary measurements. Circ Cardiovasc Interv 2011; 4: 553–61.
52. Papadopoulou SL, Girasis C, Dharampal A, et al. Tomography (MSCT) SYNTAX Score: a feasibility and reproducibility study. JACC Imaging (in press).
53. Palmerini T, Biondi-Zoccai G, Della Riva D, et al. Stent thrombosis with drug-eluting and bare-metal stents: evidence from a comprehensive network meta-analysis. Lancet 2012; 379: 1393–402.
54. Stefanini GG, Kalesan B, Serruys PW, et al. Long-term clinical outcomes of biodegradable polymer biolimus-eluting stents versus durable polymer sirolimus-eluting stents in patients with coronary artery disease (LEADERS): 4 year follow-up of a randomised non-inferiority trial. Lancet 2011; 378: 1940–48.
55. Cutlip DE, Windecker S, Mehran R, et al. Clinical end points in coronary stent trials: a case for standardized definitions. Circulation 2007; 115: 2344–51.
56. de Boer SP, Lenzen MJ, Oemrawsingh RM, et al. Evaluating the ‘all-comers’ design: a comparison of participants in two ‘all-comers’ PCI trials with non-participants. Eur Heart J 2011; 32: 2161–67.

13

Predictive performance of SYNTAX Score II in patients with left main and multivessel coronary artery disease – analysis of CREDO-Kyoto registry –

CM Campos
D van Klaveren
J Iqbal
Y Onuma
YJ Zhang
HM Garcia-Garcia
MA Morel
V Farooq
H Shiomi
Y Furukawa
Y Nakagawa
K Kadota
PA Lemos
T Kimura
EW Steyerberg
PW Serruys

Circulation Japan 2014

ABSTRACT

Background SYNTAX score II (SSII) provides individualized estimates of 4-year mortality after coronary artery bypass grafting (CABG) and percutaneous coronary intervention (PCI) in order to facilitate decision-making between these revascularization methods. The purpose of the present study was to assess SSII in a real-world multicenter registry with distinct regional and epidemiological characteristics.

Methods and Results Long-term mortality was analyzed in 3,896 patients undergoing PCI (n=2,190) or CABG (n=1,796) from the Coronary REvascularization Demonstrating Outcome Study in Kyoto (CREDO-Kyoto) PCI/CABG registry cohort-2. SSII discriminated well in both CABG and PCI patient groups (concordance index [c-index], 0.70; 95% CI: 0.68–0.72; and 0.75, 95% CI: 0.72–0.78) surpassing anatomical SYNTAX score (SS; c-index, 0.50; 95% CI: 0.47–0.53; and 0.59, 95% CI: 0.57–0.61). SSII had the best discriminative ability to separate low-, medium- and high-risk tertiles, and calibration plots showed good predictive performance for CABG and PCI groups. Use of anatomical SS as a reference improved the overall reclassification provided by SSII, with a net reclassification index of 0.5 (P<0.01).

Conclusions SSII has robust prognostic accuracy, both in CABG and in PCI patient groups and, compared with the anatomical SS alone, was more accurate in stratifying patients for late mortality in a real-world complex coronary artery disease Eastern population.

INTRODUCTION

Percutaneous coronary intervention (PCI), until recently, has been considered a class III indication (ie, potentially harmful) for patients with unprotected left main (ULMCA) and 3-vessel coronary artery disease (CAD) [1, 2]. Coronary artery bypass grafting (CABG) has been the standard treatment for these patients with complex CAD for more than 50 years. Over the last decade, PCI has undergone a number of technical and technological advancements and hence has challenged the superiority of CABG [3]. Consequently, every advance in PCI technology has been scrutinized and compared against CABG, generating debate as to whether a patient should be referred to CABG or PCI, with advantages for one or the other depending on context [4–10]. Therefore, the accurate risk estimation of multivessel CAD remains a fundamental step in the decision-making process [11].

Presently, for patients with ULMCA or complex CAD, the prevailing guidelines recommend a multidisciplinary approach referred to as the heart team [12, 13]. These guidelines also advise the heart team to use synergy between PCI with taxus and cardiac surgery (SYNTAX) score alone or combined with the Society of Thoracic Surgeons (STS) score as a tool to make an objective risk stratification [12, 13]. The SYNTAX score II (SSII) has been recently developed by applying a Cox proportional hazards model to the results of the SYNTAX trial, obtaining a combination of clinical and anatomical predictors [9, 14]. Given that the SSII has been derived from an all-comers randomized trial of PCI vs. CABG, it has the potential to assess individual risk estimation between these revascularization strategies and facilitate multidisciplinary decision-making.

SSII has been shown to provide reliable predictions of 4-year mortality for complex CAD in an external validation of the Drug Eluting stent for LefT main coronary Artery disease (DELTA) registry [14, 15]. The DELTA registry consisted of predominantly Western patients with ULMCA disease. In patients with 3-vessel disease and no left main involvement, however, SYNTAX score (SS) would represent more complex downstream coronary anatomical disease. This may be a signal of a more adverse risk profile, in patients who have evidence of systemic atherosclerosis and therefore are at greater longer-term cardiovascular risk [16]. This score has not been assessed in an Eastern population with complex 3-vessel CAD.

The purpose of the present study was therefore to assess SSII in patients with 3-vessel and/or ULMCA disease in a real-world multicenter registry with distinct regional and epidemiological characteristics.

METHODS

Subjects

The Coronary REvascularization Demonstrating Outcome Study in Kyoto (CREDO-Kyoto) PCI/CABG registry cohort-2 has been previously described in detail [17]. Briefly, this was a physician-initiated non-industry-sponsored multicenter registry enrolling consecutive patients undergoing first coronary revascularization among 26 centers in Japan between January 2005 and December 2007. The relevant ethics committees in all participating centers approved the research protocol.

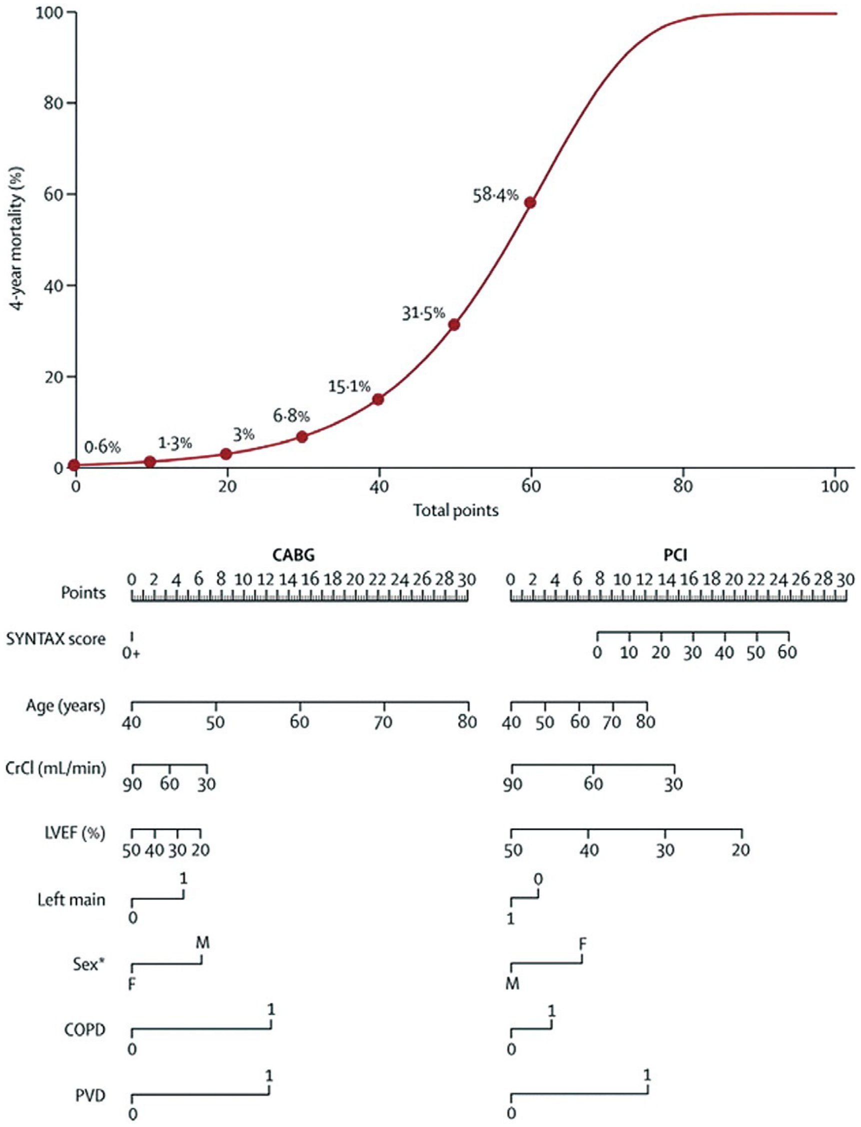


Figure 13.1 SYNTAX score II nomogram for bedside application. Total number of points for 8 factors can be used to accurately predict 4-year mortality for the individual patient preparing to undergo CABG or PCI. 3VD, 3-vessel disease; CABG, coronary artery bypass grafting; COPD, chronic obstructive pulmonary disease; CrCl, creatinine clearance; left main, unprotected left main coronary artery disease; LVEF, left ventricular ejection fraction; PCI, percutaneous coronary intervention; PVD, peripheral vascular disease. (Adapted with permission from Farooq V, et al. [14])

Because of retrospective enrolment, written informed consent from the patients was waived, excluding those patients who refused participation in the study when contacted for follow-up.

Among 15,939 patients enrolled in the registry, 3,986 participants had 3-vessel and/or ULMCA and were included in current analyses.

SSII

The SSII has been described in detail previously [14]. Briefly, SSII uses the 2 anatomical variables (anatomical SS and ULMCA disease) and 6 clinical variables (age, creatinine clearance, left ventricular ejection fraction [LVEF], sex, chronic obstructive pulmonary disease, and peripheral vascular disease) to predict 4-year mortality after revascularization with CABG or PCI.

For the present study, SSII was calculated using a nomogram, with scores assigned for the presence and magnitude of each predictor directly based on the Cox proportional hazards model coefficients (Figure 13.1), generating different scores for PCI and CABG [14]. The 4-year mortality estimates were obtained in accordance with the revascularization procedure that each patient underwent: PCI or CABG.

Statistical Analysis

Categorical variables are presented as numbers and percentages and were compared using the chi-squared test. Continuous variables are expressed as mean ± SD or median with interquartile range (IQR), and were compared using Student’s t-test or Wilcoxon rank-sum test based on their distributions.

SSII predictor data were all present in at least 90% of the patients. Multiple imputation (5×) of missing data was undertaken using an imputation strategy that takes into account the correlation between all potential predictors. To obtain 4-year mortality predictions based on anatomical SS alone, Cox logistic regression analysis was used with anatomical SS as a sole linear predictor.

SSII for PCI (in patients undergoing PCI) and for CABG (in patients undergoing CABG) was evaluated using 4 metrics: c-statistics; calibration plots; reclassification tables; and net reclassification index (NRI). Outcome was analyzed using Kaplan-Meier curves with a 4-year time horizon. Discrimination was studied with the concordance index (c-index) [18]. Calibration was assessed by plotting the observed 4-year mortality by quintiles of the predicted 4-year mortality [19]. Comparison between the anatomical and II SYNTAX scores was further quantified using a reclassification table and its NRI [20, 21]. The NRI uses reclassification tables constructed separately for participants with and without events, and quantifies the correct movement in categories: upwards for events and downwards for non-events as follows: NRI=([percentage of events moved to higher risk category in event group]−[percentage of events moved to lower risk category in event group])−([percentage of non-events moved to higher risk category in non-event group]− [percentage of non-events moved to lower risk category in non-event group]). Given that not all persons had follow-up completed to 4 years, the present reclassification was based on the expected number of case and control patients calculated using the Kaplan-Meier estimator [21]. All statistical analysis was done using IBM SPSS Statistics for Windows, version 21.0 (IBM, Armonk, NY, USA).

Table 13.1 Subject baseline characteristics.

	PCI (n=2,190)	CABG (n=1,796)	P-value
Clinical caracteristics			
Age (years)	71 (63–77)	70 (63–75)	<0.01
Male	1,554 (71)	1,336 (74.4)	0.02
BMI	23.7 (21.5–25.8)	23.3 (21.1–25.8)	<0.01
Diabetes	1,066 (48.7)	935 (52.1)	0.03
On insulin therapy	287 (13.1)	309 (17.2)	<0.01
Hypertension	1,907 (87.1)	1,514 (84.3)	0.01
Current smoking	541 (24.7)	437 (24.3)	0.79
Heart failure	454 (20.7)	387 (21.5)	0.53
Prior MI	415 (18.9)	396 (22)	0.02
Prior symptomatic stroke	346 (15.8)	248 (13.8)	0.08
Hemodialysis	124 (5.7)	119 (6.6)	0.21
COPD	70 (3.2)	60 (3.3)	1
PVD	227 (12.6)	256 (11.7)	0.36
Ejection fraction (%)	60 (50–67)	60 (49–68)	0.85
Creatinine clearance (mg/dl)	61.7 (44.2–80.9)	61.4 (43.7–78.9)	0.22
Procedural characteristics			
CAD extension			<0.01
3-vessel disease	1,825 (83.3)	1,156 (64.4)	
LM isolated	57 (3.2)	31 (1.4)	
LM and 1-vessel disease	89 (4.1)	108 (6)	
LM and 2-vessel disease	132 (6)	182 (10.1)	
LM and 3-vessel disease	113 (5.2)	293 (16.3)	
SYNTAX score	24 (17–30)	29 (23–37)	<0.01

Data given as median (IQR) or n (%). BMI, body mass index; CABG, coronary artery bypass grafting; CAD, coronary artery disease; COPD, chronic obstructive pulmonary disease; LM, left main; MI, myocardial infarction; PCI, percutaneous coronary intervention; PVD, peripheral vascular disease.

RESULTS

Patient Characteristics

Out of 3,986 patients included in the current study, 2,190 patients received PCI and 1,796 patients underwent CABG. Baseline characteristics of these patients are listed in Table 13.1.

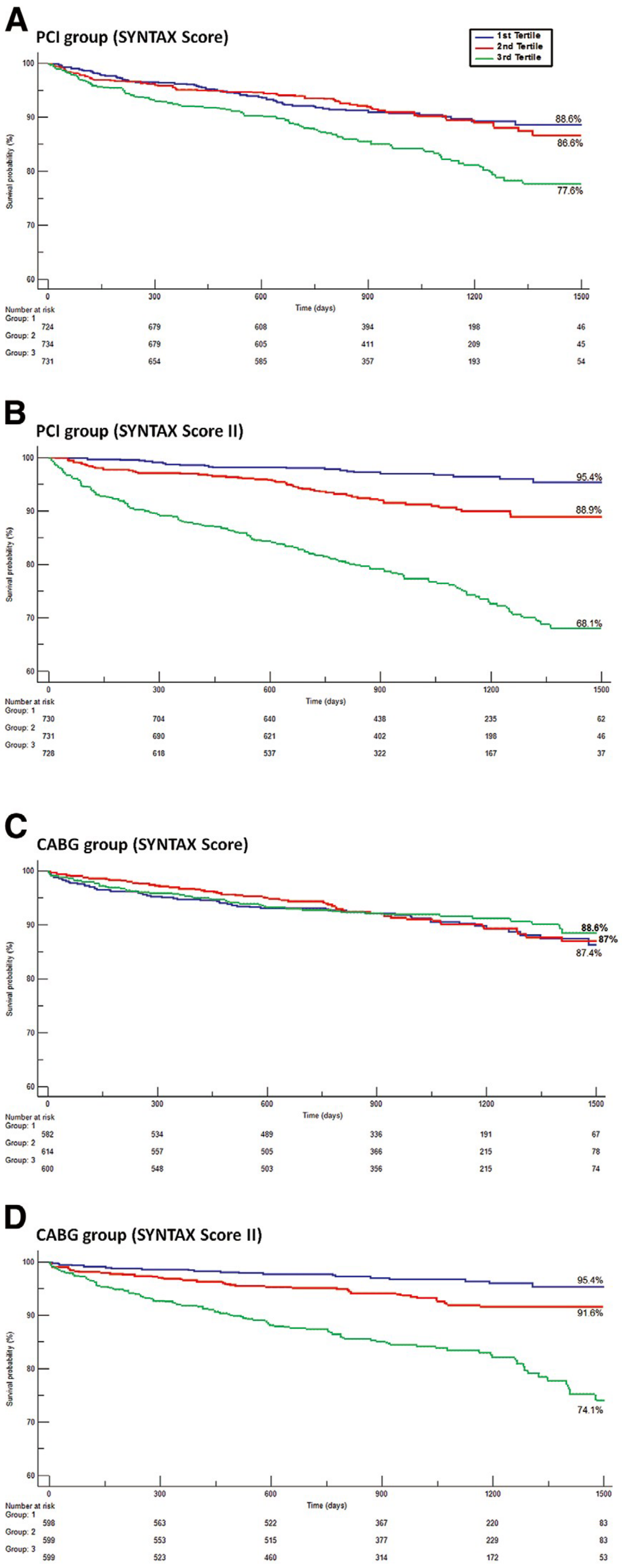


Figure 13.2 Kaplan-Meier curves for tertiles of anatomical SYNTAX score and SYNTAX score II for the percutaneous coronary intervention (PCI) and coronary artery bypass grafting (CABG) groups.

Patients in the PCI group were older, and more often had hypertension, while patients in the CABG group more often had smaller body mass index, diabetes and prior myocardial infarction. Participants treated with CABG had more complex anatomical characteristics and a higher prevalence of associated ULMCA-triple vessel disease and higher anatomical SS. Overall Kaplan-Meier estimated mortality at 4-year follow-up was 14.7% (15.9% for PCI and 12.6% for CABG).

Table 13.2 Reclassification table: 4-year mortality risk strata‡.

	Predicted mortality by SSII			
Predicted mortality by SS	**0–5%**	**5–10%**	**>10%**	**Total**
0–5%				
Persons included	0*	0	0	0
Death	0*	0	0	0
Survival	0*	0	0	0
Observed risk (%)‡‡	0*	0	0	0
5–10%				
Persons included	778	678*	1,128	2,584
Death	28.0††	57.0*	270.7†	355.7
Survival	750.0§	621.0*	857.3**	2,228.3
Observed risk (%)‡‡	3.6	8.4*	24.0	13.8
>10%				
Persons included	359	365	678*	1,402
Death	19.0††	29.2††	183.7*	232.0
Survival	340.0§	335.8§	494.3*	1,170.0
Observed risk (%)‡‡	5.3	8.0	27.1*	16.5
Total				
Persons included	1,137	1,043	1,806	3,986
Death	47.0	86.2	454.5	587.6
Survival	1,090.0	956.8	1,351.5	3,398.4
Observed risk (%)‡‡	4.1	8.3	25.2	14.7

*Patients classified as having the same risk by both scores; **patients classified as having higher risk by SSII who survived; †patients classified as having higher risk by SSII who died; ††patients classified as having lower risk by SSII who died; §patients classified as having lower risk by SSII who died survived. ‡For patients who died the SSII reclassification improved by 33% whereas in non-event patients the reclassification improved by 16.7%. The net reclassification index was 0.5 (P<0.01). ‡‡Estimated from the Kaplan-Meier curve using observations in each cell. SS, SYNTAX score; SSII, SYNTAX score II. Other abbreviations as in Table 13.1.

Predictive Performance of SSII

Discrimination The c-index of SSII was 0.70 (95% CI: 0.68–0.72) in the CABG group and 0.75 (95% CI: 0.72–078) in the PCI group. On comparison of discrimination, anatomical SYNTAX showed a significant improvement for CABG and PCI groups (c-index, 0.50; 95% CI: 0.47–0.53 and 0.59, 95% CI: 0.57–0.61; respectively). Additionally, the SSII model was able to separate low-, medium- and high-risk tertiles better than anatomical SS for both groups (Figure 13.2).

Calibration The validation plots (Figure 13.3) of SSII indicated a reasonably good agreement between the observed and predicted risks for both the CABG and PCI groups. The anatomical SS showed disparity between predicted and observed mortality.

Reclassification Reclassification for all patients (both PCI and CABG patient groups), with and without events is summarized in Table 13.2. SSII showed a significant improvement in risk stratification (NRI, 0.5; P<0.01). This was also observed when analyzing the PCI and CABG groups separately (Table S13.1).

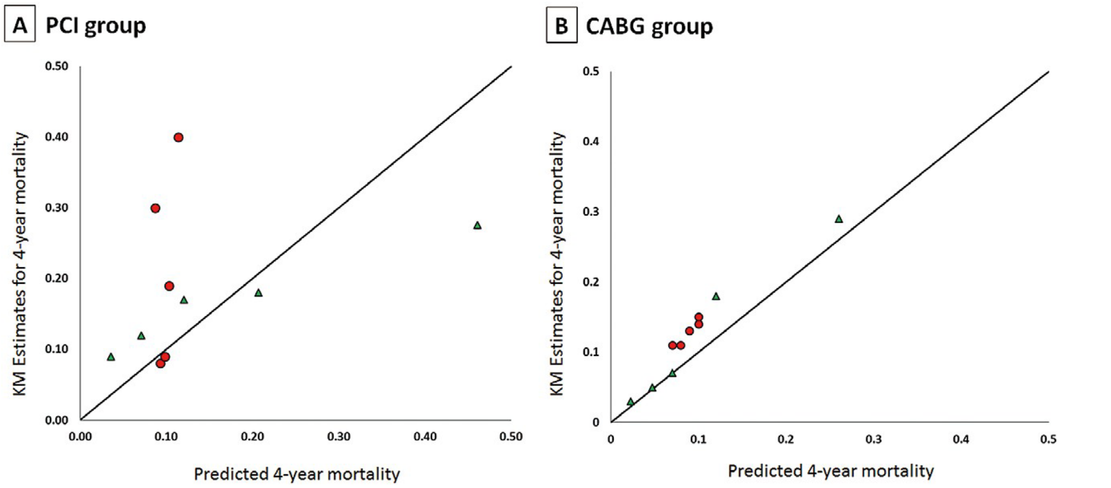


Figure 13.3 Calibration plots comparing (red circles) anatomical SYNTAX score against (green triangles) SYNTAX score II for the percutaneous coronary intervention (PCI) and coronary artery bypass grafting (CABG) groups.

DISCUSSION

In this study, SSII was assessed in a large all-comers registry of Eastern patients with predominantly high-risk CAD. The findings can be summarized as follows: (1) SSII showed agreement between observed outcomes and predictions; (2) the metrics used showed similar risk stratification for both treatment cohorts (PCI and CABG); and (3) SSII substantially improved the predictive accuracy of long-term mortality predictions if compared with the anatomical SS alone.

SSII was developed from comparison between CABG and PCI in the SYNTAX trial [14]. Its concept permits the composition of 1 single score to predict – based on randomized data – mortality if a patient is assigned to either CABG or PCI. Indeed, in the present study, SSII had similar and consistent predictive performance for both revascularization strategies in a real world population. In contrast, the current guidelines advise the heart team to use the anatomical SS alone or combined with the STS score as a tool to make objective risk stratification in the decision-making process between CABG and PCI [12, 13]. This concept, however, does not allow unified risk assessment. The anatomical SS has prognostic relevance only for patients assigned to PCI [9, 19, 22]. Despite the fact that the STS score has been widely used for risk stratification in cardiac surgery [23–25], it was not formally validated as a predictive tool for PCI.

The metrics used to perform the present analysis reinforce the importance of comprehensive assessment with a combination of angiographic and key clinical characteristics for patients with complex CAD [26]. SSII had a significantly higher accuracy compared to anatomical SS for all-cause death measured by the c-statistic. It has been argued, however, that c-statistic is insensitive to systematic errors in calibration such as differences in average outcome [20, 27]. Therefore, we studied calibration using a graphical representation where predicted risk matched observed risk. In this comparison the SSII also had a more re-fined pattern (Figure 13.2). Indeed, a better discriminating model has more spread between quintiles of predicted risk than a poorly discriminating model [20, 28].

Additionally, it is important for risk prediction as to whether a model can accurately stratify individuals into higher or lower risk categories. Therefore, we used the methodology described previously [20, 21], which balances the reclassification of a new score, subtracting, from a better risk grouping, a penalty if it lowers the estimated risk category of a patient with event or raises the estimated risk category of a patient without event. The overall NRI of 0.5 (P<0.01) indicates that 50% of patients had a net better classification for higher and lower risk categories using the SSII vs. the anatomical SS. Also, when reclassified separately for type of revascularization – PCI or CABG – the reclassification of SSII was more accurate, indicating its potential as an integrated prediction tool (Table S13.1). Grouping patients in tertiles according to SSII, a separation of the Kaplan-Meier curves for the occurrence of deaths is evident (Figure 13.2). The same approach for anatomical SS showed a poor risk stratification (Figure 13.2).

Previously, SSII was predominantly evaluated in Western patients [15]. Therefore, doubts may have existed over the utility of this tool in other populations. The present analysis confirms the potential to apply this model globally, given that we have now validated it in a population with unique epidemiological characteristics. Japan has the longest life expectancy at birth worldwide and a substantially lower proportion of mortality from cardiovascular diseases, compared with Western countries [29, 30]. Despite recent changes in the lifestyle and dietary habits of Japanese people, the incidence of myocardial infarction in Japan is still much lower than in other industrialized countries [31, 32]. Furthermore, even after revascularization – by either PCI or CABG – Japanese patients have been shown to have better long-term outcomes than US patients [33] and, regarding PCI, a significantly lower definite stent thrombosis than in Western countries [34]. All the aforementioned reasons could suggest that a score developed and validated mainly in Western populations may be less appropriate for global use. In the present cohort SSII discriminated well in both CABG and PCI patient groups (c-index, 0.70; 95% CI: 0.68–0.72 and 0.75, 95% CI: 0.72–0.78, respectively), a performance similar to its internal (c-index, 0.72) and external validations (DELTA registry; c-index, 0.71) in mainly Western patients.

Once more, the SSII predictions were consistent despite the fact that it does not include in its model diabetes mellitus. This could be questioned as a paradox because in the exclusively diabetic patients of the FREEDOM trial, CABG was superior to PCI by significantly reducing rates of death and myocardial infarction [10]. Diabetes, however, was not a useful variable in the SSII, despite medically treated diabetes being stratified at randomization in the SYNTAX trial and reported in 26% of patients. Numerous arguments might explain this apparent divergence. First, diabetes was not an independent predictor of mortality in the SYNTAX trial [35]. Second, diabetes did not have an interaction effect (P=0.67) with CABG or PCI for long-term mortality [14]. Diabetes is a systemic disease, the severity and duration of which have a specific effect on organs such as the heart (detected on complex coronary anatomy and LVEF); peripheral vascular disease (a sign of systemic atherosclerosis); kidney (detected on creatinine clearance); and age (older patients are representative of a longer diabetes multi-organ effect). These arguments may be exemplified by a large population-based cohort study and meta-analysis involving 128,505 individuals with diabetes in which patients without diabetes but with chronic kidney disease and proteinuria had a stronger association with risk of myocardial infarction, and a higher rate of mortality, compared with those with diabetes [36].

Finally, it must be acknowledged that no risk-scoring system is perfect and that careful multidisciplinary clinical reasoning remains vital for decision-making [11]. SSII, however, can be a useful instrument in this process.

Study Limitations

This study has the inherent limitations of a retrospective analysis. The ultimate goal of SSII is to assist the heart team in the decision-making process between CABG and PCI [37]. Thus, a prospective study would be needed to achieve true validation of SSII, where the decision between CABG and PCI is randomized. The present analysis, being retrospective, cannot assess the treatment recommendation based on SSII for the simple fact that the decision was likely made based on a combination of measured variables (as included in SSII) and unmeasured variables (eg, bleeding risk, duration of dual antiplatelet therapy, frailty etc). Presently, validation of SSII is a pre-specified endpoint in the ongoing randomized EXCEL trial (NCT01205776), and SYNTAX trial II, which will use SSII to recruit participants based on patient safety. In the latest trial, functional lesion assessment was added to improve late PCI outcomes and it is plausible that this approach may improve the discrimination of anatomical SS [38].

In the PCI cohort of the CREDO-Kyoto registry patients were treated mainly with first-generation drug-eluting stent (DES). It is possible that its performance will be affected by the use of newer generation DES. SSII, however, focuses on 4-year overall mortality, an outcome that, apparently, is not affected by the type of stent used. For instance, in a recent meta-analysis of 20 clinical trials that included 20,005 patients, stent type did not alter the overall mortality, unlike late-lumen loss and stent thrombosis rate [39]. Therefore, we do not expect that the type of DES prescribed will affect the predictions made by the PCI model of SSII.

CONCLUSION

SSII has robust prognostic accuracy, both in CABG and PCI patient groups and – compared with the anatomical SS alone – was able to stratify patients for late mortality in a real-world complex CAD Eastern population.

SUPPLEMENTARY FILES

An online appendix can be found at http://dx.doi.org/10.1253/circj.CJ-14-0204.

REFERENCES

1. Smith SC Jr, Feldman TE, Hirshfeld JW Jr, Jacobs AK, Kern MJ, King SB 3rd, et al. ACC/AHA/SCAI 2005 guideline update for percutaneous coronary intervention – summary article: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (ACC/AHA/SCAI Writing Committee to update the 2001 guidelines for percutaneous coronary intervention). Circulation 2006; 113: 156 – 175.
2. Kushner FG, Hand M, Smith SC Jr, King SB 3rd, Anderson JL, Antman EM, et al. 2009 focused updates: ACC/AHA guidelines for the management of patients with ST-elevation myocardial infarction (updating the 2004 guideline and 2007 focused update) and ACC/ AHA/SCAI guidelines on percutaneous coronary intervention (updating the 2005 guideline and 2007 focused update): A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol 2009; 54: 2205 – 2241.
3. Iqbal J, Gunn J, Serruys PW. Coronary stents: Historical development, current status and future directions. Br Med Bull 2013; 106:193 – 211.
4. Henderson RA, Pocock SJ, Sharp SJ, Nanchahal K, Sculpher MJ, Buxton MJ, et al. Long-term results of RITA-1 trial: Clinical and cost comparisons of coronary angioplasty and coronary-artery bypass grafting. Randomised Intervention Treatment of Angina. Lancet 1998; 352: 1419 – 1425.
5. Rodriguez A, Mele E, Peyregne E, Bullon F, Perez-Balino N, Liprandi MI, et al. Three-year follow-up of the Argentine randomized trial of percutaneous transluminal coronary angioplasty versus coronary artery bypass surgery in multivessel disease (ERACI). J Am Coll Cardiol 1996; 27: 1178 – 1184.
6. Hueb WA, Soares PR, Almeida De Oliveira S, Arie S, Cardoso RH, Wajsbrot DB, et al. Five-year follow-up of the Medicine, Angioplasty, or Surgery Study (MASS): A prospective, randomized trial of medical therapy, balloon angioplasty, or bypass surgery for single proximal left anterior descending coronary artery stenosis. Circulation 1999; 100: II107 – II113.
7. Five-year clinical and functional outcome comparing bypass surgery and angioplasty in patients with multivessel coronary disease: A multicenter randomized trial: Writing group for the Bypass Angioplasty Revascularization Investigation (BARI) investigators. JAMA 1997; 277: 715 – 721.
8. Serruys PW, Ong AT, van Herwerden LA, Sousa JE, Jatene A, Bonnier JJ, et al. Five-year outcomes after coronary stenting versus bypass surgery for the treatment of multivessel disease: The final analysis of the Arterial Revascularization Therapies Study (ARTS) randomized trial. J Am Coll Cardiol 2005; 46: 575 – 581.
9. Mohr FW, Morice MC, Kappetein AP, Feldman TE, Stahle E, Colombo A, et al. Coronary artery bypass graft surgery versus percutaneous coronary intervention in patients with three-vessel disease and left main coronary disease: 5-year follow-up of the randomised, clinical SYNTAX trial. Lancet 2013; 381: 629 – 638.
10. Farkouh ME, Domanski M, Sleeper LA, Siami FS, Dangas G, Mack MS, et al. Strategies for multivessel revascularization in patients with diabetes. N Engl J Med 2012; 367: 2375 – 2384.
11. Iqbal J, Serruys PW, Taggart DP. Optimal revascularization for complex coronary artery disease. Nat Rev Cardiol 2013; 10: 635 – 647.
12. Wijns W, Kolh P, Danchin N, Di Mario C, Falk V, Folliguet T, et al. Guidelines on myocardial revascularization. Eur Heart J 2010; 31: 2501 – 2555.
13. Fihn SD, Gardin JM, Abrams J, Berra K, Blankenship JC, Dallas AP, et al. 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS guideline for the diagnosis and management of patients with stable ischemic heart disease: A report of the American College of Cardiology Foundation/ American Heart Association Task Force on Practice Guidelines, and the American College of Physicians, American Association for Thoracic Surgery, Preventive Cardiovascular Nurses Association, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons. J Am Coll Cardiol 2012; 60: e44 – e164, doi:10.1016/j.jacc.2012.07.013.

14. Farooq V, van Klaveren D, Steyerberg EW, Meliga E, Vergouwe Y, Chieffo A, et al. Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: Development and validation of SYNTAX score II. Lancet 2013; 381: 639 – 650.
15. Chieffo A, Meliga E, Latib A, Park SJ, Onuma Y, Capranzano P, et al. Drug-eluting stent for left main coronary artery disease. The DELTA registry: A multicenter registry evaluating percutaneous coronary intervention versus coronary artery bypass grafting for left main treatment. JACC Cardiovasc Interv 2012; 5: 718 – 727.
16. Serruys PW, Farooq V, Vranckx P, Girasis C, Brugaletta S, Garcia-Garcia HM, et al. A global risk approach to identify patients with left main or 3-vessel disease who could safely and efficaciously be treated with percutaneous coronary intervention: The SYNTAX trial at 3 years. JACC Cardiovasc Interv 2012; 5: 606 – 617.
17. Kimura T, Morimoto T, Furukawa Y, Nakagawa Y, Kadota K, Iwabuchi M, et al. Long-term safety and efficacy of sirolimus-eluting stents versus bare-metal stents in real world clinical practice in Japan. Cardiovasc Interv Ther 2011; 26: 234 – 245.
18. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 1982; 247: 2543 – 2546.
19. Kappetein AP, Feldman TE, Mack MJ, Morice MC, Holmes DR, Stahle E, et al. Comparison of coronary bypass surgery with drug-eluting stenting for the treatment of left main and/or three-vessel disease: 3-year follow-up of the SYNTAX trial. Eur Heart J 2011; 32: 2125 – 2134.
20. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Stat Med 2008; 27:157 – 172; discussion 207 – 212.
21. Steyerberg EW, Pencina MJ. Reclassifi calculations for persons with incomplete follow-up. Ann Intern Med 2010; 152: 195 – 196; author reply 196 – 197.
22. Shiomi H, Tamura T, Niki S, Tada T, Tazaki J, Toma M, et al. Inter- and intra-observer variability for assessment of the synergy between percutaneous coronary intervention with TAXUS and cardiac surgery (SYNTAX) score and association of the SYNTAX score with clinical outcome in patients undergoing unprotected left main stenting in the real world. Circ J 2011; 75: 1130 – 1137.
23. Anderson RP. First publications from the Society of Thoracic Surgeons national database. Ann Thorac Surg 1994; 57: 6 – 7.
24. Ad N, Barnett SD, Speir AM. The performance of the EuroSCORE and the Society of Thoracic Surgeons mortality risk score: The gender factor. Interact Cardiovasc Thorac Surg 2007; 6: 192 – 195.
25. Handa N, Miyata H, Motomura N, Nishina T, Takamoto S. Procedure- and age-specific risk stratification of single aortic valve replacement in elderly patients based on Japan Adult Cardiovascular Surgery Database. Circ J 2012; 76: 356 – 364.
26. Park KW, Kang J, Kang SH, Ahn HS, Lee HY, Kang HJ, et al. Usefulness of the SYNTAX and clinical SYNTAX scores in predicting clinical outcome after unrestricted use of sirolimus- and everolimus-eluting stents. Circ J 2013; 77: 2912 – 2921.
27. Capodanno D. Beyond the SYNTAX score: Advantages and limitations of other risk assessment systems in left main percutaneous coronary intervention. Circ J 2013; 77: 1131 – 1138.
28. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology 2010; 21: 128 – 138.
29. Ikeda N, Saito E, Kondo N, Inoue M, Ikeda S, Satoh T, et al. What has made the population of Japan healthy? Lancet 2011; 378: 1094 – 1105.
30. Sekikawa A, Satoh T, Hayakawa T, Ueshima H, Kuller LH. Coronary heart disease mortality among men aged 35–44 years by prefecture in Japan in 1995–1999 compared with that among white

men aged 35–44 by state in the United States in 1995–1998: Vital statistics data in recent birth cohort. Jpn Circ J 2001; 65: 887 – 892.

31. Rumana N, Kita Y, Turin TC, Murakami Y, Sugihara H, Morita Y, et al. Trend of increase in the incidence of acute myocardial infarction in a Japanese population: Takashima AMI Registry, 1990–2001. Am J Epidemiol 2008; 167: 1358 – 1364.
32. Sekikawa A, Willcox BJ, Usui T, Carr JJ, Barinas-Mitchell EJ, Masaki KH, et al. Do differences in risk factors explain the lower rates of coronary heart disease in Japanese versus U.S. women? J Womens Health (Larchmt) 2013; 22: 966 – 977.
33. Kohsaka S, Kimura T, Goto M, Lee VV, Elayda M, Furukawa Y, et al. Difference in patient profiles and outcomes in Japanese versus American patients undergoing coronary revascularization (collaborative study by CREDO-Kyoto and the Texas Heart Institute Research Database). Am J Cardiol 2010; 105: 1698 – 1704.
34. Ishikawa T, Nakano Y, Endoh A, Kubota T, Suzuki T, Nakata K, et al. Significantly lower incidence of early definite stent thrombosis of drug-eluting stents after unrestricted use in Japan using ticlopidine compared to western countries using clopidogrel: A retrospective comparison with western mega-studies. J Cardiol 2009; 54: 238 – 244.
35. Farooq V, Serruys PW, Bourantas C, Vranckx P, Diletti R, Garcia Garcia HM, et al. Incidence and multivariable correlates of long-term mortality in patients treated with surgical or percutaneous revascularization in the synergy between percutaneous coronary intervention with taxus and cardiac surgery (SYNTAX) trial. Eur Heart J 2012; 33: 3105 – 3113.
36. Tonelli M, Muntner P, Lloyd A, Manns BJ, Klarenbach S, Pannu N, et al. Risk of coronary events in people with chronic kidney disease compared with those with diabetes: A population-level cohort study. Lancet 2012; 380: 807 – 814.
37. Farooq V, van Klaveren D, Steyerberg EW, Serruys PW. SYNTAX score II: Authors' reply. Lancet 2013; 381: 1899 – 1900.
38. Tanaka H, Chikamori T, Hida S, Igarashi Y, Shiba C, Usui Y, et al. Relationship of SYNTAX score to myocardial ischemia as assessed on myocardial perfusion imaging. Circ J 2013; 77: 2772 – 2777.
39. Lupi A, Gabrio Secco G, Rognoni A, Lazzero M, Fattori R, Sheiban I, et al. Meta-analysis of bioabsorbable versus durable polymer drug-eluting stents in 20,005 patients with coronary artery disease: An update. Catheter Cardiovasc Interv 2014; 83: E193 – E206, doi:10.1002/ccd.25416.

14

Long-term forecasting and comparison of mortality in the Evaluation of the Xience Everolimus Eluting Stent vs. Coronary Artery Bypass Surgery for Effectiveness of Left Main Revascularization (EXCEL) trial: prospective validation of the SYNTAX Score II

CM Campos*
D van Klaveren*
V Farooq
CA Simonton
AP Kappetein
JF Sabik 3rd
EW Steyerberg
GW Stone
PW Serruys

European Heart Journal 2015

* Equal contribution

ABSTRACT

Background To prospectively validate the SYNTAX Score II and forecast the outcomes of the randomized Evaluation of the Xience Everolimus-Eluting Stent Versus Coronary Artery Bypass Surgery for Effectiveness of Left Main Revascularization (EXCEL) Trial.

Methods and results Evaluation of the Xience Everolimus Eluting Stent vs. Coronary Artery Bypass Surgery for Effectiveness of Left Main Revascularization is a prospective, randomized multicenter trial designed to establish the efficacy and safety of percutaneous coronary intervention (PCI) with the everolimus-eluting stent compared with coronary artery bypass graft (CABG) surgery in subjects with unprotected left-main coronary artery (ULMCA) disease and low-intermediate anatomical SYNTAX scores (<33). After completion of patient recruitment in EXCEL, the SYNTAX Score II was prospectively applied to predict 4-year mortality in the CABG and PCI arms. The 95% prediction intervals (PIs) for mortality were computed using simulation with bootstrap resampling (10 000 times). For the entire study cohort, the 4-year predicted mortalities were 8.5 and 10.5% in the PCI and CABG arms, respectively [odds ratios (OR) 0.79; 95% PI 0.43 – 1.50). In subjects with low (≤22) anatomical SYNTAX scores, the predicted OR was 0.69 (95% PI 0.34 – 1.45); in intermediate anatomical SYNTAX scores (23 – 32), the predicted OR was 0.93 (95% PI 0.53 – 1.62). Based on 4-year mortality predictions in EXCEL, clinical characteristics shifted long-term mortality predictions either in favour of PCI (older age, male gender and COPD) or CABG (younger age, lower creatinine clearance, female gender, reduced left ventricular ejection fraction).

Conclusion The SYNTAX Score II indicates at least an equipoise for long-term mortality between CABG and PCI in subjects with ULMCA disease up to an intermediate anatomical complexity. Both anatomical and clinical characteristics had a clear impact on long-term mortality predictions and decision making between CABG and PCI.

INTRODUCTION

Coronary artery bypass graft (CABG) surgery was introduced in 1967 [1] with the aim of relieving angina pectoris, enhancing quality of life and improving survival. In patients with unprotected left-main coronary artery (ULMCA) disease, the superiority of CABG over optimal medical treatment has been demonstrated in multiple studies and meta-analyses [2,3] and has been the standard of care for over 30 years.

Percutaneous coronary intervention (PCI) was introduced into clinical practice in 1977 [4] and was initially considered appropriate only for single-vessel disease. With the advent of drug-eluting stents (DES), long-term outcomes after PCI have markedly improved in patients with more complex coronary artery disease. Specifically for ULMCA disease, numerous registries and three randomized trials have compared outcomes in subjects treated with either CABG or PCI [5–7]. Consequently, the prevailing international revascularization guidelines recommend revascularization of ULMCA with CABG or PCI in subjects with SYNTAX scores that are low [SYNTAX score <23: class I recommendation for CABG or PCI (level of evidence B for both)] and intermediate [SYNTAX score 23 – 32: class I for CABG and class IIa for PCI (level of evidence B for both)]. The same guidelines recommend against revascularization with PCI of ULMCA disease with high SYNTAX scores [SYNTAX score ≥33: class I for CABG and class III for PCI (level of evidence B for both)] [8]. These recommendations are based on similar 5-year mortality and myocardial infarction, with a lower incidence of stroke and increased risk of repeat revascularization with PCI compared with CABG in subjects with ULMCA disease and lower anatomical complexity [6,7].

The introduction of the newer-generation everolimus-eluting stent (EES) — with proven marked improvements in both safety and efficacy [9–13] — has prompted the design of the randomized Evaluation of Xience Everolimus-Eluting Stent Versus Coronary Artery Bypass Surgery for Effectiveness of Left-Main Revascularization (EXCEL) Trial.

Aiming to improve decision making between CABG and PCI in patients with complex coronary artery disease, the SYNTAX Score II combines anatomic and clinical factors [14,15]. Importantly, the SYNTAX Score II was developed in the landmark, all-comers, randomized SYNTAX (Synergy between PCI with Taxus and Cardiac Surgery) Trial where selection bias would have been minimal, and externally validated in two real world registries [14,16]. In addition, the SYNTAX Score II has been included in international revascularization guidelines [8].

Although numerous risk scores and prospective trials are available in the medical literature, their performances are reported when the outcomes are already known. The aim of the present study is to apply the SYNTAX Score II in the ongoing EXCEL trial, in order to prospectively validate the SYNTAX Score II before independent reporting of the outcomes of the trial, forecast the 4-year mortality outcomes in the PCI and CABG arms, and to describe how anatomical and clinical characteristics impact on the long-term mortality predictions and decision making between CABG and PCI.

METHODS

Study population

The EXCEL trial (clinicaltrials.gov identifier: NCT01205776) is an international, prospective, unblinded, randomized multicenter trial that enrolled 1905 subjects in 131 centres. Evaluation of the Xience Everolimus-Eluting Stent vs. Coronary Artery Bypass Surgery for Effectiveness of Left-Main Revascularization was designed to establish the safety and efficacy of the EES (XIENCE PRIME^{TM} or XIENCE V® or XIENCE Xpedition^{TM} or XIENCE PRO^{TM}; Abbott Vascular, Santa Clara, CA, USA)) in patients with ULMCA disease. Evaluation of the Xience Everolimus-Eluting Stent vs. Coronary Artery Bypass Surgery for Effectiveness of Left-Main Revascularization adopted an enrolment criteria of subjects with ULMCA disease up to intermediate anatomical complexity (SYNTAX Score<33), with minimal exclusion criteria to allow meaningful comparisons between revascularization modalities (Supplementary material online, Table S1). The information on the trial endpoints and sample size calculation is also available in the Supplementary material online.

Following diagnostic angiography demonstrating significant ULMCA disease and the consensus of the local Heart Team (qualified participating interventional cardiologist and cardiac surgeon), subjects were consented and randomized 1 : 1 to: (i) PCI with the EES or (ii) CABG. All randomized patients were scheduled to undergo follow-up telephone contact or office visit up to 5 years post-procedure.

The primary endpoint of the EXCEL trial is the composite measure of all-cause mortality, myocardial infarction or stroke [modified Rankin Scale (mRS) ≥1 and increase by ≥1 from baseline] at a median follow-up interval of 3 years post-index procedure.

SYNTAX Score II

The SYNTAX Score II has been described previously [14]. In brief, the SYNTAX Score II augments the purely anatomical SYNTAX score with anatomical and clinical factors that were shown to alter the threshold value of the anatomical SYNTAX score in order for equipoise to be achieved between CABG and PCI for long-term mortality. The SYNTAX Score II is composed of the anatomical SYNTAX score, presence of ULMCA disease, and six clinical characteristics [age, creatinine clearance (CrCl), left ventricular ejection fraction (LVEF), gender, chronic obstructive pulmonary disease (COPD) and peripheral vascular disease (PVD)]. The SYNTAX Score II allows for 4-year mortality predictions to be made following revascularization with CABG or PCI to aid decision making between CABG and PCI. Importantly, the SYNTAX Score II was developed in the randomized SYNTAX Trial (n = 1800), and externally validated in the multinational DELTA (n = 2891) and Credo-KYOTO (n = 3896) registries [14,16,17].

Using the actual baseline clinical and angiographic data from each enrolled patient in EXCEL, the SYNTAX Score II was calculated for each patient. Scores were assigned for the presence and magnitude of each predictor directly based on the Cox proportional hazards model coefficients generating different scores and 4-year mortality predictions for PCI and CABG [14]. To mirror conventional clinical practice, investigator reported anatomical SYNTAX Scores were used in the analysis [18].

Statistical analysis

Categorical variables are presented as numbers and percentages and are compared with the χ^2 test. Continuous variables are expressed as mean ± SD or median with interquartile range (IQR), and are compared using the Student's t-test or Wilcoxon rank-sum test based on their distributions. Within EXCEL, SYNTAX Score II predictor values were >99% complete with the exception of LVEF which was 95.1% complete. An advanced multiple imputation strategy which takes the correlation between all potential predictors (method of chained equations) was used to account for missing values as previously described [19].

Comparison of predicted 4-year mortality between CABG and PCI arms

The individual predicted mortality and the odds ratio (OR) of the two randomized revascularization strategies were calculated using the SYNTAX Score II. To determine the 95% PIs, the trial was simulated 10 000 times and generated 4-year mortality from predictions based on consecutive bootstrap samples [20] of the original SYNTAX trial (Figure 14.1) [17,21]. A prediction interval is an estimate of an interval in which future observations will fall, with a certain probability, compared with what has already been observed (SYNTAX trial) [22]. All data analyses were performed using R version 2.15.3 [23].

RESULTS

Between 29 September 2010 and 6 March 2014, 2909 patients with ULMCA disease were screened and 1905 subjects randomized to CABG (n = 957) or PCI (n = 948) (Figure 14.2). Subjects in the two randomization arms were well balanced with regards to baseline demographic and clinical characteristics included in the SYNTAX Score II (Table 14.1). Overall, the median age was 66.0 (IQR 59.0 – 73.0) years, 76.3% male, 24.7% female, 7.8% COPD and 8.6% PVD. The median LVEF was 60.0% (IQR 52.0 – 63.0%), median CrCl 85.0 mL/min (IQR 66.8 – 106.2 mL/ min) and the median anatomical SYNTAX score 21.0 (IQR 15.0 – 26.0).

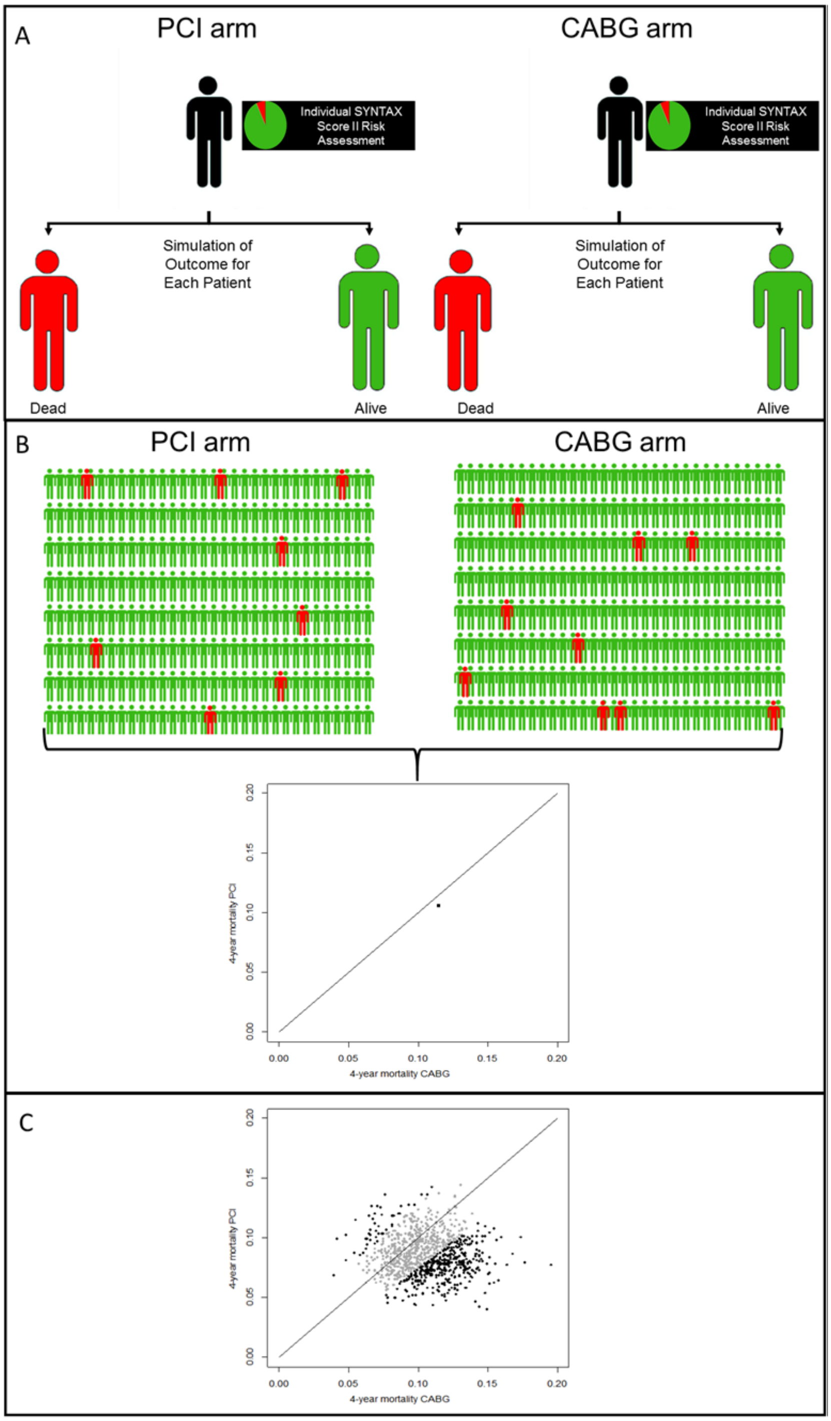


Figure 14.1 Schematic representation of the SYNTAX Score II predictions used in the Evaluation of the Xience Everolimus-Eluting Stent vs. Coronary Artery Bypass Surgery for Effectiveness of Left-Main Revascularization trial. (A) The mortality predictions of percutaneous coronary intervention and coronary artery bypass graft for each patient enrolled in the Evaluation of the Xience Everolimus-Eluting Stent vs. Coronary Artery Bypass Surgery for Effectiveness of Left-Main Revascularization trial were calculated using the SYNTAX Score II. The pie chart represents the individual risk of 4-year mortality (red slice). (B) Based on individual mortality predictions, patients' outcomes were simulated to obtain the 4-year mortality in both trial arms. (C) To determine the 95% prediction intervals steps A and B were repeated 10 000 times with 4-year mortality predictions based on consecutive bootstrap samples of the original SYNTAX trial.

Table 14.1 SYNTAX score II baseline.

	PCI N=948	CABG N=957
Anatomical SYNTAX Score (IQR)	21.0 (15-26)	20.0 (15.0-25)
Age, years (IQR)	66.0 (59.0-73.0)	66.0 (60.0-73.0)
Creatinine Clearance, ml/min (IQR)	85.7 (66.2-107.5)	84.6 (67.1-105.0)
LVEF (IQR)	60.0 (52.0-62.0)	60.0 (52.0-63.0)
Left Main Coronary Artery Disease, n (%)	948 (100)	957 (100)
Female gender, n (%)	226 (23.8)	214 (22.4)
COPD, n (%)	67 (7.1)	81 (8.5)
PVD, n (%)	97 (10.3)	84 (8.8)
Anatomical SYNTAX Score (IQR)	21.0 (15-26)	20.0 (15.0-25)

PCI, percutaneous coronary intervention; CABG, coronary artery bypass graft;
LVEF, left ventricular ejection fraction; IQR, interquartile range

SYNTAX Score II 4-year mortality predictions in the cohorts

The predicted mortality was 8.5% (95% PI 5.4 – 11.9%) in the PCI arm and 10.5% (95% PI 6.6 – 15.1%) in the CABG arm (OR 0.79; 95% PI 0.43 – 1.50%) (Table 14.2).

Table 14.2 Four-year mortality predictions comparisons between coronary artery bypass graft and percutaneous coronary.

		PCI, n (%)	CABG, n (%)	Predicted 4-year mortality PCI,% (95% PI)	Predicted 4-year mortality CABG,% (95% PI)	OR PCI:CABG (95% PI)
Overall		948	957	8.5 (5.4-11.9)	10.5 (6.6-15.1)	0.79 (0.43-1.50)
SYNTAX Score	**≤22**	563 (59%)	589 (61.5%)	7.3 (4.2-11.0)	10.3 (5.9-15.6)	0.69 (0.34-1.45)
	23-32	385 (40.6%)	368 (38.5%)	10.1 (6.2-14.6)	10.8 (6.5-15.5)	0.93 (0.53-1.62)
Age*	**≤66**	483 (50.9%)	486 (50.8%)	5.4 (2.7-8.5)	5.8 (2.7-9.5)	0.92 (0.38-2.25)
	>66	465 (49.1%)	471 (49.2%)	11.8 (7.3-16.8)	15.41 (9.3-22.5)	0.73 (0.37-1.48)
CrCl, mg/mL	**≤60**	168 (17.7%)	147 (15.4%)	19.7 (11.9-28.7)	18.4 (9.5-28.6)	1.08 (0.48-2.64)
	>60	780 (82.3%)	810 (84.6%)	6.1 (3.5-9.2)	9.1 (5.2-13.6)	0.65 (0.31-1.35)
LVEF, %	**≤50**	118 (12.4%)	119 (12.4%)	18.3 (9.3-28.0)	18.3 (9.3-28.0)	1.25 (0.49-3.27)
	>50	830 (87.6%)	838 (87.6%)	7.1 (4.3-10.4)	9.9 (5.9-14.4)	0.70 (0.36-1.39)
Gender	**Female**	226 (23.8%)	214 (22.4%)	13.1 (7.1-19.9)	8.7 (3.3-15.9)	1.59 (0.61-5.00)
	Male	722 (76.2%)	743 (77.6%)	7.1 (4.0-10.5)	11.1 (6.7-15.9)	0.61 (0.30-1.23)
COPD	**No**	881 (92.9%)	876 (91.5%)	7.9 (4.9-11.2)	9.1 (5.3-13.5)	0.86 (0.44-1.72)
	Yes	67 (7.1%)	81 (8.5%)	16.7 (6.0-29.9)	26.1 (12.4-40.7)	0.57 (0.16-1.75)
PVD	**No**	851 (89.8%)	873 (91.2%)	6.9 (4.2-10.1)	9.0 (5.3-13.3)	0.75 (0.38-1.54)
	Yes	97 (10.2%)	84 (8.8%)	22.5 (11.3-36.1)	26.5 (13.1-41.7)	0.81 (0.27-2.32)
Diabetes	**Yes**	286 (30.2%)	266 (27.8%)	9.9 (5.6-14.7)	11.4 (6.4-17.3)	0.86 (0.40-1.9)
	No	662 (69.8%)	691 (72.2%)	7.9 (4.8-11.3)	10.2 (6.2-14.8)	0.76 (0.39-1.48)

*Separated by the median

PCI percutaneous coronary intervention; CABG coronary artery bypass graft surgery; PI prediction intervals; CrCl creatinine clearance; LVEF left ventricular ejection fraction; COPD chronic obstructive pulmonary disease; PVD=peripheral vascular disease.

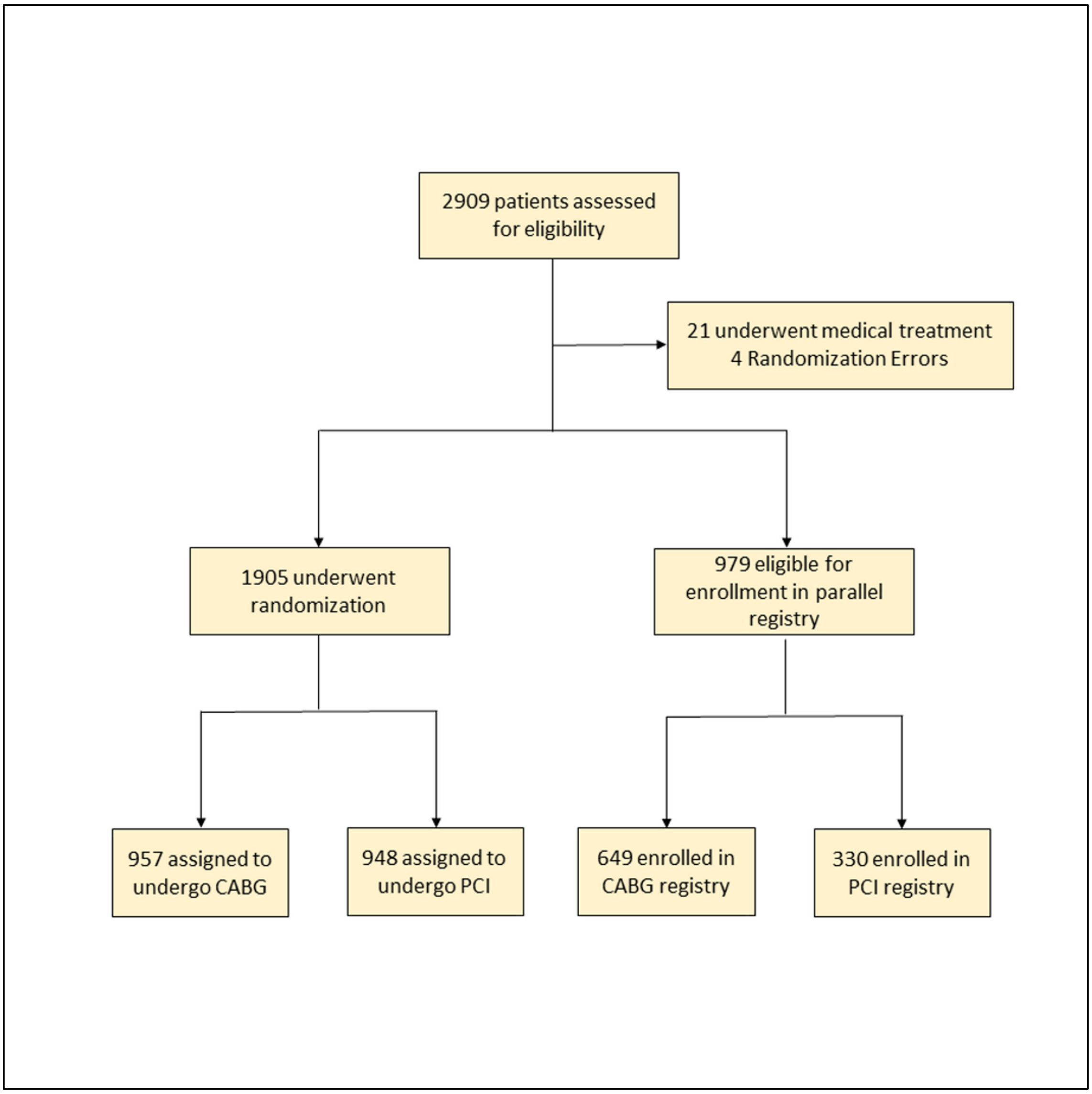

2909 patients assessed for eligibility
21 underwent medical treatment
4 Randomization Errors
1905 underwent randomization
979 eligible for enrollment in parallel registry
957 assigned to undergo CABG
948 assigned to undergo PCI
649 enrolled in CABG registry
330 enrolled in PCI registry

Figure 14.2 Enrolment and randomization of patients with previously untreated left-main coronary artery disease in the Evaluation of the Xience Everolimus-Eluting Stent vs. Coronary Artery Bypass Surgery for Effectiveness of Left-Main Revascularization trial.

Figure 14.3 demonstrates the first 1000 trial simulations. Based on numerical differences in 4-year mortality predictions, 77.9% of trial simulations (n = 7790) favoured PCI and 22.1% of trial simulations (n = 2210) favoured CABG. In 55.2% of trial simulations (n = 5520) 4-year mortality predictions between CABG and PCI could not be separated with statistical significance ($P > 0.05$). 40.4% (n = 4040) of trial simulations had mortality predictions separated with statistical significance ($P < 0.05$) in favour of PCI, and 4.4% (n = 40) had mortality predictions separated with statistical significance ($P < 0.05$) in favour of CABG.

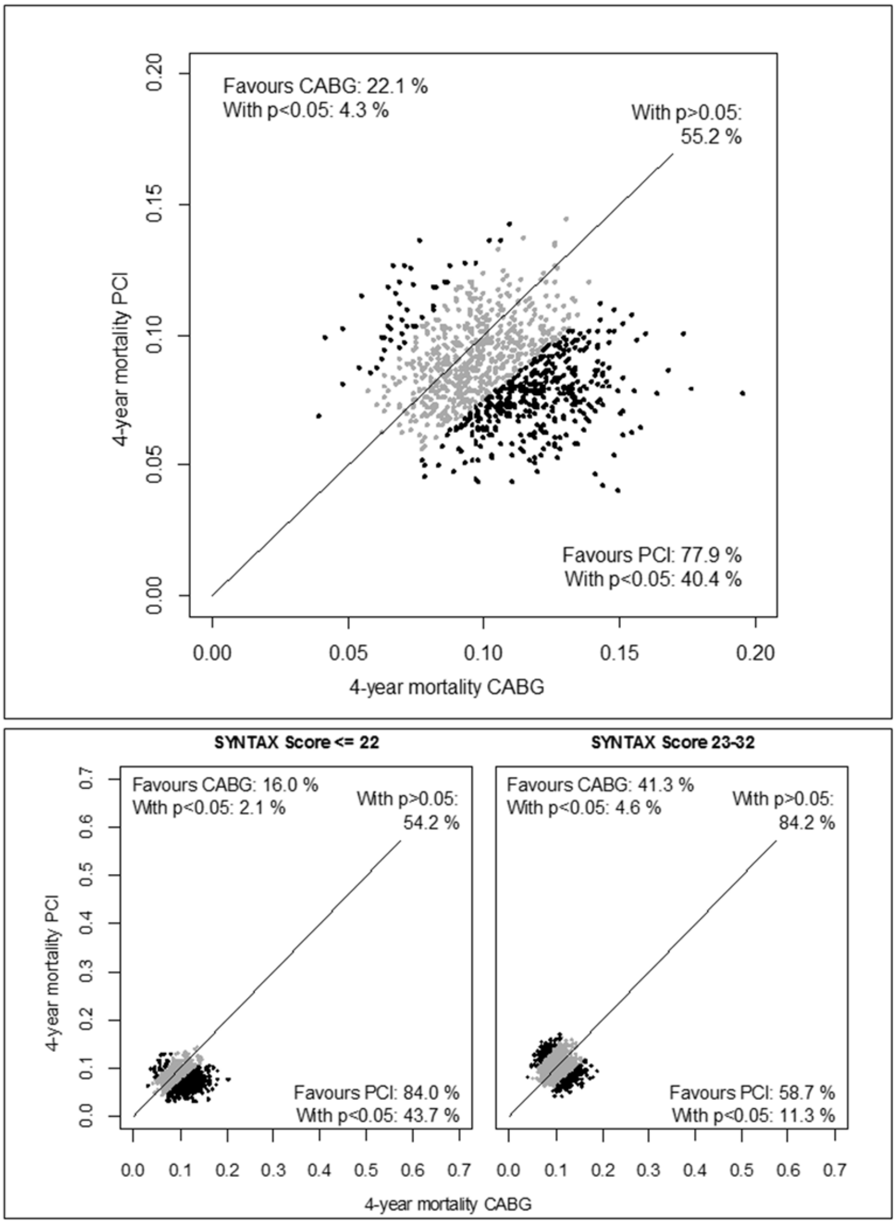


Figure 14.3 First 1000 4-year mortality simulations of the Evaluation of the Xience Everolimus-Eluting Stent vs. Coronary Artery Bypass Surgery for Effectiveness of Left-Main Revascularization trial on the SYNTAX Score II. Each dot represents one simulated trial mortality in both randomization arms based on individual predictions. The diagonal line represents identical mortality for coronary artery bypass graft and percutaneous coronary intervention. A dot plotted to the left of the diagonal line favours coronary artery bypass graft (actual percentages shown in top left corner), and to the right favours percutaneous coronary intervention (actual percentages shown in bottom right corner). Simulated trials with a significant (P ≤ 0.05) mortality difference between coronary artery bypass graft and percutaneous coronary intervention are coloured black (actual percentage shown in parentheses in respective corners). Simulated trials with a non-significant (P > 0.05) mortality difference between coronary artery bypass graft and percutaneous coronary intervention are coloured grey.

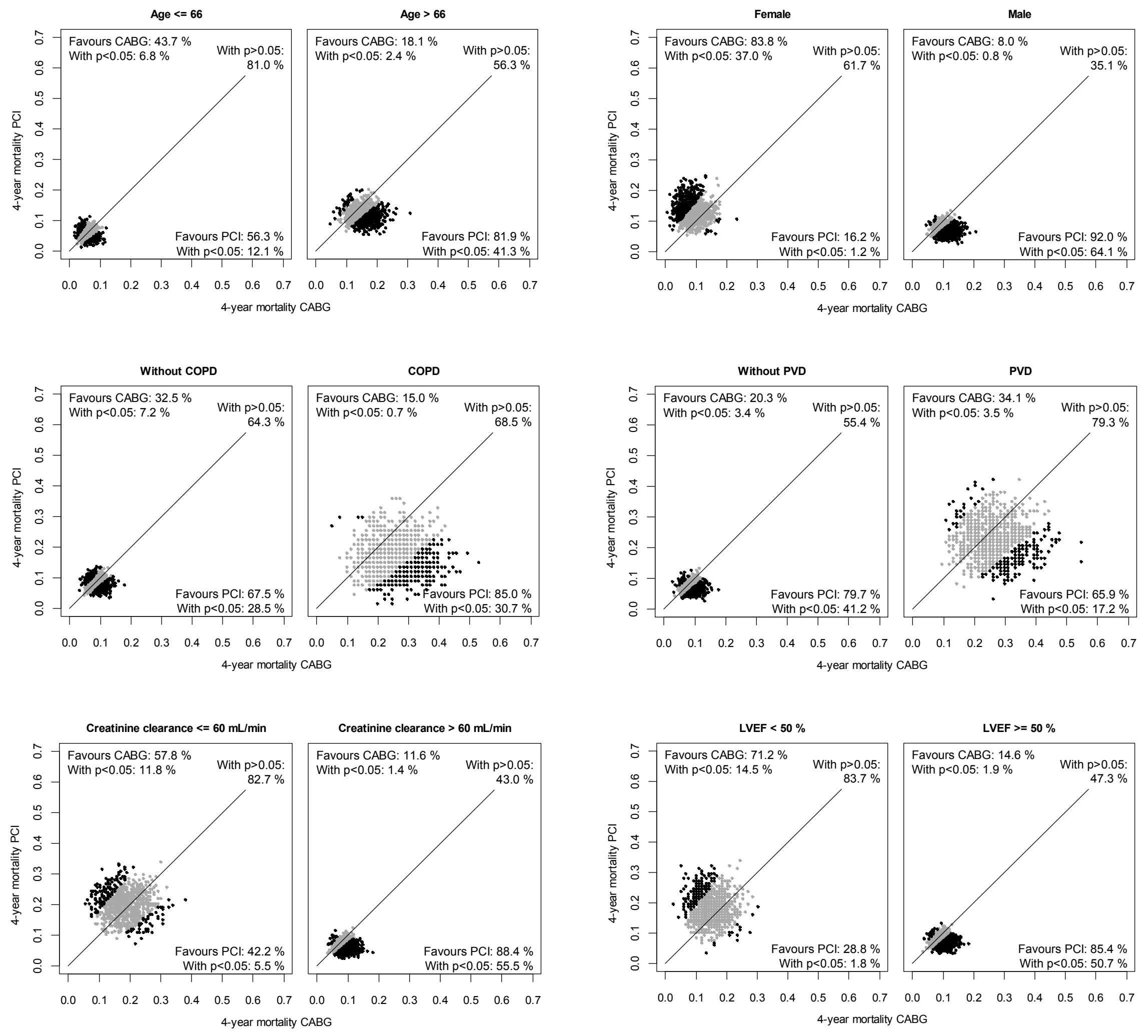


Figure 14.4 First 1000 4-year mortality simulations of the Evaluation of the Xience Everolimus-Eluting Stent vs. Coronary Artery Bypass Surgery for Effectiveness of Left-Main Revascularization trial on the SYNTAX Score II according to clinical characteristics. Each dot represents one simulated trial mortality in both randomization arms based on individual predictions. The diagonal line represents identical mortality for coronary artery bypass graft and percutaneous coronary intervention. A dot plotted to the left of the diagonal line favours coronary artery bypass graft and to the right favours percutaneous coronary intervention. Simulated trials with a significant (P ≤ 0.05) mortality difference between coronary artery bypass graft and percutaneous coronary intervention are coloured black (actual percentage shown in parentheses in respective corners). Simulated trials with a non-significant (P > 0.05) mortality difference between coronary artery bypass graft and percutaneous coronary intervention are coloured grey. Some factors are more favourable for percutaneous coronary intervention and others for coronary artery bypass graft and had different impact in the simulation pattern. CrCl, creatinine clearance; LVEF, left ventricular ejection fraction; COPD, chronic obstructive pulmonary disease; PVD, peripheral vascular disease.

Anatomical complexity

Anatomical complexity had a clear impact on mortality predictions. In subjects with low (≤22) and intermediate (23–32) anatomical SYNTAX scores the predicted OR were 0.69 (95% PI 0.34 – 1.45) and 0.93 (95% PI 0.53 – 1.62), respectively (Table 14.2).

In the low SYNTAX score group, 54.2% (n = 5420) of mortality predictions were similar (P > 0.05) between CABG and PCI; in the intermediate SYNTAX score group, 84.1% (n = 8410) of mortality predictions were similar (P > 0.05) between CABG and PCI (Figure 14.3).

Mortality predictions that were separated with statistical significance (P < 0.05) in favour of PCI were 43.7% (n = 4370) in the low SYNTAX score group, compared with 11.3% (n = 1130) in the intermediate SYNTAX score group. Conversely, mortality predictions that were separated with statistical significance (P < 0.05) in favour of CABG were 2.1% (n = 210) in the low SYNTAX score group, compared with 4.6% (n = 460) in the intermediate SYNTAX score group.

Impact of clinical characteristics

Clinical characteristics had a clear impact on 4-year mortality predictions (Table 14.2, Figure 14.4). In both arms the subgroup with the highest predicted mortalities was PVD [22.5% (95% PI 11.3 – 36.1%) in the PCI arm and 26.5% (95% PI 13.1 – 41.7 in the CABG arm)].

Based on 4-year mortality predictions, older age, male gender and COPD favoured PCI, whereas younger age, lower CrCl, impaired LVEF and female gender favoured CABG (Figure 14.4).

Diabetes

In subjects with diabetes, predicted mortality was 9.9% (95% PI 5.6 – 14.7%) in the PCI arm and 11.4% (95% PI 6.4 – 17.3%) in the CABG arm [OR 0.86 (PI 0.40 – 1.90; Table 14.2]. Similar analyses in non-diabetics yielded predicted mortalities of 7.9% (95% PI 4.8 – 11.3%) in the PCI arm and 10.2% (95% PI 6.2 – 14.8%) in the CABG arm [0.75 (PI 0.39 – 1.48)]. The presence of diabetes had a clear impact on mortality predictions (Figure 14.5). Trial simulations were separated with statistical significance (P < 0.05) in favour of PCI in 14.2% (n = 1420) of diabetics, compared with 37.5% (n = 3750) in non-diabetics. Comparatively, trial mortality predictions were separated with statistical significance (P < 0.05) in favour of CABG in 3.3% (n = 330), compared with 2.5% (n = 250) in non-diabetics.

DISCUSSION

The main findings of the study are: (i) The prospective use of a decision making and risk prediction tool (SYNTAX Score II) was feasible in a large-scale randomized trial on completion of enrolment of subjects, in which the follow-up results were unknown and blinded; (ii) based on the SYNTAX Score II, we predicted a 77.9% chance of a lower 4-year mortality in the PCI arm of the EXCEL trial, with a 40% chance that this will achieve statistical significance in favour of PCI; (iii) The interplay between angiographic and clinical characteristics has an important impact on decision-making and risk stratification of patients with ULMCA disease.

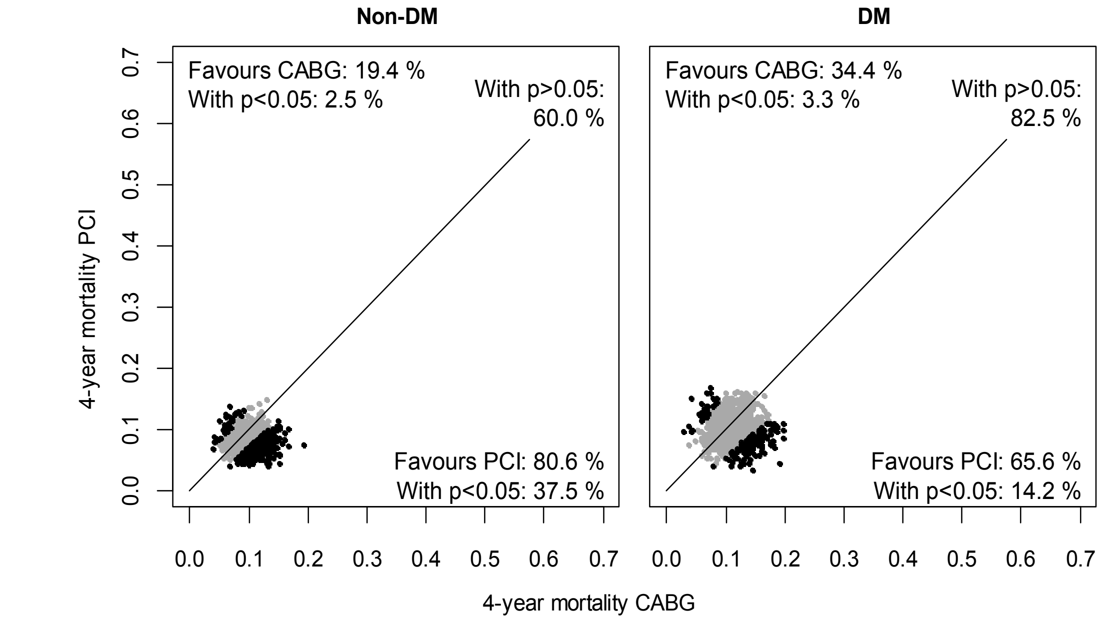


Figure 14.5 First 1000 4-year mortality simulations of the Evaluation of the Xience Everolimus-Eluting Stent vs. Coronary Artery Bypass Surgery for Effectiveness of Left-Main Revascularization trial on the SYNTAX Score II according to the presence or absence of diabetes mellitus. See Figure 14.4 legend for text in interpreting figure. As illustrated, the presence of diabetes was associated with an increase in predicted mortality within both the coronary artery bypass graft and percutaneous coronary intervention arms, with the predicted benefits of percutaneous coronary intervention being less pronounced in diabetics compared with non-diabetics. DM, diabetes mellitus.

SYNTAX Score II and prospective mortality predictions

The unprecedented aspect of the present study was to prospectively validate the SYNTAX Score II in a randomized trial that is still ongoing, despite completion of enrolment of patients, with expected reporting of the primary outcome in another 2 years. It is important to emphasize that outcomes of EXCEL are being collected, analysed and reported by an independent clinical events committee (CEC), and that the current analyses were performed with all authors completely blinded to any outcome data. A second unique aspect of the present study was to report the predicted long-term mortality of a randomized trial following completion of patient enrolment, blinded and prior to the actual reporting of the trial. This was only possible because the SYNTAX Score II was developed in the randomized SYNTAX trial — consisting of a population with complex coronary artery disease (ULMCA disease or de novo three-vessel disease)—and importantly where selection bias was minimal secondary to the unique all-comers design of SYNTAX. In addition, the SYNTAX Score II has shown consistent and solid predictive performances in two multicenter registries for CABG and PCI-treated patients with left-main and/or complex coronary artery disease [14,16].

The SYNTAX Score II predicted a 55.2% likelihood that there will not be a statistically significant difference in mortality between the PCI and CABG arms of EXCEL at 4 years. This is likely to be secondary to the clinical profile of the patients recruited in EXCEL. On average subjects in EXCEL had preserved LVEF, reasonable renal function, were predominantly male, and importantly more complex coronary artery disease (SYNTAX score ≥ 33) was a key exclusion criteria. In the SYNTAX trial, female gender, reduced LVEF, lower CrCl, higher anatomical SYNTAX scores and younger age were all shown to favour CABG [14,16,22]. The combination of these angiographic and clinical profile is therefore likely to explain the predicted favourable results for PCI, despite similar baseline clinical characteristics in the CABG and PCI arms of EXCEL. The present study therefore does not imply that PCI reduces mortality in all ULMCA revascularization, but predicts that subjects with ULMCA disease with a lower anatomical and risk profile may potentially derive a prognostic benefit from undergoing PCI, whilst more complex disease and a higher risk clinical profile would remain the domain of CABG on the grounds of prognosis.

Impact of anatomic complexity in risk predictions

Unprotected left-main coronary artery disease should be regarded as a heterogeneous pathology when considering the choice of revascularization modality. The anatomical complexity of the left main may vary from a single lesion in the shaft to distal trifurcation disease and its association with more complex downstream (three-vessel) disease. These variances may influence the capacity of PCI to achieve complete revascularization, the number of stents implanted and complexity of interventional techniques employed. Moreover, incomplete revascularization and anatomical complexity have been directly correlated to late all-cause mortality following PCI [24 – 26]. This was exemplified in the PCI arm of the left-main subgroup of SYNTAX, where the incidence of 5-year all-cause mortality was shown to markedly increase in subjects with a SYNTAX score ≥33 (5-year mortality 20.9%) compared with subjects with a SYNTAX score <33 (5-year mortality 7.9%).

Conversely, in subjects undergoing CABG, anatomical complexity has been shown to not affect long-term prognosis, as exemplified in the CABG arm of the left-main subgroup of SYNTAX, where the incidence of 5-year all-cause mortality remained almost unchanged in subjects with a SYNTAX score ≥33 (5-year mortality 14.1%) compared with subjects with a SYNTAX Score <33 (5-year mortality 15.1%) [24]. In the present analysis, although the PIs were wide, the expected mortality favoured PCI in the low (≤22) anatomical SYNTAX score group (7.3 vs. 10.3%), and was practically equipoised between PCI and CABG in the moderate (>22) anatomical SYNTAX score group (10.1 vs. 10.8%; Table 14.2).

The aforementioned reasons explain why the risk predictions in EXCEL are not at variance with results of the recent randomized comparisons between CABG and PCI [6]. Al Ali et al. pooled the results of three randomized trials of first-generation DES vs. CABG in left-main coronary artery disease and demonstrated that PCI did not reduce the overall mortality (HR 1.08; 95% CI 0.75 – 1.57), since the global outcomes reflected a composite of lower and higher mortality related to simple and more complex coronary anatomy [6]. In contrast, EXCEL set an exclusion criteria of a SYNTAX score ≥33, thereby selecting subjects with less complex coronary artery disease and therefore potentially more favourable results for PCI.

Impact of clinical characteristics in risk predictions

The predictions provided by the SYNTAX Score II displayed in the Figure 14.4 deserve detailed examination since clinical characteristics markedly affect the simulation patterns. Although it was shown that certain subsets of patients were more likely to have a mortality reduction with PCI or CABG, it is important to emphasize that the associated mortality impact was not exclusively derived from these factors alone. The underlying principle of the SYNTAX Score II being that it balances the interaction of anatomical complexity and six clinical variables that were shown to directly effect decision making on the most appropriate revascularization modality, and not each individual anatomical/clinical characteristic (Supplementary material online, Figure S1). Within the SYNTAX Score II, younger age, female gender, impaired renal function and reduced LVEF were shown to favour CABG compared with PCI on long-term prognostic grounds. As a result, patients with these specific characteristics were shown to derive a prognostic benefit from CABG, even when the anatomical complexity was lower. Conversely, older age, preserved renal and left ventricular function, and COPD were shown to favour PCI compared with CABG on long-term prognostic grounds. As a result, patients with these specific characteristics were shown to derive a prognostic benefit form PCI, even when the anatomical complexity was higher.

Diabetes

Diabetes has previously been shown not to be an independent predictor of mortality in the CABG or PCI arms of the SYNTAX trial, nor to have an interaction effect between CABG and PCI for long-term mortality when the end organ manifestations of diabetes were accounted for, as exemplified in the SYNTAX Score II [14,22]. Conversely, the FREEDOM trial demonstrated a reduction in mortality in diabetics with predominantly three-vessel disease treated by CABG compared with first-generation drug-eluting stents at a median follow-up of 3.8 years [27]. Importantly, ULMCA disease was an exclusion criteria in FREEDOM and is what prompted the presentation of mortality predictions in the diabetic subset of EXCEL. Notably in EXCEL, the presence of diabetes was associated with an increase in predicted mortality within both the CABG and PCI arms. Additionally, the predicted benefits of PCI were less pronounced in diabetics compared to non-diabetics, but remained similar to CABG (Figure 14.5; Table 14.2). In essence, despite the fact that diabetes was not contained within the SYNTAX Score II, the systemic metabolic effect of diabetes (such as age, CrCl, LVEF and other factors in the SYNTAX Score II) were associated with an increase in the patient risk profile and less favourable mortality predictions for PCI [14,28,29].

The SYNTAX Score II and medical advances

Evaluation of the Xience Everolimus-Eluting Stent vs. Coronary Artery Bypass Surgery for Effectiveness of Left-Main Revascularization was designed to study the impact of revascularization on ULMCA disease, incorporating changes in medical therapies, PCI technology and techniques, and advances in CABG that were introduced since the completion of the SYNTAX trial. For example, as the SYNTAX Score II was developed in the SYNTAX trial which— exclusively used the first-generation paclitaxel-eluting TAXUS stent, it is not inconceivable that the PCI arm of EXCEL may outperform the mortality predictions made in the present study. In EXCEL, the workhorse drug-eluting stent was the EES (XIENCE). The randomized comparisons of everolimus- vs. paclitaxel-eluting stents were designed and powered for a combination of angiographic, ischemic and safety outcomes, and have consistently shown the EES to be associated with more favourable outcomes compared with paclitaxel-eluting stents [9–12]. In addition, the largest patient level meta-analysis (n = 4989) of the SPIRIT clinical program has shown that EES was superior to paclitaxel-eluting stents in reducing all-cause mortality (3.2 vs. 5.1%, HR: 0.65, 95% CI: 0.49 – 0.86; P = 0.003 [13]. It is however important to emphasize that this difference was only driven by a lower non-cardiac mortality in the EES group and left-main revascularization was an exclusion criteria [13]. More specifically in subjects undergoing ULMCA revascularization, a recent systematic review comparing EES with first-generation DES (n = 2231) and a propensity match study (n = 344) have shown no statistically significant differences in all-cause mortality [30,31]. Furthermore, even within the SYNTAX Trial, when all stent thrombosis related deaths were removed, the impact on mortality reductions was shown to be modest (definite stent 0.5% reduction in mortality, definite to probable stent thrombosis 1.5% reduction in mortality) [32,33].

Limitations

The major limitation of the present study is also its greatest strength, namely the complete absence of the EXCEL trial outcomes (expected in the fall of 2016). We therefore cannot verify that the SYNTAX Score II predictions are accurate. However, predicting today the results of a randomized trial which will not be known for 2 years, assuming these predictions are reasonably borne out, opens the door for how future randomized trials may be considered whilst the longer-term (5 year) results of EXCEL are awaited. In addition, the present study will also enable unbiased validation of the SYNTAX Score II, fostering understanding of the multiple risk factors involved in ULMCA disease and decision making on the most appropriate revascularization modality. Although we are not using risk prediction for the primary endpoint of the trial, all-cause death is a hard, reproducible endpoint not subject to adjudication bias or definitional variation.

CONCLUSION

In the large-scale, prospective randomized EXCEL trial, the SYNTAX Score II indicated at least an equipoise for long-term mortality between CABG and PCI in subjects with ULMCA disease with low and intermediate anatomical complexity. Clinical characteristics had a clear impact on long-term mortality predictions and decision making between CABG and PCI. The accuracy of these mortality predictions will be compared with the actual individual outcome data from EXCEL in the coming years.

ACKNOWLEDGEMENTS

The EXCEL trial was supported by funds from Abbot Vascular (Santa Clara, CA, USA). The EXCEL trial design and conduct was performed by the EXCEL steering committee in which representatives of the study sponsor served. Data analysis, interpretation and writing of the manuscript were performed wholly independent from the study sponsor. All authors had unrestricted access to the full study database.

SUPPLEMENTARY MATERIAL

An online appendix can be found at http://dx.doi.org/10.1093/eurheartj/ehu518.

REFERENCES

1. Favaloro RG. Saphenous vein autograft replacement of severe segmental coronary artery occlusion: operative technique. Ann Thorac Surg 1968;5:334 – 339.
2. Yusuf S, Zucker D, Peduzzi P, Fisher LD, Takaro T, Kennedy JW, Davis K, Killip T, Passamani E, Norris R et al. Effect of coronary artery bypass graft surgery on survival: overview of 10-year results from randomised trials by the Coronary Artery Bypass Graft Surgery Trialists Collaboration. Lancet 1994;344:563 – 570.
3. Taggart DP. Thomas B. Ferguson Lecture. Coronary artery bypass grafting is still the best treatment for multivessel and left main disease, but patients need to know. Ann Thorac Surg 2006;82:1966 – 1975.
4. Gruntzig AR, Senning A, Siegenthaler WE. Nonoperative dilatation of coronary-artery stenosis: percutaneous transluminal coronary angioplasty. N Engl J Med 1979;301:61 – 68.
5. Farooq V, Serruys PW, Stone GW, Virmani R, Chieffo A, Fajadet J. Percutaneous interventional cardiovascular medicine. Left main coronary artery disease. The PCR-EAPCI Textbook. EUROPA edn. Toulouse, France: PCR Publishing; 2012. pp. 329 – 405.
6. Al Ali J, Franck C, Filion KB, Eisenberg MJ. Coronary artery bypass graft surgery versus percutaneous coronary intervention with first-generation drug-eluting stents: a meta-analysis of randomized controlled trials. JACC Cardiovasc Interv 2014; 7:497 – 506.
7. Athappan G, Patvardhan E, Tuzcu ME, Ellis S, Whitlow P, Kapadia SR. Left main coronary artery stenosis: a meta-analysis of drug-eluting stents versus coronary artery bypass grafting. JACC Cardiovasc Interv 2013;6:1219 – 1230.
8. Authors/Task Force members, Windecker S, Kolh P, Alfonso F, Collet JP, Cremer J, Falk V, Filippatos G, Hamm C, Head SJ, Juni P, Kappetein AP, Kastrati A, Knuuti J, Landmesser U, Laufer G, Neumann FJ, Richter DJ, Schauerte P, Sousa Uva M, Stefanini GG, Taggart DP, Torracca L, Valgimigli M, Wijns W, Witkowski A, Authors/ Task Force members. 2014 ESC/EACTS Guidelines on myocardial revascularization: The Task Forceon Myocardial Revascularizationofthe European Societyof Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS)Developed with the special contribution of the European Association of Percutaneous Cardiovascular Interventions (EAPCI). Eur Heart J 2014;35:2541– 2619.
9. Gada H, Kirtane AJ, Newman W, Sanz M, Hermiller JB, Mahaffey KW, Cutlip DE, Sudhir K, Hou L, Koo K, Stone GW. 5-year results of a randomized comparison of XIENCE V everolimus-eluting and TAXUS Paclitaxel-eluting stents: final results from the SPIRIT III trial (clinical evaluation of the XIENCE V everolimus eluting coronary stent system in the treatment of patients with de novo native coronary artery lesions). JACC Cardiovasc Interv 2013;6:1263 – 1266.
10. Brener SJ, Kereiakes DJ, Simonton CA, Rizvi A, Newman W, Mastali K, Wang JC, Caputo R, Smith RS Jr, Ying SW, Cutlip DE, Stone GW. Everolimus-eluting stents in patients undergoing percutaneous coronary intervention: final 3-year results of the Clinical Evaluation of the XIENCE V Everolimus Eluting Coronary Stent System in the Treatment of Subjects With de Novo Native Coronary Artery Lesions trial. Am Heart J 2013;166:1035 – 1042.
11. Stone GW, Rizvi A, Newman W, Mastali K, Wang JC, Caputo R, Doostzadeh J, Cao S, Simonton CA, Sudhir K, Lansky AJ, Cutlip DE, Kereiakes DJ, Investigators SI. Everolimus-eluting versus paclitaxel-eluting stents in coronary artery disease. N Engl J Med 2010;362:1663 – 1674.
12. Kedhi E, Joesoef KS, McFadden E, Wassing J, van Mieghem C, Goedhart D, Smits PC. Second-generation everolimus-eluting and paclitaxel-eluting stents in real-life practice (COMPARE): a randomised trial. Lancet 2010;375:201 – 209.
13. Dangas GD, Serruys PW, Kereiakes DJ, Hermiller J, Rizvi A, Newman W, Sudhir K, Smith RS Jr, Cao S, Theodoropoulos K, Cutlip DE, Lansky AJ, Stone GW. Meta-analysis of everolimus-eluting versus paclitaxel-eluting stents in coronary artery disease: final 3-year results of the SPIRIT clinical trials program (Clinical Evaluation of the Xience V Everolimus Eluting Coronary Stent

System in the Treatment of Patients With De Novo Native Coronary Artery Lesions). JACC Cardiovasc Interv 2013;6:914 – 922.

14. Farooq V, van Klaveren D, Steyerberg EW, Meliga E, Vergouwe Y, Chieffo A, Kappetein AP, Colombo A, Holmes DR Jr, Mack M, Feldman T, Morice MC, Stahle E, Onuma Y, Morel MA, Garcia-Garcia HM, van Es GA, Dawkins KD, Mohr FW, Serruys PW. Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score II. Lancet 2013;381:639 – 650.
15. Farooq V, van Klaveren D, Steyerberg EW, Serruys PW. SYNTAX score II – authors' reply. Lancet 2013;381:1899 – 1900.
16. Campos CM, van Klaveren D, Iqbal J, Onuma Y, Zhang YJ, Garcia-Garcia HM, Morel MA, Farooq V, Shiomi H, Furukawa Y, Nakagawa Y, Kadota K, Lemos PA, Kimura T, Steyerberg EW, Serruys PW. Predictive performance of SYNTAX Score II in patients with left main and multivessel coronary artery disease. Circ J 2014;78:1942 – 1949.
17. Serruys PW, Morice MC, Kappetein AP, Colombo A, Holmes DR, Mack MJ, Stahle E, Feldman TE, van den Brand M, Bass EJ, Van Dyck N, Leadley K, Dawkins KD, Mohr FW, Investigators S. Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. N Engl J Med 2009;360: 961 – 972.
18. Zhang YJ, Iqbal J, Campos CM, Klaveren DV, Bourantas CV, Dawkins KD, Banning AP, Escaned J, de Vries T, Morel MA, Farooq V, Onuma Y, Garcia-Garcia HM, Stone GW, Steyerberg EW, Mohr FW, Serruys PW. Prognostic value of site SYNTAX score and rationale for combining anatomic and clinical factors in decision making: insights from the SYNTAX trial. J Am Coll Cardiol 2014;64:423 – 432.
19. Stef van Buuren KG-O. Multivariate imputation by chained equations in R. J Stat Softw 2011;45; 1 – 67.
20. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Taylor & Francis; 1994.
21. Mohr FW, Morice MC, Kappetein AP, Feldman TE, Stahle E, Colombo A, Mack MJ, Holmes DR Jr, Morel MA, Van Dyck N, Houle VM, Dawkins KD, Serruys PW. Coronary artery bypass graft surgery versus percutaneous coronary intervention in patients with three-vessel disease and left main coronary disease: 5-year follow-up of the randomised, clinical SYNTAX trial. Lancet 2013;381: 629 – 638.
22. Farooq V, Serruys PW, Bourantas C, Vranckx P, Diletti R, Garcia Garcia HM, Holmes DR, Kappetein AP, Mack M, Feldman T, Morice MC, Colombo A, Morel MA, de Vries T, van Es GA, Steyerberg EW, Dawkins KD, Mohr FW, James S, Stahle E. Incidence and multivariable correlates of long-term mortality in patients treated with surgical or percutaneous revascularization in the synergy between percutaneous coronary intervention with taxus and cardiac surgery (SYNTAX) trial. Eur Heart J 2012;33:3105 – 3113.
23. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: the R Foundation for Statistical Computing; 2005.
24. Morice MC, Serruys PW, Kappetein AP, Feldman TE, Stahle E, Colombo A, Mack MJ, Holmes DR, Choi JW, Ruzyllo W, Religa G, Huang J, Roy K, Dawkins KD, Mohr F. Five-year outcomes in patients with left main disease treated with either percutaneous coronary intervention or coronary artery bypass grafting in the synergy between percutaneous coronary intervention with taxus and cardiac surgery trial. Circulation 2014;129:2388 – 2394.
25. Farooq V, Serruys PW, Bourantas CV, Zhang Y, Muramatsu T, Feldman T, Holmes DR, Mack M, Morice MC, Stahle E, Colombo A, de Vries T, Morel MA, Dawkins KD, Kappetein AP, Mohr FW. Quantification of incomplete revascularization and its association with five-year mortality in the synergy between percutaneous coronary intervention with taxus and cardiac surgery (SYNTAX) trial validation of the residual SYNTAX score. Circulation 2013;128:141 – 151.

26. Généreux P, Campos CM, Yadav M, Palmerini T, Caixeta A, Xu K, Francese DP, Dangas GD, Mehran R, Leon MB, Serruys PW, Stone GW. Reasonable incomplete revascularisation after percutaneous coronary intervention: the SYNTAX Revascularisation Index. EuroIntervention 2014; doi:10.4244/EIJY14M10_05.
27. Farkouh ME, Domanski M, Sleeper LA, Siami FS, Dangas G, Mack M, Yang M, Cohen DJ, Rosenberg Y, Solomon SD, Desai AS, Gersh BJ, Magnuson EA, Lansky A, Boineau R, Weinberger J, Ramanathan K, Sousa JE, Rankin J, Bhargava B, Buse J, Hueb W, Smith CR, Muratov V, Bansilal S, King S III, Bertrand M, Fuster V, Investigators FT. Strategies for multivessel revascularization in patients with diabetes. N Engl J Med 2012;367:2375 – 2384.
28. Tonelli M, Muntner P, Lloyd A, Manns BJ, Klarenbach S, Pannu N, James MT, Hemmelgarn BR, Alberta Kidney Disease N. Risk of coronary events in people with chronic kidney disease compared with those with diabetes: a population-level cohort study. Lancet 2012;380:807 – 814.
29. Mahmoodi BK, Matsushita K, Woodward M, Blankestijn PJ, Cirillo M, Ohkubo T, Rossing P, Sarnak MJ, Stengel B, Yamagishi K, Yamashita K, Zhang L, Coresh J, de Jong PE, Astor BC, Chronic Kidney Disease Prognosis C. Associations of kidney disease measures with mortality and end-stage renal disease in individuals with and without hypertension: a meta-analysis. Lancet 2012;380:1649 – 1661.
30. Capodanno D, Capranzano P, La Manna A, Tamburino C. Meta-analysis of everolimus-eluting stents versus first-generation drug-eluting stents in patients with left main coronary artery undergoing percutaneous coronary intervention. Int J Cardiol 2013;168:1718 – 1719.
31. Moynagh A, Salvatella N, Harb T, Darremont O, Boudou N, Dumonteil N, Lefevre T, Carrie D, Louvard Y, Leymarie JL, Chevalier B, Morice MC, Garot P. Two-year outcomes of everolimus vs. paclitaxel-eluting stent for the treatment of unprotected left main lesions: a propensity score matching comparison of patients included in the French Left Main Taxus (FLM Taxus) and the LEft MAin Xience (LEMAX) registries. EuroIntervention 2013;9:452 – 462.
32. Farooq V, Serruys PW. Complex coronary artery disease: would outcomes from the SYNTAX (synergy between percutaneous coronary intervention with taxus and cardiac surgery) trial have differed with newer-generation drug-eluting stents? JACC Cardiovasc Interv 2013;6:1023 – 1025.
33. Farooq V, Serruys PW, Zhang Y, Mack M, Stahle E, Holmes DR, Feldman T, Morice MC, Colombo A, Bourantas CV, de Vries T, Morel MA, Dawkins KD, Kappetein AP, Mohr FW. Short-term and long-term clinical impact of stent thrombosis and graft occlusion in the SYNTAX trial at 5 years: synergy between percutaneous coronary intervention with taxus and cardiac surgery trial. J Am Coll Cardiol 2013;62:2360 – 2369.

15

A novel risk score to predict out-of-hospital bleeding on dual antiplatelet therapy: derivation and validation of the predicting bleeding complications in patients undergoing stent implantation and subsequent dual anti platelet therapy (PRECISE-DAPT) score

F Costa*
D van Klaveren*
S James
F Feres
L Räber
T Pilgrim
MK Hong
HS Kim
A Colombo
PG Steg
T Palmerini
L Wallentin
DL Bhatt
GW Stone
S Windecker
EW Steyerberg
M Valgimigli

Submitted
* Equal contribution

ABSTRACT

Background Dual antiplatelet therapy (DAPT) with aspirin plus a $P2Y_{12}$ inhibitor prevents ischemic events after coronary stenting, but increases bleeding. Although guidelines support weighting bleeding risk prior to the selection of treatment duration, no standardised tool exists for this purpose. We sought to develop and validate a novel score to appraise out-of-hospital bleeding risk during DAPT in a large population treated with coronary stenting.

Methods A total of 14,963 patients, treated with DAPT after coronary stenting were pooled from eight contemporary multicentre randomized clinical trials with independent adjudication of events. Using Cox proportional hazards regression, we identified predictors of out-of-hospital bleeding stratified by trial, and developed a numerical bleeding risk score. The predictive performance of the novel score was assessed in the derivation cohort and validated in 13,016 patients from the PLATelet inhibition and patient Outcomes (PLATO) trial. The novel score was tested among patients randomized for DAPT duration (N=10,081) to identify those deriving a significant reduction in bleeding after a short (3-6 months) rather than long (12-24 months) treatment.

Findings After a median follow-up of one and a half-year, 218 (1.46%) patients suffered TIMI major or minor, and 124 (0.83%) TIMI major bleeding. The PRECISE-DAPT risk score (age, creatinine clearance, haemoglobin, white-blood cell count, prior spontaneous bleeding) showed a c-index of 0.73 (95%CI 0.61-0.85) in the derivation cohort and 0.68 (95%CI 0.64-0.71) in the validation cohort for out-of-hospital TIMI major or minor bleeding. A shorter DAPT duration significantly reduced out-of-hospital TIMI bleeding in patients deemed at high risk (score≥25), but not in those with a lower risk profile ($P_{interaction}$ 0.007).

Interpretation The PRECISE-DAPT is a simple risk score, which provides a standardized tool for the prediction of significant out-of-hospital bleeding during DAPT, hence supporting clinical decision-making for treatment duration.

INTRODUCTION

Dual antiplatelet therapy (DAPT) with aspirin and a $P2Y_{12}$ inhibitor reduces ischemic recurrences in patients with coronary artery disease treated with coronary stents [1, 2]. Yet this benefit is counterbalanced by higher bleeding risk, which is linearly related to the treatment duration. Both ischemic and bleeding risks have potential to negatively impact prognosis [3]. As a result, despite 12 months of DAPT after stenting has been commonly suggested, the optimal duration of treatment is still debated [4, 5]. Shortening DAPT duration to six or three months significantly reduced bleeding liability as compared to longer treatment duration [3]. However, a prolonged treatment beyond 12 months reduced non-fatal cardiovascular ischemic events in selected patients who tolerated the first year of treatment without bleeding [3, 6].

International guidelines advocate weighting bleeding risk prior to selection of treatment duration and suggest a shorter than 12 month treatment regimen in patients at high bleeding risk [4, 5]. Although this seems a reasonable strategy to at least reduce the risk for harm, no standardized tool exists to weigh bleeding risk at the time of DAPT initiation. Clinical risk scores, which integrate predictors to establish an individual's absolute risk of a condition, are often used in practice to estimate the bleeding risk in patients with atrial fibrillation and treated with oral anticoagulants [7]. Similarly, preliminary data suggested this approach might also be applied for a better selection of DAPT duration [8].

In response to this clinical need, we created a bleeding risk score for patients treated with DAPT after coronary stent implantation, in a large pooled dataset of contemporary randomized clinical trials implementing different DAPT duration strategies. We also externally validated this novel risk score in an independent, large, clinical trial cohort, which randomized patients to clopidogrel or ticagrelor, a novel $P2Y_{12}$ inhibitor, on a background therapy of aspirin.

METHODS

Study design and population

The PRECISE-DAPT (**P**RODIGY – **R**ESET – **E**XCELLENT – **C**OMFORTABLE AMI – B**I**OSCIENCE – **S**ECURITY – ZEUS – OPTIMIZ**E**) collaborative study included a total of 14,963 patients with coronary artery disease who underwent elective, urgent, or emergent percutaneous coronary intervention (PCI) with coronary stent implantation and subsequent DAPT. These were pooled at an individual patient level from eight contemporary multicentre randomized clinical trials [9-16]. Patients have been enrolled in 139 different clinical sites from 12 countries worldwide. Extensive details regarding the pooled datasets are provided in Appendix 15.1. Details regarding population type, randomization, DAPT duration and drug adherence are presented in Supplementary table 15.1. All clinical trials were approved by the ethics committees at each study centre, and all patients provided written informed consent.

Definitions of predictors and outcomes

All clinical and laboratory variables included in the current analysis were prospectively collected. The primary endpoint of this analysis was out-of-hospital bleeding defined according to the Thrombosis in Myocardial Infarction (TIMI) definition, and occurring 7 days or later after the initial invasive procedure, while bleeding occurring earlier was censored. We selected the 7-day time frame as a conservative estimate based on current hospitalization trends in acute coronary syndrome (ACS) patients, and to exclude events occurring during hospital stay, which are largely related to an invasive procedure [17]. Further details for bleeding and clinical variables definitions are provided in the appendix.

Validation cohort

An external validation of the risk score was performed within the PLATelet inhibition and patient Outcomes (PLATO) trial [2]. In brief, this included 18,421 patients with ST- or non-ST elevation ACS randomized to receive DAPT with either clopidogrel or ticagrelor in addition to aspirin for up to 12 months. In the current analysis we excluded 1,445 patients who underwent coronary artery bypass grafting (CABG) in order to keep the focus on spontaneous (i.e. non-procedural) bleeding. The novel score was calculated and assigned to each participant in a similar manner as in the derivation cohort. The information on “prior bleeding” in PLATO was related to prior gastro-intestinal bleeding, as no other prior bleeding types were prospectively collected in the study case report form. We also calculated the PARIS score in the external validation cohort in order to provide comparative assessment of the two bleeding prediction models [18]. Further details for score calculation in the validation cohort are provided in the appendix. The primary endpoint for score validation was the occurrence of TIMI major or minor and TIMI major bleeding 7 days or later after randomization. All variables included in PLATO trial were prospectively collected and a blinded clinical events committee independently adjudicated all possible protocol defined clinical events.

Statistical analysis including sensitivity analyses

A detailed description of the statistical analysis is provided in the appendix. We studied the associations between possible predictors and TIMI bleeding from day 7 onwards with a Cox regression analysis, stratified by trial. Potential predictors of bleeding were selected at univariable analysis ($p<0.10$) [19]. Independent bleeding predictors were selected with multivariable backward selection ($p<0.10$). Linear predictor values were scaled and rounded to a score with integer values between 0 and 100. Discrimination of the bleeding risk score was assessed by trial-specific Harrell's c-indexes, which were pooled with a random effects meta-analysis [20, 21]. Given the focus of this analysis on bleeding risk prediction on DAPT, we evaluated the score performance censoring patients' follow-up time and events occurring after the intended DAPT treatment duration and excluded patients who were not treated with DAPT at discharge (1.7%). The ability to separate bleeding risk categories was visualized by Kaplan Meier cumulative bleeding incidence curves. Calibration was assessed by comparing predicted probabilities with one-year Kaplan Meier bleeding incidence estimates. Furthermore, discrimination and calibration of the bleeding risk score were assessed in the external PLATO cohort. We compared the c-index of the new risk score with the c-index of the PARIS score in the external validation cohort [22]. The analyses were done in accordance with the TRIPOD statement [23]. Data were analysed with R version 3.6 (R Foundation, Austria).

Role of the funding source

No additional funding was used for current analysis. All trials included in the PRECISE-DAPT collaborative study were investigator-initiated and each sponsor had no role in the data analysis, interpretation, or writing of the report. The corresponding author had full access to the data and had final responsibility for the decision to submit for publication.

RESULTS

The study population included 14,963 patients with established coronary artery disease, treated with coronary stent implantation (Table 15.1). Dual antiplatelet therapy at discharge was implemented virtually in all patients (98.3%) with median treatment duration of 360 days (IQR 95-365). In a total of 21,963 person-years of follow-up (median follow-up 552 days, IQR 365-725), out-of-hospital TIMI major or minor bleeding occurred in 218 patients (one-year bleeding risk 12.5 /1000 patients), 124 of whom fulfilled TIMI major criteria (one-year bleeding risk 6.9 /1000 patients). The median time to first occurrence of TIMI major or minor and TIMI major bleeding was 158 days (IQR 57-333) and 150 days (62-326), respectively. The rate of bleeding stratified by clinical trial is presented in Supplementary table 15.2.

Univariable and multivariable analysis

Univariable analysis for TIMI major or minor bleeding is presented in Table 15.1. Predictors with a p<0.10 at univariable analysis were included into the multivariable model.

Table 15.1 Baseline characteristics and univariable analysis for out of hospital TIMI bleeding.

	Derivation Cohort (N=14,963)	Validation Cohort (N=13,016)	Hazard Ratio (95% CI)	p value
Age (years)	65.0 (56.9-73.0)	62.0 (54.0-71.0)	1.041 (1.028-1.055)	<0.0001
Women (vs. Man)	4,414/14,963 (29.5%)	3,793/13,016 (29.1%)	1.018 (0.755-1.373)	0.910
Hypertension	10,739/14,942 (71.9%)	8,439/13,016 (64.8%)	1.271 (0.921-1.753)	0.145
Dyslipidaemia	9,080/14,822 (61.3%)	5,918/13,015 (45.5%)	0.935 (0.711-1.229)	0.628
Current Smoking	3,757/14,871 (28.0%)	4,716/13,016 (36.2%)	0.747 (0.536-1.042)	0.085
Diabetes	4,168/14,949 (27.9%)	3,143/13,016 (24.1%)	1.198 (0.894-1.606)	0.226
Insulin dependent	797/14,888 (5.4%)	710/13,016 (5.5%)	1.652 (1.027-2.658)	0.038
PVD	714/6885 (10.4%)	804/13,016 (6.2%)	1.124 (0.630-2.006)	0.693
LVEF (%)	55.0 (45.0-61.0)	n.a.	0.982 (0.969-0.995)	0.008
History				
Prior MI	2946/14,902 (19.8%)	2,623/13,016 (20.2%)	1.141 (0.871-1.494)	0.338
Prior PCI	2392/14,933 (16.0%)	1,740/13,016 (13.4%)	1.057 (0.789-1.415)	0.710
Prior CABG	893/14,943 (6.0%)	807/13016 (6.2%)	0.893 (0.517-1.542)	0.685
Prior Stroke	473/13,061 (3.6%)	505/13,016 (3.9%)	1.160 (0.543-2.479)	0.701
Prior Bleeding †	82/4286 (1.9%)	192/13,016 (1.5%)	2.822 (0.881-9.045)	0.08
Weight (Kg)	74.0 (65-84)	80.0 (70.0-90.0)	0.982 (0.970-0.994)	0.003
Laboratory data at baseline				
CrCl (ml/min)	79.1 (60.8-98.0)	80.8 (63.3-99.5)	0.981 (0.974-0.988)	<0.0001
WBC (10^3 units/µL)	7.8 (6.3-10.2)	9.2 (7.3-11.5)	1.049 (0.992-1.109)	0.095
Haemoglobin (g/dL)	13.8 (12.7-14.9)	14.0 (12.9-14.9)	0.756 (0.684-0.835)	<0.0001
Clinical Presentation				
SCAD	6,299/14,183 (44.4%)	n.a.	Ref.	-
UA	3,215/14,183 (22.7%)	2,199/13,004 (16.9%)	0.903 (0.628-1.298)	0.582
NSTEMI	1,990/14,183 (14.0%)	5,384/13,004 (41.4%)	1.468 (0.927-2.228)	0.122
STEMI	2,679/14,183 (18.9%)	5,125/13,004 (39.4%)	1.300 (0.795-2.128)	0.296
Therapy at discharge				
$P2Y_{12}$ inhibitors (days)	360 (95-365)	280 (182-365)	1.030 (1.004-1.055)	0.021
Aspirin	14,660/14,860 (98.7%)	12,433/13,016 (95.5%)	-	-
Clopidogrel	13,028/14,849 (87.7%)	6,499/13,016 (49.9%)	Ref.	-
Prasugrel	1,123/14,849 (7.6%)	n.a.	1.265 (0.780-2.052)	0.340
Ticagrelor	578 /14,849 (3.9%)	6,517/13,016 (50.1%)	0.649 (0.314-1.344)	0.245
OAC	431/6,871 (6.3%)	28/13016 (0.2%)	1.270 (0.715-2.253)	0.415
Statin	12,038/13,467 (89.4%)	8311/13016 (63.9%)	0.705 (0.457-1.085)	0.112
Beta blocker	10,007/13,467 (74.3%)	10384/13016 (79.8%)	0.840 (0.612-1.153)	0.282
ACE/ARB	8,984/13,466 (66.7%)	9808/13016 (75.4%)	0.857 (0.667-1.100)	0.225
NSAID	103/4,868 (2.1%)	734/13016 (5.6%)	1.047 (0.257-4.262)	0.586
PPI	2,669/6,868 (38.9%)	4333/13016 (33.3%)	1.519 (1.058-2.181)	0.024

Data are n (%) or median (IQR). PVD= peripheral vascular disease. LVEF= left ventricle ejection fraction. MI= myocardial infarction. PCI= percutaneous coronary intervention. CABG= coronary artery bypass graft. SCAD= stable coronary artery disease. UA= unstable angina. NSTEMI= non-ST segment elevated myocardial infarction. STEMI= ST segment elevated myocardial infarction. OAC= oral anticoagulant. ACE/ARB: ACE inhibitor or angiotensin-II receptor blocker. NSAID= non-steroidal anti-inflammatory drug. PPI= proton pump inhibitor. N.A.= Not Avaiable. †Information regarding prior bleeding within the PLATO trial was limited to prior gastro-intestinal bleeding.

Use of proton pump inhibitor at discharge was subsequently excluded because of lack of prediction within studies where DAPT duration was randomised. Five predictors remained in the final model at $p<0.10$, including age, white-blood cell count (WBC), haemoglobin, creatinine clearance, and history of spontaneous bleeding (Table 15.2). All five independent predictors were consistently associated with bleeding during the first trimester after treatment initiation as well as beyond (data not shown). The model showed c-indexes of 0.73 (95%CI 0.62-0.85) for TIMI major or minor and 0.72 for TIMI major bleeding (95%CI 0.58-0.85). An alternative model, which has been generated after excluding WBC, is shown in the appendix (Supplementary table 15.4).

Table 15.2 Multivariable study-stratified analysis for out-of-hospital TIMI major or minor bleeding after backward selection at a significance level of 0.1.

	Hazard Ratio (95% CI)	χ^2	p
Age (ten years)	1.34 (1.11 - 1.48)	8.0	0.005
Prior bleeding	4.14 (1.22 - 14.02)	5.2	0.023
White Blood Cell Count (10^3 units/µL)	1.06 (0.99 - 1.13)	3.1	0.078
Haemoglobin at baseline (one g/dL)	0.67 (0.53 - 0.84)	11.4	0.001
Creatinine Clearance (ten ml/min)	0.90 (0.82 - 0.99)	8.2	0.004

Age was truncated below 50 years; haemoglobin at baseline was truncated above 12 and below 10 g/dl; creatinine clearance was truncated above 100 ml/min; white blood cell count was truncated above 20 and below 5 x10^3 units/µL

Risk score

From the final multivariable model we developed a five-item bleeding risk score (age, creatinine clearance, haemoglobin and WBC at baseline, and prior spontaneous bleeding – PRECISE-DAPT score) assigning points to each factor based on the magnitude of association of each predictor with bleeding. A nomogram to calculate the score and the risk of bleeding at 12 months is presented in Figure 15.1. Similar information derived from the model lacking WBC is presented in the appendix (Supplementary figure 15.1). A web-calculator and mobile App are available at www.precisedaptscore.com.

Score performance in the derivation cohort

The PRECISE-DAPT score showed a c-index of 0.73 (95%CI 0.61-0.85) for out-of-hospital TIMI major or minor bleeding and 0.71 (95%CI 0.57-0.85) for TIMI major bleeding (Table 15.3). C-Indexes for each of the included studies are presented in Supplementary table 15.3. The score's discrimination was also consistent irrespective to the clinical presentation at the time of PCI (Table 15.4). On further sensitivity analysis, the score discriminated TIMI bleeding in patients receiving clopidogrel (c-index: 0.76, 95%CI 0.65-0.86) or ticagrelor (c-index: 0.71, 95%CI 0.44-0.98), whereas lower discrimination was observed for prasugrel (c-index: 0.60, 95%CI 0.41-0.78). The performance of the score lacking WBC, is presented in Table 15.3 and Supplementary figure 15.2.

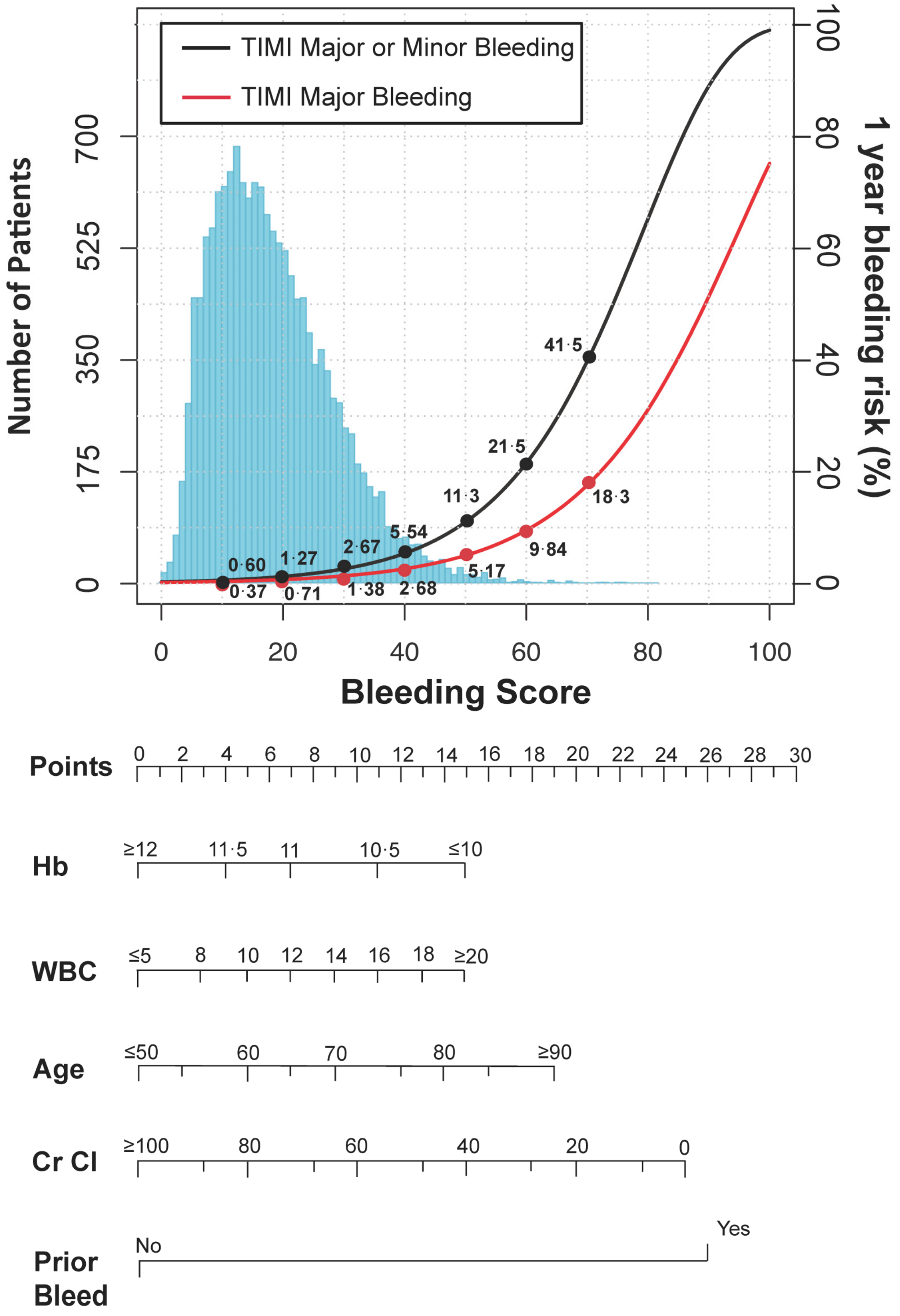


Figure 15.1 The PRECISE-DAPT score nomogram for bedside application. The figure refers to out-of-hospital TIMI major or minor and TIMI major bleeding at 12 months, while on-treatment with DAPT.

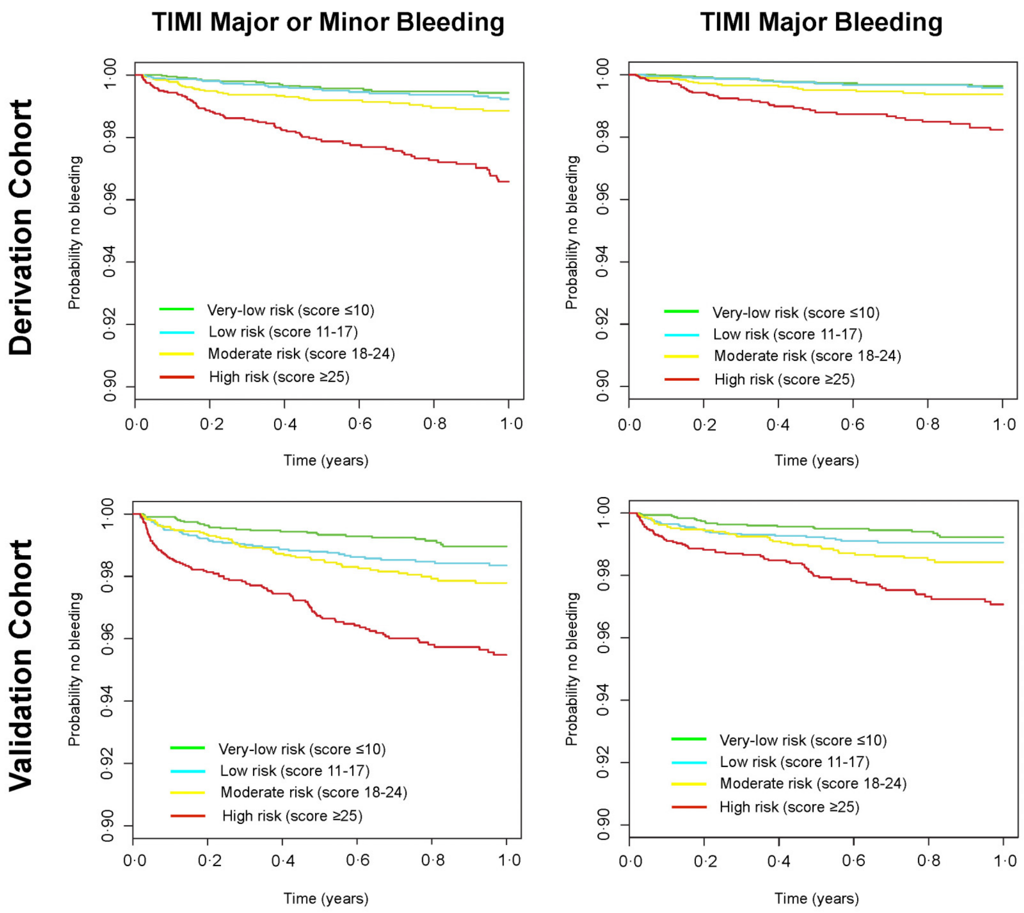


Figure 15.2 Kaplan Meier estimates of survival free from bleeding in both derivation and validation cohort stratified by score categories. Estimates for TIMI major or minor bleeding and TIMI major bleeding occurring while on-treatment with DAPT are presented.

Kaplan-Meier bleeding rates by score quartiles (very low: score ≤10; low: score 11-17; moderate: score 18-24; and high risk: score ≥25) are presented in Figure 15.2. Score calibration appeared good for both TIMI major or minor and TIMI major bleeding in the derivation cohort (Figure 15.3).

Score performance in the validation cohort

The PRECISE-DAPT score was available in 13,016 patients from the PLATO trial (Table 15.1). TIMI major or minor bleeding occurred in 232 patients (2.26%) whereas TIMI major occurred in 154 patients (1.49%). The score achieved c-indexes of 0.68 (95%CI 0.65-0.72) for out-of-hospital TIMI major or minor and 0.67 (95%CI 0.62-0.71) for TIMI major bleeding in the PLATO validation cohort (Table 15.3). Because the PLATO trial enrolled a higher-risk population, the score underestimated the absolute bleeding risk, yet it well discriminated bleeding risk status across patients (Figure 15.3). Discriminative ability and calibration remained consistent for the score lacking WBC (Table 15.3 and Supplementary figure 15.2).

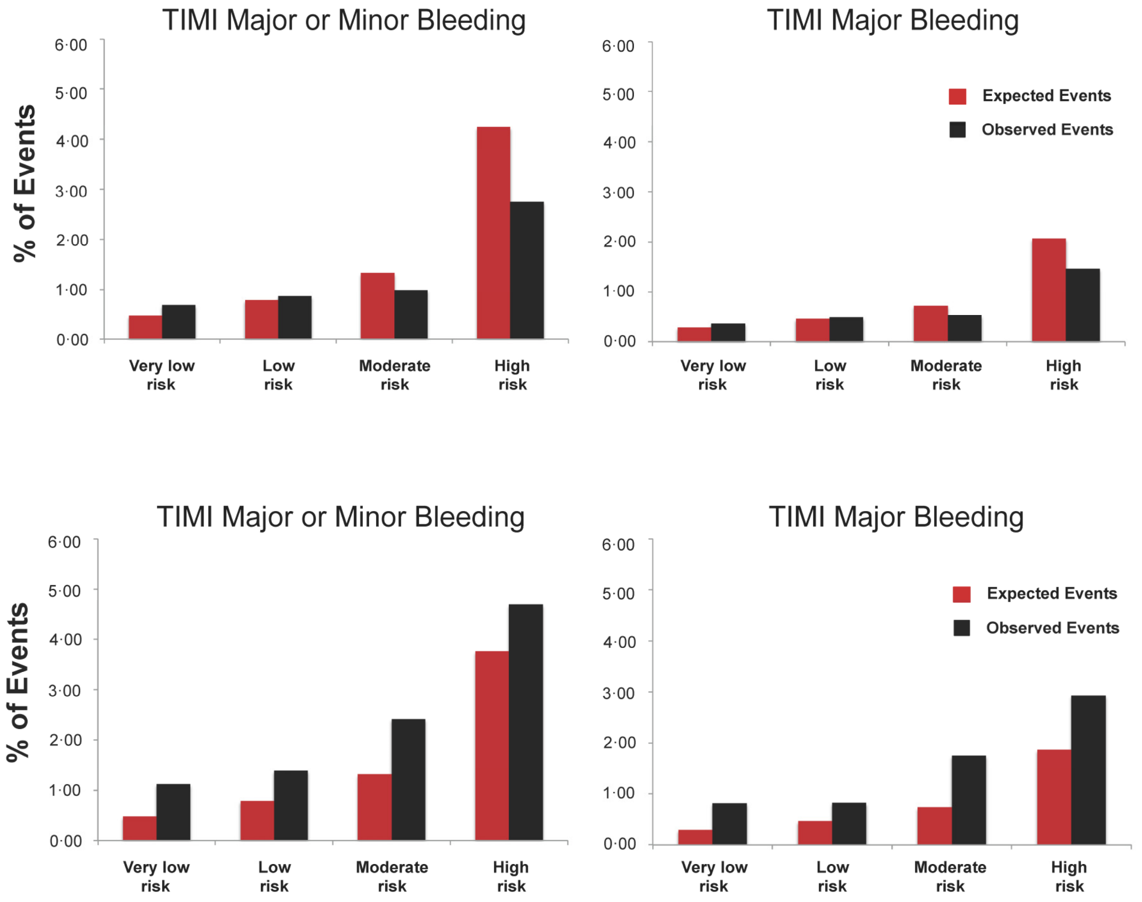


Figure 15.3 Calibration plot for the PRECISE-DAPT score in the derivation (upper panel) and validation (lower panel) cohort for TIMI major or minor bleeding and TIMI major bleeding occurring while on-treatment with DAPT.

Comparative bleeding risk prediction assessment

The PARIS bleeding risk score was used as benchmark comparator for the risk prediction offered by the PRECISE DAPT score. In the validation cohort, the PRECISE-DAPT score showed superior discrimination as compared to the PARIS score for both TIMI major or minor (0.68 vs. 0.65; p= 0.016) and for TIMI major bleeding (0.67 vs. 0.61; p= 0.003)(Table 15.3). The PRECISE DAPT score lacking WBC provided consistent superior discrimination than the PARIS score for both TIMI major or minor (0.68 vs. 0.65; p= 0.008) and TIMI major bleeding (0.66 vs. 0.61; p= 0.004)(Table 15.3).

Table 15.3 Discriminative ability of the PRECISE-DAPT score in the derivation and validation cohort for bleeding occurring while on-treatment with DAPT.

	TIMI Major or Minor	p-value for Difference*	TIMI Major	p-value for Difference*
Derivation Cohort				
Events n/ in group	218/14,963		124/14,963	
PRECISE-DAPT	0.73 (0.61-0.85)	-	0.71 (0.57-0.85)	-
PRECISE-DAPT Alternative*	0.71 (0.57-0.84)	-	0.69 (0.53-0.85)	-
Validation Cohort				
Events n/ in group	229/13,016		151/13,016	
PRECISE-DAPT	0.68 (0.65-0.72)	0.016	0.67 (0.62-0.71)	0.003
PRECISE-DAPT Alternative*	0.68 (0.65-0.72)	0.008	0.66 (0.62-0.71)	0.004
PARIS	0.65 (0.61-0.68)	Ref.	0.61 (0.57-0.66)	Ref.

Data are c-indices (95% CI) for each score. Descriptions of the scores: PRECISE-DAPT score is age, creatinine clearance, haemoglobin and white-blood cell count at baseline, and prior spontaneous bleeding; PRECISE-DAPT Alternative is age, creatinine clearance, haemoglobin at baseline, and prior spontaneous bleeding; PARIS is Age, body mass index, current smoking status, presence of anemia (haemoglobin < 12g/dl in men and 11g/dl in women), creatinine clearance < 60ml/dl and treatment with triple therapy (i.e. aspirin plus $P2Y_{12}$ inhibitor plus oral anticoagulant) at discharge. *The PARIS score has been used as comparator to test the difference in discriminative ability with the PRECISE-DAPT scores.

Impact of randomized DAPT duration among bleeding risk strata

Out of five studies where DAPT duration was randomly allocated, 5,050 patients were assigned to either 12 or 24 months of treatment and 5,031 to three or six months of treatment duration [11-15]. We observed a significant reduction in bleeding risk with a short (3-6 months) rather than a long (12-24 months) duration of treatment exclusively in patients at high bleeding risk (ARD -2.59, 95%CI -4.34 to -0.82; NNT: 38) but not in those with a lower bleeding risk profile (ARD -0.14, 95%CI -0.49 to +0.22) (P_{int}=0.007)(Figure 15.4). The bleeding risk status-by-DAPT-duration-interaction on bleeding events remained significant after censoring events occurring beyond 12 months (P_{int}=0.047).

Table 15.4 Discriminative ability of the PRECISE-DAPT score for bleeding events occurring while on-treatment with DAPT and stratified according to the clinical presentation at the time of PCI.

Clinical Presentation	TIMI Major or Minor Bleeding*	TIMI Major Bleeding**
	C-Index (95%CI)	C-Index (95%CI)
Stable coronary artery disease	0.72 (0.64-0.80)	0.72 (0.62-0.82)
Unstable Angina	0.66 (0.53-0.79)	0.80 (0.65-0.94)
Non ST-segment elevated MI	0.84 (0.72-0.96)	0.84 (0.70-0.98)
ST-segment elevated MI	0.69 (0.60-0.78)	0.62 (0.51-0.72)

MI: Myocardial Infarction

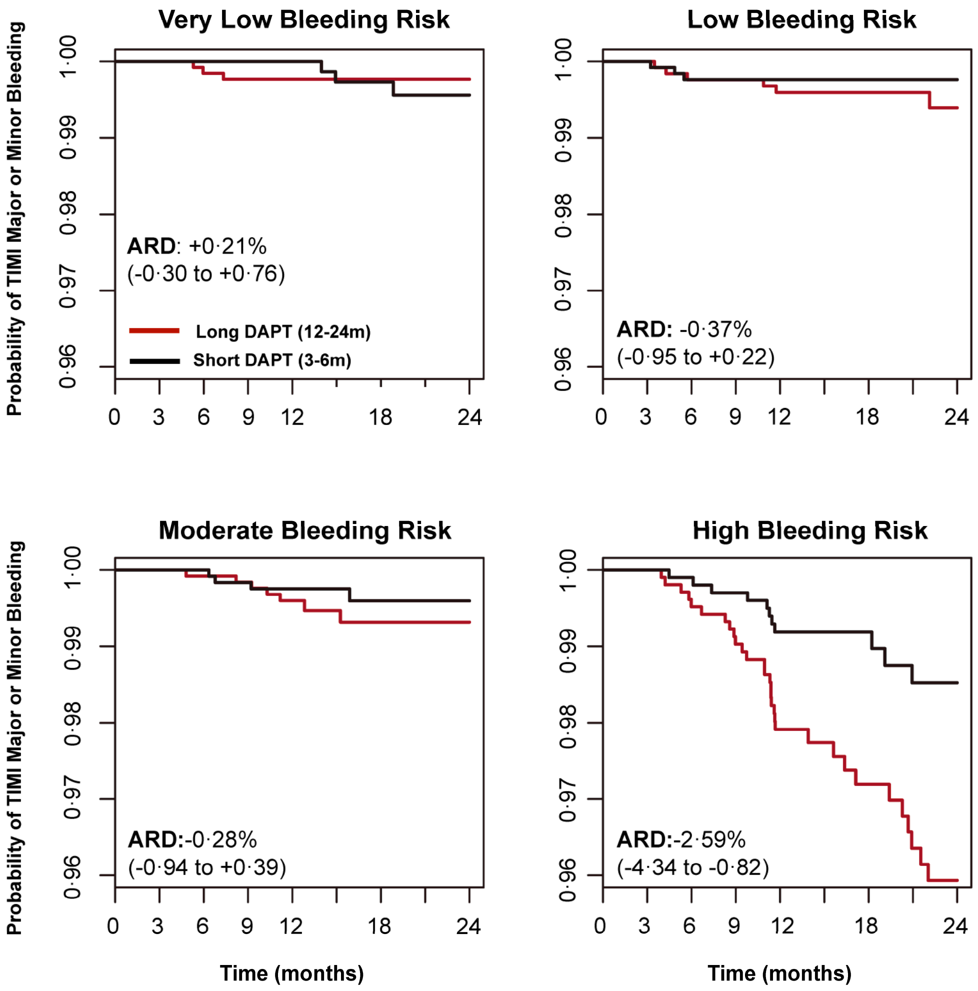


Figure 15.4 Twenty-four month Kaplan Meier estimates of survival free from TIMI major or minor bleeding among PRECISE-DAPT bleeding risk categories (i.e. very-low, low, moderate and high bleeding risk) for patients randomized to short (3-6 months) or longer (12-24 months) dual antiplatelet therapy. Absolute risk differences (ARD) are presented: a negative ARD represent the risk reduction for a shorter as compared to a longer course of DAPT.

DISCUSSION

Ischemic events after stenting dropped in the last years thanks to the introduction of novel generation stent technologies and progressive refinement of interventional techniques. However, due to more potent and prolonged platelet inhibition, the incidence of major bleeding has increased [2, 4]. DAPT related bleeding is the most common complication after coronary stent implantation in current practice, and it is associated with lower survival, lower quality of life and higher health costs [24, 25].

This study developed and validated the PRECISE-DAPT score, a new tool for the prediction of out-of-hospital bleeding risk in patients treated with DAPT. We confirmed the role of well-known risk factors associated with post-discharge bleeding such as age and haemoglobin at baseline. Similarly, risk factors, which were previously linked with in-hospital bleeding, such as renal function, and WBC, have been extended to the prediction of out-of-hospital events [25, 26]. In addition, we underlined the clinical relevance of prior bleeding, which is commonly appraised in practice despite being not routinely collected in clinical studies [27]. Recently, a contemporary real-world registry showed that a history of prior gastro-intestinal bleeding was a strong predictor for long-term bleeding [28]. We extended this finding evaluating the impact of all prior spontaneous bleeding requiring medical attention, which ultimately emerged as the most impactful predictor of bleeding in our score. A simplified score modelled without WBC was also derived and validated. This simplified four-item score showed reasonable discriminatory capability and may help providing objective bleeding risk assessment in cases where WBC might not be readily available.

International guidelines suggest individualization of the antiplatelet treatment duration [4, 5]. This recommendation was reinforced after consistent evidence from multiple randomized studies invariably showing bleeding liability associated with a prolonged as compared to shortened DAPT duration regimens [3, 6, 13]. In our analysis we observed that the increase in clinically significant bleeding related to a prolonged treatment duration with DAPT occurred almost exclusively in patients with a higher bleeding risk score at baseline, whereas the impact of a prolonged DAPT regimen on bleeding in patients at lower bleeding risk was marginal. As such, the PRECISE-DAPT score shows potential to inform decision-making for DAPT duration. Selecting upfront a shorter than 12-month treatment duration in patients deemed at high bleeding risk (PRECISE-DAPT score ≥25) may prevent exposing them to an excessive bleeding hazard. In turn, patients not at high bleeding risk might receive a standard (i.e. 12 months) or a prolonged (i.e. > 12 months) course of treatment if DAPT was well tolerated.

A recent post hoc analysis from the DAPT trial developed and validated a similar standardized tool to support decision-making on DAPT duration [29]. Still, only patients who tolerated and adhered to an initial 12-month course of treatment were included in the DAPT study [6]. Hence, at variance with our score, this prediction rule cannot be applied at the time of the initial stent placement to select DAPT duration within the first year. Earlier decision-making, especially in patients at high bleeding risk, is advisable since most bleeding occurs early after treatment initiation. With that respect our score has the potential to integrate the one developed in the context of the DAPT trial. Patients at non-high bleeding risk may be targeted with 12-month therapy and in those, in whom no bleeding events occurred during this period, the DAPT risk prediction model can better inform on benefits and risks to further continue or not the treatment.

In a recent analysis from the PARIS registry two independent risk scores have been developed to evaluate the risk benefit ratio of DAPT continuation [18]. At variance with the DAPT score, PARIS accounted for all patients that initiated the treatment with DAPT, which is consistent with our analysis. Our score ultimately showed improved discrimination as compared to the PARIS bleeding risk score. Interestingly, also the simplified score (i.e. the one lacking WBC in the model), despite some predictable loss in discrimination, showed superior performance when compared to the risk prediction model developed within PARIS. This may possibly due to the implementation of prior spontaneous bleeding in the model, which is commonly appraised in practice [27], and it consistently emerged as a major bleeding predictor in patients treated with oral anticoagulants [7, 30].

The strengths of our study include: the derivation of a risk score, which may be easily applied in everyday clinical practice; this was developed from a largely representative, prospectively investigated, cohort with rigorous event adjudication, and based on a well-standardised and accepted bleeding definition. In addition our score was validated in an independent, large cohort of patients treated with novel $P2Y_{12}$ inhibitors, which now represents the standard of care for patients with ACS [2, 4, 5].

Several limitation of our study should be acknowledged: first, emerging predictors for bleeding in patients treated with DAPT might be missing in our model [30]. Future studies might use our score as a basis to evaluate the incremental value of novel clinical, laboratory or genetic factors in an attempt to improve risk prediction. Second, information regarding prior bleeding in the validation cohort was limited to prior gastro-intestinal bleeding; this might have reduced discrimination of our score in the validation cohort. Third, because of the higher bleeding risk in the PLATO validation cohort, which included only patients with ACS, our score underestimated the absolute bleeding risk in this population. Still, we decided not to recalibrate the score on PLATO trial, because this might not be entirely representative of an all-comer, real-world, population treated with coronary stents. In addition, discrimination in patients treated with prasugrel was poorer in the derivation cohort. Whether this is a true or a chance finding remains unclear at this stage. This observation could not be verified in the validation cohort, which lacked patients treated with prasugrel. Fourth, PARIS bleeding score discrimination might have been underestimated in our validation cohort since a treatment with oral anticoagulants was an exclusion criterion in the PLATO trial. Finally, whether the routine use of the PRECISE DAPT risk score in an unselected population, significantly mitigates bleeding risk by better informing decision-making remains to be prospectively ascertained.

CONCLUSION

We developed and validated the PRECISE-DAPT score, a simple prediction algorithm for out-of-hospital bleeding in patients treated with DAPT. The PRECISE-DAPT score was able to identify patients at high bleeding risk, who may benefit from a shorter than a longer DAPT duration, and patients at lower bleeding risk, who may reasonably well tolerate 12-month or longer treatment duration. Prospective validation of this score in practice is required.

RESEARCH IN CONTEXT

Evidence before this study Spontaneous bleeding during treatment with dual antiplatelet therapy is the most common complication after coronary stenting [25], and its incidence increased with the introduction of novel and more potent antithrombotic agents [2]. Despite recommendations from international guidelines, means to gauge out-of-hospital bleeding risk in patients treated with dual antiplatelet therapy (DAPT) are limited [4, 5]. In other scenarios (e.g. oral anticoagulation for atrial fibrillation) clinical risk scores are commonly used to standardize bleeding risk, and guiding treatment accordingly [7]. A dedicated risk score specifically designed to predict spontaneous on-DAPT bleeding events might improve risk assessment and support clinicians' decisions with respect to dual antiplatelet therapy. We searched PubMed without language or date restrictions for publications up to May 2016 about bleeding risk scores in patients treated with dual antiplatelet therapy. We used the search terms "percutaneous coronary intervention", "coronary stent", "acute coronary syndrome", "stable coronary artery disease", "bleeding risk score", "bleeding", "antiplatelet therapy", "dual antiplatelet therapy", "clopidogrel", "prasugrel" and "ticagrelor". We excluded articles regarding antithrombotic treatment in atrial fibrillation, concomitant use of oral anticoagulants and risk prediction models for in-hospital bleeding. We identified two reports focused on out-of-hospital events in patients treated with DAPT [18, 29], and one was only applicable after a 12-month course with DAPT was completed without complications [29].

Added value of this study We propose a novel risk score for the prediction of out-of-hospital bleeding in patients treated with DAPT using age, creatinine clearance, white-blood cell count, haemoglobin and history of bleeding. The PRECISE-DAPT score is a simple bedside risk assessment tool which can be easily implemented in everyday clinical practice, and that was developed and internally and externally validated in two large prospective cohorts of patients with rigorous event adjudication, along with different clinical presentations and $P2Y_{12}$ inhibitors. This novel score showed superior discrimination as compared to the previously validated PARIS bleeding risk score and might be particularly useful for its applicability at the time of treatment initiation. In fact, the PRECISE-DAPT score showed potential to identify patients at high bleeding risk (score≥25) who may benefit from a shortened (i.e. less than 12 months) DAPT duration. In turn, patients not at high bleeding risk (score<25) might receive a standard (i.e. 12 months) or prolonged (i.e. > 12 months) treatment without being exposed to significant bleeding liability. As such, our score might also well integrate with decision-making tools selecting patients for a prolonged DAPT regimen beyond 12-months, if treatment was well tolerated.

Implication of all the available evidence Our study provides awareness to clinicians regarding out-of-hospital bleeding risk factors in patients treated with DAPT after coronary stent implantation and offers an objective and standardised tool to quantify such risk in clinical practice. Systematic evaluation of these predictors with the novel PRECISE-DAPT bleeding risk score has potential to support clinical decision-making with respect to the optimal duration of dual antiplatelet therapy.

REFERENCES

1. Yusuf S, Zhao F, Mehta SR, Chrolavicius S, Tognoni G, Fox KK, Clopidogrel in Unstable Angina to Prevent Recurrent Events Trial I. Effects of clopidogrel in addition to aspirin in patients with acute coronary syndromes without ST-segment elevation. *N Engl J Med* 2001; 345: 494-502.
2. Wallentin L, Becker RC, Budaj A, Cannon CP, Emanuelsson H, Held C, Horrow J, Husted S, James S, Katus H, Mahaffey KW, Scirica BM, Skene A, Steg PG, Storey RF, Harrington RA, Investigators P, Freij A, Thorsen M. Ticagrelor versus clopidogrel in patients with acute coronary syndromes. *N Engl J Med* 2009; 361: 1045-1057.
3. Navarese EP, Andreotti F, Schulze V, Kolodziejczak M, Buffon A, Brouwer M, Costa F, Kowalewski M, Parati G, Lip GY, Kelm M, Valgimigli M. Optimal duration of dual antiplatelet therapy after percutaneous coronary intervention with drug eluting stents: meta-analysis of randomised controlled trials. *BMJ* 2015; 350: h1618.
4. Authors/Task Force m, Windecker S, Kolh P, Alfonso F, Collet JP, Cremer J, Falk V, Filippatos G, Hamm C, Head SJ, Juni P, Kappetein AP, Kastrati A, Knuuti J, Landmesser U, Laufer G, Neumann FJ, Richter DJ, Schauerte P, Sousa Uva M, Stefanini GG, Taggart DP, Torracca L, Valgimigli M, Wijns W, Witkowski A. 2014 ESC/EACTS Guidelines on myocardial revascularization: The Task Force on Myocardial Revascularization of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS)Developed with the special contribution of the European Association of Percutaneous Cardiovascular Interventions (EAPCI). *Eur Heart J* 2014; 35: 2541-2619.
5. Levine GN, Bates ER, Bittl JA, Brindis RG, Fihn SD, Fleisher LA, Granger CB, Lange RA, Mack MJ, Mauri L, Mehran R, Mukherjee D, Newby LK, O'Gara PT, Sabatine MS, Smith PK, Smith SC, Jr., Focused Update Writing G. 2016 ACC/AHA Guideline Focused Update on Duration of Dual Antiplatelet Therapy in Patients With Coronary Artery Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J Am Coll Cardiol* 2016.
6. Mauri L, Kereiakes DJ, Yeh RW, Driscoll-Shempp P, Cutlip DE, Steg PG, Normand SL, Braunwald E, Wiviott SD, Cohen DJ, Holmes DR, Jr., Krucoff MW, Hermiller J, Dauerman HL, Simon DI, Kandzari DE, Garratt KN, Lee DP, Pow TK, Ver Lee P, Rinaldi MJ, Massaro JM, Investigators DS. Twelve or 30 months of dual antiplatelet therapy after drug-eluting stents. *N Engl J Med* 2014; 371: 2155-2166.
7. Pisters R, Lane DA, Nieuwlaat R, de Vos CB, Crijns HJ, Lip GY. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey. *Chest* 2010; 138: 1093-1100.
8. Costa F, Tijssen JG, Ariotti S, Giatti S, Moscarella E, Guastaroba P, De Palma R, Ando G, Oreto G, Zijlstra F, Valgimigli M. Incremental Value of the CRUSADE, ACUITY, and HAS-BLED Risk Scores for the Prediction of Hemorrhagic Events After Coronary Stent Implantation in Patients Undergoing Long or Short Duration of Dual Antiplatelet Therapy. *J Am Heart Assoc* 2015; 4.
9. Pilgrim T, Heg D, Roffi M, Tuller D, Muller O, Vuilliomenet A, Cook S, Weilenmann D, Kaiser C, Jamshidi P, Fahrni T, Moschovitis A, Noble S, Eberli FR, Wenaweser P, Juni P, Windecker S. Ultrathin strut biodegradable polymer sirolimus-eluting stent versus durable polymer everolimus-eluting stent for percutaneous coronary revascularisation (BIOSCIENCE): a randomised, single-blind, non-inferiority trial. *Lancet* 2014; 384: 2111-2122.
10. Raber L, Kelbaek H, Ostojic M, Baumbach A, Heg D, Tuller D, von Birgelen C, Roffi M, Moschovitis A, Khattab AA, Wenaweser P, Bonvini R, Pedrazzini G, Kornowski R, Weber K, Trelle S, Luscher TF, Taniwaki M, Matter CM, Meier B, Juni P, Windecker S, Investigators CAT. Effect of biolimus-eluting stents with biodegradable polymer vs bare-metal stents on cardiovascular events among patients with acute myocardial infarction: the COMFORTABLE AMI randomized trial. *JAMA* 2012; 308: 777-787.

11. Gwon HC, Hahn JY, Park KW, Song YB, Chae IH, Lim DS, Han KR, Choi JH, Choi SH, Kang HJ, Koo BK, Ahn T, Yoon JH, Jeong MH, Hong TJ, Chung WY, Choi YJ, Hur SH, Kwon HM, Jeon DW, Kim BO, Park SH, Lee NH, Jeon HK, Jang Y, Kim HS. Six-month versus 12-month dual antiplatelet therapy after implantation of drug-eluting stents: the Efficacy of Xience/Promus Versus Cypher to Reduce Late Loss After Stenting (EXCELLENT) randomized, multicenter study. *Circulation* 2012; 125: 505-513.
12. Feres F, Costa RA, Abizaid A, Leon MB, Marin-Neto JA, Botelho RV, King SB, 3rd, Negoita M, Liu M, de Paula JE, Mangione JA, Meireles GX, Castello HJ, Jr., Nicolela EL, Jr., Perin MA, Devito FS, Labrunie A, Salvadori D, Jr., Gusmao M, Staico R, Costa JR, Jr., de Castro JP, Abizaid AS, Bhatt DL, Investigators OT. Three vs twelve months of dual antiplatelet therapy after zotarolimus-eluting stents: the OPTIMIZE randomized trial. *JAMA* 2013; 310: 2510-2522.
13. Valgimigli M, Campo G, Monti M, Vranckx P, Percoco G, Tumscitz C, Castriota F, Colombo F, Tebaldi M, Fuca G, Kubbajeh M, Cangiano E, Minarelli M, Scalone A, Cavazza C, Frangione A, Borghesi M, Marchesini J, Parrinello G, Ferrari R, Prolonging Dual Antiplatelet Treatment After Grading Stent-Induced Intimal Hyperplasia Study I. Short- versus long-term duration of dual-antiplatelet therapy after coronary stenting: a randomized multicenter trial. *Circulation* 2012; 125: 2015-2026.
14. Kim BK, Hong MK, Shin DH, Nam CM, Kim JS, Ko YG, Choi D, Kang TS, Park BE, Kang WC, Lee SH, Yoon JH, Hong BK, Kwon HM, Jang Y, Investigators R. A new strategy for discontinuation of dual antiplatelet therapy: the RESET Trial (REal Safety and Efficacy of 3-month dual antiplatelet Therapy following Endeavor zotarolimus-eluting stent implantation). *J Am Coll Cardiol* 2012; 60: 1340-1348.
15. Colombo A, Chieffo A, Frasheri A, Garbo R, Masotti-Centol M, Salvatella N, Oteo Dominguez JF, Steffanon L, Tarantini G, Presbitero P, Menozzi A, Pucci E, Mauri J, Cesana BM, Giustino G, Sardella G. Second-generation drug-eluting stent implantation followed by 6- versus 12-month dual antiplatelet therapy: the SECURITY randomized clinical trial. *J Am Coll Cardiol* 2014; 64: 2086-2097.
16. Valgimigli M, Patialiakas A, Thury A, McFadden E, Colangelo S, Campo G, Tebaldi M, Ungi I, Tondi S, Roffi M, Menozzi A, de Cesare N, Garbo R, Meliga E, Testa L, Gabriel HM, Airoldi F, Ferlini M, Liistro F, Dellavalle A, Vranckx P, Briguori C, Investigators Z. Zotarolimus-eluting versus bare-metal stents in uncertain drug-eluting stent candidates. *J Am Coll Cardiol* 2015; 65: 805-815.
17. Tickoo S, Bhardwaj A, Fonarow GC, Liang L, Bhatt DL, Cannon CP. Relation Between Hospital Length of Stay and Quality of Care in Patients With Acute Coronary Syndromes (from the American Heart Association's Get With the Guidelines--Coronary Artery Disease Data Set). *Am J Cardiol* 2016; 117: 201-205.
18. Baber U, Mehran R, Giustino G, Cohen DJ, Henry TD, Sartori S, Ariti C, Litherland C, Dangas G, Gibson CM, Krucoff MW, Moliterno DJ, Kirtane AJ, Stone GW, Colombo A, Chieffo A, Kini AS, Witzenbichler B, Weisz G, Steg PG, Pocock S. Coronary Thrombosis and Major Bleeding After PCI With Drug-Eluting Stents: Risk Scores From PARIS. *J Am Coll Cardiol* 2016.
19. Steyerberg E. *Clinical prediction models.* Springer, 2008.
20. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982; 247: 2543-2546.
21. van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. *BMC Med Res Methodol* 2014; 14: 5.
22. Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. *Stat Med* 2015; 34: 685-703.
23. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; 162: W1-73.

24. Amin AP, Bachuwar A, Reid KJ, Chhatriwalla AK, Salisbury AC, Yeh RW, Kosiborod M, Wang TY, Alexander KP, Gosch K, Cohen DJ, Spertus JA, Bach RG. Nuisance bleeding with prolonged dual antiplatelet therapy after acute myocardial infarction and its impact on health status. *J Am Coll Cardiol* 2013; 61: 2130-2138.
25. Genereux P, Giustino G, Witzenbichler B, Weisz G, Stuckey TD, Rinaldi MJ, Neumann FJ, Metzger DC, Henry TD, Cox DA, Duffy PL, Mazzaferri E, Yadav M, Francese DP, Palmerini T, Kirtane AJ, Litherland C, Mehran R, Stone GW. Incidence, Predictors, and Impact of Post-Discharge Bleeding After Percutaneous Coronary Intervention. *J Am Coll Cardiol* 2015; 66: 1036-1045.
26. Gurm HS, Bhatt DL, Lincoff AM, Tcheng JE, Kereiakes DJ, Kleiman NS, Jia G, Topol EJ. Impact of preprocedural white blood cell count on long term mortality after percutaneous coronary intervention: insights from the EPIC, EPILOG, and EPISTENT trials. *Heart* 2003; 89: 1200-1204.
27. Valgimigli M, Costa F, Byrne R, Haude M, Baumbach A, Windecker S. Dual antiplatelet therapy duration after coronary stenting in clinical practice: results of an EAPCI survey. *EuroIntervention* 2015; 11: 68-74.
28. Koskinas KC, Raber L, Zanchin T, Wenaweser P, Stortecky S, Moschovitis A, Khattab AA, Pilgrim T, Blochlinger S, Moro C, Juni P, Meier B, Heg D, Windecker S. Clinical impact of gastrointestinal bleeding in patients undergoing percutaneous coronary interventions. *Circ Cardiovasc Interv* 2015; 8.
29. Yeh RW, Secemsky EA, Kereiakes DJ, Normand SL, Gershlick AH, Cohen DJ, Spertus JA, Steg PG, Cutlip DE, Rinaldi MJ, Camenzind E, Wijns W, Apruzzese PK, Song Y, Massaro JM, Mauri L, Investigators DS. Development and Validation of a Prediction Rule for Benefit and Harm of Dual Antiplatelet Therapy Beyond 1 Year After Percutaneous Coronary Intervention. *JAMA* 2016; 315: 1735-1749.
30. Hijazi Z, Oldgren J, Lindback J, Alexander JH, Connolly SJ, Eikelboom JW, Ezekowitz MD, Held C, Hylek EM, Lopes RD, Siegbahn A, Yusuf S, Granger CB, Wallentin L, Aristotle, Investigators R-L. The novel biomarker-based ABC (age, biomarkers, clinical history)-bleeding risk score for patients with atrial fibrillation: a derivation and validation study. *Lancet* 2016.

APPENDIX 15.1

Pooled Datasets included in the PRECISE-DAPT collaborative study

The PRECISE-DAPT dataset included data pooled at an individual patient level from 8 contemporary multicentre randomized clinical trials: Sirolimus-eluting Stents With Biodegradable Polymer Versus an Everolimus-eluting Stents (BIOSCIENCE) (NCT01443104)[1]; Comparison of Biomatrix Versus Gazelle in ST-Elevation Myocardial Infarction (COMFORTABLE AMI) (NCT00962416)[2]; Efficacy of Xience/Promus Versus Cypher to Reduce Late Loss After Stenting (EXCELLENT) (NCT00698607)[3]; Optimized Duration of Clopidogrel Therapy Following Treatment With the Endeavor - Optimize Trial (OPTIMIZE) (NCT01113372)[4]; Prolonging Dual-Antiplatelet Treatment After Grading Stent-Induced Intimal Hyperplasia (PRODIGY) (NCT00611286)[5]; Real Safety and Efficacy of 3-Month Dual Antiplatelet Therapy Following Zotarolimus-Eluting Stent Implantation (RESET)(NCT01145079)[6]; Second Generation Drug-Eluting Stents Implantation Followed by Six Versus Twelve-Month - Dual Antiplatelet Therapy (SECURITY) (NCT00944333)[7]; Zotarolimus-eluting Endeavor Sprint Stent in Uncertain DES Candidates (NCT01385319)[8].

Thrombosis in Myocardial Infarction (TIMI) bleeding definition

Thrombosis in Myocardial Infarction (TIMI) bleeding were categorized into Major and Minor [9]. TIMI major bleeding are defined as any intracranial bleeding (excluding micro-haemorrhages <10 mm evident only on gradient-echo MRI), clinically overt signs of haemorrhage associated with a drop in haemoglobin of ≥5 g/dL, or the occurrence of a fatal bleeding (bleeding that directly results in death within 7 days). TIMI Minor bleeding was defined as clinically overt bleeding (including imaging), resulting in haemoglobin drop of 3 to <5 g/dL.

Clinical variables and bleeding predictors definitions

Prior bleeding was defined as a history of previous clinically significant bleeding requiring medical attention. Laboratory data have been considered as values at baseline or in any case in close proximity with the interventional procedure. Age was considered in years at the time of enrolment. White blood cell count was considered as 10^3 units/ µL and haemoglobin in g/dL. Creatinine clearance was considered as ml/min and was calculated using the Cockrouft-Gault formula in the PRODIGY, RESET, EXCELLENT, ZEUS, OPTIMIZE and SECURITY trials, whereas the Modification of Diet in Renal Diseases (MDRD) formula was used in BIOSCIENCE and COMFORTABLE-AMI trials. Clinical events committees within each trial independently and blinded adjudicated all adverse events.

Score validation and comparison in the Validation cohort

External validation was conducted in a population of 13,016 patients from the PLATelet inhibition and patient Outcomes (PLATO) trial which were not treated with coronary artery bypass graft neither during index hospitalization or afterwards. The PRECISE-DAPT score was calculated and assigned to each patient in the validation cohort in a similar manner as in the derivation cohort with the exception of a history of prior spontaneous bleeding requiring medical attention, which was not totally available in the PLATO dataset, and was instead limited to a history of prior gastro-intestinal bleeding. Age was considered in years at the time on enrolment. White blood cell count was considered as 10^3 units/μL and haemoglobin in g/dL. Creatinine clearance was calculated using the Cockrouft-Gault formula and considered as ml/min. The PARIS bleeding risk score was also calculated in the PLATO validation cohort [10]. This is a bleeding prediction tool for patients treated with dual antiplatelet therapy including clinical variables (age, body mass index, current smoking status at the time of enrolment) and laboratory variables (presence of anemia defined as a haemoglobin level of less than 12g/dL in men and 11g/dL in women, creatinine clearance defined as a renal filtrate of less than 60ml/dL) and a treatment with triple therapy (i.e. aspirin plus $P2Y_{12}$ inhibitor plus oral anticoagulant) at discharge. At difference with the original report of the PARIS score, a treatment with oral anticoagulants at discharge and in general configuring triple therapy as previously defined, was rare in the PLATO trial, which excluded patients treated with oral anticoagulants at the time of randomization.

Detailed description of the statistical analysis

Baseline clinical and procedural characteristics were compared by the presence or absence of TIMI major or minor bleeding using Student t tests and χ^2 tests for continuous and categorical variables, respectively. We studied the associations between possible predictors and TIMI major or minor bleeding from day 7 onwards with a Cox regression analysis, stratified by trial. Potential predictors of significant bleeding were first selected at univariable analysis ($p<0.10$) [11]. To make efficient use of the available data we used an advanced multiple imputation of missing values strategy (5 imputations) [12]. Independent bleeding predictors were selected with multivariable backward selection ($p<0.10$). Linear predictor values (i.e. the sum of truncated predictor values times their predictor effects) were scaled and rounded to a score with integer values between 0 and 100. To translate score values into probabilities of two different bleeding outcomes (TIMI bleeding while on DAPT treatment between day 7 and day 365; and TIMI major bleeding while on DAPT treatment between day 7 and day 365), the association between the score and these outcomes was quantified with Cox regression analysis [11]. The resulting cumulative baseline hazard together with the predictive effect per point of the bleeding risk score were used to plot the absolute probability of each of the two bleeding outcomes against the

bleeding risk score. Discrimination of the bleeding risk score was assessed by trial-specific Harrell's c-indexes, which were pooled with a random effects meta-analysis [13, 14]. Since DAPT duration was not equal in all patients included in our study, and in order to focus on event discrimination specifically during DAPT treatment, we evaluated the performance of the score censoring patients' follow-up time at the end of the intended treatment with DAPT and excluding patients who were not treated with DAPT at discharge (1.7%). The ability to separate high bleeding risk patients from lower bleeding risk patients was also visualized by Kaplan Meier cumulative bleeding incidence curves in quartiles of the bleeding risk score. Calibration was assessed by comparing predicted bleeding probabilities with 1-year Kaplan Meier bleeding incidence estimates by using quartiles of the risk score. Furthermore, both discrimination and calibration of the bleeding risk score were assessed in the external PLATO database. Since DAPT duration was not randomized in the PLATO trial, no censoring for follow-up time at the end of DAPT treatment was necessary in the validation cohort, and all patients have been considered as on-intended-treatment with DAPT during study follow-up.
We compared the discriminative ability among bleeding risk scores (i.e. PRECISE-DAPT vs. PARIS) in the external validation cohort using the recently introduced R library CompareC, comparing the two correlated C indices with right-censored survival outcome as suggested by Kang et al. [15].

We tested the impact of a randomized DAPT duration among the PRECISE-DAPT bleeding risk strata in a large sub-group of patients where DAPT duration was randomized (n= 10,081). Randomization scheme was similar within each of the study appraised, and included a shorter treatment duration (i.e. three or six months) versus a standard treatment with 12 months of DAPT or 24 months in one case (i.e. PRODIGY trial). In this analysis events were counted in the two study arms after treatment divergence occurred within each trial included. For instance, if in one of the trials included, DAPT was randomized to 3 vs. 12 months, events occurring in the first three months, where treatment was identical in the two study arms, were censored. The absolute risk difference was calculated to evaluate the difference in bleeding after a long vs. short DAPT duration in each bleeding strata according to the PRECISE-DAPT score. The 95% confidence interval for the absolute risk difference was calculated according to Newcombe & Altman. Finally, interaction testing was performed to determine the effect of DAPT duration among bleeding risk strata assigned by the bleeding risk score. The analyses were done in accordance with the TRIPOD statement [16]. All analyses were performed with R package version 3.6-3 (2013).

Supplementary table 15.1 Population type, design and dual antiplatelet treatment in the original trials forming the PRECISE-DAPT pooled dataset.

Trial Name (Year)	N	Inclusion		Median Follow-up (days)	Population Type	Randomization	DAPT Duration* (months)	DAPT Adherence (%)
		Start	End					
BIOSCIENCE (2014)	2,119	2009	2014	731	All PCI	Biodegradable polymer stent vs. durable polymer stent	12	>80
COMFORTABLE AMI (2012)	1,157	2009	2011	737	STEMI	Biolimus eluting stent vs. BMS	12	≅90
EXCELLENT (2012)	1,443	2008	2009	367	All PCI	6 months vs. 12 months DAPT after stenting	6-12	≅80
OPTIMIZE (2013)	3,119	2010	2012	555	All PCI	3 months vs. 12 months DAPT after stenting with ZES	3-12	≅99
PRODIGY (2012)	2,003	2007	2009	720	All PCI	6 months vs. 24 months DAPT after stenting	6-24	≅97
RESET (2012)	2,117	2009	2010	374	All PCI	6 months vs. 12 months DAPT after stenting	3-12	N.A.
SECURITY (2014)	1,399	2009	2014	720	SCAD	3 months vs. 12 months DAPT after stenting	6-12	≅97
ZEUS (2015)	1,606	2010	2012	360	All PCI	ZES vs. BMS with tailored DAPT	1-12	≅95

PCI= Percutaneous Coronary Intervention. STEMI= ST-segment elevated myocardial infarction. BMS= Bare Metal Stent. DAPT= Dual Antiplatelet Therapy. ZES= Zotarolimus Eluting Stent. N.A.= Not Available.

* As intended in the study protocol

Supplementary table 15.2 Bleeding events stratified by clinical trial of origin.

Trial Name	N	TIMI Major or Minor Bleeding	Median Time to Event (IQR)	TIMI Major Bleeding	Median Time to Event (IQR)
BIOSCIENCE	2,099	46 (2.2%)	166 (62-398)	33 (1.6%)	170 (65-431)
COMFORTABLE AMI	1,125	23 (2.0%)	183 (62-393)	17 (1.5%)	183 (76-397)
EXCELLENT	1,439	11 (0.8%)	81 (7-280)	4 (0.3%)	236 (56-277)
OPTIMIZE	3,088	53 (1.7%)	69 (32-164)	17 (0.6%)	98 (34-147)
PRODIGY	1,981	43 (2.2%)	333 (115-483)	23 (1.2%)	142 (65-521)
RESET	2,112	12 (0.6%)	162 (71-313)	6 (0.3%)	179 (143-278)
SECURITY	1,394	15 (1.1%)	144 (57-296)	12 (0.9%)	100 (46-271)
ZEUS	1,568	15 (1.0%)	242 (68-294)	12 (0.8%)	223 (41-284)

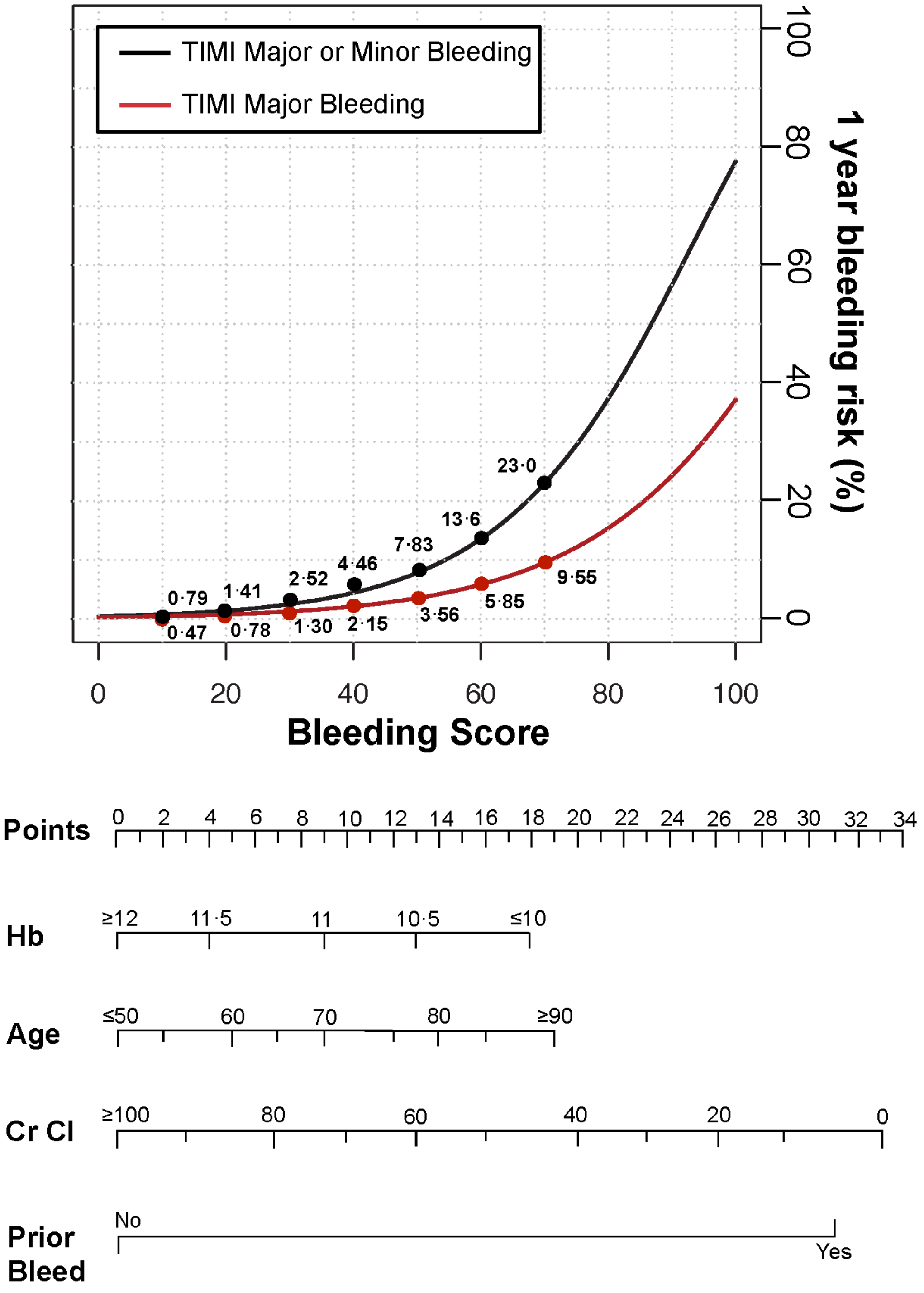
Supplementary table 15.3 Harrell's c-index (95% confidence interval) of the bleeding risk model and bleeding risk score stratified by clinical trial of origin.

Trial Name	TIMI Major/Minor On-Treatment		TIMI Major On-Treatment	
	Model	Score	Model	Score
BIOSCIENCE	0.66 (0.54-0.78)	0.66 (0.54-0.77)	0.62 (0.46-0.78)	0.62 (0.47-0.77)
COMFORTABLE AMI	0.69 (0.53-0.84)	0.69 (0.57-0.81)	0.60 (0.42-0.78)	0.60 (0.47-0.74)
EXCELLENT	0.77 (0.60-0.93)	0.75 (0.58-0.92)	0.83 (0.57-1.00)	0.82 (0.51-1.00)
OPTIMIZE	0.68 (0.55-0.82)	0.68 (0.55-0.81)	0.65 (0.47-0.84)	0.64 (0.46-0.83)
PRODIGY	0.78 (0.69-0.87)	0.78 (0.69-0.86)	0.76 (0.66-0.85)	0.75 (0.66-0.84)
RESET	0.55 (0.25-0.85)	0.55 (0.25-0.85)	0.51 (0.15-0.86)	0.50 (0.14-0.87)
SECURITY	0.64 (0.43-0.85)	0.63 (0.41-0.84)	0.69 (0.48-0.90)	0.68 (0.46-0.90)
ZEUS	0.85 (0.81-1.00)	0.86 (0.82-1.00)	0.85 (0.81-1.00)	0.86 (0.82-1.00)
OVERALL	0.73 (0.62-0.85)	0.73 (0.61-0.85)	0.72 (0.58-0.85)	0.71 (0.57-0.85)

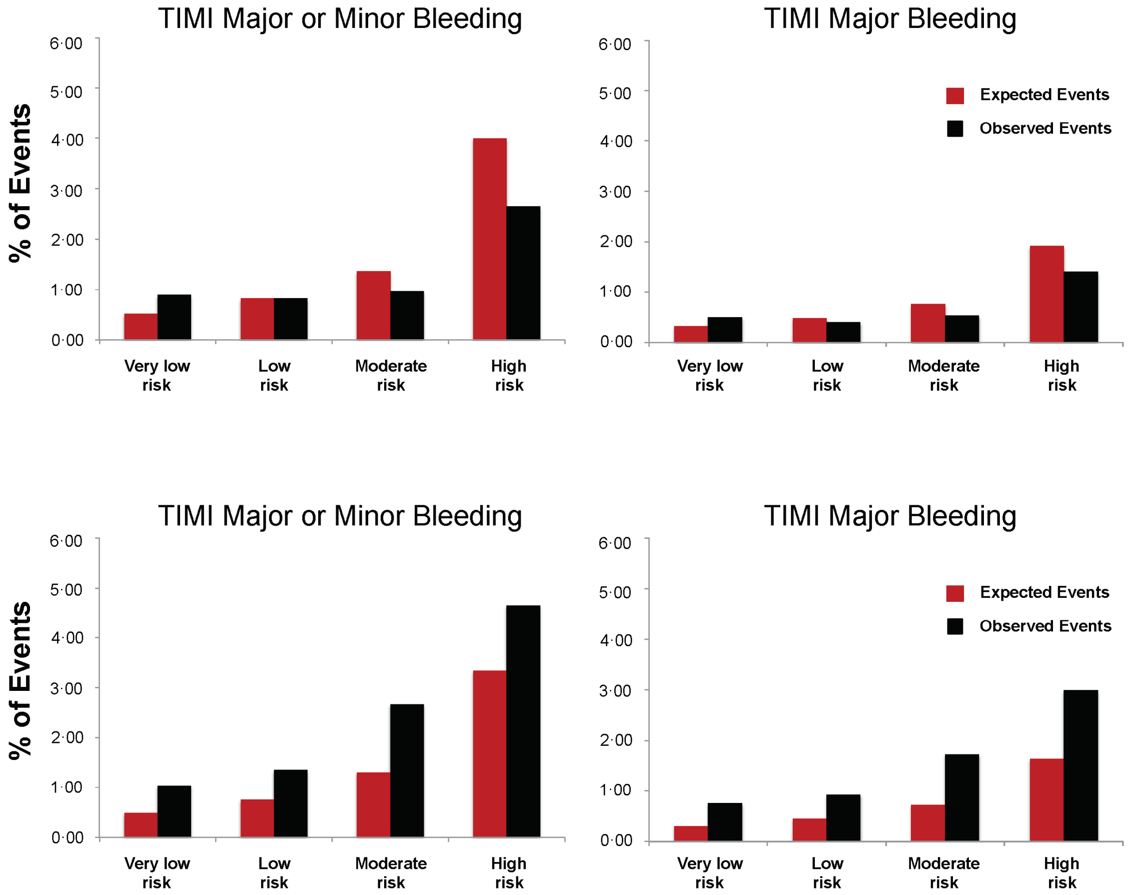
Supplementary table 15.4 Alternative multivariable model for out-of-hospital TIMI major or minor bleeding after excluding white blood cell count.

	Hazard Ratio (95% CI)	χ^2	p
Age (ten years)	1.23 (1.01 - 1.59)	6.3	0.012
Prior bleeding	4.13 (1.19 - 14.37)	5.0	0.026
Haemoglobin at baseline (one g/dL)	0.67 (0.53 - 0.84)	11.9	0.001
Creatinine Clearance (ten ml/min)	0.90 (0.82 - 0.99)	10.3	0.001

Age was truncated below 50 years; haemoglobin at baseline was truncated above 12 and below 10 g/dl; creatinine clearance was truncated above 100 ml/min.



Supplementary figure 15.1 Alternative version of the PRECISE-DAPT score excluding white blood cell count from the calculation. The figure refers to out-of-hospital TIMI major or minor and TIMI major bleeding at 12 months, while on-treatment with DAPT.



Supplementary figure 15.2 Calibration plot for the alternative version of the PRECISE-DAPT score excluding white blood cell count from the calculation, in the derivation (upper panel) and validation (lower panel) study cohort for TIMI major or minor bleeding and TIMI major bleeding occurring while on-treatment with DAPT.

SUPPLEMENTARY REFERENCES

1. Pilgrim T, Heg D, Roffi M, et al. Ultrathin strut biodegradable polymer sirolimus-eluting stent versus durable polymer everolimus-eluting stent for percutaneous coronary revascularisation (BIOSCIENCE): a randomised, single-blind, non-inferiority trial. *Lancet* 2014; 384(9960): 2111-22.
2. Raber L, Kelbaek H, Ostojic M, et al. Effect of biolimus-eluting stents with biodegradable polymer vs bare-metal stents on cardiovascular events among patients with acute myocardial infarction: the COMFORTABLE AMI randomized trial. *Jama* 2012; 308(8): 777-87.
3. Gwon HC, Hahn JY, Park KW, et al. Six-month versus 12-month dual antiplatelet therapy after implantation of drug-eluting stents: the Efficacy of Xience/Promus Versus Cypher to Reduce Late Loss After Stenting (EXCELLENT) randomized, multicenter study. *Circulation* 2012; 125(3): 505-13.
4. Feres F, Costa RA, Abizaid A, et al. Three vs twelve months of dual antiplatelet therapy after zotarolimus-eluting stents: the OPTIMIZE randomized trial. *Jama* 2013; 310(23): 2510-22.
5. Valgimigli M, Campo G, Monti M, et al. Short- versus long-term duration of dual-antiplatelet therapy after coronary stenting: a randomized multicenter trial. *Circulation* 2012; 125(16): 2015-26.
6. Kim BK, Hong MK, Shin DH, et al. A new strategy for discontinuation of dual antiplatelet therapy: the RESET Trial (REal Safety and Efficacy of 3-month dual antiplatelet Therapy following Endeavor zotarolimus-eluting stent implantation). *Journal of the American College of Cardiology* 2012; 60(15): 1340-8.
7. Colombo A, Chieffo A, Frasheri A, et al. Second-generation drug-eluting stent implantation followed by 6- versus 12-month dual antiplatelet therapy: the SECURITY randomized clinical trial. *Journal of the American College of Cardiology* 2014; 64(20): 2086-97.
8. Valgimigli M, Patialiakas A, Thury A, et al. Zotarolimus-eluting versus bare-metal stents in uncertain drug-eluting stent candidates. *Journal of the American College of Cardiology* 2015; 65(8): 805-15.
9. Chesebro JH, Knatterud G, Roberts R, et al. Thrombolysis in Myocardial Infarction (TIMI) Trial, Phase I: A comparison between intravenous tissue plasminogen activator and intravenous streptokinase. Clinical findings through hospital discharge. *Circulation* 1987; 76(1): 142-54.
10. Baber U, Mehran R, Giustino G, et al. Coronary Thrombosis and Major Bleeding After PCI With Drug-Eluting Stents: Risk Scores From PARIS. *Journal of the American College of Cardiology* 2016.
11. Steyerberg E. Clinical prediction models: Springer; 2008.
12. van Buuren G-O. MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 2011; 45(3): 1-67.
13. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *Jama* 1982; 247(18): 2543-6.
14. van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. *BMC medical research methodology* 2014; 14: 5.
15. Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. *Statistics in medicine* 2015; 34(4): 685-703.
16. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine* 2015; 162(1): W1-73.

16

Prediction of survival in patients with esophageal or junctional cancer: impact of neoadjuvant chemoradiotherapy on conventional prognostic factors

J Shapiro
D van Klaveren
SM Lagarde
ELA Toxopeus
A van der Gaast
MCCM Hulshof
BPL Wijnhoven
MI van Berge Henegouwen
EW Steyerberg
JJB van Lanschot

British Journal of Surgery 2016

ABSTRACT

Background The value of conventional prognostic factors is unclear in the era of multimodality treatment for esophageal cancer. This study aimed to quantify the impact of neoadjuvant chemoradiotherapy (nCRT) on well-established prognostic factors and to develop and validate a prognostic model.

Methods Patients treated with surgery alone or with nCRT plus surgery were included. Multivariable Cox modeling was used to identify prognostic factors for overall survival, with treatment included as an interaction. A prediction model for individual survival was developed using stepwise backward selection in nCRT plus surgery patients. The model was internally and cross-validated and a nomogram was designed for use in clinical practice.

Results In total, 1017 patients were included, 391 in the surgery alone group and 626 in the nCRT plus surgery group. Independent prognostic factors in the surgery alone group were age, tumor histology, surgical approach, radicality, pT-stage and pN-stage. Whereas, in the nCRT plus surgery group, only cN-stage and pN-stage remained. Tumor histology, surgical approach and pT-stage were significantly less prognostic, while cN-stage was significantly more prognostic in patients treated with nCRT plus surgery. The final prognostic model included cN-stage, pT-stage and pN-stage and had moderate discrimination (c-index at internal validation 0.63).

Conclusion In nCRT plus surgery patients, only pretreatment cN-stage and posttreatment pN-stage remain as independent prognostic factors. The final prediction model, based on cN-stage, pT-stage and pN-stage, has moderate discriminatory ability. These results strengthen the need for new prognostic factors to improve survival prediction in the era of multimodality treatment for esophageal cancer.

INTRODUCTION

In patients with cancer of the esophagus or esophagogastric junction, several pretreatment prognostic factors for long-term survival after primary surgical resection have been identified. These include age [1,2], gender [2-5], weight loss [6,7], histological tumor subtype [8,9], tumor location [10-12], tumor length [13-15] and clinical TNM-stage [16,17]. Well-established prognostic factors which become available after esophagectomy include surgical approach [18,19], surgical radicality [20-22], tumor grade [8,23] and pathological TNM-stage [12,24]. Prediction models have been developed to predict overall survival in individual patients, based on these prognostic factors [25,26]. However, most of these well-established prognostic factors have been identified and validated in the era of primary surgical resection.

Recent studies show that the addition of neoadjuvant chemoradiotherapy (nCRT) to surgery substantially improves locoregional control and long-term survival as compared to surgery alone [27,28]. In many countries, nCRT plus surgery is now standard of care for these patients. However, the value of conventional prognostic factors and the accuracy of models for individual survival prediction are still unclear in the context of current multimodality treatment strategies and have not yet been investigated in a large patient cohort.

We aim (I) to quantify the impact of nCRT on several well-established prognostic factors, and (II) to develop and validate a prognostic model in patients treated with nCRT plus surgery for esophageal or junctional cancer.

METHODS

Patient selection

Patients were included, who were treated with surgery alone as standard of care (pre-CROSS, 1993–2001), with nCRT plus surgery as part of the single-center non-randomized CROSS-I trial [29] (2001–2004), with surgery alone or nCRT plus surgery as part of the multicenter randomized controlled CROSS-II trial [28] (2004–2009), or with nCRT plus surgery as standard of care at the Erasmus MC, Rotterdam, The Netherlands or at the Academic Medical Center, Amsterdam, The Netherlands (post-CROSS, 2009–2013). Both squamous cell carcinoma (SCC) and adenocarcinoma (AC) histologies were included. Patients who did not receive at least 80% of the planned dose of chemoradiotherapy, who received a nCRT regimen other than CROSS or in whom surgical resection was not completed were excluded. Inclusion criteria of the randomized CROSS-II trial [28] were retrospectively applied to patients treated with surgery alone as part of standard of care, *i.e.* only those patients were included, who underwent the complete staging protocol and who had locally advanced disease (cT2-T4a or cT1N+).

Clinical staging

In all patients, pretreatment staging included endoscopy with biopsy, endoscopic ultrasonography (with fine needle aspiration [FNA] when indicated), CT scan of the neck, chest and abdomen and external ultrasonography of the neck (with FNA when indicated). PET scans were not routinely performed during this study period but were performed in some patients when available and when conventional EUS and CT showed signs of extensive lymph node involvement (in order to get further assurance of absence of distant dissemination). Tumor location and tumor length were determined by pretreatment endoscopy. Clinical T-stage and N-stage were determined by endoscopic ultrasonography and CT-scanning with or without FDG-PET-scanning according to the Union for International Cancer Control (UICC) TNM Cancer Staging, 6th edition [30]. Many of the earlier patients were staged prior to the introduction of TNM Cancer Staging, 7th edition [31], whereby the exact number of clinically suspected nodes was not recorded and could not be determined retrospectively (especially for EUS). Therefore, clinical nodal staging could only be determined for all patients according to UICC TNM Cancer Staging, 6th edition [30] (i.e. suspected nodal involvement [cN1] versus no suspected nodal involvement [cN0]).

Neoadjuvant and surgical treatment

Neoadjuvant chemoradiotherapy (nCRT) was given according to the CROSS regimen [28,29]. For carcinomas at or above the level of the carina a transthoracic esophageal resection (TTE) with a two-field lymph node dissection was performed. For carcinomas located well below the level of the carina, either TTE with two-field lymph node dissection or a transhiatal esophageal resection (THE) was performed depending on fitness of the patient and preference of the surgeon. For carcinomas involving the esophagogastric junction, THE was the preferred technique. In both the transthoracic and the transhiatal approach, an upper abdominal lymphadenectomy was performed, including resection of nodes along the hepatic artery, splenic artery and origin of the left gastric artery. Open- as well as minimally invasive techniques were used.

Pathological assessment

All histopathological parameters were prospectively collected. Tumor histology was determined based on the pretreatment biopsy, while tumor grade was determined in the resection specimen only. In the absence of residual tumor in the resection specimen, tumor grade was scored as ‘non determinable’ (Gx). A microscopically radical resection (R_0) was defined as a tumor-free resection margin ≥1 mm. R_1 was defined as a macroscopically radical resection, with a microscopically tumor-free resection margin <1 mm. Pathological T-stage and N-stage were (re)scored according to the UICC TNM Cancer Staging, 7th edition [31].

The tumor regression grade (TRG) was scored using the system as reported by Chirieac *et al.* [32,33].

Follow-up and data collection

Clinical and surgical characteristics were collected from prospectively maintained institutional databases. Overall survival was assessed based on all-cause mortality and was determined using hospital records and municipal registers. Overall survival was limited at five years to reduce the effect of death by other causes.

Data analysis

Data were described as medians with an interquartile range in case of continuous variables and frequencies with percentages in case of categorical variables. Grouped data were compared using Student's t-test and Pearson's chi-squared test. An advanced multiple imputation approach was used to impute missing data, resulting in five separate datasets [34]. Categories with less than 20 cases were combined with related categories. Weight loss was truncated at 10 kilograms (95th percentile). During imputation pT0 was set to combine with TRG1 and Gx. In the total patient cohort, hazard ratios (HRs), with corresponding 95% confidence intervals (CIs) were calculated using a multivariable Cox proportional-hazards model, with treatment (*i.e.* surgery alone or nCRT plus surgery) included as an interaction in the analysis. The following characteristics were included: age, gender, weight loss, tumor histology, tumor location, tumor length, clinical T-stage (cT), clinical N-stage (cN), surgical approach, radicality of resection, tumor grade in the resection specimen, pathological T-stage (pT), pathological N-stage (pN) and tumor regression grade (TRG). Overall survival was calculated from the end of therapy (day of surgery in both treatment groups) until death (from any cause) or end of follow-up. Differences in prognostic impact of the various characteristics were quantified by including statistical interaction terms for the two treatment groups ('treatment interaction'), defined as the HR associated with a characteristic among patients undergoing nCRT plus surgery divided by the HR for the same characteristic among patients undergoing surgery alone (*i.e.* HR_{nCRT+S}/HR_S). Significance was set to $p<0.05$.

Development, validation and visualization of prognostic model

A prognostic model was developed in the nCRT plus surgery group, using stepwise backward selection, where variables were excluded from the model in a stepwise manner, testing for the significance of elimination per variable [35], until no further improvement was achieved. The prognostic model was internally validated by correcting for optimism and cross-validated by dividing the total nCRT plus surgery cohort into a cohort with patients from the

Erasmus MC, Rotterdam, and a cohort with patients from all other centers. Where the prognostic was developed in one cohort and validated in the other, and *vice versa*.

Table 16.1 Clinical, surgical and histopathological characteristics in 1017 patients with potentially curable carcinoma of the esophagus or esophagogastric junction, treated with surgery alone or neoadjuvant chemoradiotherapy (nCRT) plus surgery.

	Total		**Surgery alone**		**nCRT plus surgery**		
	(n= 1017)		**(n= 391)**		**(n= 626)**		
	1993 – 2013		**1993 – 2009**		**2001 – 2013**		
	n	**(%)***	**n**	**(%)***	**n**	**(%)***	**p ****
Age [years]							0.275
median (p25 – p75)	63 (56 – 69)		63 (55 – 69)		63 (56 – 69)		
Gender							0.578
Female	212	(21)	78	(20)	129	(22)	
Male	805	(79)	313	(80)	451	(78)	
Weight loss [kilograms]							**0.001**
median (p25 – p75)	3 (0 – 6)		2 (0 – 5)		3 (0 – 6)		
missing (%)	42	(4)	23	(6)	19	(3)	
Tumor histology€							0.866
Squamous cell carcinoma	224	(22)	85	(22)	139	(22)	
Adenocarcinoma	783	(77)	302	(77)	481	(78)	
Undeterminable	10	(1)	4	(1)	6	(1)	
Tumor location§							**0.045**
Cervical	1	(0)	1	(0)	–	–	
Upper third esophagus	12	(1)	9	(2)	3	(1)	
Middle third esophagus	129	(13)	48	(13)	81	(13)	
Lower third esophagus	653	(65)	240	(63)	413	(66)	
Esophagogastric junction	211	(21)	86	(22)	125	(20)	
missing	11		7		4		
Tumor length§ [centimeters]							
median (p25 – p75)	5 (3 – 6)		4 (3 – 6)		5 (3 – 6)		0.262
missing (%)	85	(9)	38	(10)	47	(8)	
cT-stage‡							0.739
cT1	20	(2)	8	(2)	12	(2)	
cT2	193	(20)	80	(21)	113	(19)	
cT3	757	(77)	281	(75)	476	(78)	
cT4	16	(2)	6	(2)	10	(2)	
missing	31		16		15		
cN-stage‡							**<0.001**
cN0	387	(39)	205	(55)	182	(30)	
cN1	599	(61)	169	(45)	430	(70)	
missing	31		17		14		
Surgical approach							<0.001
Transhiatal approach	487	(48)	263	(67)	224	(36)	
Transthoracic approach	525	(52)	128	(33)	397	(63)	
Other	5	(1)	–	–	5	(1)	
Radicality◊							**<0.001**
R0	851	(84)	262	(67)	589	(94)	
R1	163	(16)	126	(32)	37	(6)	
R2	3	(0)	3	(1)	–	–	

Table 16.1 Continued.

	Total		**Surgery alone**		**nCRT plus surgery**		
	n	**(%)***	**n**	**(%)***	**n**	**(%)***	**p ****
Tumor grade[£]							**<0.001**
Gx (undeterminable)	171	(20)	–	–	171	(36)	
G1	40	(5)	31	(8)	9	(2)	
G2	327	(38)	194	(51)	133	(28)	
G3	313	(37)	155	(41)	158	(34)	
missing	166		11		155		
pT-stage[Δ]							**<0.001**
pT0	187	(19)	–	–	187	(30)	
pT1, includes pTis	128	(13)	39	(10)	89	(14)	
pT2	169	(17)	63	(16)	106	(17)	
pT3	518	(51)	278	(72)	240	(38)	
pT4	9	(1)	5	(1)	4	(1)	
missing	6		6		–		
pN-stage[Δ]							**<0.001**
pN0	523	(52)	123	(32)	400	(64)	
pN1	247	(24)	101	(26)	146	(23)	
pN2	149	(15)	89	(23)	60	(10)	
pN3	93	(9)	73	(19)	20	(3)	
missing	5		5		–		
Tumor regression grade[¥]							–
TRG1	187	(30)	–	–	187	(30)	
TRG2	135	(22)	–	–	135	(22)	
TRG3	175	(28)	–	–	175	(28)	
TRG4	124	(20)	–	–	124	(20)	
missing	5		–		5		

* Data presented as median (interquartile range) or number (%). Percentages may not add up to 100 due to rounding.

** Data were compared between the surgery alone and nCRT plus surgery groups using Student's t-test for continuous variables and Pearson's chi-squared test for categorical variables.

€ Tumor histology was determined in the pretreatment biopsy.

§ Tumor location and tumor length were determined by endoscopy.

‡ Clinical T-stage and N-stage were determined by endoscopic ultrasonography and CT-scanning with or without FDG-PET-scanning according to the Union for International Cancer Control (UICC) TNM Cancer Staging, 6th edition [30]. cT1: (sub)mucosal involvement, cT2: proper muscle layer involvement, cT3: surrounding stroma involvement.

◊ R0 was defined as a tumor-free resection margin ≥ 1 mm. R1 was defined as a macroscopically radical resection, with a microscopically tumor-free resection margin < 1 mm.

£ Tumor grade was determined in the resection specimen only. Histological tumor grade was not determined in the pretreatment biopsy.

Δ Pathological T-stage and N-stage, as measured in the resection specimen were (re)scored according to UICC TNM Cancer Staging, 7th edition [31]; pT1: (sub)mucosal involvement, pT2: proper muscle layer involvement, pT3: surrounding stroma involvement; pN0: no lymph node positivity, pN1: 1-2 lymph nodes positive, pN2: 3-6 lymph nodes positive, pN3: ≥7 lymph nodes positive.

¥ Tumor regression grade was scored as defined by Chirieac *et al.* [32,33] : TRG1: no residual tumor cells found; TRG2: 1-10% residual tumor cells; TRG3: 11-50% residual tumor cells; TRG4: > 50% residual tumor cells.

The model was tested for prognostic accuracy, using Harrell's concordance-index (c-index) [36]. The c-index determines for two randomly chosen subjects the probability that the model predicts a higher risk for the subject with poorer outcome. Analyses were performed using the following R-packages [37]: 'multivariate imputation by chained equations' (mice) [34], 'regression modeling strategies' (rms) [38].

The prognostic strength of individual risk factors in the prognostic model were visualized in a nomogram. The weights for each category within an individual risk factor were calculated by multiplying the original coefficients of the multivariable Cox model with ten and rounding the result to the lowest whole number. The total number of points derived from all predictors was used to calculate the expected one-year and five-year overall survival rates.

RESULTS

Patient, tumor and treatment related characteristics

In total, 1017 patients were included, of whom 391 were treated with surgery alone and 626 were treated with nCRT plus surgery. Median age at diagnosis was 63 years (Table 16.1). Most patients were male (79%), had an adenocarcinoma (77%), most often clinically staged as cT3 (77%). Significant differences were found between surgery alone patients and nCRT plus surgery patients for weight loss (p=0.001), tumor location (p=0.045) and clinical N-stage (p<0.001) (Table 16.1). Also, a transhiatal approach was performed significantly more often in the surgery alone group as compared to the nCRT plus surgery group (p<0.001).

Prognostic factors in patients treated with surgery alone or nCRT plus surgery

In patients treated with surgery alone, independent prognostic factors for overall survival were age (HR per decade= 1.21 95%CI 1.05–1.40 p=0.009), tumor histology (HR SCC vs. AC= 1.93 95%CI 1.36–2.75 p<0.001), surgical approach (HR TTE or other vs. THE= 0.74 95%CI 0.55–0.98 p=0.036), radicality (HR R1-R2 vs. R0= 1.63 95%CI 1.25–2.13 p<0.001), pT-stage (HR pT3-pT4 vs. pT1= 3.35 95%CI 1.64–6.85 p=0.001) and pN-stage (HR pN1 vs. pN0= 2.07 95%CI 1.41–3.04 p<0.001, HR pN2 vs. pN0= 2.99 95%CI 2.01–4.46 p<0.001 and pN3 vs. pN0= 4.57 95%CI 3.01–6.95 p<0.001) (Table 16.2). Whereas, in patients treated with nCRT plus surgery, the only independent prognostic factors were cN-stage (HR cN1 vs. cN0= 1.46 95%CI 1.09–1.95 p=0.012) and pN-stage (HR pN1 vs. pN0= 1.78 95%CI 1.32–2.39 p<0.001, HR pN2 vs. pN0= 1.98 95%CI 1.29–3.02 p<0.001 and pN3 vs. pN0= 4.34 95%CI 2.38–7.93 p<0.001). Specifically, TRG was not prognostic for survival (HR TRG1 vs. TRG2= 0.77 95%CI 0.52–1.12, HR TRG3 vs. TRG2= 1.21 95%CI 0.85–1.72 and TRG4 vs. TRG2= 1.03 95%CI 0.69–1.54). This was also not the case when different groupings of TRG [39-41] were tested (data not shown).

Table 16.2 Prognostic factors for overall survival in 1017 patients with potentially curable carcinoma of the esophagus or esophagogastric junction, treated with surgery alone or neoadjuvant chemoradiotherapy (nCRT) plus surgery.

	Surgery alone			nCRT plus surgery					
	HR	**95% CI**	**p**	**HR**	**95% CI**	**p**	**HR_{nCRT+S} / HR_S**	**95% CI**	**p_{int}**
Age (per 10 years)	1.21	(1.05 – 1.40)	**0.009**	1.12	(0.98 – 1.28)	0.099	**0.93**	**(0.76 – 1.13)**	0.444
Gender									
Female	0.80	(0.57 – 1.13)	0.198	0.77	(0.55 – 1.08)	0.129	0.96	(0.59 – 1.56)	0.879
Male	1 (ref)			1 (ref)			1 (ref)		
Weight loss (per kg)	1.01	(0.97 – 1.05)	0.617	1.01	(0.98 – 1.05)	0.462	1	(0.95 – 1.06)	0.872
Tumor histology									
Squamous cell carcinoma	1.93	(1.36 – 2.75)	**<0.001**	0.77	(0.53 – 1.12)	0.173	0.4	(0.24 – 0.67)	**0.001**
Adenocarcinoma	1 (ref)			1 (ref)			1 (ref)		
Tumor location									
Cervical-to-middle third esophagus	1 (ref)			1 (ref)			1 (ref)		
Lower third esophagus	1.05	(0.69 – 1.61)	0.807	0.93	(0.60 – 1.44)	0.740	0.88	(0.48 – 1.62)	0.681
Esophagogastric junction	0.79	(0.48 – 1.31)	0.359	0.72	(0.42 – 1.23)	0.232	0.92	(0.44 – 1.90)	0.812
Tumor length (per cm)	1.00	(0.95 – 1.05)	0.942	0.97	(0.92 – 1.04)	0.414	0.98	(0.90 – 1.06)	0.579
cT–stage									
cT1 – cT2	1 (ref)			1 (ref)			1 (ref)		
cT3 – cT4	0.90	(0.63 – 1.29)	0.566	1.16	(0.81 – 1.67)	0.415	1.29	(0.78 – 2.14)	0.325
cN–stage									
cN0	1 (ref)			1 (ref)			1 (ref)		
cN1	0.94	(0.71 – 1.24)	0.653	1.46	(1.09 – 1.95)	**0.012**	1.55	(1.04 – 2.31)	**0.030**
Surgical approach									
Transhiatal approach	1 (ref)			1 (ref)			1 (ref)		
Transthoracic approach or other	0.74	(0.55 – 0.98)	**0.036**	1.15	(0.87 – 1.52)	0.333	1.56	(1.04 – 2.33)	**0.030**
Radicality									
R0	1 (ref)			1 (ref)			1 (ref)	–	
R1 – R2	1.63	(1.25 – 2.13)	**<0.001**	1.35	(0.85 – 2.15)	0.202	0.83	(0.48 – 1.41)	0.486

Table 16.2 Continued.

	Surgery alone			nCRT plus surgery					
	HR	**95% CI**	**p**	**HR**	**95% CI**	**p**	**HR_{nCRT+S} / HR_S**	**95% CI**	**p_{int}**
Tumor grade									
Gx				1.94	(0.36 – 8.61)	0.385			
G1	1 (ref)			1 (ref)			1 (ref)		
G2	1.47	(0.72 – 3.00)	0.289	2.78	(0.58 – 13.25)	0.202	1.89	(0.36 – 9.93)	0.454
G3	1.66	(0.80 – 3.46)	0.176	2.80	(0.67 – 11.74)	0.159	1.69	(0.35 – 8.13)	0.514
pT-stage									
pT0				0.76	(0.49 – 1.17)	0.318			
pT1	1 (ref)			1 (ref)			1 (ref)		
pT2	1.87	(0.87 – 4.03)	0.110	1.02	(0.64 – 1.61)	0.945	0.54	(0.22 – 1.33)	0.182
pT3 – pT4	3.35	(1.64 – 6.85)	**0.001**	1.15	(0.76 – 1.73)	0.499	0.34	(0.15 – 0.78)	**0.011**
pN-stage									
pN0	1 (ref)			1 (ref)			1 (ref)		
pN1	2.07	(1.41 – 3.04)	**<0.001**	1.78	(1.32 – 2.39)	**<0.001**	0.86	(0.53 – 1.40)	0.538
pN2	2.99	(2.01 – 4.46)	**<0.001**	1.98	(1.29 – 3.02)	**<0.001**	0.66	(0.37 – 1.18)	0.161
pN3	4.57	(3.01 – 6.95)	**<0.001**	4.34	(2.38 – 7.93)	**<0.001**	0.95	(0.46 – 1.97)	0.890
Tumor regression grade *									
TRG1				0.77	(0.52 – 1.12)	0.173			
TRG2				1 (ref)					
TRG3				1.21	(0.85 – 1.72)	0.302			
TRG4				1.03	(0.69 – 1.54)	0.895			

HR: hazard ratio. CI: confidence interval. HRnCRT+S/HRS: HR associated with a characteristic among patients undergoing nCRT plus surgery divided by the HR for the same characteristic among patients undergoing surgery alone. p_{int}: p-value for the treatment interaction.

A significant difference in prognostic value between the two treatment groups (*i.e.* treatment interaction) was identified for tumor histology (HR_{nCRT+S}/HR_S SCC vs. AC= 0.40 95%CI 0.24–0.67, $p_{interaction}$=0.001), indicating that tumor histology significantly decreased in prognostic value in the nCRT plus surgery group (*i.e.* the HR between subgroups decreased). Significant treatment interaction was also identified for cN-stage (HR_{nCRT+S}/HR_S cN1 vs. cN0= 1.55 95%CI 1.04–2.31, $p_{interaction}$=0.030), surgical approach (HR_{nCRT+S}/HR_S TTE vs. THE= 1.56 95%CI 1.04–2.33, $p_{interaction}$=0.030) and pT-stage (HR_{nCRT+S}/HR_S pT3-pT4 vs. pT1= 0.34 95%CI 0.15–0.78, $p_{interaction}$=0.011). These results indicate that clinical N-stage significantly improved in prognostic value in the nCRT plus surgery group, while surgical approach and pT-stage significantly decreased in prognostic value.

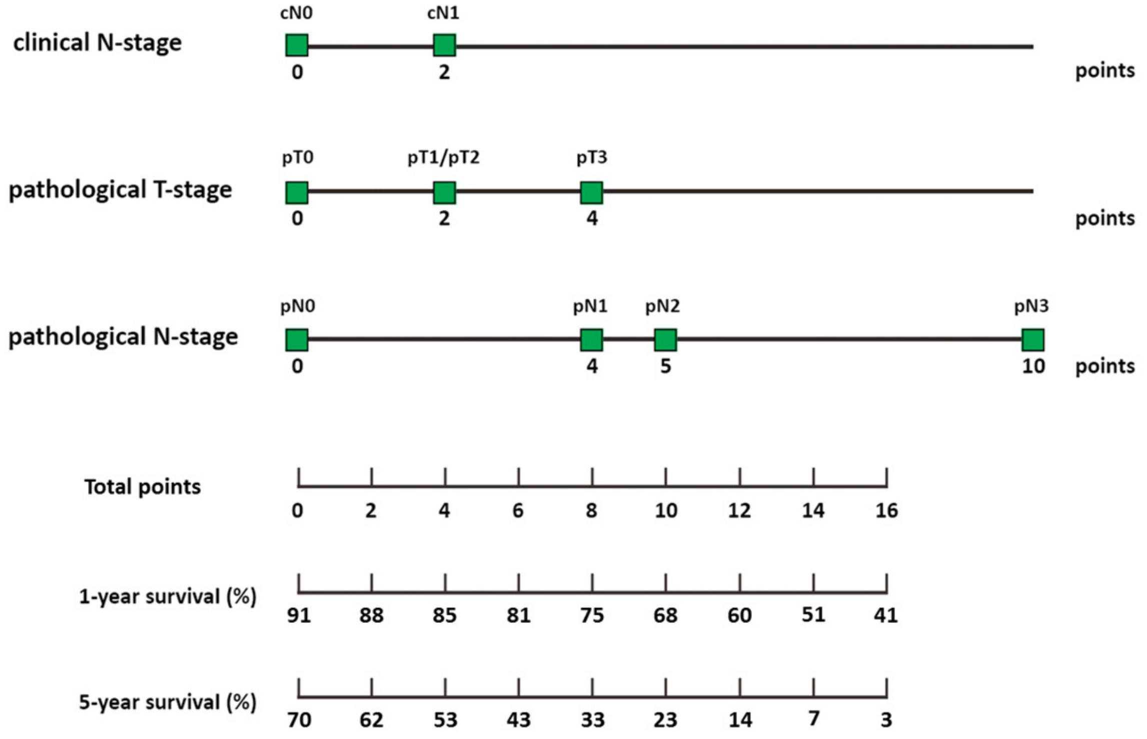


Figure 16.1 Nomogram for overall survival as developed in 626 patients with potentially curable carcinoma of the esophagus or esophagogastric junction, treated with neoadjuvant chemoradiotherapy (nCRT) plus surgery. From the total points axis, a straight line down through the survival axes shows survival probabilities at one-and five years. Clinical N-stage according to UICC TNM Cancer Staging, 6th edition30; cN0: no clinical suspicion of pretreatment lymph node involvement, cN1: clinical suspicion of pretreatment lymph node involvement. Pathological T-stage according to UICC TNM Cancer Staging, 7th edition31; pT0: no residual tumor at the primary tumor site, pT1: (sub)mucosal involvement, pT2: proper muscle layer involvement, pT3: surrounding stroma involvement. Pathological N-stage according to UICC TNM Cancer Staging, 7th edition31; pN0: no lymph node positivity, pN1: 1-2 lymph nodes positive, pN2: 3-6 lymph nodes positive, pN3: ≥7 lymph nodes positive.

Prediction model for survival in patients treated with nCRT plus surgery

After stepwise backward selection, the final prediction model included cN-stage, pT-stage and pN-stage. Discrimination of the prediction model was moderate (c-index at internal validation 0.63). Cross-validation between the Erasmus MC cohort (n= 246) and the other centers (n=380) was comparable (c-index 0.62 and 0.63, resp.). Discrimination of the prediction model was higher in surgery alone patients (c-index 0.66). Finally, a nomogram was constructed (Figure 16.1) to allow for individual one-year and five-year overall survival estimations, based on the three variables included in the final prediction model. As an example, patients with pretreatment suspicion of nodal disease (cN1) and a complete response in the resection specimen (pT0, pN0) would have a total of two points, which corresponds with an estimated one-year and five-year survival rate of 88% and 62%, respectively.

DISCUSSION

In this large and comprehensive study on patients with esophageal or junctional cancer, only clinical N-stage (cN-stage) and pathological N-stage (pN-stage) remained as independent prognostic factors in patients treated with neoadjuvant chemoradiotherapy (nCRT) plus surgery. Pathological T-stage (pT-stage) was added to the final prediction model for these patients after backward selection. Tumor histology, surgical approach and pT-stage were significantly less prognostic, while cN-stage was significantly more prognostic in patients treated with nCRT plus surgery as compared to patients who underwent surgery alone.

By using a single statistical model, which included both treatment groups, specific effects of nCRT on prognostic factors could be determined, independent from the effects of surgery. Results indicate an overall decrease in significance and in number of independent prognostic factors in patients treated with nCRT plus surgery. Interestingly, there was no overlap in significant independent prognostic factors between the two treatment groups, except for pN-stage, confirming previous reports [42,43], and thus underlining the continued significance of pN-stage as an important prognostic factor in the era of multimodality treatment [42-45].

Surprisingly, pretreatment clinical N-stage (cN-stage) increased in prognostic value in patients treated with nCRT plus surgery. In patients treated with surgery alone, cN-stage and pathological N-stage (pN-stage) are different estimations of the same disease state (*i.e.* pretreatment clinical estimation and posttreatment pathological estimation, resp.). Clinical N-stage is known to be relatively inaccurate [46,47] and therefore has little additional prognostic value (on top of pN-stage) in patients treated with surgery alone. However, in patients treated with nCRT plus surgery, cN-stage is no longer necessarily similar to pN-stage. By definition, cN-stage is an estimation of nodal involvement before nCRT, while pN-stage is an estimation of nodal involvement after nCRT. Therefore, cN-stage and pN-stage represent different disease states in these patients and cN-stage, although relatively inaccurate, had additional prognostic value, as was seen in these analyses.

Another important finding is that surgical approach lost prognostic value in patients treated with nCRT plus surgery. In patients treated with surgery alone, a transthoracic approach was associated with a significantly more favorable prognosis as compared to a transhiatal approach, whereas in patients treated with nCRT plus surgery, a transthoracic approach was associated with a non-significantly less favorable prognosis. These findings suggest that in patients treated with nCRT plus surgery, the benefit of a transthoracic approach is at best limited and the necessity for maximization of surgical lymph node retrieval should be questioned. However, only a new randomized trial, comparing these two surgical approaches (with their inherent differences in extent of lymphadenectomy) after neoadjuvant treatment will offer a more definitive answer.

The final prognostic model had moderate discriminatory ability in patients treated with nCRT plus surgery, which is lower than what is generally reported for other tumor types after neoadjuvant treatment [48-50]. Interestingly, the model (although developed in patients who underwent nCRT plus surgery) performed better in patients who underwent surgery alone. This indicates that from a prognostic perspective, neoadjuvant chemoradiotherapy has a strong equalizing effect on patients, making individual survival predictions in the era of nCRT less reliable.

Unfortunately, this study could not identify any additional factors outside of the already well-established TNM-staging system that might contribute to more accurate prognostication in the era of multimodality treatment. Even the much studied and widely applied tumor regression grading (TRG) systems [32,33,39-41] were not significantly associated with survival in this large and homogeneous patient cohort, thus questioning the usefulness of TRG as an independent prognostic factor in esophageal or junctional cancer patients. These results, therefore, strengthen the need for new prognostic factors, such as genetic and molecular markers, to improve the accuracy of individual survival prediction in the era of multimodality treatment for esophageal and junctional cancer.

Limitations

Although this study only included parameters that have been collected prospectively, the time period was relatively long (1993-2013), which might have caused bias in the comparison of patients treated with surgery alone or with nCRT plus surgery despite the selection of all patients according to the same inclusion criteria as applied in the CROSS-I and CROSS-II trials. A further limitation is that PET-CT was applied only in patients with extensive clinical lymph node involvement, in order to get further assurance that there are indeed no signs of distant dissemination. Such a selective use of PET-CT might have introduced a selection bias into our data. Specifically, it might have caused an improved survival of the clinically node-positive patient group due to the additional exclusion of patients with disseminated disease. Another limitation is that the exact number of clinically suspected nodes was not recorded in many of the earlier patients in this cohort (before the introduction of UICC TNM7). Therefore, clinical nodal staging could only be determined for all patients according to UICC TNM6 (i.e. suspected nodal involvement [cN1] versus no suspected nodal involvement [cN0]). This necessary categorization possibly restricts the prognostic significance of pretreatment clinical nodal staging. Furthermore, in this study a microscopically radical resection (R0), was defined as a tumor-free resection margin ≥1 mm (according to the Royal College of Pathologists criteria). However, it is unclear whether our conclusions apply to the also commonly used radicality criteria according to the College of American Pathologists, where an R0 resection is defined as a tumor-free resection margin ≥0 mm [51]. A final limitation is that not all recognized prognostic factors in esophageal and junctional cancer could be included in this study, such as extracapsular lymph node involvement [52-54], signet cell features in esophageal adenocarcinomas [55,56] and genetic and molecular markers [57,58].

CONCLUSION

Most conventional prognostic factors lose their prognostic significance in patients with potentially curable esophageal or junctional cancer, when treated with neoadjuvant chemoradiotherapy (nCRT) plus surgery. In the era of nCRT, clinical N-stage and pathological N-stage remain as independent prognostic factors. Surgical approach, which is of prognostic relevance in patients treated with surgery alone, loses its prognostic significance after nCRT, thus questioning the necessity of maximization of surgical lymph node retrieval. Furthermore, tumor regression grading is not independently associated with survival in patients treated with nCRT plus surgery. The final prediction model, based on clinical N-stage, pathological T-stage and pathological N-stage, has moderate discriminatory ability. These results strengthen the need for new prognostic factors to improve survival prediction in the era of multimodality treatment for esophageal and junctional cancer.

ACKNOWLEDGEMENTS

The authors are grateful to Mrs. Conny M. Vollebregt-Uiterwijk from the Department of Surgery, Erasmus MC—University Medical Center, Rotterdam, for her dedicated prospective data collection. The authors thank Dr. Hester F. Lingsma from the Department of Public Health, Erasmus MC—University Medical Center Rotterdam, and Prof. Patrick M.M. Bossuyt from the Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center Amsterdam, for their assistance in study design, statistical analysis and interpretation of the data. Moreover, the authors acknowledge the prior work by the members of the CROSS study group: Maarten C.C.M. Hulshof, M.D., Ph.D., Dept. of Radiation Oncology, Academic Medical Center, Amsterdam, Pieter van Hagen, M.D., Dept. of Surgery, Erasmus MC - University Medical Center, Rotterdam, Hanneke W. M. van Laarhoven, M.D., Ph.D., Dept. of Medical Oncology, Academic Medical Center, Amsterdam, Grard A.P. Nieuwenhuijzen, M.D., Ph.D., Dept. of Surgery, Catharina Hospital, Eindhoven, Geke A.P. Hospers, M.D., Ph.D., Dept. of Medical Oncology, University Medical Center Groningen, Groningen, Johannes J. Bonenkamp, M.D., Ph.D., Dept. of Surgery, Radboud University Medical Center, Nijmegen, Miguel A. Cuesta, M.D., Ph.D., formerly Dept. of Surgery, VU Medical Center, Amsterdam, Reinoud J.B. Blaisse, M.D., Dept. of Medical Oncology, Rijnstate Hospital, Arnhem, Olivier R.C. Busch, M.D., Ph.D., Dept. of Surgery, Academic Medical Center, Amsterdam, Fiebo J.W. ten Kate, M.D., Ph.D. formerly Dept. of Pathology, Academic Medical Center, Amsterdam, Geert-Jan Creemers, M.D., Ph.D., Dept. of Medical Oncology, Catharina Hospital, Eindhoven, Cornelis J.A. Punt, M.D., Ph.D., Dept. of Medical Oncology, formerly Radboud University Medical Center, Nijmegen, currently Dept. of Medical Oncology, Academic Medical Center, Amsterdam, John T.M. Plukker, M.D., Ph.D., Dept. of Surgery, University Medical Center Groningen, Groningen, Henk M.W. Verheul, M.D., Ph.D., Dept. of Medical Oncology, VU Medical Center, Amsterdam, Ernst J. Spillenaar Bilgen, M.D., Ph.D., Dept. of Surgery, Rijnstate Hospital, Arnhem, Herman van Dekken, M.D., Ph.D., formerly Dept. of Pathology, Erasmus MC - University Medical Center, Rotterdam, Maurice J.C. van der Sangen, M.D., Ph.D., Dept. of Radiation Oncology, Catharina Hospital, Eindhoven, Tom Rozema, M.D., Dept. of Radiation Oncology, Radboud University Medical Center, Nijmegen and Verbeeten Institute Tilburg, Tilburg, Katharina Biermann, M.D., Ph.D. Dept. of Pathology, Erasmus MC - University Medical Center, Rotterdam, Jannet C. Beukema, M.D., Dept. of Radiation Oncology, University Medical Center Groningen, Groningen, Anna H.M. Piet, M.D., Dept. of Radiation Oncology, VU Medical Center, Amsterdam, Caroline M. van Rij, M.D., Dept. of Radiation Oncology, Erasmus MC - University Medical, Center, Rotterdam, Janny G. Reinders, M.D., Arnhem Radiotherapeutic Institute ARTI, Arnhem, Hugo W. Tilanus, M.D., Ph.D. formerly Dept. of Surgery, Erasmus MC - University Medical Center, Rotterdam.

REFERENCES

1. Markar SR, Karthikesalingam A, Thrumurthy S, Ho A, Muallem G, Low DE. Systematic review and pooled analysis assessing the association between elderly age and outcome following surgical resection of esophageal malignancy. Dis Esophagus 2013;26:250-62.
2. Chen MF, Yang YH, Lai CH, Chen PC, Chen WC. Outcome of patients with esophageal cancer: a nationwide analysis. Ann Surg Oncol 2013;20:3023-30.
3. Micheli A, Mariotto A, Giorgi Rossi A, Gatta G, Muti P. The prognostic role of gender in survival of adult cancer patients. EUROCARE Working Group. Eur J Cancer 1998;34:2271-8.
4. Ferri LE, Law S, Wong KH, Kwok KF, Wong J. The influence of technical complications on postoperative outcome and survival after esophagectomy. Ann Surg Oncol 2006;13:557-64.
5. Micheli A, Ciampichini R, Oberaigner W, et al. The advantage of women in cancer survival: an analysis of EUROCARE-4 data. Eur J Cancer 2009;45:1017-27.
6. Polee MB, Hop WC, Kok TC, et al. Prognostic factors for survival in patients with advanced oesophageal cancer treated with cisplatin-based combination chemotherapy. Br J Cancer 2003;89:2045-50.
7. D'Journo XB, Ouattara M, Loundou A, et al. Prognostic impact of weight loss in 1-year survivors after transthoracic esophagectomy for cancer. Dis Esophagus 2012;25:527-34.
8. Khan OA, Alexiou C, Soomro I, Duffy JP, Morgan WE, Beggs FD. Pathological determinants of survival in node-negative oesophageal cancer. Br J Surg 2004;91:1586-91.
9. Gertler R, Stein HJ, Langer R, et al. Long-term outcome of 2920 patients with cancers of the esophagus and esophagogastric junction: evaluation of the New Union Internationale Contre le Cancer/American Joint Cancer Committee staging system. Ann Surg 2011;253:689-98.
10. Christein JD, Hollinger EF, Millikan KW. Prognostic factors associated with resectable carcinoma of the esophagus. Am Surg 2002;68:258-62.
11. Omloo JM, Lagarde SM, Hulscher JB, et al. Extended transthoracic resection compared with limited transhiatal resection for adenocarcinoma of the mid/distal esophagus: five-year survival of a randomized clinical trial. Ann Surg 2007;246:992-1000.
12. Rice TW, Rusch VW, Ishwaran H, Blackstone EH. Cancer of the esophagus and esophagogastric junction: data-driven staging for the seventh edition of the American Joint Committee on Cancer/International Union Against Cancer Cancer Staging Manuals. Cancer 2010;116:3763-73.
13. Eloubeidi MA, Desmond R, Arguedas MR, Reed CE, Wilcox CM. Prognostic factors for the survival of patients with esophageal carcinoma in the U.S.: the importance of tumor length and lymph node status. Cancer 2002;95:1434-43.
14. Bollschweiler E, Baldus SE, Schroder W, Schneider PM, Holscher AH. Staging of esophageal carcinoma: length of tumor and number of involved regional lymph nodes. Are these independent prognostic factors? J Surg Oncol 2006;94:355-63.
15. Griffiths EA, Brummell Z, Gorthi G, Pritchard SA, Welch IM. Tumor length as a prognostic factor in esophageal malignancy: univariate and multivariate survival analyses. J Surg Oncol 2006;93:258-67.
16. Mariette C, Balon JM, Maunoury V, Taillier G, Van Seuningen I, Triboulet JP. Value of endoscopic ultrasonography as a predictor of long-term survival in oesophageal carcinoma. Br J Surg 2003;90:1367-72.
17. Omloo JM, Sloof GW, Boellaard R, et al. Importance of fluorodeoxyglucose-positron emission tomography (FDG-PET) and endoscopic ultrasonography parameters in predicting survival following surgery for esophageal cancer. Endoscopy 2008;40:464-71.
18. Lerut T, De Leyn P, Coosemans W, Van Raemdonck D, Scheys I, LeSaffre E. Surgical strategies in esophageal carcinoma with emphasis on radical lymphadenectomy. Ann Surg 1992;216:583-90.
19. Altorki N, Kent M, Ferrara C, Port J. Three-field lymph node dissection for squamous cell and adenocarcinoma of the esophagus. Ann Surg 2002;236:177-83.

20. Wijnhoven BP, Tran KT, Esterman A, Watson DI, Tilanus HW. An evaluation of prognostic factors and tumor staging of resected carcinoma of the esophagus. Ann Surg 2007;245:717-25.
21. O'Farrell NJ, Donohoe CL, Muldoon C, et al. Lack of independent significance of a close (<1 mm) circumferential resection margin involvement in esophageal and junctional cancer. Ann Surg Oncol 2013;20:2727-33.
22. Wu J, Chen QX, Teng LS, Krasna MJ. Prognostic significance of positive circumferential resection margin in esophageal cancer: a systematic review and meta-analysis. Ann Thorac Surg 2014;97:446-53.
23. Thompson SK, Ruszkiewicz AR, Jamieson GG, et al. Improving the accuracy of TNM staging in esophageal cancer: a pathological review of resected specimens. Ann Surg Oncol 2008;15:3447-58.
24. Kim HI, Cheong JH, Song KJ, et al. Staging of adenocarcinoma of the esophagogastric junction: comparison of AJCC 6th and 7th gastric and 7th esophageal staging systems. Ann Surg Oncol 2013;20:2713-20.
25. Lagarde SM, Reitsma JB, de Castro SM, Ten Kate FJ, Busch OR, van Lanschot JJ. Prognostic nomogram for patients undergoing oesophagectomy for adenocarcinoma of the oesophagus or gastro-oesophageal junction. Br J Surg 2007;94:1361-8.
26. Lagarde SM, Reitsma JB, Ten Kate FJ, et al. Predicting individual survival after potentially curative esophagectomy for adenocarcinoma of the esophagus or gastroesophageal junction. Ann Surg 2008;248:1006-13.
27. Sjoquist KM, Burmeister BH, Smithers BM, et al. Survival after neoadjuvant chemotherapy or chemoradiotherapy for resectable oesophageal carcinoma: an updated meta-analysis. Lancet Oncol 2011;12:681-92.
28. van Hagen P, Hulshof MC, van Lanschot JJ, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. N Engl J Med 2012;366:2074-84.
29. van Meerten E, Muller K, Tilanus HW, et al. Neoadjuvant concurrent chemoradiation with weekly paclitaxel and carboplatin for patients with oesophageal cancer: a phase II study. Br J Cancer 2006;94:1389-94.
30. Sobin LH, Wittekind C, et al. TNM classification of malignant tumors. 6th ed. New York: Wiley-Liss; 2002.
31. Sobin LH, Gospodarowicz MK, Wittekind C, International Union against Cancer., ebrary Inc. TNM classification of malignant tumors. 7th ed. New York: Wiley-Blackwell; 2009.
32. Mandard AM, Dalibard F, Mandard JC, et al. Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. Clinicopathologic correlations. Cancer 1994;73:2680-6.
33. Chirieac LR, Swisher SG, Ajani JA, et al. Posttherapy pathologic stage predicts survival in patients with esophageal carcinoma receiving preoperative chemoradiation. Cancer 2005;103:1347-55.
34. van Buuren S, Groothuis K, Robitzsch A, Vink G, Doove L, Shahab J. Mice: multivariate imputation by chained equations in R. R package version 2.22 ed2011.
35. Steyerberg EW. Clinical prediction models. A practical approach to development, validation, and updating. New York: Springer-Verlag; 2009.
36. Harrell Jr. FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.
37. Team RC. R: a language and environment for statistical computing. 3.0.2 ed. Vienna, Austria: R Foundation for Statistical Computing; 2013.
38. Harrell Jr. FE. Rms: regression modeling strategies. R package version 3.6-3 ed2013.
39. Donington JS, Miller DL, Allen MS, Deschamps C, Nichols FC, 3rd, Pairolero PC. Tumor response to induction chemoradiation: influence on survival after esophagectomy. Eur J Cardiothorac Surg 2003;24:631-6; discussion 6-7.

40. Schneider PM, Baldus SE, Metzger R, et al. Histomorphologic tumor regression and lymph node metastases determine prognosis following neoadjuvant radiochemotherapy for esophageal cancer: implications for response classification. Ann Surg 2005;242:684-92.
41. Donohoe CL, O'Farrell NJ, Grant T, et al. Classification of pathologic response to neoadjuvant therapy in esophageal and junctional cancer: assessment of existing measures and proposal of a novel 3-point standard. Ann Surg 2013;258:784-92.
42. Gu Y, Swisher SG, Ajani JA, et al. The number of lymph nodes with metastasis predicts survival in patients with esophageal or esophagogastric junction adenocarcinoma who receive preoperative chemoradiation. Cancer 2006;106:1017-25.
43. Rizk NP, Venkatraman E, Bains MS, et al. American Joint Committee on Cancer staging system does not accurately predict survival in patients receiving multimodality therapy for esophageal adenocarcinoma. J Clin Oncol 2007;25:507-12.
44. Ajani JA, Correa AM, Swisher SG, Wu TT. For localized gastroesophageal cancer, you give chemoradiation before surgery, but then what happens? J Clin Oncol 2007;25:4315-6.
45. Holscher AH, Drebber U, Schmidt H, Bollschweiler E. Prognostic classification of histopathologic response to neoadjuvant therapy in esophageal adenocarcinoma. Ann Surg 2014;260:779-84; discussion 84-5.
46. van Vliet EP, Eijkemans MJ, Kuipers EJ, Poley JW, Steyerberg EW, Siersema PD. Publication bias does not play a role in the reporting of the results of endoscopic ultrasound staging of upper gastrointestinal cancers. Endoscopy 2007;39:325-32.
47. van Vliet EP, Heijenbrok-Kal MH, Hunink MG, Kuipers EJ, Siersema PD. Staging investigations for oesophageal cancer: a meta-analysis. Br J Cancer 2008;98:547-57.
48. Rouzier R, Pusztai L, Delaloge S, et al. Nomograms to predict pathologic complete response and metastasis-free survival after preoperative chemotherapy for breast cancer. J Clin Oncol 2005;23:8331-9.
49. Quah HM, Chou JF, Gonen M, et al. Pathologic stage is most prognostic of disease-free survival in locally advanced rectal cancer patients after preoperative chemoradiation. Cancer 2008;113:57-64.
50. Kim MS, Lee SY, Lee TR, et al. Prognostic nomogram for predicting the 5-year probability of developing metastasis after neo-adjuvant chemotherapy and definitive surgery for AJCC stage II extremity osteosarcoma. Ann Oncol 2009;20:955-60.
51. Harvin JA, Lahat G, Correa AM, et al. Neoadjuvant chemoradiotherapy followed by surgery for esophageal adenocarcinoma: significance of microscopically positive circumferential radial margins. J Thorac Cardiovasc Surg 2012;143:412-20.
52. Lerut T, Coosemans W, Decker G, et al. Extracapsular lymph node involvement is a negative prognostic factor in T3 adenocarcinoma of the distal esophagus and gastroesophageal junction. J Thorac Cardiovasc Surg 2003;126:1121-8.
53. D'Annoville T, D'Journo XB, Loundou A, et al. Prognostic impact of the extracapsular lymph node involvement on disease-free survival according to the 7th edition of American Joint Committee on Cancer Staging System. Eur J Cardiothorac Surg 2013;44:e207-11.
54. Nafteux P, Lerut T, De Hertogh G, et al. Can extracapsular lymph node involvement be a tool to fine-tune pN1 for adenocarcinoma of the oesophagus and gastro-oesophageal junction in the Union Internationale contre le Cancer (UICC) TNM 7th edition?dagger. Eur J Cardiothorac Surg 2014;45:1001-10.
55. Enlow JM, Denlinger CE, Stroud MR, Ralston JS, Reed CE. Adenocarcinoma of the esophagus with signet ring cell features portends a poor prognosis. Ann Thorac Surg 2013;96:1927-32.
56. Nafteux PR, Lerut TE, Villeneuve PJ, et al. Signet ring cells in esophageal and gastroesophageal junction carcinomas have a more aggressive biological behavior. Ann Surg 2014 [Epub ahead of print].

57. Izzo JG, Wu TT, Wu X, et al. Cyclin D1 guanine/adenine 870 polymorphism with altered protein expression is associated with genomic instability and aggressive clinical biology of esophageal adenocarcinoma. J Clin Oncol 2007;25:698-707.
58. Ong CA, Shapiro J, Nason KS, et al. Three-gene immunohistochemical panel adds to clinical staging algorithms to predict prognosis for patients with esophageal adenocarcinoma. J Clin Oncol 2013;31:1576-82.

17

A Dutch prediction tool to assess the risk of additional axillary non-sentinel node involvement in sentinel node positive breast cancer patients

I van den Hoven
D van Klaveren
A Voogd
Y Vergouwe
V Tjan-Heijnen
RMH Roumen

Clinical Breast Cancer 2016

ABSTRACT

Background Multiple predictive systems have previously been developed in order to identify the sentinel lymph node (SLN) positive patients that are at low risk for additional axillary non-sentinel lymph node (non-SLN) involvement and for whom a completion axillary lymph node (ALND) could be avoided. However, previous studies showed that these tools had a poor performance in Dutch breast cancer patients.The aim of this study was to develop a predictive tool for the risk of non-SLN involvement in a Dutch population of SLN positive breast cancer patients.

Methods Data of 513 SLN positive breast cancer patients of 10 participating hospitals, that underwent a completion axillary lymph node dissection between January 2007 and December 2008 were studied. The uni- and multivariable associations of predictors for non-SLN metastases were analysed and a predictive model was developed. The discriminative ability of the model was measured by the area under the receiver operating characteristic curve (AUC) and the agreement between predicted probabilities and observed frequencies was visualized by a calibration plot.

Results A predictive model was developed including the two strongest predictors; the size of the SLN metastases in mm and the presence of a negative sentinel node. The model showed good discriminative ability (AUC 0.75) and good calibration over the complete range of predicted probabilities.

Conclusion We have developed a tool for predicting the risk of additional non-SLN metastases in Dutch SLN positive breast cancer patients that is easy to use in the daily clinical breast cancer practice.

INTRODUCTION

Breast cancer treatment is changing rapidly over the years. This applies not only to the systemic treatment, but also to the locoregional strategies like surgery or radiotherapy. After the introduction of breast conserving therapy (BCT) the sentinel lymph node biopsy (SLNB) procedure was implemented for axillary staging, which had a tremendous impact on the treatment of the axilla and has spared a lot of women a completion axillary lymph node dissection(ALND) with its associated morbidity.

Since the implementation of the SLNB for staging the axilla, a completion ALND was the standard treatment when SLN metastases were discovered. However, the benefit of ALND for all sentinel lymph node (SLN) positive patients has been questioned for some years now. It is suggested that there is a subgroup of SLN positive patients with only a low risk of additional non-SLN involvement, knowing that in up to 70% of patients, the SLN is the only metastatically involved lymph node [1-5]. Moreover, the results of the American College of Surgeons Onoclogy Group (ACOSOG) Z0011 trial, already showed that there is a subgroup of SLN positive patients who seem to have no benefit from completion ALND at all, in terms of recurrence or overall survival [6].

The identification of patients at low risk for additional non SLN involvement, however, is difficult. Over the years, various predictive systems have been developed to define such a low risk subgroup of SLN positive patients for whom a completion ALND might safely be omitted. In general, the purpose of these predictive systems or tools, is to provide the evidence-based input, to help clinicians and patients in their clinical decision-making process [7]. However, validation studies of the different systems have shown that a good performance in one population, does not necessarily mean that it is as reliable in another population. Variations in pathology settings and differences in baseline population characteristics due to specific selection practices [8] seem to play an important role in this [9-16]. Therefore, the generalization and applicability of the different predictive tools is often limited. Next, the variability of outcomes, using different predictive systems for the same individual patient, are impractically large and therefore not acceptable nor helpful in the decision making process. This has clearly been shown for individual patients of our own (Dutch) population [17]. The aim of the present study was to develop a tool to predict the risk of non-SLN involvement for Dutch SLN positive breast cancer patients.

PATIENTS AND METHODS

Study population

Patients with primary breast cancer were identified from the Dutch National Cancer Registration of the South region of the Netherlands in which 10 hospitals participated. Patients were eligible for the study if they met the following conditions: (1) treated between

January 2007 and December 2008, (2) had a successful SLNB with histopathological proven metastases, and (3) a completion ALND was performed. Patients were excluded from the study when axillary metastasis was proven by ultrasound guided lymph node biopsy (UGLNB) or within the neo-adjuvant treatment setting.

Data accrual

The following clinical data were collected: age at time of diagnosis, site and lateralization of the tumor and type of surgery (mastectomy or breast conserving therapy). Histopathological data of the tumor included: tumor size in millimeters, tumor type, tumor grade (using the modified Bloom and Richardson classification), presence of lymphovascular invasion; multifocality, presence of estrogen or progesterone sensitivity and Her2Neu status. Histopathological data of the sentinel node and other harvested lymph nodes included: total number of resected lymph nodes and total number of positive nodes, size of largest metastases in the SLN as a continuous variable and categorized as macrometastases (>2mm), micrometastases (> 0.2 but ≤ 2 mm) or submicrometastases/isolated tumor cells (ITC) (≤ 0.2mm) and finally the presence of extracapsular extension. There were some missing values concerning lymphovascular invasion, tumor grade, categorized size of the SLN metastasis, the size of the SLN metastasis as a continuous value, hormonal receptor and Her2Neu status. We assumed that lymphovascular invasion was absent, if values for this variable were missing and it was not reported in the pathology report. Other missing values were multiply imputed with the *mice* algorithm in R software allowing all observed values to be analysed [18,19].

Predictors and model development

We included candidate predictors for non SLN involvement in the statistical analysis based on the literature and previously reported models including the Memorial Sloan Kettering Cancer Center (MSKCC), the Mayo, the Cambridge, the Stanford, the Masaryk Memorial Cancer Institute (MOU nomogram), a model developed by Gur et al. and the MD Anderson, the Tenon and the Saïdi scores [2,14,20-25]. The univariable associations with non-SLN metastases were analysed with logistic regression.

We evaluated the contribution of each predictive factor by its multivariable odds ratio together with its likelihood ratio χ^2 test statistic minus twice the degrees of freedom (number of regression coefficients used to model a predictive factor). The latter is consistent with Akaike's Information Criterion (AIC) which balances the goodness-of-fit of a model with its complexity and gives a fair assessment of a factor's predictiveness [26]. Following a backward selection approach, we eliminated predictors with negative additional AIC from the model. We assessed the discriminative ability of the model by the area under the receiver operating characteristic curve (AUC).

Table 17.1 Univariable associations of predictors for non-sentinel lymph node metastases.

Predictors	Level	N	Non-SLN+	%	OR	95% CI
Mastectomy	No	322	90	28	1.00	
	Yes	191	53	28	0.99	0.66-1.48
LVI	No	437	112	26	1.00	
	Yes	76	31	41	2.00	1.21-3.31
Morphology IDC	No	423	118	28	1.00	
	Yes	90	25	28	0.99	0.60-1.65
Tumour grade	1	163	36	22	1.00	
	2	232	64	28	1.34	0.84-2.15
	3	112	40	36	1.96	1.15-3.35
	Missing	6	3	50		
ER positive	No	53	16	30	1.00	
	Yes	460	127	28	0.88	0.47-1.64
PR positive	No	96	32	33	1.00	
	Yes	381	102	27	0.73	0.45-1.18
	Missing	36	9	25		
Her2Neu positive	No	451	120	27	1.00	
	Yes	59	22	37	1.64	0.93-2.89
	Missing	3	1	33		
Multifocal	No	457	125	27	1.00	
	Yes	56	18	32	1.26	0.69-2.29
ECE	No	401	94	23	1.00	
	Yes	112	49	44	2.54	1.64-3.94
Laterality (Left)	No	258	75	29	1.00	
	Yes	255	68	27	0.89	0.60-1.31
Tumour size (mm)	≤15	184	35	19	1.00	
	>15 ≤25	229	68	30	1.80	1.13-2.86
	>25	100	40	40	2.84	1.65-4.89
Age at diagnosis (years)	≤ 50	160	45	28	1.00	
	>50 ≤ 65	203	62	31	1.12	0.71-1.77
	>65	150	36	24	0.81	0.49-1.34
SLNs positive	1	423	107	25	1.00	
	2+	90	36	40	1.97	1.22-3.17
SLNs negative	0	268	93	35	1.00	
	1	150	32	21	0.51	0.32-0.81
	2+	95	18	19	0.44	0.25-0.78
Macrometastases	No	148	24	16	1.00	
	Yes	223	73	33	2.51	1.50-4.22
	Missing	142	46	32		
Size of SLN Metastases (mm)	≤ 2	105	15	14	1.00	
	>2 ≤5	80	9	11	0.76	0.31-1.84
	>5	72	34	47	5.37	2.62-10.99
	Missing	256	85	33		
Overall		513	143	28		

CI = confidence interval; ECE = extracapsular extension; ER = estrogen receptor status; ILC = invasive lobular carcinoma; LVI = lymphovascular invasion; OR = odds ratio; PR = progesterone receptor; SLN = sentinel lymph node.

The AUC is equal to the probability that a randomly selected patient with the outcome (being non-SLN metastasis) has a higher risk prediction than a randomly selected patient without the outcome. A useless prediction model, such as a coin flip, would result in an area of 0.5. When the AUC is 1 the model discriminates perfectly [27,28]. We used a bootstrap procedure to correct the AUC for optimism [27]. A calibration plot was used to visualize the agreement between predicted probabilities and observed frequencies. The final model was presented as a score chart. All analyses were performed with R (version 2.13.1; R foundation for Statistical Computing, Vienna, Austria).

Table 17.2 Multivariable associations of predictors for non-sentinel lymph node metastases, including the presence of macrometastasis in the sentinel lymph node.

Predictors	OR	95% CI	χ^2	df	p-value
Full set of predictors					
Mastectomy	0.85	0.54-1.36	0.4	1	0.5072
LVI	1.63	0.90-2.93	2.7	1	0.1038
Morphology IDC	1.12	0.62-2.05	0.2	1	0.7011
Tumour grade (2:1)	1.47	0.88-2.48	3.4	2	0.1824
Tumour grade (3:1)	1.82	0.93-3.54			
ER positive	1.29	0.54-3.04	0.3	1	0.5683
PR positive	0.67	0.35-1.27	1.5	1	0.2197
Her2Neu positive	1.32	0.69-2.54	0.7	1	0.4022
Multifocal	1.29	0.65-2.56	0.5	1	0.4739
ECE	1.87	1.13-3.08	6.0	1	0.0143
Laterality (left)	0.89	0.58-1.36	0.3	1	0.5864
Tumour size (10 mm)	1.32	1.03-1.68	4.9	1	0.0264
Age (10 years)	0.93	0.78-1.11	0.6	1	0.4327
SLN positive (1 node)	1.55	0.91-2.63	2.6	1	0.1083
SLN negative (1 node)	0.45	0.29-0.70	12.9	1	0.0003
SLN macrometastases	1.87	1.06-3.31	4.6	1	0.0320
Selected model					
Tumour grade (2:1)	1.63	0.98-2.69	8.0	2	0.0186
Tumour grade (3:1)	2.29	1.28-4.09			
ECE	1.94	1.19-3.16	7.2	1	0.0075
Tumour size (10 mm)	1.32	1.05-1.65	5.9	1	0.0153
SLN negative (1 node)	0.46	0.30-0.71	12.6	1	0.0004
SLN macrometastases	1.99	1.15-3.43	6.1	1	0.0138

CI = confidence interval; df = degrees of freedom; ECE = extracapsular extension; ER = estrogen receptor; ILC = invasive lobular carcinoma; LVI = lymphovascular invasion; OR = odds ratio; PR = progesterone receptor; SLN = sentinel lymph node.

RESULTS

The study population consisted of 513 sentinel node positive breast cancer patients with a mean age of 58 years. Of these, 322 (62,8%) patients were treated with BCT and 191 (37,2%)

underwent a mastectomy. Most patients (n=405) had an invasive ductal carcinoma which accounts for 78,9%, 90 patients had invasive lobular carcinoma (17,5%) and 18 patients had a different or mixed type breast carcinoma(3,5%). In 143 (28%) patients additional axillary metastases were found by completion ALND with a mean retrieval of 12 lymph nodes.

Table 17.3 Multivariable associations of predictors for non-sentinel lymph node metastases, including the size of the largest sentinel lymph node metastasis as a continuous variable.

Predictors	OR	95% CI	χ^2	df	p-value
Full set of predictors					
Mastectomy	0.84	0.50-1.43	0.4	1	0.5296
LVI	1.35	0.68-2.68	0.8	1	0.3860
Morphology IDC	1.02	0.52-2.02	0.0	1	0.9466
Tumour grade (2:1)	1.24	0.70-2.21	1.0	2	0.6050
Tumour grade (3:1)	1.47	0.68-3.20			
ER positive	1.09	0.41-2.88	0.0	1	0.8633
PR positive	0.67	0.32-1.39	1.2	1	0.2800
Her2Neu positive	1.59	0.76-3.34	1.5	1	0.2183
Multifocal	1.14	0.51-2.55	0.1	1	0.7537
ECE	0.91	0.49-1.71	0.1	1	0.7759
Laterality (left)	0.86	0.52-1.40	0.4	1	0.5384
Tumour size (10 mm)	1.20	0.91-1.58	1.6	1	0.2051
Age (10 years)	1.02	0.83-1.24	0.0	1	0.8742
SLN positive (1 node)	0.97	0.48-1.98	0.0	1	0.9325
SLN negative (1 node)	0.47	0.28-0.77	8.9	1	0.0028
Size of SLN metastases (mm)	1.22	1.14-1.31	30.9	1	<.0001
Selected model					
SLN negative (1 node)	0.49	0.30-0.79	8.6	1	0.0033
Size SLN metastases (mm)	1.22	1.15-1.29	46.7	1	<.0001

CI = confidence interval; df = degrees of freedom; ECE = extracapsular extension; ER = estrogen receptor; ILC = invasive lobular carcinoma; LVI = lymphovascular invasion; OR = odds ratio; PR = progesterone receptor; SLN = sentinel lymph node.

Significant univariable predictors were: the presence of lymphovascular invasion (OR 2.0 95% CI 1.21-3.31), extracapsular extension (OR 2.54 95%CI 1.64-3.94), tumour grade 3 (OR 1.96 95%CI 1.15-3.35), size of the tumour >25 mm (OR 2.84 95%CI 1.65-4.89), ≥ 2 positive SLNs (OR 1.97 95%CI 1.22-3.17), ≥ 2 negative SLNs (OR 0.44 95%CI 0.25-0.78), SLN macrometastases (OR 2.51 95%CI 1.50-4.22) and a SLN metastases size > 5 mm (OR 5.37 95%CI 2.62-10.99). Predictor effects of: multifocality of the tumour (OR 1.26 95%CI 0.69-2.29), ER positive (OR 0.88 95%CI 0.47-1.64), PR positive (OR 0.73 95%CI 0.45-1.18) and Her2Neu positivity (OR 1.64 95%CI 0.93-2.89) were non-significant but in the expected direction (Table 17.1). When we modelled the size of the largest SN metastases by a binary predictor (macrometastases yes or no), the strongest predictors in the multivariable regression model were tumour size in mm, tumor grade, presence of macrometastases in the SLN, extracapsular extension of the SLN metastasis and the presence of a negative SLN. However, the continuous size of the largest SN metastasis (in mm) proved to be a much stronger predictor than the presence of macrometastases (Table 17.3). As a consequence of adding the size of the largest SN metastasis as a continuous predictor to the multivariable regression model, all other factors, except for the presence of a negative SLN, lost their predictiveness. The final model included the size of the SLN metastasis in mm and the presence of a negative SLN as the sole predictors.

The AUC after optimism correction increased from 0.68 to 0.75 for the final model, when substituting the continuous metastasis size for the categorical size of the SLN metastases. Predicted probabilities of the patients were grouped into quintiles and for each quintile the mean predicted probability was compared with the actual proportion of non-SLN metastases. Calibration was good over the complete range of predicted probabilities (Figure 17.1).

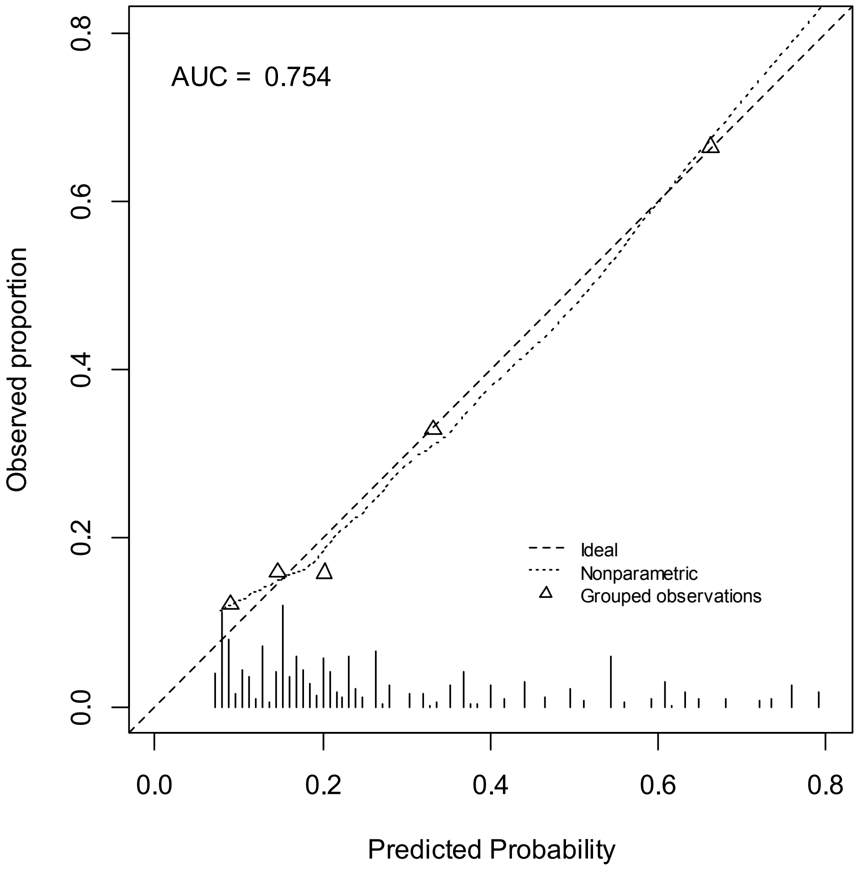


Figure 17.1 Validity of the model for prediction of non- sentinel lymph node metastases expressed as calibration. The distribution of predicted risks is shown at the bottom of the graphs, by non-sentinel lymph node metastasis. The triangles indicate the observed proportions by quintiles of predicted risks.

The final model was presented as a score chart (figure 17.2) and can easily be used for calculation of individual risk predictions. For example, a patient with a largest SLN metastases of 4 mm (4 points) with 1 negative SLN removed (0 points) has a sum score of 4, which amounts to a predicted probability of 15% for additional non-SLN metastases.

Sum score =
Size largest SLN metastases (mm)
+ 4 if no negative SLN were found

Sum score	Probability non-SLN+ (%)
<1	7.6
1	9.2
2	11
3	13
4	15
5	18
6	21
7	25
8	28
9	33
10	37
11	42
12	47
13	52
14+	56

Figure 17.2 Score chart to predict the risk of additional non-sentinel lymph node metastases in breast cancer patients after positive sentinel lymph node biopsy. SLN: sentinel lymph node. non-SLN+ = non-sentinel lymph node metastases

DISCUSSION

This study was conducted because previously developed predictive systems showed to have a poor performance in our Dutch cohort of SLN positive breast cancer patients [10]. The generalization and applicability of these tools was limited probably due to variations in pathology settings and differences in baseline population characteristics. Moreover, selection of patients as for instance by ultrasound guided lymph node biopsy varies per country [8]. Also the variability of outcomes, when using different predictive systems for the same individual patient were too large resulting in non-applicability in the daily decision making process [17].

We have now developed a predictive system that incorporates only two variables that are significantly associated with the presence of additional non-SLN metastases. These variables: the size of the SLN metastasis and the presence of a negative SLN, are both known predictors for additional non-SLN metastases. The size of the SLN metastasis is often categorized as macrometastases (>2mm), micrometastases (> 0.2 but ≤ 2 mm) or submicrometastases/isolated tumor cells (ITC) (≤ 0.2mm). However, when using a cutoff above 5 mm for the SLN metastases size, or using the actual SLN metastases size in mm, this is highly associated with the risk of non-SLN metastases (table 17.1 and 17.3). The size of the SLN metastases was already used in some other previously published predictive systems [11,14,21-23,29]. However, none of these systems showed that the size of the SLN metastases was so important that it can almost predict the risk for non-SLN metastases by itself. Moreover, some of the existing predictive systems have previously been developed specifically for patients with micrometastases or isolated tumor cells, which means that these models can only be applied to a certain small subgroup of patients [9,30]. Besides that, Cserni et al. showed that up to 90% of breast cancer patients with micrometastatic disease in the SLN did a priori not have additional non-SLN involvement [31]. Also Bilimoria et al. showed that the 5-year overall survival and axillary recurrence rates was not different between breast cancer patients with micrometastases that received an ALND in comparison to those without a completion ALND after a median follow-up of 5 years [32]. Many centres therefore already stopped performing an ALND for patients with micrometastatic disease in the axilla. Besides the metastases size, the presence of a negative SLN was also a significant predictor and therefore included in the model. Other models often incorporated the SLN ratio into their models, which is the ratio of the number of positive SLNs and the total number of SLNs that is removed. This also reflects the negative SLNs removed but in a different manner, which seems less intuitive.

Several previously developed models were often based on single centre patient series, using data from more than 10 to 15 years ago and included up to 8 variables in the model. An important goal of the present study was to develop a user friendly model with only a few predictors, but with still a good discriminative ability. We have developed a model based on a patient population from ten different participating centres, used more recent data after the introduction of the revised TNM classification [33] and included only two variables into the model which makes it easy to use. Our model shows a good overall performance and is well calibrated as compared with previously published predictive systems and validation studies which show a wide range in discriminative ability measured by an AUC ranging from 0.68-0.86 [11-15,34-39]. Therefore we conclude that the present model can safely be applied for risk prediction in our own patient population. Nonetheless, we expect that the two predictors included in our model will hold their predictiveness also in other populations and settings.

Of course this study also has some limitations. We used a retrospectively collected database and therefore had to deal with some missing values. The size of the SLN metastasis as a continuous variable appeared to be the strongest predictor for non-SLN metastases. However at present, the actual size of the SLN metastases is not always reported in the pathology reports and the Dutch guidelines only recommend the documentation of the SLN metastases size categorized as macrometastases (>2mm), micrometastases (> 0.2 but ≤ 2 mm) or submicrometastases/isolated tumor cells (ITC) (≤ 0.2mm) [40]. Perhaps, these guidelines should be adjusted whereas the actual size of the SLN metastases will be reported to facilitate accurate predictions and treatment recommendations in the future. Furthermore, although the present AUC showed a good discriminative ability, it did not outstand the AUC's of some previously developed models in other patient populations.

The axillary treatment for breast cancer changed rapidly over the years and following the results of the ACOSOG Z0011 trial many centres have already abandoned a completion ALND for low risk patients and even so the predictive systems to identify them. It is obvious that a paradigm shift is developing in the axillary treatment of breast cancer, from a concept: to treat all,- except towards a concept of: treat none,- unless. However, an important concern is that premature implementation of the results of the Z0011 trial might result in hazardous under treatment in a considerable number of patients. Güth et al calculated in how many patients of their own population a completion ALND would have been omitted if the inclusion criteria of the Z0011 trial had been applied. Their results show that only 35 (9%) of the patients that underwent SLNB, met the inclusion criteria of the Z0011 trial and thus would not have had a completion ALND. This, on the other hand, represented only 5,9% of all surgically treated patients [41].

The results of the Z0011 trial therefore must be interpreted cautiously, as they apply only to a small subgroup of women that meet the exact inclusion criteria. A substantial amount of patients (in our population 37% of all patients) for example is treated with a mastectomy without additional radiotherapy. It has been shown that these patients may be at higher risk of regional recurrence [42]. Whether the results of the Z0011 trial can be generalized also to this subgroup of patients is questioned.

To our opinion there remains a substantial group of patients with axillary metastases that do benefit from some sort of treatment of the axilla. Instead of looking for those with a low risk for further non-SLN involvement, we should maybe focus on finding the high risk population with extensive nodal disease (more than 3 or 4 positive nodes) who still might benefit from axillary treatment by either more extensive surgery like ALND or radiotherapy [43]. This concerns than the patients after BCT, but also other patient categories like those treated by mastectomy, or patients in the neo-adjuvant treatment setting, or those with axillary metastases proven by UGLNB may be candidates for risk prediction, since they do not fit in the Z0011 selected group. Katz et al. already published a nomogram for the prediction of having four or more involved nodes in SLN positive breast cancer to provide information for clinicians and patients making decisions with respect to type and extent use of adjuvant systemic and radiation therapy [44]. At present, we are investigating whether a comparable predictive model could be developed to identify breast cancer patients with a high risk of extensive nodal disease that have three or more positive axillary lymph nodes as in Z0011. Such a development will be the next step in the paradigm shift in axillary breast cancer treatment: from “treat all, - except, towards “treat none, - unless”.

CONCLUSION

We have presented a multicenter predictive tool for predicting the risk of additional non-SLN metastases in patients with SLN-positive breast cancer. Although we find ourselves in a paradigm shift concerning axillary treatment after positive SLNB, our model will find its use in patients comparable to our population who do not meet the strictly defined Z0011 inclusion criteria. Our model has good discriminative ability and is easy to use in daily clinical breast cancer practice. However, it must be validated before incorporating it at other centers.

ACKNOWLEDGEMENTS

We thank the Dutch National Cancer Registration and the 10 participating centers for the collection and use of the data. This research was supported and facilitated by Stichting Heelkundig Kankeronderzoek Zuid-Nederland. D. van Klaveren and Y. Vergouwe are funded by the Netherlands Organization for Scientific Research (grant 917.11.383).

REFERENCES

1. Chu KU, Turner RR, Hansen NM, Brennan MB, Bilchik A, Giuliano AE. Do all patients with sentinel node metastasis from breast carcinoma need complete axillary node dissection? Ann Surg 1999 Apr;229(4):536-541.
2. Hwang RF, Krishnamurthy S, Hunt KK, Mirza N, Ames FC, Feig B, et al. Clinicopathologic factors predicting involvement of nonsentinel axillary nodes in women with breast cancer. Ann Surg Oncol 2003 Apr;10(3):248-254.
3. Nos C, Harding-MacKean C, Freneaux P, Trie A, Falcou MC, Sastre-Garau X, et al. Prediction of tumour involvement in remaining axillary lymph nodes when the sentinel node in a woman with breast cancer contains metastases. Br J Surg 2003 Nov;90(11):1354-1360.
4. Kim T, Giuliano AE, Lyman GH. Lymphatic mapping and sentinel lymph node biopsy in early-stage breast carcinoma: a metaanalysis. Cancer 2006 Jan 1;106(1):4-16.
5. Kamath VJ, Giuliano R, Dauway EL, Cantor A, Berman C, Ku NN, et al. Characteristics of the sentinel lymph node in breast cancer predict further involvement of higher-echelon nodes in the axilla: a study to evaluate the need for complete axillary lymph node dissection. Arch Surg 2001 Jun;136(6):688-692.
6. Giuliano AE, McCall L, Beitsch P, Whitworth PW, Blumencranz P, Leitch AM, et al. Locoregional recurrence after sentinel lymph node dissection with or without axillary dissection in patients with sentinel lymph node metastases: the American College of Surgeons Oncology Group Z0011 randomized trial. Ann Surg 2010 Sep;252(3):426-32; discussion 432-3.
7. Laupacis A, Wells G, Richardson WS, Tugwell P. Users' guides to the medical literature. V. How to use an article about prognosis. Evidence-Based Medicine Working Group. JAMA 1994 Jul 20;272(3):234-237.
8. Verheuvel NC, van den Hoven I, Ooms HW, Voogd AC, Roumen RM. The role of ultrasound-guided lymph node biopsy in axillary staging of invasive breast cancer in the post-ACOSOG Z0011 trial era. Ann Surg Oncol 2015 Feb;22(2):409-415.
9. Alran S, De Rycke Y, Fourchotte V, Charitansky H, Laki F, Falcou MC, et al. Validation and limitations of use of a breast cancer nomogram predicting the likelihood of non-sentinel node involvement after positive sentinel node biopsy. Ann Surg Oncol 2007 Aug;14(8):2195-2201.
10. van den Hoven I, Kuijt GP, Voogd AC, van Beek MW, Roumen RM. Value of Memorial Sloan-Kettering Cancer Center nomogram in clinical decision making for sentinel lymph node-positive breast cancer. Br J Surg 2010 Nov;97(11):1653-1658.
11. Degnim AC, Reynolds C, Pantvaidya G, Zakaria S, Hoskin T, Barnes S, et al. Nonsentinel node metastasis in breast cancer patients: assessment of an existing and a new predictive nomogram. Am J Surg 2005 Oct;190(4):543-550.
12. Klar M, Jochmann A, Foeldi M, Stumpf M, Gitsch G, Stickeler E, et al. The MSKCC nomogram for prediction the likelihood of non-sentinel node involvement in a German breast cancer population. Breast Cancer Res Treat 2008 Dec;112(3):523-531.
13. Kocsis L, Svebis M, Boross G, Sinko M, Maraz R, Rajtar M, et al. Use and limitations of a nomogram predicting the likelihood of non-sentinel node involvement after a positive sentinel node biopsy in breast cancer patients. Am Surg 2004 Nov;70(11):1019-1024.
14. Pal A, Provenzano E, Duffy SW, Pinder SE, Purushotham AD. A model for predicting non-sentinel lymph node metastatic disease when the sentinel lymph node is positive. Br J Surg 2008 Mar;95(3):302-309.
15. Smidt ML, Kuster DM, van der Wilt GJ, Thunnissen FB, Van Zee KJ, Strobbe LJ. Can the Memorial Sloan-Kettering Cancer Center nomogram predict the likelihood of nonsentinel lymph node metastases in breast cancer patients in the Netherlands? Ann Surg Oncol 2005 Dec;12(12):1066-1072.
16. Soni NK, Carmalt HL, Gillett DJ, Spillane AJ. Evaluation of a breast cancer nomogram for prediction of non-sentinel lymph node positivity. Eur J Surg Oncol 2005 Nov;31(9):958-964.

17. van den Hoven I, Kuijt GP, Voogd AC, Roumen RM. High intersystem variability for the prediction of additional axillary non-sentinel lymph node involvement in individual patients with sentinel node-positive breast cancer. Ann Surg Oncol 2012 Jun;19(6):1841-1849.
18. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med 1991 Apr;10(4):585-598.
19. van Buuren S, Groothuis K. Multivariate Imputation by Chained Equations. Available at: http://www.jstatsoft.org/v45/i03/paper. Accessed April/13, 2015.
20. Van Zee KJ, Manasseh DM, Bevilacqua JL, Boolbol SK, Fey JV, Tan LK, et al. A nomogram for predicting the likelihood of additional nodal metastases in breast cancer patients with a positive sentinel node biopsy. Ann Surg Oncol 2003 Dec;10(10):1140-1151.
21. Kohrt HE, Olshen RA, Bermas HR, Goodson WH, Wood DJ, Henry S, et al. New models and online calculator for predicting non-sentinel lymph node status in sentinel lymph node positive breast cancer patients. BMC Cancer 2008 Mar 4;8:66-2407-8-66.
22. Coufal O, Pavlik T, Fabian P, Bori R, Boross G, Sejben I, et al. Predicting non-sentinel lymph node status after positive sentinel biopsy in breast cancer: what model performs the best in a Czech population? Pathol Oncol Res 2009 Dec;15(4):733-740.
23. Gur AS, Unal B, Ozbek U, Ozmen V, Aydogan F, Gokgoz S, et al. Validation of breast cancer nomograms for predicting the non-sentinel lymph node metastases after a positive sentinel lymph node biopsy in a multi-center study. Eur J Surg Oncol 2010 Jan;36(1):30-35.
24. Barranger E, Coutant C, Flahault A, Delpech Y, Darai E, Uzan S. An axilla scoring system to predict non-sentinel lymph node status in breast cancer patients with sentinel lymph node involvement. Breast Cancer Res Treat 2005 May;91(2):113-119.
25. Degnim AC, Griffith KA, Sabel MS, Hayes DF, Cimmino VM, Diehl KM, et al. Clinicopathologic features of metastasis in nonsentinel lymph nodes of breast carcinoma patients. Cancer 2003 Dec 1;98(11):2307-2315.
26. Harrell FE,Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984 Apr-Jun;3(2):143-152.
27. Harrell FE,Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996 Feb 28;15(4):361-387.
28. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982 Apr;143(1):29-36.
29. Mittendorf EA, Hunt KK, Boughey JC, Bassett R, Degnim AC, Harrell R, et al. Incorporation of sentinel lymph node metastasis size into a nomogram predicting nonsentinel lymph node involvement in breast cancer patients with a positive sentinel lymph node. Ann Surg 2012 Jan;255(1):109-115.
30. Saidi RF, Dudrick PS, Remine SG, Mittal VK. Nonsentinel lymph node status after positive sentinel lymph node biopsy in early breast cancer. Am Surg 2004 Feb;70(2):101-5; discussion 105.
31. Cserni G, Gregori D, Merletti F, Sapino A, Mano MP, Ponti A, et al. Meta-analysis of non-sentinel node metastases associated with micrometastatic sentinel nodes in breast cancer. Br J Surg 2004 Oct;91(10):1245-1252.
32. Bilimoria KY, Bentrem DJ, Hansen NM, Bethke KP, Rademaker AW, Ko CY, et al. Comparison of sentinel lymph node biopsy alone and completion axillary lymph node dissection for node-positive breast cancer. J Clin Oncol 2009 Jun 20;27(18):2946-2953.
33. Sobin L, Wittekind C editors. International Union Against Cancer (UICC) TNM Classification Of Malignant Tumors. 6th edition ed. New York: NY: Wiley-Liss; 2002.
34. Coutant C, Olivier C, Lambaudie E, Fondrinier E, Marchal F, Guillemin F, et al. Comparison of models to predict nonsentinel lymph node status in breast cancer patients with metastatic

sentinel lymph nodes: a prospective multicenter study. J Clin Oncol 2009 Jun 10;27(17):2800-2808.

35. Lambert LA, Ayers GD, Hwang RF, Hunt KK, Ross MI, Kuerer HM, et al. Validation of a breast cancer nomogram for predicting nonsentinel lymph node metastases after a positive sentinel node biopsy. Ann Surg Oncol 2006 Mar;13(3):310-320.
36. Ponzone R, Maggiorotto F, Mariani L, Jacomuzzi ME, Magistris A, Mininanni P, et al. Comparison of two models for the prediction of nonsentinel node metastases in breast cancer. Am J Surg 2007 Jun;193(6):686-692.
37. Specht MC, Kattan MW, Gonen M, Fey J, Van Zee KJ. Predicting nonsentinel node status after positive sentinel lymph biopsy for breast cancer: clinicians versus nomogram. Ann Surg Oncol 2005 Aug;12(8):654-659.
38. Zgajnar J, Perhavec A, Hocevar M, Podkrajsek M, Hertl K, Frkovic-Grazio S, et al. Low performance of the MSKCC nomogram in preoperatively ultrasonically negative axillary lymph node in breast cancer patients. J Surg Oncol 2007 Dec 1;96(7):547-553.
39. van la Parra RF, Ernst MF, Bevilacqua JL, Mol SJ, Van Zee KJ, Broekman JM, et al. Validation of a nomogram to predict the risk of nonsentinel lymph node metastases in breast cancer patients with a positive sentinel node biopsy: validation of the MSKCC breast nomogram. Ann Surg Oncol 2009 May;16(5):1128-1135.
40. Integraal Kanker Centrum Nederland. Richtlijn mammacarcinoom 2.0. Available at: http://www.oncoline.nl/mammacarcinoom, 2015.
41. Guth U, Myrick ME, Viehl CT, Schmid SM, Obermann EC, Weber WP. The post ACOSOG Z0011 era: does our new understanding of breast cancer really change clinical practice? Eur J Surg Oncol 2012 Aug;38(8):645-650.
42. van Wely BJ, van den Wildenberg FJ, Gobardhan P, van Dalen T, Borel Rinkes IH, Theunissen EB, et al. "Axillary recurrences after sentinel lymph node biopsy: a multicentre analysis and follow-up of sentinel lymph node negative breast cancer patients". Eur J Surg Oncol 2012 Oct;38(10):925-931.
43. Donker M, van Tienhoven G, Straver ME, Meijnen P, van de Velde CJ, Mansel RE, et al. Radiotherapy or surgery of the axilla after a positive sentinel node in breast cancer (EORTC 10981-22023 AMAROS): a randomised, multicentre, open-label, phase 3 non-inferiority trial. Lancet Oncol 2014 Nov;15(12):1303-1310.
44. Katz A, Smith BL, Golshan M, Niemierko A, Kobayashi W, Raad RA, et al. Nomogram for the prediction of having four or more involved nodes for sentinel lymph node-positive breast cancer. J Clin Oncol 2008 May 1;26(13):2093-2098.

18

Predicting the extent of nodal involvement for node positive breast cancer patients: development and validation of a novel tool

I van den Hoven*
D van Klaveren*
N Verheuvel
R van la Parra
A Voogd
W de Roos
K Bosscha
E Heuts
V Tjan-Heijnen
RMH Roumen
EW Steyerberg

Submitted

* Equal contribution

ABSTRACT

Purpose The axillary treatment for breast cancer patients with a positive sentinel lymph node (SLN) biopsy has undergone some important changes, resulting in a more complex clinical decision making process. The aim of this study was to develop an easy to use prediction model for the extent of nodal involvement of node positive breast cancer patients.

Patients and Methods Data of 911 SLN positive breast cancer patients, from 10 participating centers in the Netherlands were used for model development. Predictors associated with axillary lymph node involvement were evaluated by univariable and multivariable analysis. The model was validated externally in an independent patient population of 180 patients with SLN positive breast cancer.

Results A novel model was developed to predict the risk of having a total of 1-2, ≥3 or ≥4 positive axillary lymph nodes, for patients with SLN positive breast cancer. Final pathology after axillary lymph node dissection (ALND) showed additional positive lymph nodes for 259 (28%) of the patients. 726 (81%) out of 911 patients had a total of 1-2 positive nodes, whereas 175 (19%) had ≥ 3 positive lymph nodes. The model included three predictors: the tumour size (in mm), the presence of a negative SLN and the size of the SLN metastases (in mm). At external validation, the model showed a good discriminative ability (AUC 0.82, 95% CI 0.74-0.90) and good calibration over the full range of predicted probabilities.

Conclusion We have developed and validated a new model that predicts the extent of nodal involvement in node positive breast cancer patients. This new tool will particularly be useful for counselling patients regarding their personalized axillary treatment.

INTRODUCTION

The axillary treatment of breast cancer has changed significantly over time. Originally, all patients were treated with an axillary lymph node dissection (ALND) for locoregional control and for further treatment recommendations. After the implementation of the sentinel lymph node biopsy (SLNB) procedure, only patients with sentinel lymph node (SLN) metastases were selected for treatment by a completion ALND. In 2013, the International Breast Cancer Study Group (IBCSG) 23-01 study showed that a completion ALND could be omitted for patients with micrometastases [1]. Furthermore, the American College Of Surgeons Oncology Group Z0011 trial showed that for a selected subgroup of patients, a small-volume disease left behind in the axilla does not compromise the oncological safety, in terms of recurrence and disease free and overall survival [2]. Also, the results of the AMAROS (After Mapping of the Axilla: Radiotherapy or Surgery?) trial, published in 2014, changed our perspective on axillary treatment showing that both radiotherapy as well as surgery can provide excellent regional control [3]. These studies have had a significant impact on the management of the axilla [4].

It is obvious that the trend is heading towards a less invasive surgical treatment of the axilla. ALND has lost its importance for determining the need of adjuvant systemic treatment and gradually seems to lose its importance for locoregional control. However, there remain several subgroups of breast cancer patients for whom treatment of the axilla may still be necessary. These include patients who were found to be node positive with ultrasound-guided lymph node biopsy (UGLNB). This appears to be a different group of node positive patients with less favourable disease characteristics and a worse disease free and overall survival as compared to those with SLN positive disease [6, 7]. Another group may be the patients who are treated with a mastectomy rather than breast conserving surgery (BCS), as radiotherapy may partially include the axilla when used as adjuvant treatment after BCS [8].

In the last decade, the focus was set on finding patients with SLN positive breast cancer with a *low risk* of additional nodal involvement, for whom a completion ALND could be omitted. Several predictive systems have been developed to help identifying such patients [9-17]. Now that also a *low risk* of *limited* nodal involvement is increasingly accepted to omit further axillary treatment, it is time to search for the patients at *high risk* for *extensive* nodal involvement who may still benefit from additional treatment of the axilla. Presently, three predictive models have been proposed, that predict the risk of having four or more positive axillary lymph nodes [18-20]. The main purpose of these models is to help decide on the extent of radiation and/or systemic therapy and whether an immediate breast reconstruction can be offered to these patients [18-20]. To our knowledge there is no model, that predicts the extent of nodal involvement. According to the Z0011 trial results, a completion ALND can be omitted for patients with limited nodal involvement (1-2 positive

SLNs) who underwent BCS [2]. However, there is no evidence that additional axillary treatment can also be omitted in patients with a high risk of having more than 2 positive lymph nodes, resulting in a new cut-off point for “extensive nodal involvement”.

The aim of the present study was to develop a tool for predicting the extent of nodal involvement in node positive breast cancer patients. Such a tool may then be used for counselling in the clinical decision making process, in the present “treat none – unless” era, regarding the additional axillary treatment strategies.

METHODS

Study population

The study population consisted of three groups. The original patient series for model development were identified from the Netherlands Cancer Registry of the South region of the Netherlands in which 10 hospitals participated. The dataset included breast cancer patients with SLN metastases who were treated between January 2007 and December 2008. For two of these hospitals, the Máxima Medical Center (MMC) and the Jeroen Bosch Hospital, data from the years 2000-2006 were also available as were data of MMC from the additional years of 2009-2011. The second group consisted of patients with SLN positive disease from the Gelderse Vallei Hospital and was used for external validation of the developed prediction model. The third group of patients were those found to have node positive disease by ultrasound-guided lymph node biopsy and who were treated at the MMC between January 2006 and December 2011 [6]. In accordance with Dutch guidelines [21] all patients had sonographic evaluation of the axilla after mammography and clinical evaluation. Ultrasound-guided lymph node biopsies (with cytological and/or histological sampling) were performed on suspicious axillary lymph nodes as previously described [6]. All patients included in the present study underwent a completion ALND. Patients receiving neo-adjuvant treatment, those with stage IV breast cancer and patients with a clinical N_{2-3} axillary status or without a completion ALND, were excluded from the study.

Data accrual

Data were collected from the existing database of the South region of the Netherlands Cancer Registry and from the patients’ medical charts and pathology reports. The following data were entered into the database: age at time of diagnosis, lateralization of the tumour, type of surgery (breast conserving surgery (BCS) or mastectomy), tumour morphology, tumour size in millimetres, histological grade (conform the modified Bloom and Richardson classification), the presence of lymphovascular invasion, multifocality of the tumour, estrogen and progesterone receptor status and Her2Neu status. Histopathological data of the lymph nodes included: total number of resected and total number of positive/negatives nodes for both the SLNB procedure and the ALND, the size of the largest metastases of the SLN as a continuous variable (in millimetres) and categorised as macrometastases (>2mm), micrometastases (>0.2 but ≤2 mm) or isolated tumor cells (ITC) (≤ 0.2mm) and the presence of extracapsular extension in the SLN.

The total number of axillary lymph nodes was computed by adding the total number of lymph nodes harvested during SLNB to the number of lymph nodes that was found by ALND. The total number of positive axillary lymph nodes was divided into 3 categories: 1-2, ≥ 3 or ≥ 4 positive lymph nodes. For tumour grade, lymphovascular invasion, hormonal receptor and Her2Neu status, the size of the SLN metastases and the presence of extracapsular extension there were some missing values. When values for lymphovascular invasion were missing and were not reported in the pathology report, it was assumed to be absent. Other missing values were multiply imputed with the *mice* algorithm in R software allowing all observed values to be analysed [22, 23]. The imputation model included all predictors and the total number of positive axillary lymph nodes.

Predictors and model development

Candidate predictors for axillary lymph node involvement were included in the statistical analysis based on the literature and previously reported models. These included the models that predict the risk of having four or more involved axillary lymph nodes [9-20]. We developed a model to predict the number of positive non-SLNs, and thus the total number of positive lymph nodes. We used proportional odds logistic regression analysis to model the association between predictors and the number of positive non-SLNs. The proportional odds model assumes that each predictor affects the log odds of having at least 1 positive non-SLN equally as it affects the log odds of having at least 2 (or any other number of) positive non-SLNs. We checked the proportional odds assumption graphically for all potential predictors. We evaluated the strength of each predictor by its univariable odds ratio, and by its multivariable odds ratio together with its likelihood ratio χ^2 test statistic minus twice the degrees of freedom (number of regression coefficients used to model a predictor). The latter is consistent with Akaike's Information Criterion (AIC) which balances the goodness-of-fit of a model with its complexity, and was also used to select predictors into a final model [24]. For the patients in the ultrasound guided lymph node biopsy group, the overall probabilities of having a total of 1-2, ≥3 or ≥4 positive lymph nodes were derived from the data.

Presentation of the prediction model

For easy calculation of the probability of having a particular number of positive lymph nodes we presented the final model with a score chart [24, 25]. The score chart is based on the regression coefficients of the final proportional odds model. Predictor values were translated into a sum score that can be used to read the probability of having a total of 1-2, ≥3 or ≥4 positive lymph nodes from a table, given the number of positive SLNs.

Validation of the new prediction model

We validated predictions of having ≥3 positive lymph nodes for patients with less than 3 positive SLNs within the development data and within the external validation data. We used validation plots to visualise the performance of the model to predict the extent of nodal involvement. A validation plot represents the actual proportion vs. the predicted probability by quartiles of increasing risk predictions [26]. The ability of our model to predict ≥4 positive lymph nodes was compared with the previously developed nomograms by Katz et al., Meretoja et al. and Chagpar et al. [18-20]. We assessed calibration of our prediction model by the calibration slope (the regression slope in a logistic regression model with solely the linear predictor from the prediction model) and the calibration-in-the-large (the logistic regression model intercept given that the calibration slope equals 1) [25]. We assessed discriminative ability of our prediction model by the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. Since we are assessing predictions of binary outcomes, the AUC is equal to the c-index, which estimates the probability that the risk prediction of a randomly chosen patient with the outcome (e.g. ≥3 positive lymph nodes) is higher than the risk prediction of a randomly chosen patient without the outcome (1-2 positive lymph nodes).

For proportional odds regression analysis and validation of prediction models we used R package rms. (version 2.13.1; R foundation for Statistical Computing, Vienna, Austria)

RESULTS

A total of 1230 patients with node positive breast cancer were included in this study. The model development population consisted of 911 patients, the external validation population of 180 and the ultrasound guided lymph node biopsy group of 139 patients. Of the 911 patients from the model development population 349 (38%) underwent a mastectomy and 562 (62%) were treated with breast conserving surgery. For 259 (28%) patients, the final pathology showed additional positive lymph nodes after completion ALND. Of these, 175 (19%) had a total of 3 or more involved axillary lymph nodes (Table 18.1). The observed overall proportions of patients in the ultrasound guided lymph node biopsy group with a total of 1-2, ≥3 or ≥4 positive lymph nodes were 37%, 63% and 51% respectively.

Table 18.1 Characteristics of total study population.

	Development group		External validation group		UGLNB group	
Variables	**N**	**%**	**N**	**%**	**N**	**%**
Age at diagnosis (years)						
≤ 50	297	32.6	59	32.8	42	30.2
>50 ≤65	363	39.8	68	37.8	38	27.3
>65	251	27.6	53	29.4	59	42.4
Laterality						
Left	473	51.9	93	51.7	78	56.1
Right	437	48.0	87	48.3	61	43.9
Missing	1	0.1	0	0	0	0
Type of surgery						
BCS	562	61.7	97	53.9	49	35.3
Mastectomy	349	38.3	83	46.1	90	64.7
Tumor size (mm)						
≤5	20	2.2	3	1.7	2	1.4
6-10	72	7.9	12	6.7	3	2.2
11-20	446	49.0	99	55.0	23	16.5
21-30	270	29.6	50	27.8	102	73.4
31-50	90	9.9	14	7.8	9	6.5
>50	13	1.4	2	1.1	0	0
Multifocal						
Yes	112	12.3	24	13.3	24	17.3
No	799	87.7	156	86.7	111	79.9
Missing	0	0	0	0	4	2.9
Morphology						
IDC	722	79.3	130	72.2	108	72.7
ILC	156	17.1	30	6.7	23	16.5
Other	33	3.6	20	11.1	8	5.8
Tumor grade						
1	263	28.8	34	18.9	23	16.5
2	421	46.2	89	49.4	75	54.0
3	155	17.0	52	28.9	38	27.3
Missing	72	7.9	5	2.8	3	2.2
LVI						
Yes	179	19.6	26	14.4	34	24.5
No	421	46.2	121	67.2	80	57.6
Missing	311	34.1	33	18.3	25	18.0
ER status						
Positive	780	85.6	146	81.1	100	71.9
Negative	114	12.5	34	18.9	39	28.1
Missing	17	1.9	0	0	0	0

Table 18.1 Continued.

Variables	Development group N	%	External validation group N	%	UGLNB group N	%
PR status						
Positive	673	73.9	120	66.7	80	57.6
Negative	183	20.1	60	33.3	59	42.4
Missing	55	6.0	0	0	0	0
Her2Neu status						
Positive	80	8.8	24	13.3	26	18.7
Negative	613	67.3	125	69.4	113	81.3
Missing	218	23.9	31	17.2	0	0
SLNs positive						
1	755	82.9	152	84.4		
2	124	13.6	23	12.8		
3	21	2.3	4	2.2		
>3	11	1.2	1	0.6		
SLN Metastases (mm)						
ITC (<0.2)	37	4.1	9	5.0		
Micro (0.2-2.0)	242	26.6	50	27.8		
Macro (>2.0)	485	53.2	118	65.6		
Missing	147	16.1	3	1.7		
ECE						
Yes	184	20.2	29	16.1		
No	650	71.3	86	47.8		
Missing	77	8.5	65	36.1		
Total number of positive lymph nodes						
1-2	736	80.8	148	82.2	51	36.7
≥3	175	19.2	32	17.8	88	64.0
≥4	111	12.2	24	13.3	71	51.1
Overall	911		180		139	

BCS = breast conserving surgery. IDC = invasive ductal carcinoma. ILC = invasive lobular carcinoma. SLN(s) = sentinel lymph node(s). LVI = lymphovascular invasion. ER = estrogen receptor status. PR = progesterone receptor status. ITC = isolated tumor cells. ECE = extracapsular extension.

Predictors and model development

The univariable analysis showed the following significant predictors for additional axillary lymph node involvement (Odds Ratio and 95% Confidence Interval): tumour size in mm, tumour grade, lymphovascular invasion, the presence of a negative SLN, >1 positive SLN, SLN macrometastases, a SLN metastases size >5 mm and the presence of extracapsular extension. Predictor effects of age >65 years mulitfocality of the tumour, invasive lobular carcinoma morphology, a positive ER status, a positive PR status and a positive Her2Neu status were non-significant but in the expected direction (Table 18.2).

Table 18.2 Univariable associations of the number of additional positive lymph nodes ≥ 1, 2, 3, respectively with predictors.

Predictor	Level	N	Frequency			%			Odds ratio	95% CI
			Additional positive lymph nodes							
			≥1	≥2	≥3	≥1	≥2	≥3		
Age at diagnosis (years)	≤ 50	297	89	51	29	30	17	10	1	
	>50 ≤65	363	104	56	40	29	15	11	0.94	0.68-1.31
	>65	251	66	31	18	26	12	7	0.81	0.56-1.17
Laterality (left)	No	437	130	70	42	30	016	10	1	
	Yes	473	129	68	45	27	14	10	0.89	0.67-1.18
Mastectomy	No	562	157	82	53	28	15	9	1	
	Yes	349	102	56	34	29	16	10	1.07	0.80-1.43
Tumor size (mm)	≤15	327	65	30	19	20	9	6	1	
	<15 ≤25	400	114	48	30	29	12	8	1.56	1.10-2.20
	>25	184	80	60	38	43	33	21	3.52	2.38-5.22
Multifocal	No	799	223	117	72	28	15	9	1	
	Yes	112	36	21	15	32	19	13	1.26	0.83-1.92
Morphology (ILC)	No	755	210	104	65	28	14	9	1	
	Yes	156	49	34	22	31	22	14	1.29	0.89-1.87
Tumor grade	1	263	61	30	21	23	11	8	1	
	2	421	134	73	43	32	17	10	1.54	1.09-2.18
	3+	155	54	30	21	35	19	14	1.84	1.21-2.80
LVI	No	732	196	100	60	27	14	8	1	
	Yes	179	63	38	27	35	21	15	1.55	1.10-2.18
ER positive	No	114	36	22	14	32	19	12	1	
	Yes	780	223	116	73	29	15	9	0.84	0.55-1.27
PR positive	No	183	56	31	21	31	17	11	1	
	Yes	673	194	104	64	29	15	10	0.91	0.64-1.29
Her2Neu positive	No	613	158	88	54	26	14	9	1	
	Yes	80	30	14	8	38	18	10	1.60	1.00-2.56
SLNs negative	0	518	179	101	65	35	19	13	1	
	1	260	54	24	16	21	9	6	0.49	0.34-0.69
	2+	133	26	13	6	20	10	5	0.45	0.29-0.72
SLNs positive	1	755	195	96	58	26	13	8	1	
	2+	156	64	42	29	41	27	19	2.15	1.52-3.05
Macrometastases	No	279	36	12	5	13	4	2	1	
	Yes	485	177	102	65	36	21	13	4.05	2.73-6.00
Size of SLN metastases (mm)	≤2	218	24	7	3	11	3	1	1	
	>2 ≤5	183	36	13	9	20	7	5	1.99	1.14-3.47
	>5	207	97	58	32	47	28	15	7.56	4.58-12.47
ECE	No	651	156	77	47	24	12	7	1	
	Yes	184	85	51	35	46	28	19	2.78	2.00-3.86
Overall		911	259	138	87	28	15	10		

ILC = invasive lobular carcinoma. SLN(s) = sentinel lymph node(s). LVI = lymphovascular invasion. ER = estrogen receptor status. PR = progesterone receptor status. ECE = extracapsular extension.

The proportional odds assumption was well satisfied upon graphical inspection. The effects of the predictors were reasonably constant across any cut-off level for the extent of lymph node positivity (Supplementary figure 18.1, constant horizontal distance between any two of the three symbols).

The most important predictors in the multivariable analysis were (Odds Ratio and 95% Confidence Interval): tumor size in mm (OR 1.04 95%CI 1.02-1.05) the presence of a negative SLN (OR 0.48 95%CI 0.35-0.67), the size of the SLN metastases (in mm) (OR 1.17 95%CI 1.13-1.22) and the presence of extracapsular extension (OR 1.50 95%CI 1.01-2.25) as shown in Table 18.3.

Table 18.3 Multivariable associations of predictors for having a total of 1-2, ≥3 or ≥4 positive lymph nodes and model selection.

Predictor	Odds ratio	95% CI	χ^2	df	p-value
Full model					
Age at diagnosis (10 years)	0.94	0.83-1.07	0.8	1	0.3638
Laterality (left)	0.96	0.70-1.31	0.1	1	0.8057
Mastectomy	0.77	0.55-1.08	2.2	1	0.1362
Tumor size (mm)	1.04	1.02-1.05	20.7	1	<.0001
Multifocal	1.18	0.73-1.91	0.5	1	0.4947
Morphology (ILC)	1.16	0.75-1.81	0.4	1	0.5068
Tumor grade 2:1	1.42	0.96-2.11	4.2	2	0.1216
Tumor grade 3:1	1.65	0.98-2.79			
LVI	1.14	0.78-1.69	0.5	1	0.4946
ER positive	0.92	0.52-1.65	0.1	1	0.7883
PR positive	1.09	0.68-1.77	0.1	1	0.7128
Her2Neu positive	1.35	0.82-2.23	1.4	1	0.2437
SLNs negative	0.47	0.34-0.66	19.3	1	<.0001
SLNs positive	1.41	0.95-2.10	2.9	1	0.0899
Size of SLN metastases (mm)	1.15	1.10-1.21	41.1	1	<.0001
ECE	1.50	1.01-2.25	4.0	1	0.0466
Selected model					
Tumor size (mm)	1.04	1.02-1.05	26.6	1	<.0001
SLNs negative	0.48	0.35-0.67	19.3	1	<.0001
Size of SLN metastases (mm)	1.17	1.13-1.22	63.5	1	<.0001

ILC = invasive lobular carcinoma. SLN(s) = sentinel lymph node(s). LVI = lymphovascular invasion. ER = estrogen receptor status. PR = progesterone receptor status. ECE = extracapsular extension.

Presentation of the prediction model

In the final model the following three predictors were selected: the size of the tumor in mm, the presence of a negative SLN and the size of the SLN metastases in mm (Table 18.3). The model is presented as a simple score chart (Figure 18.1) and is also provided as an easy to use online (excel –based) calculator (Figure 18.2). For example, a patient with a SLN metastasis size of 8 mm, a tumour size of 25 mm and no negative SLN, has predicted probabilities of 73%, 27% and 17% of having a total of 1-2, ≥3 or ≥4 positive lymph nodes, respectively.

Sum score =

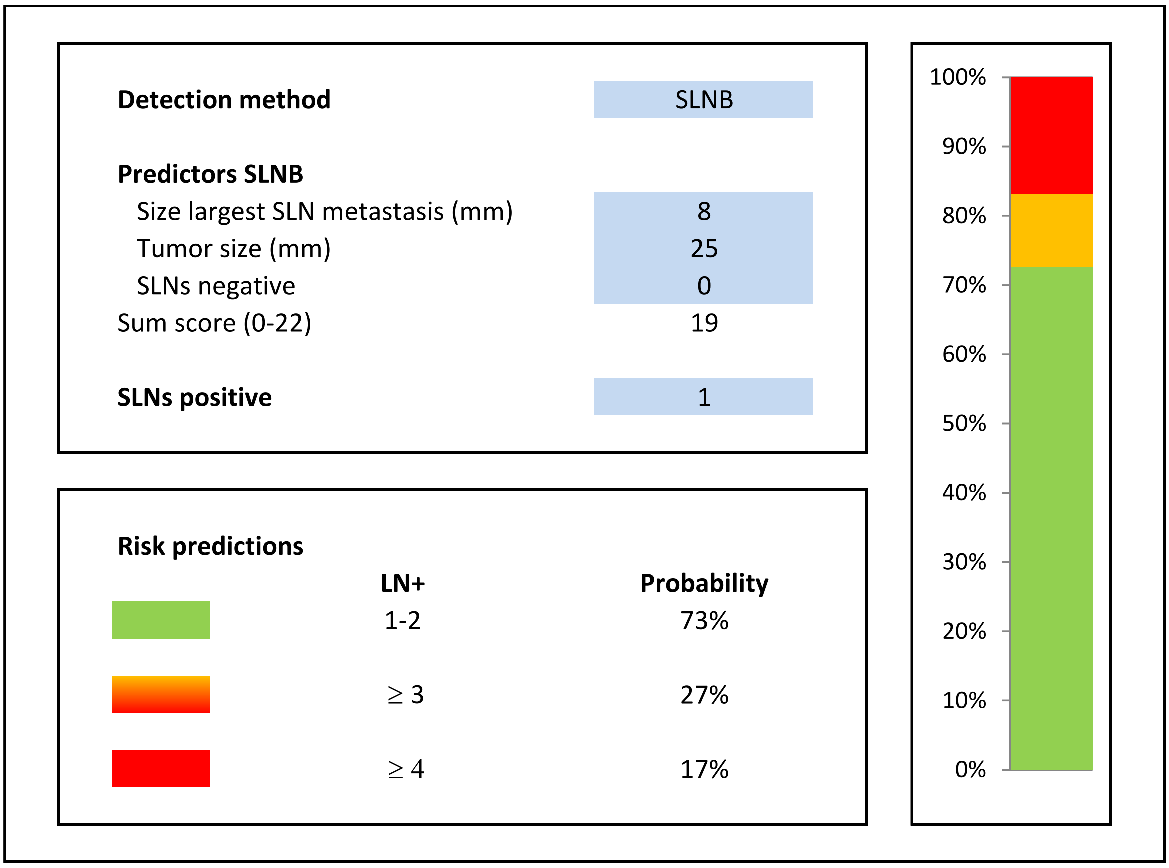
Size largest SLN metastasis (mm)
+ 0.25 x tumor size (mm)
+ 5 if no negative SLN was found

Sum	LN+ when 1 SLN+			LN+ when 2 SLN+			LN+ when 3 SLN+		
score	**1-2**	**≥3**	**≥4**	**1-2**	**≥3**	**≥4**	**1-2**	**≥3**	**≥4**
<1	98	1.8	1.0	95	4.6	1.8	0	100	4.6
1	98	2.1	1.2	95	5.4	2.1	0	100	5.4
2	98	2.5	1.4	94	6.2	2.5	0	100	6.2
3	97	2.9	1.6	93	7.2	2.9	0	100	7.2
4	97	3.4	1.8	92	8.3	3.4	0	100	8.3
5	96	3.9	2.1	90	9.6	3.9	0	100	9.6
6	95	4.6	2.5	89	11	4.6	0	100	11
7	95	5.3	2.9	87	13	5.3	0	100	13
8	94	6.1	3.4	86	14	6.1	0	100	14
9	93	7.1	3.9	83	17	7.1	0	100	17
10	92	8.2	4.6	81	19	8.2	0	100	19
11	91	9.5	5.3	79	21	9.5	0	100	21
12	89	11	6.2	76	24	11	0	100	24
13	88	12	7.1	73	27	12	0	100	27
14	86	14	8.2	70	30	14	0	100	30
15	84	16	9.5	66	34	16	0	100	34
16	81	19	11	63	37	19	0	100	37
17	79	21	13	59	41	21	0	100	41
18	76	24	14	55	45	24	0	100	45
19	73	27	16	51	49	27	0	100	49
20	70	30	19	47	53	30	0	100	53
21	67	33	21	44	56	33	0	100	56
22+	63	37	24	40	60	37	0	100	60

Figure 18.1 Score chart for the probability of finding a total of 1-2, ≥3 or ≥4 positive lymph nodes. LN+ = lymph node(s) positive. SLN = sentinel lymph node.

Validation of the prediction model

The model predictions of having ≥3 positive lymph nodes (for patients with less than 3 positive SLNs) were validated within the development data and the external validation data (Figure 18.3). In both the apparent and external validation, the model showed a very good discriminative ability with AUCs of 0.80 (95% CI 0.76-0.84) and 0.82 (95% CI 0.74-0.90), respectively. Calibration was good over the complete range of predicted probabilities in both the apparent and external validation (Figure 18.3). When the actual size of the SLN metastases (in mm) is not provided, assigning 8 points for macrometastases gives a good approximation. The performance of the model then remained satisfactory with an AUC of 0.79 (95% CI 0.75-0.83) in apparent validation and 0.80 (95% CI 0.72-0.88) at external validation (Supplementary figure 18.2).



Detection method	SLNB
Predictors SLNB	
Size largest SLN metastasis (mm)	8
Tumor size (mm)	25
SLNs negative	0
Sum score (0-22)	19
SLNs positive	1

Risk predictions

	LN+	**Probability**
🟩	1-2	73%
🟧	≥ 3	27%
🟥	≥ 4	17%

Figure 18.2 Print screen of online tool to predict the extent of nodal involvement. Example of predicted risks for having a total of 1-2, ≥3 or ≥4 lymph nodes positive for a patient with a sentinel node metastasis size of 8 mm, a tumor size of 25 mm and no negative sentinel node.

Comparison with previously developed models

When predicting the probability of having ≥4 positive lymph nodes (for patients with less than 4 positive SLNs), the discriminative ability of our new model was equally good or even better in the external validation data (AUC 0.82 95%CI 0.74-0.90) as compared to the three previously developed predictive systems (AUC's of 0.82, 0.80 and 0.66 respectively) [18-20]. Furthermore, calibration was also superior for the newly developed prediction model. (Supplementary Figure 18.3)

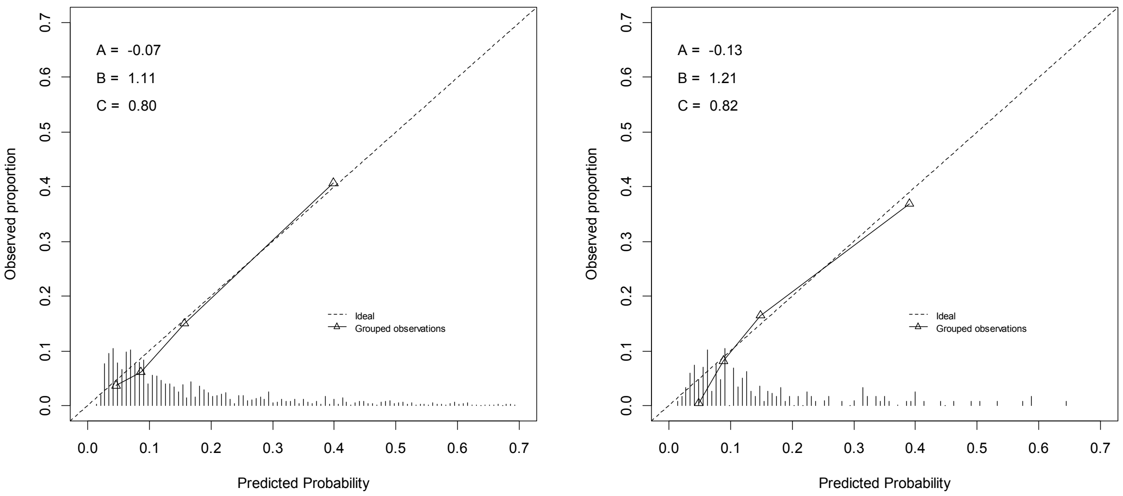


Figure 18.3 Internal (left) and external (right) validation of the predicted probability of ≥3 positive lymph nodes (when the number of positive SLNs is <3). The distribution of predicted risks for ≥3 or more positive axillary lymph nodes is shown at the bottom of the graphs. The triangles indicate the observed proportions by quartiles of predicted risks.

DISCUSSION

Early stage breast cancer patients, with limited nodal involvement are no longer subjected to a completion ALND, based on the results of the IBCSG 23-01, AMAROS and Z0011 trials [1-3]. Because the evidence for omitting further axillary treatment in patients with extensive nodal involvement is lacking, it is useful to predict the extent of nodal involvement. This will enable recommendations concerning a personalized treatment plan for the axilla.

We have developed a novel model that predicts the risk of having a total of 1-2, ≥3 or ≥4 positive lymph nodes using only three predictors: tumor size (in mm), the presence of a negative SLN and the size of the SLN metastases (in mm). Although the presence of extracapsular extension also showed to be a significant predictor in both univariable and multivariable analysis we chose to incorporate only the three strongest predictors. This did not affect model performance and was in line with the aim of this study to keep the model as simple and user-friendly as possible. The discriminative ability of the model was good (AUC 0.80) and it showed adequate calibration over the complete range of predicted probabilities. Furthermore, the model was validated in an independent external patient population and showed good discrimination (AUC 0.82) and calibration.

Because we wanted to provide risk predictions for all breast cancer patients who were found node positive, regardless of the method of detection, a group of patients that were found positive by ultrasound guided lymph node biopsy was analyzed. However, when the lymph node metastases are detected by ultrasound, this staging method by itself seems to be the most important predictor for extensive nodal involvement given the fact that 63% of these patients had ≥3 positive lymph nodes. These findings are in concordance with the results of a study from Schipper et al, that showed that the finding of suspicious nodes with ultrasound resulted in pN2-pN3 disease in 41.2% of the patients [27]. For this group of patients we found no additional predictors that could further discriminate between low risk and high risk patients. Consequently, we decided that the development of a separate model for this group was not relevant. Because these patients also need counseling regarding the axillary treatment strategy, the overall risk estimates of having a total of 1-2, ≥3 or ≥4 positive lymph nodes are also visualized in our prediction tool.

Previously published models were mostly developed prior to the publication of the Z0011 trial results and are intended to identify patients at low risk for additional nodal involvement for whom a completion ALND could be omitted [9-14, 16, 17]. The few models that have been designed to predict the risk of having ≥ 4 involved lymph nodes were used to guide decisions on the extent of radiation and systemic therapy regimens [18-20]. To our knowledge our model is the first that actually predicts the extent of nodal involvement, therefore easily classifying patients to have limited nodal involvement (1-2 positive lymph nodes, corresponding to the conclusions of Z0011) or extensive lymph node involvement (≥3 positive lymph nodes). Although it was not a primary goal of this study, our model also predicts the risk of having ≥4 positive lymph nodes. This is another cut-off point for extensive nodal involvement, that is used to decide about the need for additional axillary irradiation. The new model was compared with the existing three models that predict ≥4 positive lymph nodes and outperformed the other models in terms of discriminative ability and calibration for our population.

The variables included in our new prediction tool already proved to be strongly associated with nodal involvement in other prediction models and validation studies. Furthermore, other previously reported models included up to 9 variables, resulting in more complex calculations and perhaps more importantly, in a less user-friendly model [15]. The present study shows that model performance can still be very good when only a few but strongly prognostic variables are included.

Our study has limitations. A retrospectively collected database was used that contained some missing values. The size of the SLN metastases (in mm) is a strong predictor for extensive nodal involvement, however, the actual size was not always provided. Although our model also works based on the presence or absence of a macrometastasis in the SLN, with acceptable model performance, we recommend that the actual size of the SLN metastases should be reported consistently by pathologists to enable more accurate predictions. The risk predictions for patients that were found node positive by ultrasound guided lymph node biopsy are based on a relatively small sample size. Therefore, we are currently investigating whether these risks will be similar in a larger population.

Some clinicians have already abandoned the use of ALND for SLN positive patients and even so the prediction models to identify these patients. However, we must be aware for generalization of the conclusions of the Z0011-trial, as these are only applicable to about 6% of the total breast cancer population [28]. An advantage of our new prediction model is that the online tool can visualize the predicted risk of the extent of nodal disease for each individual patient. Therefore the model can be a useful tool in counseling patients, to help them understand their risks. Together with their doctor they can subsequently decide what the best personalized axillary treatment would be for them. Because the model gives no actual treatment recommendations, or a given cut-off point, the risks and benefits of further axillary treatment need to be weighed individually. Following the results of the Z0011 and AMAROS trials, it appears reasonable to give no further axillary treatment to patients that are very likely to have limited nodal involvement (1-2 positive lymph nodes) provided that adjuvant systemic treatment is offered, and consider radiation therapy of the axilla or a completion ALND when they are at risk for extensive nodal involvement (≥3 positive lymph nodes). We strongly advise against the omission of further axillary treatment for patients with a high risk of having ≥4 positive lymph nodes.

CONCLUSION

We have developed and validated a new model that predicts the extent of nodal involvement in node positive breast cancer patients. This new tool “Maxy-Risk Score (Maxima Axillary Risk score) will particularly be useful for counseling patients regarding their personalized axillary treatment.

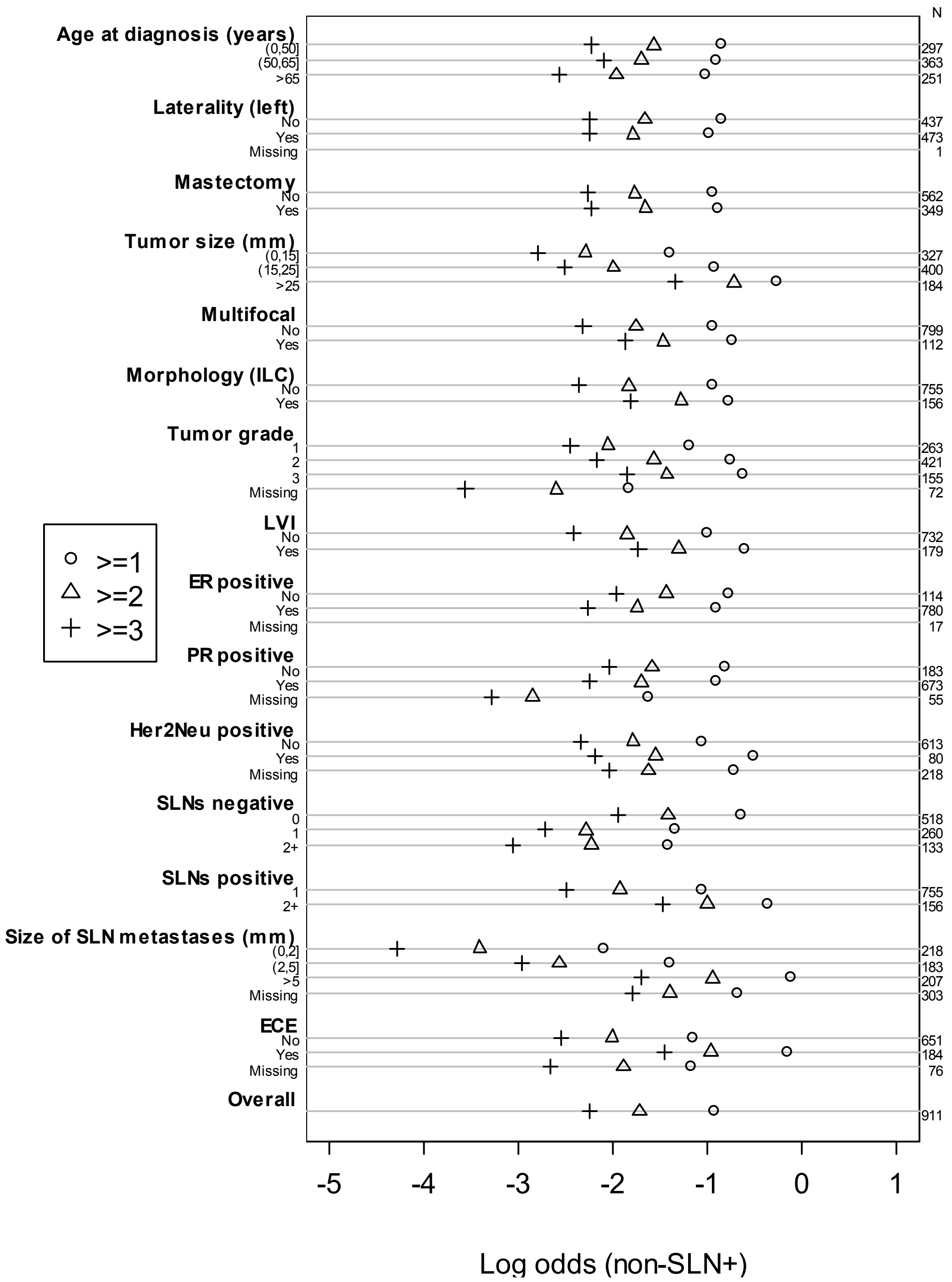
ACKNOWLEDGEMENTS

D. van Klaveren is funded by the Netherlands Organization for Scientific Research (Grant 917.11.383).

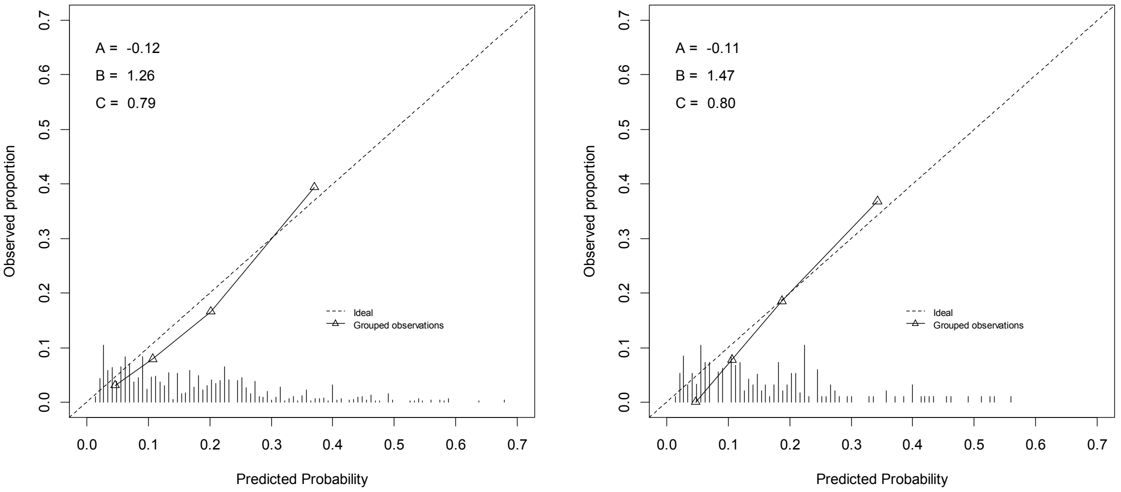
REFERENCES

1. Galimberti V, Cole BF, Zurrida S, Viale G, Luini A, Veronesi P, et al. Axillary dissection versus no axillary dissection in patients with sentinel-node micrometastases (IBCSG 23-01): a phase 3 randomised controlled trial. Lancet Oncol. 2013 Apr;14(4):297-305.
2. Giuliano AE, McCall L, Beitsch P, Whitworth PW, Blumencranz P, Leitch AM, et al. Locoregional recurrence after sentinel lymph node dissection with or without axillary dissection in patients with sentinel lymph node metastases: the American College of Surgeons Oncology Group Z0011 randomized trial. Ann Surg. 2010 Sep;252(3):426,32; discussion 432-3.
3. Donker M, van Tienhoven G, Straver ME, Meijnen P, van de Velde CJ, Mansel RE, et al. Radiotherapy or surgery of the axilla after a positive sentinel node in breast cancer (EORTC 10981-22023 AMAROS): a randomised, multicentre, open-label, phase 3 non-inferiority trial. Lancet Oncol. 2014 Nov;15(12):1303-10.
4. Beek MA, Verheuvel NC, Luiten EJT, Klompenhouwer EG, Rutten HJT, Roumen RMH, et al. Two decades of axillary management in breast cancer. Br J Surg. 2015;102:1658-64.
5. van Roozendaal LM, Schipper RJ, Van de Vijver KK, Haekens CM, Lobbes MB, Tjan-Heijnen VC, et al. The impact of the pathological lymph node status on adjuvant systemic treatment recommendations in clinically node negative breast cancer patients. Breast Cancer Res Treat. 2014 Feb;143(3):469-76.
6. Verheuvel NC, van den Hoven I, Ooms HW, Voogd AC, Roumen RM. The role of ultrasound-guided lymph node biopsy in axillary staging of invasive breast cancer in the post-ACOSOG Z0011 trial era. Ann Surg Oncol. 2015 Feb;22(2):409-15.
7. Caudle AS, Kuerer HM, Le-Petross HT, Yang W, Yi M, Bedrosian I, et al. Predicting the Extent of Nodal Disease in Early-Stage Breast Cancer. Ann Surg Oncol. 2014 May 24.
8. van Wely BJ, van den Wildenberg FJ, Gobardhan P, van Dalen T, Borel Rinkes IH, Theunissen EB, et al. "Axillary recurrences after sentinel lymph node biopsy: a multicentre analysis and follow-up of sentinel lymph node negative breast cancer patients". Eur J Surg Oncol. 2012 Oct;38(10):925-31.
9. Van Zee KJ, Manasseh DM, Bevilacqua JL, Boolbol SK, Fey JV, Tan LK, et al. A nomogram for predicting the likelihood of additional nodal metastases in breast cancer patients with a positive sentinel node biopsy. Ann Surg Oncol. 2003 Dec;10(10):1140-51.
10. Degnim AC, Reynolds C, Pantvaidya G, Zakaria S, Hoskin T, Barnes S, et al. Nonsentinel node metastasis in breast cancer patients: assessment of an existing and a new predictive nomogram. Am J Surg. 2005 Oct;190(4):543-50.
11. Hwang RF, Krishnamurthy S, Hunt KK, Mirza N, Ames FC, Feig B, et al. Clinicopathologic factors predicting involvement of nonsentinel axillary nodes in women with breast cancer. Ann Surg Oncol. 2003 Apr;10(3):248-54.
12. Barranger E, Coutant C, Flahault A, Delpech Y, Darai E, Uzan S. An axilla scoring system to predict non-sentinel lymph node status in breast cancer patients with sentinel lymph node involvement. Breast Cancer Res Treat. 2005 May;91(2):113-9.
13. Coufal O, Pavlik T, Fabian P, Bori R, Boross G, Sejben I, et al. Predicting non-sentinel lymph node status after positive sentinel biopsy in breast cancer: what model performs the best in a Czech population? Pathol Oncol Res. 2009 Dec;15(4):733-40.
14. Gur AS, Unal B, Ozbek U, Ozmen V, Aydogan F, Gokgoz S, et al. Validation of breast cancer nomograms for predicting the non-sentinel lymph node metastases after a positive sentinel lymph node biopsy in a multi-center study. Eur J Surg Oncol. 2010 Jan;36(1):30-5.
15. Meretoja TJ, Leidenius MH, Heikkila PS, Boross G, Sejben I, Regitnig P, et al. International multicenter tool to predict the risk of nonsentinel node metastases in breast cancer. J Natl Cancer Inst. 2012 Dec 19;104(24):1888-96.

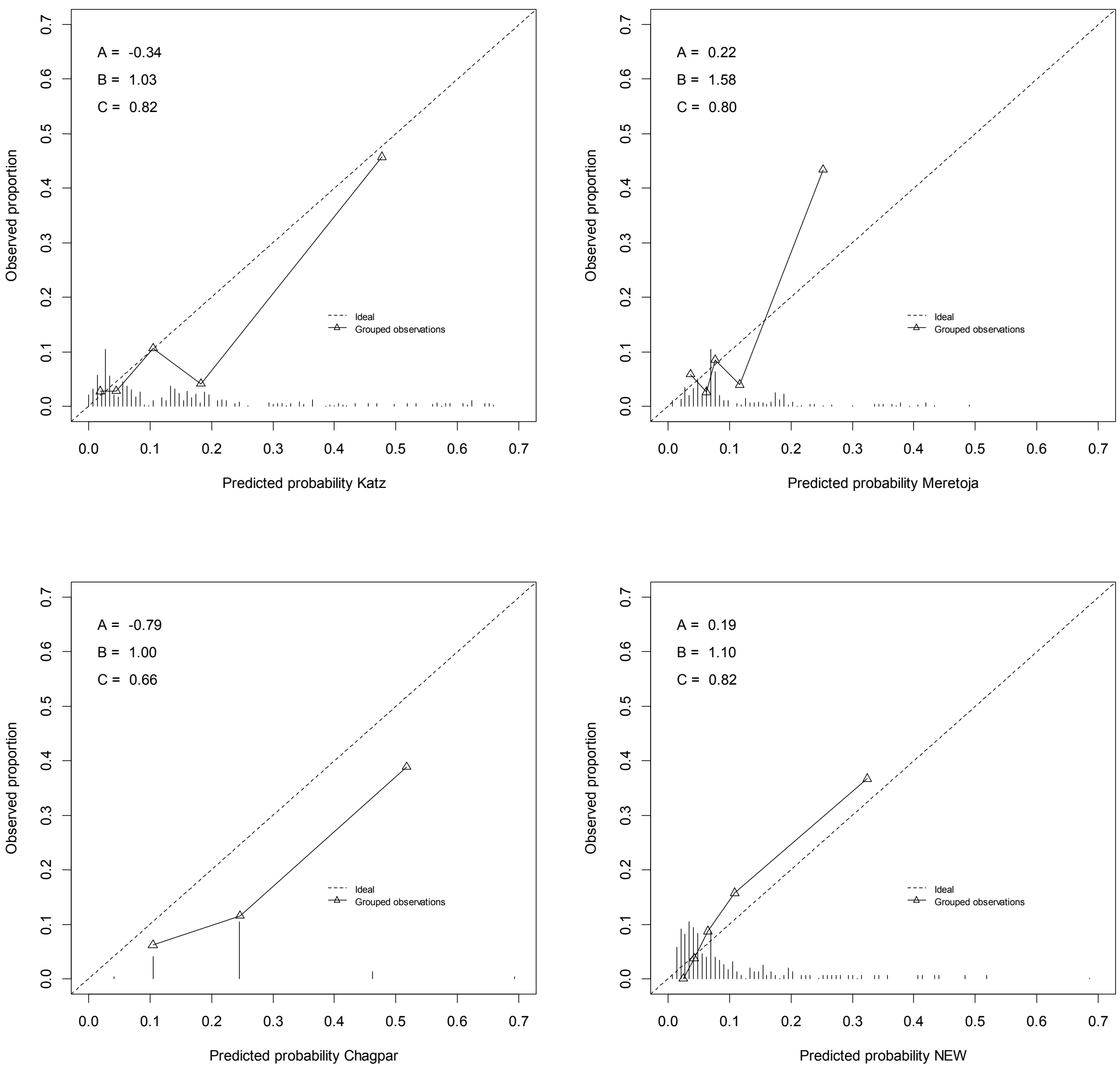
16. Pal A, Provenzano E, Duffy SW, Pinder SE, Purushotham AD. A model for predicting non-sentinel lymph node metastatic disease when the sentinel lymph node is positive. Br J Surg. 2008 Mar;95(3):302-9.
17. Kohrt HE, Olshen RA, Bermas HR, Goodson WH, Wood DJ, Henry S, et al. New models and online calculator for predicting non-sentinel lymph node status in sentinel lymph node positive breast cancer patients. BMC Cancer. 2008 Mar 4;8:66,2407-8-66.
18. Katz A, Smith BL, Golshan M, Niemierko A, Kobayashi W, Raad RA, et al. Nomogram for the prediction of having four or more involved nodes for sentinel lymph node-positive breast cancer. J Clin Oncol. 2008 May 1;26(13):2093-8.
19. Meretoja TJ, Audisio RA, Heikkila PS, Bori R, Sejben I, Regitnig P, et al. International multicenter tool to predict the risk of four or more tumor-positive axillary lymph nodes in breast cancer patients with sentinel node macrometastases. Breast Cancer Res Treat. 2013 Apr;138(3):817-27.
20. Chagpar AB, Scoggins CR, Martin RC,2nd, Cook EF, McCurry T, Mizuguchi N, et al. Predicting patients at low probability of requiring postmastectomy radiation therapy. Ann Surg Oncol. 2007 Feb;14(2):670-7.
21. Richtlijn mammacarcinoom 2.0 [Internet]. Available from: http://www.oncoline.nl/mammacarcinoom.
22. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med. 1991 Apr;10(4):585-98.
23. Multivariate Imputation by Chained Equations. [Internet]. Available from: http://www.jstatsoft.org/v45/i03/paper.
24. Harrell F. Regression Modeling Strategies: With applications to linear models, logistic regression, and survival analysis. Springer Series in Statistics; 2001.
25. Steyerberg EW, editor. Clinical prediction models A practical approach to development, validation, and updating. 1th ed. Springer; 2009.
26. Harrell FE,Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996 Feb 28;15(4):361-87.
27. Schipper RJ, van Roozendaal LM, de Vries B, Pijnappel RM, Beets-Tan RG, Lobbes MB, et al. Axillary ultrasound for preoperative nodal staging in breast cancer patients: is it of added value? Breast. 2013 Dec;22(6):1108-13.
28. Guth U, Myrick ME, Viehl CT, Schmid SM, Obermann EC, Weber WP. The post ACOSOG Z0011 era: does our new understanding of breast cancer really change clinical practice? Eur J Surg Oncol. 2012 Aug;38(8):645-50.



Supplementary figure 18.1 Checking the proportional odds assumption. The circle, triangle, and plus sign correspond to the number of additional positive lymph nodes ≥ 1, 2, 3, respectively. The proportional assumption holds when the distances between any two of these three symbols are vertically constant. ILC = invasive ductal carcinoma. LVI = lymphovascular invasion. ER = estrogen receptor status. PR = progesterone receptor status. SLN = sentinel lymph node. ECE = extracapsular extension.



Supplementary figure 18.2 Assigning 8 points to macrometastases instead of using continuous metastases size: internal (left) and external (right) validation of the predicted probability of ≥ 3 positive lymph nodes (when the number of positive SLNs is <3). The distribution of predicted risks for ≥3 or more positive axillary lymph nodes is shown at the bottom of the graphs. The triangles indicate the observed proportions by quartiles of predicted risks.



Supplementary figure 18.3 External validation of models predicting the risk of ≥4 positive lymph nodes (when the number of positive SLNs is < 4) (n=180). The distribution of predicted risks for ≥4 or more positive axillary lymph nodes is shown at the bottom of the graphs. The triangles indicate the observed proportions by quartiles of predicted risks.

19

Prediction of Chlamydia trachomatis infection to facilitate selective screening on population and individual level: a cross-sectional study of a population-based screening programme

D van Klaveren*
HM Götz*
ELM Op de Coul
EW Steyerberg
Y Vergouwe

Sexually Transmitted Infections 2016
* Equal contribution

ABSTRACT

Objectives To develop prediction models for *Chlamydia trachomatis* (Ct) infection with different levels of detail in information, i.e. from readily available data in registries and from additional questionnaires.

Methods All inhabitants of Rotterdam and Amsterdam aged 16-29 were invited yearly from 2008 until 2011 for home-based testing. Their registry data included gender, age, ethnicity and neighbourhood-level socioeconomic status (SES). Participants were asked to fill in a questionnaire on education, STI history, symptoms, partner information and sexual behaviour. We developed prediction models for Ct infection using first-time participant data – including registry variables only and with additional questionnaire variables – by multilevel logistic regression analysis to account for clustering within neighbourhoods. We assessed the discriminative ability by the area under the receiver operating characteristic curve (AUC).

Results Four percent (3,540/80,385) of the participants was infected. The strongest registry predictors for Ct infection were young age (especially for women) and Surinamese, Antillean or Sub-Sahara-African ethnicity. Neighbourhood-level SES was of minor importance. Strong questionnaire predictors were: low to intermediate education level, ethnicity of the partner (non-Dutch), and having sex with casual partners. When using a prediction model including questionnaire risk factors (AUC 0.74, 95%CI 0.736-0.752) for selective screening, 48% of the participating population needed to be screened to find 80% (95%CI 78.4-81.0%) of Ct infections. The model with registry risk factors only (AUC 0.67, 95%CI 0.656-0.675) required 60% to be screened to find 78% (95%CI 76.6-79.4%) of Ct infections.

Conclusions A registry based prediction model can facilitate selective Ct screening at population level, with further refinement at the individual level by including questionnaire risk factors.

KEY MESSAGES

- We developed easily applicable Ct prediction models with data from a large Chlamydia screening project in urban areas (>80,000 participants).
- The newly developed Ct prediction models performed substantially better than a previously proposed Ct prediction tool.
- A registry based prediction model can facilitate selective Ct screening at population level, with further refinement at the individual level by including questionnaire risk factors.

BACKGROUND

Chlamydia trachomatis (Ct) infection is the most common bacterial sexually transmitted infection (STI) in Europe and other Western countries, especially in young people [1]. Repeated infections occur due to no or limited development of immunity, untreated sexual partners or new sexual exposure after treatment. This mostly asymptomatic infection is a public health threat, in some cases leading to serious adverse events, such a pelvic inflammatory disease, tubal pathology and infertility [2], and premature labour [3]. To detect asymptomatic infections for prevention of potential adverse events, screening is the intervention of choice although good evidence to support the cost-effectiveness of screening is still lacking [4]. In the US opportunistic screening is advised for women under the age of 25, likewise in the UK where the National Chlamydia Screening Programme advises screening for men and women under 25, and screening trials have been performed or are ongoing in other countries [5-7]. Selective systematic screening, i.e. screening of subjects identified to be at high risk, may be favourable for cost effectiveness, and results in fewer individuals undergoing an unnecessary test [7, 8]. To prevent unacceptable high proportions of missed infections it is crucial that the prediction model performs well.

In 2005 a prediction rule was developed as a tool for selective screening of Ct [9]. In 2008 this model was used in a large population based screening program to select participants at high risk in a less urban region [7, 10-12]. In urban areas, they chose to invite all sexually active individuals in the target group (men and women aged 16-29 years) to participate without further selection, because of the high Ct prevalence that was previously found in highly urban areas (4.2%) [9].

In the current study, we validate the Ct prediction model with the data from the screening program as a benchmark for predictive performance in urban areas [12, 13]. Furthermore, we aim to develop improved Ct prediction models with different levels of detail in information, i.e. with readily available registries only and with additional detailed questionnaires.

METHODS

Study Population

The data of this study were collected in the Chlamydia Screening Implementation (CSI) program. The CSI program was approved by a Medical Ethics Committee of the VUmc in Amsterdam (METc number: 2007/239) and described in detail before [12]. In summary, all inhabitants aged 16-29 of Rotterdam, Amsterdam and selected municipalities of South Limburg were invited yearly from 2008 until 2011 for home-based testing (men: urine sample; women: vaginal swab or urine sample). In the first screening round 261,025 individuals were invited of whom 41,638 effectively participated. We selected all first-time participants from Rotterdam and Amsterdam, resulting in 80,385 unique participants of whom 3,440 were infected with Ct. All participants were asked to fill in a questionnaire on education, STI history, symptoms, partner information and sexual behaviour (variable definitions in Supplementary table 1). Information on gender, age, country of birth of both the participants and their parents and residential postcode was gathered from communal registries. Ethnicity was based on the country of birth of the participant and the participant's parents, consistent with ethnicity definitions used in Dutch STI clinics. In case of a regular partner, we defined the variable 'ethnicity mixing' by all four combinations of Dutch and non-Dutch ethnicity of the participant and the regular partner. Neighbourhood-level socioeconomic status (SES) scores were based on 4-digit postcode as provided by the Netherlands Institute for Social Research (available at www.scp.nl). We used the SES score of 2010.

Multiple imputation of missing values

We used an advanced multiple imputation strategy (method of chained equations) to account for missing values [14]. General questionnaire information (education, partner information and sexual behaviour) was available in approximately 57% of the participants and more specific questionnaire information on STI history and symptoms in approximately 20% (Table 19.1). We compared results without and with imputation of missing values. We used R package mice for multiple imputation [15].

Development of improved Chlamydia prediction models

We used logistic regression analysis for the development of three prediction models: registry risk factors gender, age and ethnicity only (model 1); model 1 with neighbourhood-level SES (model 2); and model 1 with additional questionnaire variables (model 3; Supplementary table 1). Regular partnership status, ethnicity mixing with regular partner, condom use with either regular or casual partners were newly studied variables in comparison with the pilot study [9].

Table 19.1 Univariable associations between Chlamydia trachomatis infection and risk factors for the Chlamydia Screening Implementation.

	Participants	Positive	Prevalence	Odds ratio	95% CI
Gender	80,385	3,540	4.4		
Men	25,840	1,045	4.0	1.0	
Women	54,545	2,495	4.6	1.1	1.1-1.2
Age group (years)	80,385	3,540	4.4		
15-19	11,623	844	7.3	2.6	2.4-2.9
20-24	31,042	1,603	5.2	1.8	1.7-2.0
25-29	37,720	1,093	2.9	1.0	
Ethnicity	80,383	3,540	4.4		
Dutch	47,474	1,488	3.1	1.0	
Turkish/North-African	4,972	193	3.9	1.2	1.1-1.5
Surinamese/Antillean	10,101	1,017	10.1	3.5	3.2-3.8
African other	2,944	252	8.6	2.9	2.5-3.3
Non-western other	7,524	320	4.3	1.4	1.2-1.6
Western other	7,368	270	3.7	1.2	1.0-1.3
SES (postcode)	80,380	3,540	4.4		
High	23,279	738	3.2	1.0	
Average to high	17,665	701	4.0	1.3	1.1-1.4
Low to average	22,699	1,176	5.2	1.7	1.5-1.8
Low	16,737	925	5.5	1.8	1.6-2.0
Education	48,398	2,476	5.1		
Low	2,103	192	9.1	2.9	2.5-3.4
Intermediate	11,873	1,139	9.6	3.1	2.8-3.4
High	34,422	1,145	3.3	1.0	
Partners in previous 6 months	46,508	2,384	5.1		
1	32,479	1,222	3.8	1.0	
2-5	12,741	1,026	8.1	2.2	2.1-2.4
>=6	1,288	136	10.6	3.0	2.5-3.6
New contacts in previous 2 months	42,487	2,232	5.3		
Yes	10,077	807	8.0	1.9	1.7-2.1
Regular partner	47,687	2,449	5.1		
No	19,268	1,216	6.3	2.4	2.1-2.7
Yes, living separately	15,924	888	5.6	2.1	1.8-2.4
Yes, living together	12,495	345	2.8	1.0	
Ethnicity mixing with regular partner	28,074	1,217	4.3		
Dutch, Dutch partner	15,458	342	2.2	1.0	
Dutch, non-Dutch partner	2,067	147	7.1	3.4	2.8-4.1
Non-Dutch, Dutch partner	5,192	207	4.0	1.8	1.5-2.2
Non-Dutch, non-Dutch partner	5,357	521	9.7	4.8	4.1-5.5
Condom use with regular partner	27,798	1,207	4.3		
No	23,090	1,039	4.5	1.3	1.1-1.5
Casual partners	46,927	2,412	5.1		
Yes	17,445	1,328	7.6	2.2	2.0-2.3
Condom use with casual partner	17,338	1,324	7.6		
No	9,179	838	9.1	1.6	1.4-1.8
History of self-reported STI	13,525	696	5.1		
Yes	4,219	309	7.3	1.8	1.6-2.1
Women's complaints	12,208	629	5.2		
Yes	6,019	388	6.4	1.7	1.4-2.0
Men's complaints	4,714	205	4.3		
Yes	389	42	10.8	3.1	2.2-4.4

In the pilot study lifetime sexual partners were included, whereas in the current study only partners of the last 6 months were assessed. An interaction of gender with age was included in the analysis based on the a priori hypothesis that the age effect on Ct prevalence may be different for males and females. We modelled possible non-linearity of the age effect with restricted cubic splines [16]. We evaluated the contribution of each predictive factor by its multivariable odds ratio together with its likelihood ratio χ^2 test statistic minus twice the degrees of freedom (number of regression coefficients used to model a predictive factor),which balances the goodness-of-fit of a model with its complexity and gives a fair assessment of a factor's predictiveness [16]. We applied a backward selection approach to delete variables without predictive contribution i.e. when the χ^2 test statistic minus twice the degrees of freedom is negative. Since there may be differences in Ct prevalence across neighbourhoods that cannot be fully explained by individual and neighbourhood-level predictors, we added an extra neighbourhood-level to the logistic regression models [17]. For quantification of a model's unexplained heterogeneity in Ct prevalence across neighbourhoods we used the Median Odds Ratio (MOR), i.e. the odds ratio of the neighbourhood at highest risk versus the neighbourhood at lowest risk when randomly picking out two neighbourhoods [18]. We assessed the discriminative ability of each model by the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. Although we developed prediction models with a high number of events per variable, we used a bootstrap procedure to correct the AUC for a too optimistic presentation of model performance in new settings [19]. For easy calculation of an individual's risk score and the corresponding probability of having a Ct infection we present the prediction models in score charts [16, 20]. For multilevel regression analysis and construction of prediction models we used R packages lme4 and rms respectively [15].

Potential of new Chlamydia prediction models

The practical potential of the prediction models with and without questionnaire information can be assessed by their usefulness for selective screening strategies. A selective screening strategy is defined by a desirable risk score threshold: only the individuals with a risk score equal to or above the threshold will be screened for Ct infection. To estimate the impact of using the prediction models for selective screening, we present for all possible risk score thresholds: the proportion of the population that would be screened (screening eligibility); the proportion of the Ct positive population that would be screened (sensitivity); the proportion of the Ct-negative population that would not be screened (specificity); the proportion of the screened population that would be Ct positive (positive predicted value).

Table 19.2 Multivariable associations between Chlamydia trachomatis and risk factors for 3 levels of information.

	Registry data			+ SES postcode data			+ Questionnaire data		
	OR	95% CI	χ^2-2df	OR	95% CI	χ^2-2df	OR	95% CI	χ^2-2df
Gender and age			55			56			36
Female(17 years)	2.81	2.5-3.2		2.81	2.5-3.2		2.20	1.9-2.5	
Female(22 years)	1.88	1.7-2.1		1.87	1.7-2.0		1.67	1.5-1.8	
Female(27 years)	1.00			1.00			1.00		
Male (17 years)	1.45	1.2-1.7		1.44	1.2-1.7		1.12	0.9-1.4	
Male (22 years)	1.65	1.5-1.8		1.64	1.5-1.8		1.39	1.2-1.6	
Male (27 years)	1.13	1.0-1.3		1.13	1.0-1.3		1.02	0.9-1.2	
Ethnicity			868			629			207
Dutch	1.00			1.00			1.00		
Turkish/North-African	1.32	1.1-1.5		1.27	1.1-1.5		0.91	0.8-1.1	
Surinamese/Antillean	3.05	2.8-3.3		2.97	2.7-3.2		2.16	1.9-2.5	
African other	2.54	2.2-2.9		2.47	2.1-2.8		1.87	1.6-2.2	
Non-western other	1.41	1.2-1.6		1.40	1.2-1.6		1.32	1.1-1.5	
Western other	1.25	1.1-1.4		1.25	1.1-1.4		1.23	1.0-1.4	
SES (postcode)						21			
High				1.00					
Average to high				1.19	1.0-1.4				
Low to average				1.34	1.2-1.5				
Low				1.39	1.2-1.6				
Education									141
Low							1.97	1.6-2.4	
Intermediate							1.96	1.8-2.2	
High							1.00		
Partners in previous 6 months									11
1							1.00		
2+							1.36	1.2-1.6	
New contacts in previous 2 months									7
Yes							1.19	1.1-1.3	
Ethnicity mixing with regular partner									121
Dutch, Dutch partner							1.00		
Dutch, non-Dutch partner							2.52	2.0-3.2	
Non-Dutch, Dutch partner							0.90	0.8-1.1	
Non-Dutch, non-Dutch partner							1.70	1.5-2.0	
Condom use last sexual contact with regular partner									7
Yes							0.74	0.6-0.9	
Casual partners									41
Yes							2.02	1.6-2.5	
Condom use last sexual contact with casual partner									38
Yes							0.66	0.6-0.8	
History of self-reported STI									18
Yes							1.46	1.2-1.7	
Women's complaints									4
Yes							1.19	1.0-1.4	
Men's complaints									6
Yes							1.49	1.1-2.0	
Total			1105			1160			2120
MOR neighbourhood		**1.27**			**1.22**			**1.15**	
AUC		**0.67**			**0.67**			**0.74**	

OR=odds ratio; 95% CI is 95% confidence interval; χ^2 is the likelihood ratio test statistic; df=degrees of freedom; MOR=median odds ratio; AUC=area under the curve.

RESULTS

Chlamydia prevalence

Overall prevalence in the Rotterdam and Amsterdam population was 4.4% (95% confidence interval 4.26-4.55%) compared to a prevalence of 4.2% in highly urban regions of the pilot study.

Development of improved Chlamydia prediction models

Strong registry based predictors for Ct infection were young age, especially for women, and either Surinamese, Antillean or Sub-Sahara-African ethnicity (Table 19.2). The non-linear interaction between age and gender indicated that the risk for men and women decreased similarly after the age of 24 (Supplementary figure 1), was stable below the age of 24 for men but further increased for women below the age of 24. Neighbourhood-level SES was of minor importance. From the individual questionnaires, low to intermediate education level, ethnicity mixing with the regular partner (non-Dutch) and having sex with casual partners showed strong associations with Ct infection. Note that the additional risk for a non-Dutch participant with a non-Dutch regular partner was lower than for a Dutch participant with a non-Dutch regular partner since the non-Dutch ethnicity of the participant was already an important risk factor. The median odds ratio (MOR) of the random neighbourhood effect decreased from 1.27 to 1.22 when neighbourhood-level SES was added and to 1.15 when questionnaire information was added to the model (Table 19.2). The AUC at internal validation was 0.67 (95% CI 0.656-0.675) based on registry risk factors only (model 1), stayed at 0.67 (95% CI 0.657-0.677) when neighbourhood-level SES was added (model 2) and increased substantially to 0.74 (95% CI 0.736-0.752) when questionnaire risk factors were added (model 3). Model 3 also performed substantially better than the Ct prediction model developed in the pilot study in 2005 (Supplementary figure 2). Results were fairly similar when only complete cases were analysed (Supplementary table 2).

We presented the newly developed prediction models 1 and 3 by score charts (Table 19.3). To calculate an individual's probability of having a Ct infection, first determine her or his risk factor values (e.g. female, 17 years of age; Surinamese ethnicity), second look up risk points for each risk factor (6 and 4 points in Table 19.3, respectively), third add the points up – including an optional constant (7 point for the registry information based model) – to obtain a risk sum score (17 points). The probability of having a Ct infection is subsequently read from the 2 right-hand columns of Table 19.3 (14%; Supplementary table 3). The score charts were visualised by nomograms in Supplementary figure 3.

Table 19.3 Rounded score charts based on registry data and registry plus questionnaire data.

Predictor	Registry		+ Questionnaire	
	Women	Men	Women	Men
Start	7	7		
Age group (years)				
15-19	6	4	5	2
20-24	4	4	3	3
25-29	2	2	1	1
Ethnicity				
Dutch	0	0	0	0
Turkish/North-African	1	1	0	0
Surinamese/Antillean	4	4	3	3
African other	4	4	3	3
Non-western other	1	1	1	1
Western other	1	1	1	1
Education				
Low			3	3
Intermediate			3	3
High			0	0
Partners in previous 6 months				
0-1			0	0
2			2	2
3+			2	2
New contacts in previous 2 months				
Yes			1	1
No			0	0
Ethnicity mixing with regular partner				
No partner			0	0
Dutch, Dutch partner			0	0
Dutch, non-Dutch partner			4	4
Non-Dutch, Dutch partner			0	0
Non-Dutch, non-Dutch partner			3	3
Condom use last sexual contact with regular partner				
Yes			0	0
No			1	1
Casual partners				
Yes			3	3
No			0	0
Condom use last sexual contact with casual partner				
Yes			0	0
No			2	2
History of self-reported STI				
Yes			2	2
No			0	0
Complaints*				
Yes			1	2
No			0	0

Total points	Probability Ct (%)
0	0.2
1	0.3
2	0.4
3	0.5
4	0.6
5	0.8
6	1.1
7	1.3
8	1.7
9	2.2
10	2.8
11	3.5
12	4.5
13	5.7
14	7.2
15	9.0
16	11
17	14
18	17
19	21
20	26
21	31
22	36
23	42
24	48
25	54

LPS = -5.9427 + 0.2481 * total points

Probability Ct = 1 / (1 + exp (-LPS))

Potential of new Chlamydia prediction models

The estimated impact of using the prediction models for selective screening is reported in Table 19.4 for all possible risk sum score thresholds. With a sum score threshold of 10 (predicted Ct probability of 2.8%) the registry based prediction model leads to 87% (95%CI 85.4-87.6%) sensitivity and 73% of the population eligible for screening and (positive predicted value 5.2%; specificity 28%), while the prediction model including questionnaire information reaches 80% (95%CI 78.4-81.0%) sensitivity with 48% of the population eligible for screening (positive predicted value 7.3%; specificity 53%). The difference in predictive performance of the 2 prediction models was visualised with decision curves (Supplementary figure 4), with plots of screening eligibility by sensitivity (Supplementary figure 5), and with ROC curves (Supplementary figure 6).

Table 19.4 Implications of applying the prediction models.

	Registry				Registry + questionnaire			
Threshold sum score	Eligible for screening (%)	Prevalence in screened population (%)	Sensitivity TPR (%)	Specificity TNR (%)	Eligible for screening (%)	Prevalence in screened population (%)	Sensitivity TPR (%)	Specificity TNR (%)
3					100	4.4	100	0.0
4					99	4.4	100	1.0
5					92	4.7	99	8.6
6					85	5.0	97	15
7					77	5.4	94	24
8					66	6.0	90	35
9	100	4.4	100	0.0	57	6.6	86	44
10	73	5.2	87	28	48	7.3	80	53
11	60	5.8	78	41	38	8.3	72	63
12	34	7.4	56	67	30	9.3	64	71
13	24	8.4	46	77	23	11	56	79
14	12	11	30	89	16	12	47	85
15	10	12	27	91	12	14	39	89
16					8.0	17	30	93
17	2.3	16	8	98	5.2	19	23	96
18					3.3	21	16	97
19					1.9	24	10	98
20					1.0	26	6.2	99
21					0.6	29	3.9	100
22					0.3	34	2.4	100
23					0.1	38	1.1	100
24					0.1	41	0.5	100
25					0.0	40	0.1	100

TPR=True positive rate; TNR=True negative rate

DISCUSSION

Main findings

We developed easily applicable Chlamydia prediction models with data from a large Chlamydia screening project. The prediction model based on readily available registry data may serve as a simple tool for selective screening at the population level. With detailed questionnaire information the predictive performance increased substantially and was better than the performance of a previously proposed Ct prediction model. With less than half of the participating population needed to be screened to find 80% of the infections, the detailed prediction model allows for better screening decisions at the individual level.

Risk factors identified in relation to other studies

Most of the predictors for Ct infection found previously were also associated with Ct infection in our study population: young age; Surinamese or Antillean ethnicity; low or intermediate education; urogenital symptoms, especially for men; multiple sexual partners; new partner in previous 2 months; no condom use at last sexual contact. With respect to these risk factors we found additionally that the age effect was stronger for women and that Sub Saharan Africans have a similar risk as Surinamese and Antillean individuals. Number of sexual partners in the last 6 months was included in our model instead of the previously modelled lifetime partners, since it is easier to assess in practice, is less dependent of age, and less prone to recall bias. We could not include address density in our model as our study population was almost entirely very highly urban (AAD 1). Application of our prediction model in less urban areas probably requires recalibration, which could be based on the lower risk for lower address density regions in the previously reported prediction model [9]. Furthermore, we were able to add some additional predictors: sexual contact with casual partners in the last 6 months, non-Dutch ethnicity of the regular partner especially for Dutch participants, and history of self-reported STI. The latter two are in line with previous findings [21-23]. Finally, neighbourhood-level SES was of minor importance especially when questionnaire predictors were added. Apparently the registry and questionnaire data at the individual level were more predictive than the SES at neighbourhood level.

Strengths & limitations

Main strengths of this study are the large study population of 80,385 screening participants (3,540 positive) living in 2 large urban areas; the availability of high quality outcome data and objective registry data (gender, age, ethnicity derived from country of birth of participants or their parents, neighbourhood-level SES); extensive questionnaire information on education, STI history, symptoms, partner information and sexual behaviour; and advanced analysis allowing for targeting of screening with only registry data or the combination of registry and questionnaire data.

The prediction models were developed in a specific population sample, i.e. first-time responders to a Ct test invitation of all individuals aged 16-29 living in the Dutch cities of Rotterdam and Amsterdam. However, we anticipate the models to be useful for selecting high-risk individuals in other populations as well since most of the predictors are universal. Still, recalibration of the prediction models may be necessary to match the overall prevalence of a particular population, e.g. individuals in other cities or countries, or individuals who are specifically seeking care. Similarly, recalibration to match the Ct risk for repeatedly participating individuals may be required, although we included self-reported STI history as a risk factor in the questionnaire-based model. Furthermore, the absence of large communities of Surinamese and Antillean may attenuate the discriminative ability of the models in other populations. Local data would be required to extend our models with risk levels of specific ethnic groups.

The self-assessed questionnaire variables may be considered a limitation of this study. However, history of STI and STI complaints are commonly used risk factors for Ct infection [24]. Recent partner change and number of partners in the last six months as well as condom use at last sexual contact may be expected to be remembered well.

A high number of missing values in questionnaires, possibly more often for subjects with low education, is a severe limitation of this study. However, with the substantial amount of available data we were able to develop prediction models based on multiple imputation of the missing values. Fairly similar but less reliable results were noted in a complete case analysis. Furthermore, the prediction model's coefficients hardly changed when we forced an extra one third of the imputations to an intermediate or low education.

Ethnicity mixing with the partner was only available in case of regular partners. Although the effect of ethnicity mixing could only be quantified for regular partners, we suspect this effect applies to casual partners as well. This should however be validated in future studies.

Ethnicity of the participant and of the regular partner were confirmed to be important predictors for Ct. A cross-sectional study among adolescents in the US also found that race-ethnicity (either of tested individual or of partner) affected algorithm performance [25]. Presently triaging systems in Dutch Public Health STI clinics include assessing ethnicity and this is well accepted. However, ethnicity may be hard to obtain in future practice when both the participants and their parents are already born in the Netherlands (3rd generation), or when ethnicity is considered to be sensitive information. Although Ct prediction models without ethnicity show lower predictive performance – the AUC decreases from 0.665 to 0.610 when using registry data only and from 0.744 to 0.721 when using questionnaire data – they could still be useful, especially when using detailed questionnaire information.

Application of prediction models

We presented the Ct prediction models by risk score charts, which can either be implemented in paper forms or in internet-based apps. The score charts may be useful to select high-risk individuals as part of systematic screening programmes in urban areas, similar to the selection of individuals at high risk in less urban regions [10]. Furthermore, the score charts may be used in guiding STI clinicians and GPs whom to offer opportunistic Ct testing – more selectively than using age group alone. Although the prediction models are developed from general population data, we anticipate the risk factors to hold for those at higher risk. Especially the score chart based on questionnaire data would allow for better identifying individuals at high risk for Ct infection, and can be used for (internet) triaging systems. Scoring questionnaires may encourage test uptake by increasing risk awareness in those who may be reluctant to be tested [26].

One may argue that selective screening is less effective or even unethical, since it will miss Ct infections that would have been detected in a screen-all strategy. Although prediction models are imperfect, using them for selective screening may still be very helpful since the harm of missing infections needs to be balanced with the burden – including costs – of unnecessary diagnostic testing. We illustrated the helpfulness of our prediction model by choosing a particular risk threshold that implies a huge benefit (and cost-savings) of screening only half instead of the full population against the burden of missing 20% of the Ct infections. The choice of the appropriate risk threshold for a selective screening strategy – balancing the benefits and the harms – is up to decision-makers.

Implications for future research

We recommend further validation studies of our Chlamydia prediction models – with recalibration and updating of predictors or predictor effects when necessary – in different countries and in different selective screening settings, both for first-time participants and for repeatedly tested participants. Moreover, we encourage studies that focus on the impact on clinical practice, ideally by trials incorporating Chlamydia prediction models in selective screening settings [27]. Other types of studies can also be used to analyse the impact of targeted screening versus untargeted screening, as was done for HIV screening in an emergency department based on a validated HIV risk score [28]. Finally, the hypothesis that scoring questionnaires encourages test uptake deserves further analysis.

CONCLUSION

A registry based prediction model can facilitate selective Ct screening at population level, with further refinement at the individual level by including questionnaire risk factors.

ACKNOWLEDGMENTS

The authors express their gratitude to all of the principal investigators (IVF van den Broek, JEAM van Bergen, JSA Fennema, HM Götz, CJPA Hoebe, ELM Op de Coul) of the Chlamydia Screening Implementation for providing the data. David van Klaveren and Yvonne Vergouwe were supported by the Netherlands Organisation for Scientific Research (grant 917.11.383)

SUPPLEMENTARY DATA

An online appendix can be found at http://dx.doi.org/10.1136/sextrans-2015-052048.

REFERENCES

1. ECDC. European Centre for Disease Prevention and Control. Chlamydia control in Europe: literature review. In European Centre for Disease Prevention and Control. Chlamydia control in Europe: literature review, Editor (ed)^(eds). City, 2014.
2. Oakeshott P, Kerry S, Aghaizu A, Atherton H, Hay S, Taylor-Robinson D, Simms I, Hay P. Randomised controlled trial of screening for Chlamydia trachomatis to prevent pelvic inflammatory disease: the POPI (prevention of pelvic infection) trial. *BMJ* 2010; **340**: c1642.
3. Rours GI, Duijts L, Moll HA, Arends LR, de Groot R, Jaddoe VW, Hofman A, Steegers EA, Mackenbach JP, Ott A, Willemse HF, van der Zwaan EA, Verkooijen RP, Verbrugh HA. Chlamydia trachomatis infection during pregnancy associated with preterm delivery: a population-based prospective cohort study. *Eur J Epidemiol* 2011; **26**: 493-502.
4. Low N, Bender N, Nartey L, Shang A, Stephenson JM. Effectiveness of chlamydia screening: systematic review. *Int J Epidemiol* 2009; **38**: 435-448.
5. Andersen B, Olesen F, Moller JK, Ostergaard L. Population-based strategies for outreach screening of urogenital Chlamydia trachomatis infections: a randomized, controlled trial. *J Infect Dis* 2002; **185**: 252-258.
6. Hocking JS, Sparks S, Temple-Smith M, Fairley CK, Kaldor J, Donovan B, Law M, Gunn J, Low N. The Australian chlamydia control effectiveness polot (ACCEPt): first results from a randomised controlled trial of annual chlamydia screening in general practice. *Sex Transm Infect* 2012; **88**: A2-A4.
7. van den Broek IV, van Bergen JE, Brouwers EE, Fennema JS, Gotz HM, Hoebe CJ, Koekenbier RH, Kretzschmar M, Over EA, Schmid BV, Pars LL, van Ravesteijn SM, van der Sande MA, de Wit GA, Low N, Op de Coul EL. Effectiveness of yearly, register based screening for chlamydia in the Netherlands: controlled trial with randomised stepped wedge implementation. *BMJ* 2012; **345**: e4316.
8. Richardus JH, Gotz HM. Risk selection and targeted interventions in community-based control of chlamydia. *Curr Opin Infect Dis* 2007; **20**: 60-65.
9. Gotz HM, van Bergen JE, Veldhuijzen IK, Broer J, Hoebe CJ, Steyerberg EW, Coenen AJ, de Groot F, Verhooren MJ, van Schaik DT, Richardus JH. A prediction rule for selective screening of Chlamydia trachomatis infection. *Sex Transm Infect* 2005; **81**: 24-30.
10. van den Broek IV, Brouwers EE, Gotz HM, van Bergen JE, Op de Coul EL, Fennema JS, Koekenbier RH, Pars LL, van Ravesteijn SM, Hoebe CJ. Systematic selection of screening participants by risk score in a Chlamydia screening programme is feasible and effective. *Sex Transm Infect* 2012; **88**: 205-211.
11. van den Broek IV, Hoebe CJ, van Bergen JE, Brouwers EE, de Feijter EM, Fennema JS, Gotz HM, Koekenbier RH, van Ravesteijn SM, de Coul EL. Evaluation design of a systematic, selective, internet-based, Chlamydia screening implementation in the Netherlands, 2008-2010: implications of first results for the analysis. *BMC Infect Dis* 2010; **10**: 89.
12. van Bergen JE, Fennema JS, van den Broek IV, Brouwers EE, de Feijter EM, Hoebe CJ, Koekenbier RH, de Coul EL, van Ravesteijn SM, Gotz HM. Rationale, design, and results of the first screening round of a comprehensive, register-based, Chlamydia screening implementation programme in the Netherlands. *BMC Infect Dis* 2010; **10**: 293.
13. Op de Coul EL, Gotz HM, van Bergen JE, Fennema JS, Hoebe CJ, Koekenbier RH, Pars LL, van Ravesteijn SM, van der Sande MA, van den Broek IV. Who participates in the Dutch Chlamydia screening? A study on demographic and behavioral correlates of participation and positivity. *Sex Transm Dis* 2012; **39**: 97-103.
14. Van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 2011; **45**: 1-67.
15. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/, 2011.

16. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer-Verlag New York, 2001.
17. Gelman A, Hill J. *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press: Cambridge, 2007.
18. Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, Rastam L, Larsen K. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *J Epidemiol Community Health* 2006; **60**: 290-297.
19. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; **15**: 361-387.
20. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* Springer: New York, 2009.
21. Gotz HM, van den Broek IV, Hoebe CJ, Brouwers EE, Pars LL, Fennema JS, Koekenbier RH, van Ravesteijn S, Op de Coul EL, van Bergen J. High yield of reinfections by home-based automatic rescreening of Chlamydia positives in a large-scale register-based screening programme and determinants of repeat infections. *Sex Transm Infect* 2013; **89**: 63-69.
22. Woodhall SC, Atkins JL, Soldan K, Hughes G, Bone A, Gill ON. Repeat genital Chlamydia trachomatis testing rates in young adults in England, 2010. *Sex Transm Infect* 2013; **89**: 51-56.
23. Morgan J, Woodhall S. Repeat chlamydia testing across a New Zealand district: 3 years of laboratory data. *Sex Transm Infect* 2013; **89**: 28-31.
24. La Montagne DS, Patrick LE, Fine DN, Marrazzo JM. Re-evaluating selective screening criteria for chlamydial infection among women in the U S Pacific Northwest. *Sex Transm Dis* 2004; **31**: 283-289.
25. Stein CR, Kaufman JS, Ford CA, Leone PA, Feldblum PJ, Miller WC. Screening young adults for prevalent chlamydial infection in community settings. *Ann Epidemiol* 2008; **18**: 560-571.
26. Gotz HM, van Klaveren D. Use of prediction rules in control of sexually transmitted infections: challenges and chances. *Sex Transm Dis* 2014; **41**: 331-332.
27. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG, Group P. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013; **10**: e1001381.
28. Haukoos JS, Hopkins E, Bender B, Sasson C, Al-Tayyib AA, Thrun MW, Denver Emergency Department HIVTRC. Comparison of enhanced targeted rapid HIV screening using the Denver HIV risk score to nontargeted rapid HIV screening in the emergency department. *Ann Emerg Med* 2013; **61**: 353-361.

20

Use of prediction rules in control of sexually transmitted infections: challenges and chances

HM Götz
D van Klaveren

Sexually Transmitted Diseases 2014

Control of sexually transmitted infections (STIs) is a public health challenge and includes a variety of diseases such as Chlamydia trachomatis, gonorrhea, syphilis, and HIV with different epidemiology and risk groups. Well-performing clinical prediction rules (CPRs) can potentially support STI control with prioritization of testing [1]. We welcome the effort of Falasinnu et al. [2] to provide a critical appraisal of existing CPRs in sexual health contexts. They identified 16 studies reporting on CPRs for STIs and gave a broad overview of the methodological quality of the studies (Table 3) [2] and of the performance of the CPRs (Table 4) [2]. Here, we discuss and prioritize the performance measures and quality items that were used with the aim to enable identification of valid CPRs for specific STIs.

Successful external validation, including assessment of calibration (agreement between predicted probabilities and observed outcome frequencies) and discrimination (ability to distinguish between individuals with and without the outcome), is considered the most important proof of generalizability of a CPR [1,3]. External validation assesses if a CPR works in individuals other than those from whose data it was derived, in contrast with internal validation where a CPR's performance is assessed in the same individuals than those from whose data it was derived. In our view, external validation should be part of the assessment of both the methodological quality of the studies in which CPRs are derived and the performance of the derived CPRs. Haukoos et al. [4] is a good example where a CPR for HIV risk was derived from patients in a sexually transmitted disease clinic and was consecutively externally validated in an emergency department among the general population. When external validation is lacking, CPR performance within the development data (internal validation) and other methodological quality criteria become more important. As for CPR performance assessment, discrimination may be the most important measure of a CPR's internal validity. Calibration is usually good in the same individuals whose data the CPR was derived from and is particularly meaningful to assess in external validation data. Because the discrimination of a CPR within the individuals whose data were used to derive the CPR may be a too optimistic reflection of the discrimination in other individuals, it should be assessed with either cross-validation or bootstrap techniques, especially when sample size (or actually number of positives) is small [5]. This may be exemplified by the study of Verhoeven et al. [6] who predicted chlamydia infection in general practice with combinations of predictors and found excellent discrimination with an area under the receiver operating characteristic curve of 0.88: that is, an 88% probability of predicting a higher risk for a positive than for a negative individual. The study sample size was, however, quite small (n = 774; 39 positives), and no internal validation techniques were used. The area under the receiver operating characteristic curve at internal or external validation will likely be substantially smaller. Falasinnu et al. define a CPR to perform well in terms of efficiency and sensitivity if at least 90% of the infections are detected while testing at most 60% of the patients. This benchmark is taken from a study assessing selective screening criteria for

Chlamydia trachomatis in an opportunistic screening program in the UK [7]. Assessing the quality of CPRs with this benchmark for various STIs is arguable. From a decision-making viewpoint, the required sensitivity depends on the burden of missing an infection, which may be different for HIV and chlamydia infections and which needs to be balanced with the burden of unnecessary diagnostic testing. Furthermore, another benchmark efficiency may be based on costs. A prediction rule that needs 65% of the patients to be tested to detect 90% of the infections may well be cost-effective.

Falasinnu et al. assessed the methodological quality with 16 items (Tables 1 and 3) [2], similar to other methodological review studies of CPRs [8,9]. We agree with their discussion of the limitation that equal weights assigned to items on the quality checklist ignore the possibility that some items are more important than others. A description of study design and study sample will generally be given in peer-reviewed publications and does not add substantially to the quality of a CPR. Variable definitions and details of methods of assessment for predictors and outcomes should certainly be described, but some gradation in scoring may be useful. Objective predictors like sex, age, area of living, and country of birth can be derived from registries, whereas questionnaire data like education, marital status, sexual behavior data, history of STI, and symptoms are subjective and potentially subject to bias and misclassification. Although ethnicity or race is commonly reported as predictor, it may be controversial to assess. Reporting of missing values is certainly necessary as well as a description of how missing values were dealt with. However, this is less relevant when the percent- age of missing values is small. We do agree that multivariable statistical methods should be used to examine the associations between potential predictors and STI outcomes. Yet, assessing the selection of predictors is a vital first step: are they clinically meaningful and easy to use? However, this requires a clear definition of ‘‘clinically meaningful.’’ In addition, we recommend to always give a structured presentation of the number of times a predictor is used among the studies (e.g., a table representing the presence of each predictor among the CPR studies in Kulik et al. [10]) and, ideally, a measure of the strength of each predictor (e.g., a table with odds ratios for each predictor in Leushuis et al. [11]).

Summarizing, we encourage studies to identify which CPRs or predictors perform well and meet the necessary quality standards including internal and external validation, which can often be done relatively easy on existing datasets. Successful external validation of CPRs, when necessary with updating of predictors, should be followed by an impact analysis that shows whether clinical practice is changed with beneficial consequences [12]. The clinical impact of applying a CPR with a chosen risk threshold (cutoff value) is determined by the cost of missing infections versus the benefit of less unnecessary testing and by the financial costs.

Clinical impact is ideally analyzed with a randomized controlled trial, but other types of impact comparison can also be informative: Haukoos et al. [13] validated an HIV risk score in data from an emergency department and applied this risk score in a prospective, before-after design in another emergency department, comparing the targeted screening with untargeted screening for HIV in earlier periods of time. In our own studies, we externally validated our chlamydia prediction rule in 2 different populations [14,15]. The prediction rule was further implemented in a large-scale screening program, where selection was attained by a score derived from a short questionnaire. It proved highly effective in yielding high positivity rates in an area with lower prevalence of chlamydia [16]. An intermediate way of impact analysis assesses how many STIs one would miss using a CPR with different cutoff values. This is exemplified by algorithms developed for detection of chlamydial and gonococcal infections at an emergency department setting [17]. Likewise, the clinical impact of using CPRs with or without controversial predictors like race can be compared [18].

Although only few, some promising CPRs for different STIs are available, and we do think using CPRs is the way for- ward for targeted STI screening. Apart from using chlamydia prediction rules in systematic chlamydia screening, they may be used in guiding STI clinicians in whom to offer opportunistic chlamydia testing – more selectively than using age group alone. Other applications of prediction rules may be (Internet) triaging systems in STI clinics, both for targeting which STI to test for (e.g., only chlamydia or testing also for other STI) and for targeting HIV testing in low-prevalence populations, as well as for individual risk assessment by using scoring questionnaires, which may encourage test uptake by increasing risk awareness.

REFERENCES

1. Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. New York: Springer, 2009.
2. Falasinnu T, Gustafson P, Hottes TS, et al. A critical appraisal of risk models for predicting sexually transmitted infections. Sex Transm Dis 2014:321Y331.
3. Altman DG, Royston P. What do we mean by validating a prog- nostic model? Stat Med 2000; 19:453Y73.
4. Haukoos JS, Lyons MS, Lindsell CJ, et al. Derivation and validation of the Denver Human Immunodeficiency Virus (HIV) risk score for targeted HIV screening. Am J Epidemiol 2012; 175:838Y46.
5. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996; 15:361Y87.
6. Verhoeven V, Avonts D, Meheus A, et al. Chlamydial infection: an accurate model for opportunistic screening in general practice. Sex Transm Infect 2003; 79:313Y7.
7. La Montagne DS, Patrick LE, Fine DN, et al. Re-evaluating selec- tive screening criteria for chlamydial infection among women in the US Pacific Northwest. Sex Transm Dis 2004; 31:283Y9.
8. Maguire JL, Kulik DM, Laupacis A, et al. Clinical prediction rules for children: A systematic review. Pediatrics 2011; 128:e666Ye677.
9. Bouwmeester W, Zuithoff NP, Mallett S, et al. Reporting and methods in clinical prediction research: A systematic review. PLoS Med 2012; 9:1Y12.
10. Kulik DM, Uleryk EM, Maguire JL. Does this child have appen- dicitis? A systematic review of clinical prediction rules for children with acute abdominal pain. J Clin Epidemiol 2013; 66:95Y104.
11. Leushuis E, van der Steeg JW, Steures P, et al. Prediction models in reproductive medicine: A critical appraisal. Hum Reprod Update 2009; 15:537Y52.
12. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Re- search Strategy (PROGRESS) 3: Prognostic model research. PLoS Med 2013; 10:e1001381.
13. Haukoos JS, Hopkins E, Bender B, et al. Comparison of enhanced targeted rapid HIV screening using the Denver HIV risk score to nontargeted rapid HIV screening in the emergency department. Ann Emerg Med 2013; 61:353Y61.
14. Gotz HM, van Bergen JE, Veldhuijzen IK, et al. A prediction rule for selective screening of Chlamydia trachomatis infection. Sex Transm Infect 2005; 81:24Y30.
15. Gotz HM, Veldhuijzen IK, Habbema JD, et al. Prediction of Chlamydia trachomatis infection: Application of a scoring rule to other populations. Sex Transm Dis 2006; 33:374Y80.
16. van den Broek IV, Brouwers EE, Gotz HM, et al. Systematic se- lection of screening participants by risk score in a chlamydia screen- ing programme is feasible and effective. Sex Transm Infect 2012; 88:205Y11.
17. Al-Tayyib AA, Miller WC, Rogers SM, et al. Evaluation of risk score algorithms for detection of chlamydial and gonococcal in- fections in an emergency department setting. Acad Emerg Med 2008; 15:126Y35.
18. Stein CR, Kaufman JS, Ford CA, et al. Screening young adults for prevalent chlamydial infection in community settings. Ann Epidemiol 2008; 18:560Y71.

21

A prediction model for neonatal mortality in low and middle income countries: an analysis of data from population surveillance sites in South Asia

TAJ Houweling *
D van Klaveren*
S Das
K Azad
P Tripathy
D Manandhar
M Neuman
E de Jonge
J van Been
EW Steyerberg
A Costello

Submitted
* Equal contribution

ABSTRACT

Importance Neonatal mortality remains high in many countries. The ability to predict which infants are at a high risk of death is important for improved management of pregnancy, delivery, and the newborn period. Yet, to our knowledge, prediction models for neonatal mortality in the general population in low and middle income countries have not previously been reported.

Objective To develop and validate a prediction model for neonatal mortality in the general population in low and middle income countries.

Design, Setting, Participants We used prospectively collected data from demographic surveillance sites in rural Nepal and Bangladesh and rural and urban India, including all live births (49,632) and neonatal deaths (1,742) between 2001 and2011. We developed logistic regression models to predict the risk of death in the first 28 days of life, based on characteristics known at (i) the start of pregnancy, (ii) the start of delivery, and (iii) five minutes post-partum. We assessed the models' discriminative ability by the area under the receiver operating characteristic curve (AUC), using cross-validation between sites. We presented newly developed prediction models with nomograms.

Main outcome Neonatal mortality

Results At the start of pregnancy, statistically significant predictors of neonatal death were: low maternal education and economic status, short birth interval, primigravida, and young and advanced maternal age. Predictive ability at the start of pregnancy was moderate (AUC 0.58 [95%CI 0.57-0.59]). At the start of delivery, predictive ability was considerably better (AUC 0.72 [95%CI 0.69-0.75]), and prematurity and multiple pregnancy were strong predictors of death. At five minutes post-partum, predictive ability was good (AUC 0.81 [95%CI 0.79-0.83]), and very strong predictors were multiple birth, prematurity, and a poor condition of the infant at five minutes.

Conclusions and relevance We developed good performing prediction models for neonatal mortality in the general population in low and middle income countries. Based on our models, we conclude that improved management of high-risk infants can substantially reduce neonatal mortality. Population-level strategies to raise awareness about important risk factors and empower community-based care givers to deal with these should be an integral part of a high-risk approach when health systems are weak.

INTRODUCTION

Worldwide, every year nearly three million infants do not survive the first 28 days of life [1]. Nearly all (99%) of these deaths occur in low and middle income countries [2]. In poorer parts of India and Bangladesh, 35 to 65 babies in 1,000 live births die in the neonatal period [3]. For public health policy making and management of pregnancy, delivery, and the newborn period, including proper risk selection and institution of selective care pathways for high risk pregnancies, it is important to be able to predict which infants are at a high risk of neonatal death.

Prediction models of neonatal mortality are largely restricted to high-income countries, which account for only 1% of neonatal deaths. These models focus on infants in neonatal intensive care units [4-6]. Existing models for poorer countries are few, and again focus on neonatal intensive care patients [7]. In poor settings, where many births occur at home without skilled care [8], prediction models of neonatal mortality in the general population, rather than for a selective high-risk group only, can aid public health policy-making and decision-making by family members and community health workers (e.g. through early recognition of potential problems). To our knowledge, no such models have been published in English-language international peer-reviewed journals.

While prediction models for neonatal mortality are scarce, there is quite a good understanding of the causes of and risk factors for neonatal death in low and middle income countries. Preterm birth, neonatal infections, and birth asphyxia account for around 80% of neonatal deaths [1,2]. Direct risk factors include young and relatively advanced maternal age, maternal undernutrition, primiparity and high parity, short pregnancy interval, multiple pregnancy, maternal health problems during pregnancy, malpresentation, problems during delivery, male infant sex (with exceptions in settings with strong son preference), low birth weight, and exposure to infections [2,9,10]. Low socioeconomic position of the mother is an important underlying risk factor for neonatal death [11].

The advantage of prediction models is that they formally combine risk factors, allowing for more accurate risk estimation [4]. Yet, as many births in poor settings occur at home without skilled care, good data on neonatal mortality and its risk factors remain scarce. Demographic surveillance sites in South Asia, in which the full population is followed-up and all women were interviewed post-partum, do provide such data, offering a unique opportunity to develop a prediction model for neonatal mortality in poor settings.

We aimed to develop and validate a prediction model for neonatal mortality in the general population in low and middle income countries, with specific reference to South Asia, using data from four surveillance sites.

METHODS

We used prospectively collected data from surveillance sites in rural Nepal (Makwanpur district, surveillance population of 170,000) and Bangladesh (Moulvibazar, Bogra, and Faridpur district, 500,000) and rural (five districts in Odisha and Jharkhand state, 228,000) and urban (informal slum settlements in Mumbai, 283,000) India [12-16]. At each site, the full population in a geographically defined area was followed-up, and all births and birth outcomes were recorded. All women who had given birth, or a family member in case the woman had died, were interviewed at around 6 weeks post-partum, and detailed information about the mother, and the pregnancy, delivery and newborn period, was recorded. The sites were set-up for controlled trials of community-based interventions. We only included data from the control arms. The data were collected between 2001 and 2011 (Bangladesh 2005-2011, rural India 2005-2009, urban India 2006-2009, Nepal 2001-2003).

Our outcome of interest was neonatal death, i.e. death in the first 28 days of life among live born infants. All characteristics known to influence neonatal mortality as reported in the Lancet Neonatal Survival series [2,17], when available in our dataset, were included as predictor in our initial models. We also included season of birth, a predictor of neonatal death in at least one of our sites [18]. All variables were based on mothers' report, or report of a family member in the event of her death. Included characteristics at the start of pregnancy were: maternal age, maternal education (no school, primary, secondary, BSc/MSc) and literacy (can read, cannot read), household economic status (wealth tertiles, based on Principal Component Analysis) [19], and pregnancy interval (using birth interval as proxy, categorized as: <15 months, 15-26, 27-68, >68, or primigravida) [10]. We included the following characteristics known at the start of delivery: at least 1 antenatal care (ANC) visit (y/n), 4+ ANC visits (y/n), tetanus vaccination during pregnancy (y/n), premature birth (y/n, defined as gestational age of ≤8 months; gestational age in weeks not available), season of birth (warm-dry, rainy, cold), and pregnancy complications (y/n). Pregnancy complications were defined as any one of: reduced/no fetal movement, jaundice, fits/seizures/ convulsions/lost consciousness. These complications were identified as the strongest independent predictors of neonatal mortality in a preliminary logistic regression analysis which also included: excessive vomiting, felt weak/tired, swollen feet/legs/face, severe stomach pain, looked pale, malaria, severe headache/dizziness/fainting, breathless when doing household tasks, blurred vision/spots before eyes, anaemia. Multiple birth (y/n) may or may not have been known at the start of delivery, depending on the quality of antenatal care. The following characteristics known five minutes post-partum were included: presentation/mode of delivery (normal, breech, caesarean section (C-section)), place of delivery (home, facility), labor duration (≤ or >24hrs), delivery complications (y/n), maternal death (y/n), sex of baby, size of baby at birth (small, normal, large), looking abnormal (y/n), breathing/crying immediately after birth (y/n), condition of arms and legs of baby after birth (normal, floppy, stiff), and condition of baby at five minutes (“crying well, breathing well, pink and active”, "poor or no cry, poor breathing, blue limbs or body, poorly active/no movement"). Delivery complications were defined as any one of the following: fever in three days prior to labor, retained placenta, and hemorrhage (“vaginal bleeding so much that you thought you were going to die”). Looking abnormal was mostly based on the question “how did the baby look at birth, normal/abnormal?”

Most predictors were available for over 90% of deliveries (Table 21.1). Some variables were not available or had many missing values for some sites. We used an advanced multiple imputation of missing values strategy (method of chained equations) to make efficient use of the available data [20]. The imputation model included all potential predictors, the site, and the outcome variable. We used R package ‘mice’ for multiple imputation [21]. We developed three logistic regression models to predict the risk of death in the first 28 days of life at the individual level, based on characteristics known at (i) the start of pregnancy, (ii) the start of delivery, (iii) five minutes post-partum.

We modelled possible non-linearity of the association between mother’s age and the risk of neonatal death with restricted cubic splines [22]. We expressed the strength of the association between predictors and neonatal death by crude and adjusted odds ratios. We evaluated the contribution of each predictor by the difference in Akaike’s information criterion (ΔAIC) between multivariable models with and without the predictive factor, balancing the improvement in goodness-of-fit of a model with its increased complexity [22]. We deleted variables with negligible predictive contribution, i.e. when the χ^2 test statistic minus twice the degrees of freedom was relatively small (below 10).

We assessed the discriminative ability of each model by the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The AUC can be interpreted as the probability that the risk prediction of a randomly chosen neonatal death is higher than the risk prediction of a randomly chosen neonatal survivor. We determined the AUC of the models within each of the four sites (“apparent AUC”). We also used a cross-validation approach between sites to obtain a more realistic presentation of the AUC in independent settings (“cross-validated AUC”). Cross-validation means that the model is consecutively fitted in three of the four sites and validated – with the AUC – in the site that was left out when fitting the model. To obtain overall AUCs – both apparent and cross-validated – we used random-effects meta-analyses of the four site-specific AUCs [23].

For calculation of an individual’s probability of neonatal death, we present the prediction models with nomograms [22,24]. For regression analysis and construction of nomograms we used R package ‘rms’ [21].

Table 21.1 Distribution of live births and neonatal deaths across risk factors, by study site.

		Rural Bangladesh		Rural India		Urban India		Rural Nepal	
		deliveries (%)	nnd	deliveries (%]	nnd	deliveries (%]	nnd	deliveries (%]	nnd
Total per site		**30,115**	**1,041**	**8,817**	**518**	**7,478**	**64**	**3,222**	**119**
time (years)	1	4,923 (16.3)	199	2,920 (33.1)	153	2,643 (35.3)	22	1,762 (54.7)	71
	2	5,041 (16.7)	203	2,972 (33.7)	177	2,598 (34.7)	23	1,460 (45.3)	48
	3	5,234 (17.4)	175	2,925 (33.2)	188	2,237 (29.9)	19		
	4	4,773 (15.8)	156						
	5	4,204 (14.0)	133						
	6	5,940 (19.7)	175						
age	<18	986 (3.3)	40	214 (2.6)	25	58 (0.8)	1	53 (1.6)	4
	19-20	7,591 (25.2)	306	1,610 (19.4)	134	1,273 (17.1)	11	361 (11.2)	13
	21-23	5,799 (19.3)	195	1,660 (20.0)	90	2,006 (26.9)	12	733 (22.7)	23
	24-26	6,406 (21.3)	164	1,721 (20.8)	90	2,005 (26.9)	15	556 (17.3)	20
	27-30	3,556 (11.8)	117	1,121 (13.5)	64	1,069 (14.3)	13	448 (13.9)	13
	30-32	3,030 (10.1)	109	1,081 (13.0)	52	675 (9.0)	6	383 (11.9)	18
	33-35	1,470 (4.9)	51	546 (6.6)	34	241 (3.2)	4	251 (7.8)	13
	>35	1,271 (4.2)	58	337 (4.1)	15	138 (1.8)	2	437 (13.6)	15
	missing	6	1	527	14	13	0		
birth interval (months)	primi gravida	10,090 (36.6)	372	2,446 (28.2)	200	2,367 (65.7)	17	609 (100.0)	23
	<15	610 (2.2)	45	314 (3.6)	25	86 (2.4)	0		
	15-26	2,699 (9.8)	89	1,730 (20.0)	88	368 (10.2)	0		
	27-68	9,731 (35.3)	279	3,986 (46.0)	178	638 (17.7)	0		
	>68	4,441 (16.1)	136	183 (2.1)	10	144 (4.0)	0		
	missing	2,544	120	158	17	3,875	47	2,613	96
education	no school	7,107 (23.6)	320	5,974 (67.8)	372	2,094 (28.9)	26	2,769 (86.0)	103
	primary	10,076 (33.5)	369	448 (5.1)	26	397 (5.5)	5	302 (9.4)	9
	secondary	12,582 (41.9)	345	2,317 (26.3)	118	4,037 (55.8)	29	146 (4.5)	6
	BSc/MSc	297 (1.0)	5	78 (0.9)	2	706 (9.8)	1	3 (0.1)	1
	missing	53	2			244	3	2	0
illiterate	no	21,516 (71.5)	654	2,709 (30.7)	141	5,328 (73.7)	42	710 (22.0)	25
	yes	8,585 (28.5)	386	6,108 (69.3)	377	1,906 (26.3)	19	2,510 (78.0)	94
	missing	14	1			244	3	2	0
household wealth (tertiles)[a]	poorest	10,046 (33.4)	413	1,565 (17.8)	110	1,745 (23.3)	23		
	middle	10,839 (36.0)	369	3,667 (41.6)	225	3,534 (47.3)	27		
	least poor	9,228 (30.6)	259	3,584 (40.7)	182	2,199 (29.4)	14		
	missing	2	0	1	1			3,222	119
1 ANC visit	no	12,875 (42.8)	488	2,541 (28.8)	169	2,057 (27.5)	22	2,676 (83.3)	98
	yes	17,236 (57.2)	553	6,273 (71.2)	349	5,421 (72.5)	42	535 (16.7)	21
	missing	4	0	3	0			11	0
4+ ANC visits	no	25,350 (84.2)	897	6,784 (76.9)	419	2,483 (33.2)	33	3,079 (95.6)	113
	yes	4,764 (15.8)	144	2,033 (23.1)	99	4,995 (66.8)	31	141 (4.4)	6
	missing	1	0					2	0
tetanus vaccination	no	11,759 (39.0)	426	1,485 (16.8)	110	474 (6.3)	13	215 (21.4)	8
	yes	18,354 (61.0)	615	7,332 (83.2)	408	7,004 (93.7)	51	790 (78.6)	33
	missing	2	0					2,217	78
premature	no	28,332 (94.7)	672	8,290 (94.9)	382	7,082 (95.0)	27	3,140 (97.5)	84
	yes	1,600 (5.3)	362	445 (5.1)	130	374 (5.0)	15	82 (2.5)	35
	missing	183	7	82	6	22	22		
pregnancy complications	no	25,495 (84.7)	775	6,860 (77.8)	372	7,334 (98.9)	0		
	yes	4,588 (15.3)	265	1,957 (22.2)	146	80 (1.1)	0		
	missing	32	1			64	64	3,222	119

Table 21.1 Continued.									
		Rural Bangladesh		Rural India		Urban India		Rural Nepal	
		deliveries (%)	nnd	deliveries (%]	nnd	deliveries (%]	nnd	deliveries (%)	nnd
season[b]	warm	7,307 (24.3)	233	3,106 (35.2)	157	1,923 (25.7)	15	1,407 (43.7)	44
	rainy	12,106 (40.2)	416	2,944 (33.4)	158	2,380 (31.8)	20	1,013 (31.4)	30
	cold	10,702 (35.5)	392	2,767 (31.4)	203	3,175 (42.5)	29	802 (24.9)	45
delivery location	home	23,487 (78.6)	773	7,031 (79.9)	428	952 (12.7)	17	3,162 (98.1)	115
	institutional	6,403 (21.4)	260	1,769 (20.1)	90	6,526 (87.3)	47	60 (1.9)	4
	missing	225	8	17	0				
labour duration > 24h	no	23,994 (79.7)	770	7,511 (85.2)	405	7,271 (97.3)	57	2,618 (81.3)	81
	yes	6,106 (20.3)	270	1,305 (14.8)	113	202 (2.7)	2	604 (18.7)	38
	missing	15	1	1	0	5	5		
delivery complications	no	27,898 (92.9)	881	7,252 (82.3)	393	7,364 (98.5)	64	1,782 (55.3)	51
	yes	2,123 (7.1)	157	1,558 (17.7)	125	114 (1.5)	0	1,439 (44.7)	68
	missing	94	3	7	0			1	0
presentation	breech	559 (1.9)	75	96 (1.1)	29			18 (0.6)	3
	normal	25,955 (86.8)	868	8,509 (97.6)	475			3,186 (99.4)	115
	caesarean	3,396 (11.4)	81	117 (1.3)	5	1,127 (100.0)	0	2 (0.1)	0
	missing	205	17	95	9	6,351	64	16	1
mother died	no	30,065 (99.8)	1,034	8,774 (99.5)	510	7,475 (100.0)	63	3,209 (99.6)	115
	yes	50 (0.2)	7	43 (0.5)	8	3 (0.0)	1	13 (0.4)	4
sex baby	male	15,536 (51.6)	615	4,469 (50.7)	302	3,901 (52.6)	0	1,692 (52.5)	75
	female	14,579 (48.4)	426	4,348 (49.3)	216	3,513 (47.4)	0	1,530 (47.5)	44
	missing					64	64		
multiple birth	no	29,551 (98.1)	884	8,613 (97.7)	449	7,353 (98.3)	58	3,162 (98.1)	110
	yes	564 (1.9)	157	204 (2.3)	69	125 (1.7)	6	60 (1.9)	9
size at birth	small	5,400 (17.9)	394	606 (6.9)	146	877 (11.9)	0	121 (3.8)	38
	normal	22,119 (73.5)	509	8,150 (92.4)	366	4,304 (58.5)	0	3,042 (94.4)	74
	large	2,595 (8.6)	138	61 (0.7)	6	2,178 (29.6)	0	59 (1.8)	7
	missing	1	0			119	64		
looking abnormal	no	21,973 (92.4)	592	8,422 (95.6)	437	7,433 (99.4)	54	3,149 (97.7)	95
	yes	1,810 (7.6)	254	392 (4.4)	80	45 (0.6)	10	73 (2.3)	24
	missing	6,332	195	3	1				
breathed & cried immediately	no	3,977 (13.2)	402	41 (0.5)	8	117 (1.6)	19		
	yes	26,138 (86.8)	639	8,776 (99.5)	510	7,359 (98.4)	43		
	missing					2	2	3,222	119
condition at 5 mins.	poor	1,826 (6.1)	387	281 (3.2)	145				
	good	27,923 (93.9)	622	8,441 (96.8)	351				
	missing	366	32	95	22	7,478	64	3,222	119
condition arms & legs	normal	23,598 (99.1)	788	8,677 (98.4)	438				
	floppy	174 (0.7)	47	112 (1.3)	71				
	stiff	39 (0.2)	10	28 (0.3)	9				
	missing	6,304	196			7,478	64	3,222	119

a. Household wealth indicators included in the Principal Components Analysis were as follows: Bangladesh (electricity, radio/tape recorder, fan, television, telephone, generator, bicycle, fridge), rural India (electricity, radio/tape recorder, fan, television, generator, bicycle, fridge), urban India (electricity, radio/tape recorder, fan tv fridge telephone bicycle fridge), Nepal (not available, imputed based on other sites).

b. Season was defined as follows: Bangladesh (rainy: June - October, cold: November - February, warm: March - May), rural India (rainy: July - October, cold: November - February, warm: March - June), urban India (rainy: June - September, cold: November - March, warm: October, April - May), Nepal: (rainy: June - mid-September; cold: mid-November – mid-February; warm: mid-September – mid-November & mid-February - May).

RESULTS

1,742 neonatal deaths occurred in 49,632 live births across the sites, with the NMR varying from 58.8/1000 in rural India, to 36.9/1000 in Nepal, 34.6/1000 in Bangladesh, and 8.6/1000 in urban India (Table 21.1).

The following characteristics were very strongly associated neonatal death (univariable odds ratios (ORs); Supplementary table 21.1): breech delivery, premature birth, mother died, multiple birth, small size at birth, looking abnormal, not immediately crying or breathing, poor condition at five minutes, and infant had floppy or stiff arms and legs. The other included characteristics were also associated, though less strongly, with neonatal death in most sites.

Table 21.2 presents the prediction models. At the start of pregnancy, a high educational attainment was associated with a lower odds of death and low economic status was associated with a higher odds of death (Table 21.2). Also, a very short birth interval, and births to primigravid, younger (especially <18 years), and older (35+) women were associated with a higher odds of death. Socioeconomic (ΔAIC education: 36; economic status: 9) and demographic characteristics (ΔAIC birth interval: 39; maternal age: 15) were equally strong predictors of neonatal death. At the start of pregnancy, predictive ability of the model was moderate (apparent AUC: 0.59 [95%CI 0.58-0.61]; cross-validated AUC 0.58 [95%CI 0.57-0.59]).

At the start of delivery, prematurity was a very strong predictor of neonatal death (ΔAIC: 1642; OR 10.93 [95%CI 9.74-12.27]). Less strong, but still predictive, were health problems during pregnancy, and delivery in the cold season. Low maternal socioeconomic position and short birth interval were also important predictors. Predictive ability at the start of delivery was considerably better than at the start of pregnancy (AUC: 0.71 [95%CI 0.67-0.74]). Multiple pregnancy was a strong predictor of neonatal death (ΔAIC: 514; OR 7.70 [95%CI 6.46-9.18]). When information about multiple pregnancy was available at start of delivery, predictive ability improved substantively (apparent AUC: 0.73 [95%CI 0.69-0.76]; cross-validated AUC 0.72 [95%CI 0.69-0.75]).

At five minutes post-partum, prematurity (ΔAIC: 788; OR 7.63 [6.63-8.80]), a poor condition of the infant (ΔAIC: 1052; OR 10.04 [95%CI 8.74-11.54]), and multiple birth were (ΔAIC: 341; OR 6.96 [95%CI 5.67-8.54]) highly predictive of neonatal death. Less predictive, but still important, were low maternal education, short birth interval, floppy or stiff arms and legs of the baby, small or large infant size at birth, breech delivery, male infant, health problems during pregnancy, and delivery in the cold season. Predictive ability of this model was high (apparent AUC: 0.82 [95%CI 0.80-0.84]; cross-validated AUC 0.81 [95%CI 0.79-0.83]). A substantial proportion of deaths was associated with the three risk factors with the highest ΔAIC at time of delivery (62.4% of deaths and 12.9% of births had any one of these risk factors).

Table 21.2 Multivariable associations between neonatal mortality and risk factors at start of pregnancy, start of delivery and five minutes after birth. ΔAIC is reported behind the predictors in bold font; odd ratios (95% confidence intervals) are reported behind predictor levels in regular font.

Predictor	Level	Start pregnancy	Start delivery	After birth	Start delivery (incl multiple birth)
site		**198**	**165**	**136**	**171**
	Rural Bangladesh	1	1	1	1
	Rural India	1.66 (1.46,1.88)	1.61 (1.40,1.84)	2.01 (1.73,2.33)	1.62 (1.41,1.86)
	Urban India	0.27 (0.20,0.34)	0.27 (0.21,0.35)	0.38 (0.28,0.53)	0.25 (0.19,0.33)
	Rural Nepal	0.79 (0.64,0.98)	0.93 (0.75,1.16)	1.19 (0.94,1.52)	0.95 (0.76,1.18)
age		**15**			
	<18	1.55 (1.26,1.92)			
	19-20	1.30 (1.14,1.48)			
	21-23	1.11 (1.05,1.17)			
	24-26	1			
	27-30	1.00 (0.96,1.03)			
	30-32	1.05 (0.98,1.13)			
	33-35	1.15 (1.01,1.30)			
	>35	1.28 (1.05,1.56)			
birth interval (months)		**39**	**45**	**23**	**60**
	primi gravida	1.34 (1.14,1.58)	1.40 (1.23,1.59)	1.26 (1.10,1.46)	1.50 (1.32,1.72)
	<15	2.24 (1.76,2.84)	1.95 (1.53,2.48)	1.89 (1.43,2.49)	2.04 (1.60,2.60)
	15-26	1.11 (0.93,1.32)	1.00 (0.84,1.20)	1.01 (0.84,1.22)	1.03 (0.86,1.24)
	27-68	1	1	1	1
	>68	1.15 (0.96,1.39)	1.07 (0.88,1.29)	1.03 (0.83,1.27)	1.04 (0.85,1.27)
education		**36**	**52**	**44**	**51**
	no school	1	1	1	1
	primary	0.84 (0.73,0.97)	0.79 (0.69,0.92)	0.80 (0.69,0.94)	0.81 (0.70,0.93)
	secondary	0.65 (0.57,0.75)	0.60 (0.52,0.69)	0.61 (0.53,0.70)	0.61 (0.53,0.70)
	BSc/MSc	0.37 (0.19,0.73)	0.29 (0.14,0.57)	0.34 (0.17,0.69)	0.24 (0.12,0.48)
household wealth (tertiles)		**9**	**9**		**9**
	1	1.28 (1.12,1.48)	1.30 (1.12,1.50)		1.31 (1.13,1.51)
	2	1.12 (0.99,1.27)	1.09 (0.96,1.24)		1.10 (0.96,1.25)
	3	1	1		1
premature			**1642**	**788**	**1359**
	no		1	1	1
	yes		10.93 (9.74,12.27)	7.63 (6.63,8.80)	9.50 (8.43,10.70)
pregnancy complications			**46**	**21**	**40**
	no		1	1	1
	yes		1.53 (1.36,1.73)	1.40 (1.22,1.60)	1.50 (1.33,1.70)
season			**13**	**23**	**15**
	warm		1	1	1
	rainy		1.00 (0.88,1.14)	1.06 (0.92,1.22)	1.01 (0.89,1.15)
	cold		1.24 (1.09,1.41)	1.38 (1.20,1.59)	1.27 (1.11,1.44)

Table 21.2 Continued.					
Predictor	**Level**	**Start pregnancy**	**Start delivery**	**After birth**	**Start delivery (incl multiple birth)**
presentation				**45**	
	caesarean			0.51 (0.39,0.66)	
	breech			1.85 (1.42,2.42)	
	normal			1	
sex baby				**28**	
	male			1.37 (1.23,1.54)	
	female			1	
multiple birth				**341**	**514**
	no			1	1
	yes			6.96 (5.67,8.54)	7.70 (6.46,9.18)
size at birth				**61**	
	small			1.36 (1.18,1.57)	
	normal			1	
	large			2.12 (1.74,2.57)	
condition at 5m				**1052**	
	poor			10.04 (8.74,11.54)	
	good			1	
condition arms				**87**	
	normal			1	
	floppy			5.25 (3.69,7.47)	
	stiff			1.94 (0.99,3.79)	
AUC apparent validation					
Rural Bangladesh		0.59 (0.58,0.61)	0.73 (0.71,0.75)	0.83 (0.81,0.84)	0.75 (0.74,0.77)
Rural India		0.60 (0.57,0.63)	0.68 (0.65,0.71)	0.80 (0.78,0.82)	0.71 (0.68,0.73)
Urban India		0.61 (0.53,0.69)	0.70 (0.62,0.78)	0.82 (0.75,0.89)	0.70 (0.61,0.78)
Rural Nepal		0.54 (0.48,0.60)	0.73 (0.66,0.80)	0.84 (0.80,0.89)	0.74 (0.68,0.80)
Pooled average		0.59 (0.58,0.61)	0.71 (0.67,0.74)	0.82 (0.80,0.84)	0.73 (0.69,0.76)
AUC cross validation					
Rural Bangladesh		0.58 (0.56,0.60)	0.72 (0.70,0.74)	0.82 (0.80,0.83)	0.74 (0.72,0.76)
Rural India		0.59 (0.56,0.61)	0.67 (0.65,0.70)	0.79 (0.77,0.82)	0.70 (0.67,0.72)
Urban India		0.60 (0.52,0.68)	0.70 (0.61,0.78)	0.80 (0.73,0.88)	0.70 (0.61,0.78)
Rural Nepal		0.54 (0.48,0.60)	0.73 (0.66,0.79)	0.84 (0.80,0.88)	0.73 (0.67,0.79)
Pooled average		0.58 (0.57,0.59)	0.70 (0.67,0.73)	0.81 (0.79,0.83)	0.72 (0.69,0.75)

The prognostic nomograms corresponding to the three models are presented in the nomograms of Figures 21.1-3 (see explanation underneath Figure 21.1). Using Figure 21.3, for example, a singleton male infant (0.7 points), with a small size at birth (0.7 points), who presented normally (1.5 points), but was born prematurely (4.4 points) in the cold season (0.7 points) in rural India (3.6 points), to a primigravid (0.5 points) mother with no schooling (2.3 points), had an estimated mortality risk of 348/1,000 if the infant was in good condition at five minutes, with arms/legs in normal condition. If the same infant was in a poor condition at five minutes (5 points), but with arms/legs in normal condition, the mortality risk amounted to 843/1,000.

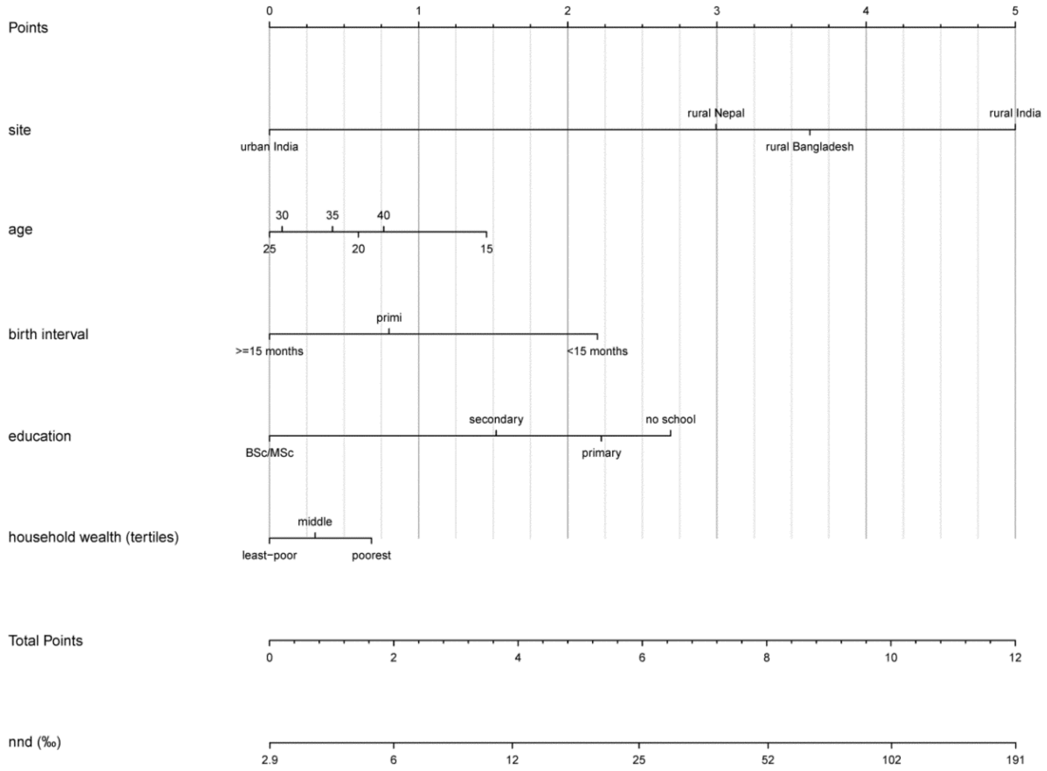


Figure 21.1 Nomogram of the prediction of neonatal mortality at start of pregnancy. To estimate an infant's probability of neonatal death, first determine all its risk factor characteristics (educational attainment of its mother, (estimated) birth interval etc.). Second, read the risk points associated with each risk factor by drawing a line up from the predictor value to the "Points"-axis. Third, add up the points for all risk factors to obtain the total points for that infant. The probability of neonatal death can be read by moving vertically from the "Total Points" axis to the "nnd" axis. The predictor 'site' can be used to take regional differences in NMR into account. When using the nomograms outside our study populations, readers are advised to use the site with an NMR closest to their own study population.

DISCUSSION

We developed and validated prognostic models for neonatal mortality in the general population in low and middle income countries, with specific reference to South Asia, on the basis of risk factors known at (i) the start of pregnancy, (ii) the start of delivery, and (iii) five minutes post-partum. At the start of pregnancy, prediction of neonatal death was difficult, although infants born to women of lower socioeconomic position and to women with certain demographic characteristics (young or advanced age, very short birth interval, primigravida) were at a higher risk of neonatal death. Predictive ability improved at the start of delivery, where multiple pregnancy and a premature start of delivery were highly predictive of neonatal death. Predictive ability was high at five minutes post-partum, where prematurity, multiple birth, and a poor condition of the infant were strong predictors of death. The models can be used to inform population-based prevention and more narrowly targeted interventions for high-risk infants.

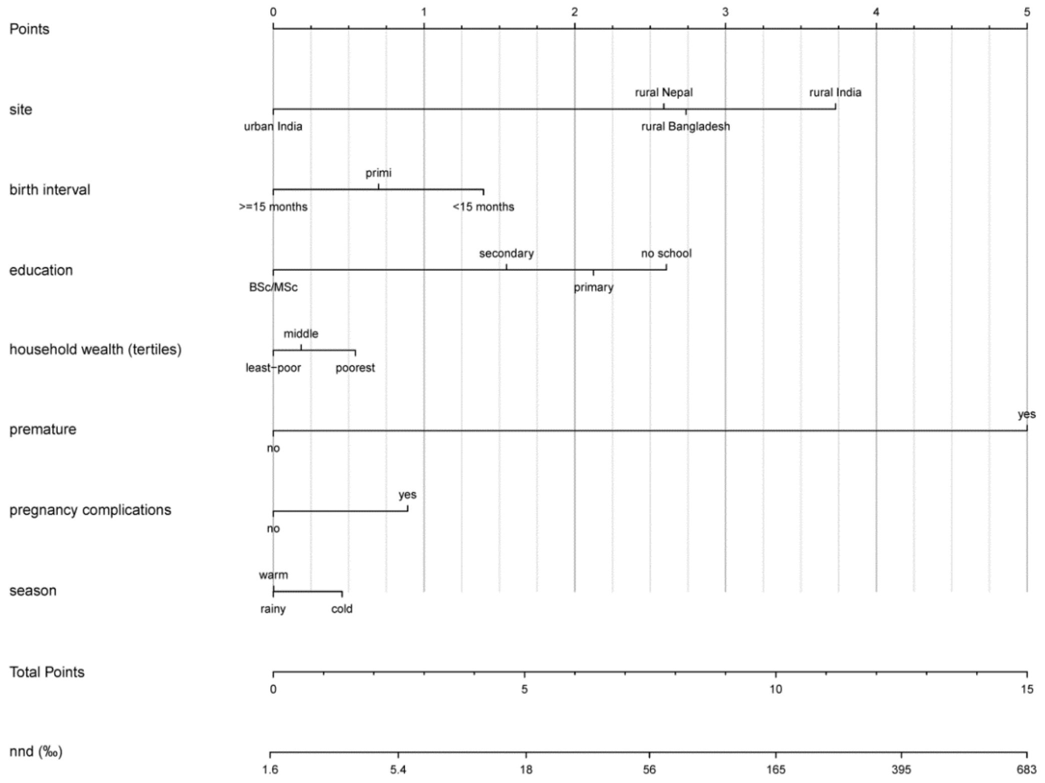


Figure 21.2a Nomogram of the prediction of neonatal mortality at start of delivery (without information on singleton/multiple pregnancy).

Methodological issues

Our models are based on large datasets from sites in which the full population was prospectively followed-up and detailed information on predictors of neonatal death was collected, allowing for precise prediction. Yet, recall bias is a potential problem, as information was based on mother’s report at approximately 6 weeks post-partum. While we reduced this problem by using broad categories for variables like size at birth, random error may remain substantial for such variables. Furthermore, mother’s report may have been biased by the outcome (death/survival), with worse conditions reported for neonatal deaths, leading to inflated odds ratios for characteristics that mothers associate with death (e.g. infant condition at five minutes). Yet, for other predictors, like multiple birth, such recall bias is probably minimal. Finally, while the high number of missing values in some predictors in particular sites may be considered a limitation, we were able to develop our models based on multiple imputation of missing values using the substantial amount of available data. Nevertheless, this may have led to an underestimation of the discriminative ability of the models. Despite these problems, we arguably used the best data available for general populations in poor settings, where home births without skilled care are common and reliable vital registration systems are non-existent.

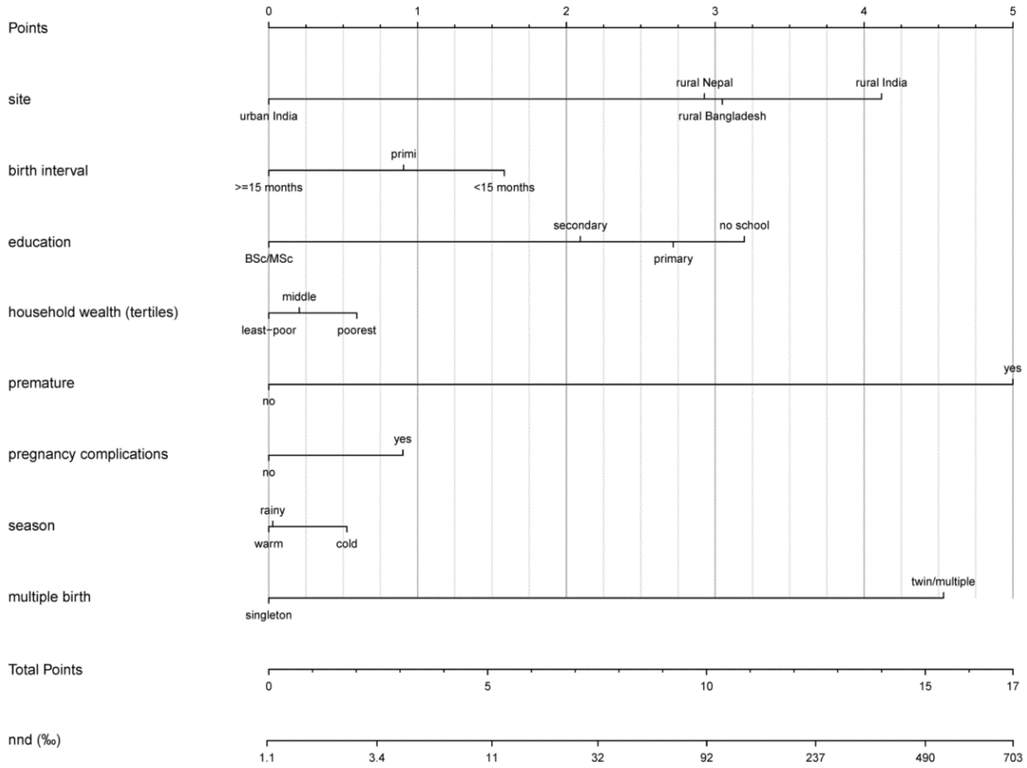


Figure 21.2b Nomogram of the prediction of neonatal mortality at start of delivery (with information on singleton/multiple pregnancy).

Our models are arguably generalizable to rural and poor urban South Asia. Our study sites ranged from informal settlements in megacity Mumbai, with a comparatively low NMR, to tribal areas in some of the poorest states in India, with a high NMR. The discriminative ability of the models – measured by the apparent and cross-validated AUC – was stable across sites, implying that the models are generally applicable across our study population.

Our models are possibly less applicable to the top-layer of South Asian society with a different cause-of-death pattern. Furthermore, their wider generalizability to, for example the African context, needs further examination.

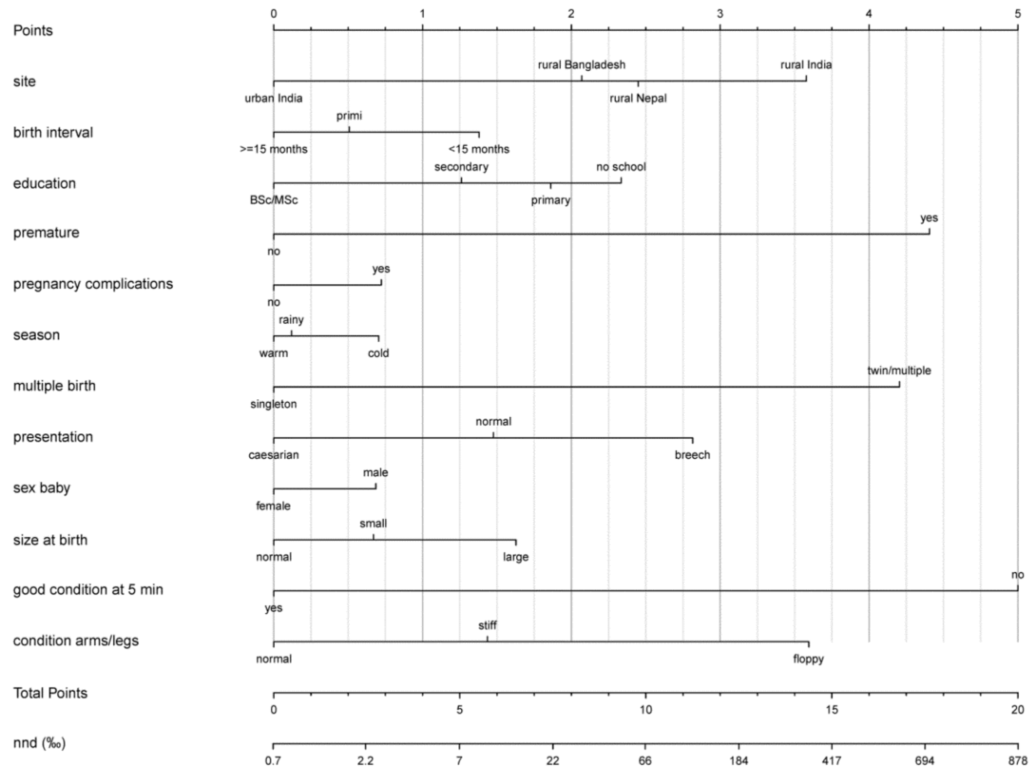


Figure 21.3 Nomogram of the prediction of neonatal mortality at five minutes after delivery.

Comparison with the literature and implications

To our knowledge, our study is the first to formally combine known risk factors for neonatal mortality into a prediction model for the general population in low and middle income countries. We developed models for three time points, i.e. onset of pregnancy, onset of delivery, immediately after birth, something we rarely encountered in the literature.

We found that three risk factors, - preterm birth, multiple birth, and poor condition at five minutes post-partum - were associated with a very high risk of neonatal death. A substantial proportion of deaths was associated with these risk factors. Tertiary prevention (improving outcomes among infants with these risk factors, rather than reducing risk factor prevalence) can play an important role in preventing these deaths. Facility-based interventions to improve management of high-risk infants exist for poor settings [25,26]. While timely access to skilled care can be critical, it is often problematic in poor rural areas.

Health system strengthening to improve quality and availability of care, and demand-side interventions (e.g. conditional cash transfers) to reduce care-seeking delays, are therefore important. Interventions also exist for community settings, including participatory women's groups and home-based neonatal care by village health workers [26,27]. Community-based management requires that care-givers are aware of important risk factors and react pro-actively to danger signs [28]. This means anticipating potential problems in women with a multiple pregnancy and/or premature start of delivery where there is still time to travel to a facility, and early recognition and home-management of problems among preterm infants and babies in a poor condition (e.g. bag-and-mask ventilation, kangaroo care, delayed bathing) [29,30]. Raising awareness about the importance of the above risk factors within community-based interventions, and empowering families and communities to address these problems, is therefore recommended. Similarly, these strategies can be used for the other described risk factors, including breech delivery (timely recognition and care seeking) and delivery in the cold season (thermal care). Also, while infants are at the highest risk of death on the day of birth,1 these strategies are equally important for the late neonatal period (comprising 20-50% of deaths in our sites). So, rather than being competing strategies, population-level interventions to raise awareness and empower communities to act are a prerequisite for effective tertiary prevention in settings where home births without professional care are the norm.

Combining the above strategies with population-level primary prevention to reduce the incidence of risk factors, e.g. by improving maternal nutrition, reducing indoor pollution, and family planning, will help further reduce neonatal mortality [1,31]. Similarly, measures to improve living conditions and hygienic practices are important. 62.4% of deaths in our sites occurred among infants without the three main risk factors; infections may have played an important role in these deaths, as well as in the death of high-risk infants [1].

CONCLUSION

We developed good performing prediction models for neonatal mortality in the general population in low and middle income countries. Based on our models, we conclude that improved management of high-risk infants can substantially reduce neonatal mortality. Population-level strategies to raise awareness about important risk factors and empower community-based care givers to take action should be an integral part of a high-risk approach in settings with weak health systems. This should be complemented with health system strengthening and action on the social determinants of health to reduce mortality in low-risk, as well as high-risk, infants.

ACKNOWLEDGEMENTS

This work was supported by the Economic and Social Research Council and the Department for International Development (grant number ES/I033572/1) and a Wellcome Trust Strategic Award (award number: 085417MA/Z/08/Z). TAJH was also supported by an EUR Research Excellence Initiative grant. DvK was supported by the Netherlands Organization for Scientific Research (grant number 917.11.383). JVB was supported by fellowship grants from the Netherlands Lung Foundation and the Erasmus University Medical Center. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. TAJH and DvK had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

REFERENCES

1. Lawn JE, Blencowe H, Oza S, et al. Every Newborn: progress, priorities, and potential beyond survival. *Lancet.* 2014;384(9938):189-205.
2. Lawn JE, Cousens S, Zupan J. 4 million neonatal deaths: when? Where? Why? *Lancet.* 2005;365(9462):891-900.
3. Houweling TA, Looman CW, Azad K, et al. The equity impact of community women's groups to reduce neonatal mortality: a meta-analysis of four cluster randomised trials. submitted.
4. Schuit E, Hukkelhoven CW, Manktelow BN, et al. Prognostic models for stillbirth and neonatal death in very preterm birth: a validation study. *Pediatrics.* 2012;129(1):e120-127.
5. Draper ES, Manktelow B, Field DJ, James D. Prediction of survival for preterm births by weight and gestational age: retrospective population based study. *BMJ.* 1999;319(7217):1093-1097.
6. Medlock S, Ravelli AC, Tamminga P, Mol BW, Abu-Hanna A. Prediction of mortality in very premature infants: a systematic review of prediction models. *PLoS One.* 2011;6(9):e23441.
7. Li L, Yu J, Wang J, et al. A prediction score model for risk factors of mortality in neonate with pulmonary hemorrhage: the experience of single neonatal intensive care unit in Southwest China. *Pediatr Pulmonol.* 2008;43(10):997-1003.
8. Houweling TA, Ronsmans C, Campbell OM, Kunst AE. Huge poor-rich inequalities in maternity care: an international comparative study of maternity and child care in developing countries. *Bull World Health Organ.* 2007;85(10):745-754.
9. Dhaded SM, Somannavar MS, Vernekar SS, et al. Neonatal mortality and coverage of essential newborn interventions 2010 - 2013: a prospective, population-based study from low-middle income countries. *Reprod Health.* 2015;12 Suppl 2:S6.
10. Conde-Agudelo A, Rosas-Bermudez A, Kafury-Goeta AC. Birth spacing and risk of adverse perinatal outcomes: a meta-analysis. *JAMA.* 2006;295(15):1809-1823.
11. Houweling TA, Kunst AE. Socio-economic inequalities in childhood mortality in low and middle income countries: a review of the international evidence. *British Medical Bulletin.* 2010;93(1):7-26.
12. Barnett S, Nair N, Tripathy P, Borghi J, Rath S, Costello A. A prospective key informant surveillance system to measure maternal mortality - findings from indigenous populations in Jharkhand and Orissa, India. *BMC Pregnancy Childbirth.* 2008;8:6.
13. Tripathy P, Nair N, Barnett S, et al. Effect of a participatory intervention with women's groups on birth outcomes and maternal depression in Jharkhand and Orissa, India: a cluster-randomised controlled trial. *Lancet.* 2010;375(9721):1182-1192.
14. Manandhar DS, Osrin D, Shrestha BP, et al. Effect of a participatory intervention with women's groups on birth outcomes in Nepal: cluster-randomised controlled trial. *Lancet.* 2004;364(9438):970-979.
15. More NS, Bapat U, Das S, et al. Community mobilization in Mumbai slums to improve perinatal care and outcomes: a cluster randomized controlled trial. *PLoS Med.* 2012;9(7):e1001257.
16. Fottrell E, Azad K, Kuddus A, et al. The Effect of Increased Coverage of Participatory Women's Groups on Neonatal Mortality in Bangladesh: A Cluster Randomized Trial. *JAMA Pediatr.* 2013;167(9):816-825.
17. Darmstadt GL, Bhutta ZA, Cousens S, Adam T, Walker N, de Bernis L. Evidence-based, cost-effective interventions: how many newborn babies can we save? *Lancet.* 2005;365(9463):977-988.
18. Roy SS, Mahapatra R, Rath S, et al. Improved neonatal survival after participatory learning and action with women's groups: a prospective study in rural eastern India. *Bull World Health Organ.* 2013;91(6):426-433B.
19. Filmer D, Pritchett LH. Estimating wealth effects without expenditure data--or tears: an application to educational enrolments in states of India. *Demography.* 2001;38(1):115-132.

20. Van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations *R. Journal of Statistical Software.* 2011;45(3):1-67.
21. *R: A Language and Environment for Statistical Computing, Version 2.15.3* [computer program]. Vienna R Foundation for Statistical Computing; 2013.
22. Harrell F. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* New York Springer-Verlag 2001.
23. van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. *BMC Med Res Methodol.* 2014;14:5.
24. Steyerberg E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* New York: Springer; 2009.
25. Lee AC, Katz J, Blencowe H, et al. National and regional estimates of term and preterm babies born small for gestational age in 138 low-income and middle-income countries in 2010. *The Lancet. Global health.* 2013;1(1):e26-36.
26. March of Dimes, PMNCH, Save the Children, WHO. *Born Too Soon: The Global Action Report on Preterm Birth.* Geneva: World Health Organization;2012.
27. Houweling TA, Tripathy P, Nair N, et al. The equity impact of participatory women's groups to reduce neonatal mortality in India: secondary analysis of a cluster-randomised trial. *Int J Epidemiol.* 2013;42(2):520-532.
28. Mesko N, Osrin D, Tamang S, et al. Care for perinatal illness in rural Nepal: a descriptive study with cross-sectional and qualitative components. *BMC Int Health Hum Rights.* 2003;3(1):3.
29. Wall SN, Lee AC, Niermeyer S, et al. Neonatal resuscitation in low-resource settings: what, who, and how to overcome challenges to scale up? *Int J Gynaecol Obstet.* 2009;107 Suppl 1:S47-62, S63-44.
30. Lawn JE, Mwansa-Kambafwile J, Barros FC, Horta BL, Cousens S. 'Kangaroo mother care' to prevent neonatal deaths due to pre-term birth complications. *Int J Epidemiol.* 2011;40(2):525-528.
31. Ramakrishnan U, Grant FK, Goldenberg T, Bui V, Imdad A, Bhutta ZA. Effect of multiple micronutrient supplementation on pregnancy and infant outcomes: a systematic review. *Paediatr Perinat Epidemiol.* 2012;26 Suppl 1:153-167.

Supplementary table 21.1 Univariable odds ratios (95% confidence interval) for all risk factors, by study site.

Predictor	Level	Start pregnancy	Start delivery	After birth	Start delivery (incl multiple birth)
time (years)	1	1	1	1	1
	2	1.00 (0.82,1.22)	1.15 (0.92,1.43)	1.06 (0.59,1.91)	0.81 (0.56,1.18)
	3	0.82 (0.67,1.01)	1.24 (1.00,1.55)	1.02 (0.55,1.89)	
	4	0.80 (0.65,0.99)			
	5	0.78 (0.62,0.97)			
	6	0.72 (0.59,0.89)			
age	<18	1.61 (1.13,2.29)	2.40 (1.50,3.83)	2.33 (0.30,17.92)	2.19 (0.72,6.66)
	19-20	1.60 (1.32,1.94)	1.65 (1.25,2.17)	1.16 (0.53,2.53)	1.00 (0.49,2.04)
	21-23	1.32 (1.07,1.64)	1.04 (0.77,1.40)	0.80 (0.37,1.71)	0.87 (0.47,1.60)
	24-26	1	1	1	1
	27-30	1.29 (1.02,1.65)	1.10 (0.79,1.53)	1.63 (0.77,3.45)	0.80 (0.39,1.63)
	30-32	1.42 (1.11,1.82)	0.92 (0.65,1.30)	1.19 (0.46,3.08)	1.32 (0.69,2.53)
	33-35	1.37 (0.99,1.88)	1.20 (0.80,1.81)	2.24 (0.74,6.80)	1.46 (0.72,2.99)
	>35	1.82 (1.34,2.47)	0.84 (0.48,1.48)	1.95 (0.44,8.62)	0.95 (0.48,1.88)
birth interval (months)	primi gravida	1.30 (1.11,1.52)	1.91 (1.55,2.35)		
	<15	2.70 (1.95,3.74)	1.85 (1.20,2.86)		
	15-26	1.16 (0.91,1.47)	1.15 (0.88,1.49)		
	27-68	1	1		
	>68	1.07 (0.87,1.32)	1.24 (0.64,2.38)		
education	no school	1	1	1	1
	primary	0.81 (0.69,0.94)	0.93 (0.62,1.40)	1.01 (0.39,2.66)	0.80 (0.40,1.59)
	secondary	0.60 (0.51,0.70)	0.81 (0.65,1.00)	0.58 (0.34,0.98)	1.11 (0.48,2.57)
	BSc/MSc	0.36 (0.15,0.89)	0.40 (0.10,1.62)	0.11 (0.02,0.83)	12.94 (1.16,143.88)
illiterate	no	1	1	1	1
	yes	1.50 (1.32,1.71)	1.20 (0.98,1.46)	1.27 (0.74,2.18)	1.07 (0.68,1.67)
household wealth	1	1.48 (1.27,1.74)	1.41 (1.11,1.80)	2.08 (1.07,4.06)	
	2	1.22 (1.04,1.43)	1.22 (1.00,1.49)	1.20 (0.63,2.30)	
	3	1	1	1	
1 ANC visit	no	1.19 (1.05,1.35)	1.21 (1.00,1.46)	1.38 (0.82,2.33)	0.93 (0.58,1.50)
	yes	1	1	1	1
4+ ANC visits	no	1.18 (0.98,1.41)	1.29 (1.03,1.61)	2.16 (1.32,3.53)	0.86 (0.37,1.98)
	yes	1	1	1	1
tetanus vaccination	no	1.08 (0.96,1.23)	1.36 (1.09,1.69)	3.84 (2.08,7.12)	0.89 (0.40,1.95)
	yes	1	1	1	1
premature	no	1	1	1	1
	yes	12.0 (10.5,13.8)	8.5 (6.8,10.7)	10.9 (5.8,20.7)	27.1 (16.6,44.2)
pregnancy complications	no	1	1		
	yes	1.96 (1.69,2.26)	1.41 (1.15,1.71)		
season	warm	1	1	1	1
	rainy	1.08 (0.92,1.27)	1.07 (0.85,1.34)	1.08 (0.55,2.11)	0.95 (0.59,1.51)
	cold	1.15 (0.98,1.36)	1.49 (1.20,1.84)	1.17 (0.63,2.19)	1.84 (1.20,2.82)

Supplementary table 21.1 Continued.

Predictor	Level	Start pregnancy	Start delivery	After birth	Start delivery (incl multiple birth)
delivery location	home	1	1	1	1
	institutional	1.24 (1.08,1.44)	0.83 (0.65,1.04)	0.40 (0.23,0.70)	1.89 (0.67,5.31)
labour duration	No	1	1	1	1
>24h	yes	1.40 (1.21,1.61)	1.66 (1.34,2.07)	1.27 (0.31,5.22)	2.10 (1.42,3.12)
delivery	no	1	1	1	1
complications	yes	2.45 (2.05,2.92)	1.52 (1.24,1.88)		1.68 (1.16,2.44)
presentation	no	4.48 (3.48,5.76)	7.32 (4.69,11.43)		5.34 (1.52,18.71)
	yes	1	1		1
	caesarean	0.71 (0.56,0.89)	0.76 (0.31,1.86)		
mother died	no	1	1	1	1
	yes	4.57 (2.05,10.2)	3.70 (1.71,8.03)	58.8 (5.27,657)	12.0 (3.63,39.4)
sex baby	male	1.37 (1.21,1.55)	1.39 (1.16,1.66)		1.57 (1.07,2.29)
	female	1	1		1
multiple birth	no	1	1	1	1
	yes	12.5 (10.3,15.2)	9.29 (6.85,12.6)	6.34 (2.68,15.0)	4.90 (2.35,10.20)
size at birth	small	3.34 (2.92,3.82)	6.75 (5.45,8.36)		18.4 (11.7,28.7)
	normal	1	1		1
	large	2.38 (1.97,2.89)	2.32 (0.99,5.42)		5.40 (2.37,12.28)
looking	no	1	1	1	1
abnormal	yes	5.90 (5.05,6.89)	4.69 (3.60,6.10)	39.0 (18.4,82.8)	15.8 (9.27,26.7)
breathed & cried	no	4.49 (3.94,5.11)	3.93 (1.81,8.55)	33.0 (18.6,58.7)	
immediately	yes	1	1	1	
condition	poor	11.8 (10.3,13.5)	24.6 (19.0,31.8)		
at 5 mins.	good	1	1		
condition	normal	1	1		
arms & legs	floppy	10.7 (7.6,15.1)	32.6 (21.9,48.4)		
	stiff	9.98 (4.9,20.6)	8.91 (4.0,19.8)		

22

General discussion

William Osler noted in 1893 that “If it were not for the great variability between individuals, medicine might as well be a science, not an art” [1]. In contrast, this thesis is based on the scientific paradigm that well performing prediction models have the potential to guide medical decisions by exploiting identifiable heterogeneity across individual patients. In PART I we proposed several methods to measure prediction model performance within clusters of patients and to quantify heterogeneity of prediction model performance across clusters. In PART II we proposed methods for development and validation of models that exploit heterogeneity of treatment benefit among patients, with an extension to modeling individualized cost-effectiveness. In PART III we applied methods for development and validation of prediction models to several case studies of guiding clinical and public health decisions. The major findings are discussed along the lines of these three parts, followed by overall conclusions.

PART I RISK HETEROGENEITY IN CLUSTERED DATA

While measuring prediction model performance has been of continuing interest, the interpretation of performance heterogeneity across study populations has only more recently received particular attention [2-9]. In PART I of this thesis we proposed several methods to measure prediction model performance within clusters of patients and to quantify heterogeneity of prediction model performance across clusters. These methods can be summarized into a framework for assessing model performance in clustered data (Box 22.1). The four steps of this framework will be discussed here.

Box 22.1 Framework for assessing model performance in clustered data

1. **Variation in model performance**

   Explore the variation in model performance by random-effects meta-analysis of cluster-specific estimates of:

   a. Calibration intercepts and slopes (chapter 3)

   b. Concordance probabilities: with Harrell’s c-index, Uno’s concordance measure or the calibrated model-based concordance (*c-mbc*; chapters 2 and 3)

   Summarize the results of the random-effects meta-analysis with a pooled estimate and a 95% prediction interval (chapter 2).

2. **Variation in case-mix**

   Explore the variation in the discriminative ability across clusters due to case-mix differences by the model-based concordance **(**chapters 4 and 5).

3. **Variation in overall regression coefficient validity**

   Explore the variation in the discriminative ability across clusters due to overall regression coefficient validity: compare the *mbc* and the *c-mbc* (chapter 5).

4. **Clusters of limited sample size**

   When clusters are small, use multilevel regression to quantify variation in:

   a. Calibration intercepts and slopes (chapter 6)

   b. Concordance probabilities: with the *c-mbc* based on random effect estimates of calibration intercepts and slopes (chapter 7).

Variation in model performance

The discriminative ability of a model may differ substantially – beyond chance – between hospitals, as illustrated by the case study of predicting mortality after traumatic brain injury (chapter 2). Because of variation beyond chance, we propose to pool cluster-specific concordance probability estimates (and other estimates of model performance) with a random effects meta-analysis, similar to pooling of study-specific treatment effect estimates [10, 11]. It has been suggested to average cluster-specific c-indexes into an overall estimate of the concordance probability, weighted by the number of comparable within-cluster patient pairs [12, 13]. This corresponds to a fixed-effect meta-analysis which assumes equal true concordance probabilities within clusters. We argue that random effects meta-analysis give better insight into the variation of the concordance probability across clusters – summarized by a variance estimate and a 95% prediction interval of the true concordance probability [14, 15]. Our approach was recently advocated for external validation of clinical prediction models [9]. We showed that the variation in performance across hospitals of a model that predicts mortality of heart failure patients is an indication of a model's geographic generalizability (transportability, chapter 3) [2]. The moderate variation across hospitals supported widespread application of the overall model.

Variation in case-mix

The estimator proposed by Gönen and Heller yields concordance probability estimates that differ between the development population and an external validation population only due to differences in patient case-mix (chapter 4) [16]. We exploit this property of model-based estimators in the model-based concordance (*mbc*), which is a closed-form alternative for the resampling-based case-mix corrected c-index [17]. We define the *mbc* for proportional hazards regression models and logistic regression models (chapter 5). The *mbc* improves our understanding of a difference between the c-statistic at model development versus the observed c-statistic at external validation. The difference between the *mbc* at model development and the *mbc* at external validation indicates the change in discriminative ability attributable to a difference in case-mix.

Variation in overall regression coefficient validity

We also propose to calculate the *mbc* for a recalibrated model (*c-mbc*), hence assessing the influence of overall regression coefficient validity on discriminative ability (chapter 5) [18]. The *c-mbc* is comparable to c-statistics in independent data and is robust to censoring of time-to-event outcomes. The difference between the *c-mbc* and the *mbc* in external validation data expresses the change in discriminative ability due to the (in)correctness of the regression coefficients. Thanks to their censoring-robustness, the *mbc* and the *c-mbc* facilitate measurements of concordance that are not influenced by differences in censoring distributions between development and the external validation settings.

Clusters of limited sample size

Cluster-specific c-indexes may be unstable in case of limited sample size, similar to extreme (fixed effect) estimates of cluster-specific calibration intercepts and slopes. We propose fitting a multilevel regression model to obtain stable estimates of cluster-level intercepts and slopes (chapters 6 and 7) [8, 19]. We use these estimates to calculate the calibrated model-based concordance (*c-mbc*), that is the expected concordance under the assumption that the random effect estimates of calibration intercepts and slopes are true (chapter 7). In simulations, the root mean squared error (rmse) of the *c-mbc* was lower than the rmse of the c-index. We basically compare an unbiased fixed effect estimator (c-index) with a less variable random effect estimator (*c-mbc*) of the concordance probability. We and others argue that unbiasedness is not the only property of an estimator that is important, and that much could be gained by compromising unbiasedness to improve the precision of an estimator [20, 21].

Limitations

The methods that constitute the proposed framework for assessing prediction model performance in clustered data were applied to a limited number of case studies. More experience with the framework – when patients are clustered in different geographical regions, hospitals, practitioners or trials – is necessary to understand its usefulness in practice.

The *mbc* and the *c-mbc* may be limited by their underlying assumptions. First, they are model-based, that is, they are based on the assumption that the true risks fit into the framework of a logistic regression model or a proportional hazards regression model. This may be a limitation compared with Harrell's c-index and Uno's concordance measure, because these pure rank-order statistics are applicable to any risk scoring system. However, because logistic regression and proportional hazards regression are commonly used to model binary outcomes and time-to-event outcomes, respectively, the *mbc* and the *c-mbc* may often be valuable. Second, the *c-mbc* assumes a linear relationship (represented by the calibration slope) between linear predictors and either the log hazard for time-to-event outcomes or the log odds for binary outcomes. Although the *c-mbc* was robust to violation of the linearity assumption in simulations, further research is necessary to understand the importance of this assumption. An alternative *c-mbc* that allows for potential non-linear relations between linear predictors and outcomes could be considered. Third, the *(c)-mbc* of proportional hazards regression models is sensitive to a violation of the proportional hazards assumption. In the presence of time-varying effects, it may be better to assess discriminative ability for a limited follow-up period. This approach was beyond the scope of our research, but we provided formulas for an *mbc* truncated at a fixed follow-up time (appendix 3 of chapter 5). When time-varying coefficients are modelled, the *c-mbc* could alternatively be based on more sophisticated conditional probabilities (extending equation 6 of chapter 5).

The methods in the fourth step of the framework – exploring variation of model performance across clusters with limited numbers of patients – depend on the ability of multilevel regression model to correctly estimate the between-cluster variances of the intercept and the slope. The minimum number of clusters needed to reliably estimate these variances is in the order of 10, but depends on the specific setting [22].

PART II HETEROGENEITY OF TREATMENT EFFECT

Large differences in risk predictions imply large differences in absolute treatment benefit between patients, even when relative treatment effects are (considered) constant [23-27]. In PART II of this thesis we proposed novel methods for development and validation of models that exploit heterogeneity of treatment benefit across patients, with an extension to modeling of individualized cost-effectiveness. These methods are summarized in Box 22.2 and will be discussed below.

Box 22.2 New methods for analyzing heterogeneity of treatment effect

1. **Estimating individualized treatment benefit**
   Estimates of absolute treatment benefit for individual patients should take carefully modeled treatment interactions into consideration. In a comparison of different modeling approaches to estimate the individual survival benefit of treatment with either CABG or PCI for patients with complex coronary artery disease, we found that omitting interactions may result in considerably different estimates of absolute treatment benefit for individual patients (chapter 8).

2. **Validating estimates of individualized treatment benefit**
   Discrimination measures for treatment selection should assess how well a model separates patients who benefit from those who do not, something that is not captured by conventional performance metrics. We proposed a treatment benefit c-statistic, which measures a prediction model's ability to predict treatment benefit. Hereto we defined observed treatment benefit by the outcome difference in *pairs* of patients matched on predicted benefit but discordant for treatment assignment (chapter 10).

3. **Estimating individualized cost-effectiveness**
   Individualized cost-effectiveness analysis should be based on estimates of long-term life expectancy at the same level of granularity as the estimates of the short-term risk reduction. Individualized estimates of cost-effectiveness of aggressive thrombolysis for patients with an acute myocardial infarction were biased when models of short-term risk reduction and life expectancy were of unequal granularity (chapter 11).

Estimating individualized treatment benefit

We compared modeling approaches to estimate the individual survival benefit of treatment with either coronary artery bypass graft surgery (CABG) or percutaneous coronary intervention (PCI) for patients with complex coronary artery disease in the SYNTAX trial (chapter 8). Treatment interactions with each of the prognostic factors fitted much better to the data compared to the treatment interaction with predicted prognosis as a single prognostic index. Penalized regression, specifically shrinking treatment interactions to the average treatment effect led mostly to similar decisions for the individual patients, although absolute risk differences between CABG and PCI predictions were smaller [28, 29]. These interactions were largely confirmed at external validation in the DELTA registry [30].

Although sub-group analysis based on interactions may be considered superior to classical sub-group analysis of single factors separately, it has similar pitfalls, such as a risk of false-positive findings if large numbers of interactions are assessed, and lack of power to detect relatively small interaction effects [31-36]. Similar to classical subgroup effect testing, our approach requires a clear biological motivation for differential mechanisms of treatment effects. In our study, more complex anatomy of the vessel makes PCI treatment a relatively less attractive treatment option. In other studies, when treatment modalities are less different or sample size is small, there may be less potential for predicting differential treatment effect. Ideally, the analysis of differential treatment effect focuses on confirming pre-specified interactions, but exploratory analyses of differential treatment effects could be considered if sample size is large. Further validation and prospective evaluation of this approach across different settings are required.

One example of potentially differential treatment effect was related to statins prescription in the American guidelines for prevention of cardiovascular disease [37]. The guidelines' assumption of a constant relative risk reduction and its subsequent focus on absolute (baseline) risk predictions were criticized, because the relative treatment effect was found to be heterogeneous across trial populations with different baseline risk [38-40]. We proposed to estimate individualized relative risk reductions in a re-analysis of these trial data, by adding a statistical treatment interaction with absolute risk predictions (chapter 9).

Validating estimates of individualized treatment benefit

For assessing a model’s decision making potential, the relevance of the conventional risk c-statistic has been questioned, because performance measures for treatment selection should assess how well a model discriminates patients who benefit from those who do not [27, 41, 42]. However, the actual benefit for individual patients is inherently unobservable, since their potential (counterfactual) outcome under the alternative therapy is not known [43, 44]. We proposed measures to validate predictions of treatment benefit, defining observed treatment benefit as the difference in outcomes between 2 patients with the same predicted benefit but discordant on treatment assignment (chapter 10). The proposed methodology gave interpretable measures of discrimination - treatment benefit c-statistics – and calibration – treatment benefit calibration plots. The increase in the treatment benefit c-statistic when comparing SYNTAX Score II with a model without treatment interactions indicated a major improvement in discriminative ability, in contrast with a negligible increase of the risk c-statistic (the classical c-statistic). Benefit graphs (chapter 8) and decision curves are alternative methods to assess the ability of a model to predict treatment benefit [45, 46]. The advantage of the methods we proposed is that they leverage simple and widely used metrics of model performance, i.e. measures for discrimination and calibration and might therefore be more easily understood by non-experts. The applicability of the benefit c-statistic needs further study across different disease domains.

Estimating individualized cost-effectiveness

Cost-effectiveness of a treatment has increasingly been recognized to be heterogeneous across individual patients [47-49]. The perceived benefit of a treatment depends on an individual valuation of the benefits and harms (including costs) of the treatment alternatives, ideally over a lifetime horizon [50]. The individualized long-term treatment benefit depends on the post-trial life expectancy, which is usually derived from less granular population life tables. However, the underlying assumption of equal post-trial life expectancy for low and high risk patients of the same sex and age is unrealistic [51]. Therefore, we studied two examples. Both had a negative correlation between short-term mortality benefit and post-trial life expectancy (chapter 11). Myocardial infarction patient at high age who were treated with aggressive thrombolysis had a high short-term mortality risk reduction but a low life expectancy, and similarly, patients with a severe myocardial infarction (Killip class IV) had a high risk reduction but a low life expectancy. We found a substantial impact on individualized cost-effectiveness estimates of modeling individualized instead of average long-term life expectancy. Consequently, when individualizing cost-effectiveness estimates, we recommended modeling short-term benefit predictions and life expectancies at equal levels of granularity to avoid overenthusiasm about the heterogeneity in cost-effectiveness. Rather than limiting the individualization of short-term risk we would like to promote a better understanding of the consequences of unequally granular risk models by means of extensive sensitivity analyses.

Limitations

The approaches that we proposed for analyzing heterogeneity of treatment effect were applied to a limited number of case studies. Further validation and prospective evaluation of these approaches across different settings is required. Specifically, new benchmarks are needed to interpret treatment benefit c-statistics. The treatment benefit c-statistics observed in our case studies were in a range that would typically be considered only weakly predictive for conventional risk c-statistics. Because improvements in discrimination for benefit are more relevant than improvements in performance measures for risk, these relatively small improvements might be of great clinical importance.

The causal effect of treatment for an individual patient was defined as the difference between outcomes under different treatment regimens [44]. To estimate the difference in outcome when treatment choices are changed, we used randomized data throughout PART II. In contrast, the use of observational data may produce biased treatment effect estimates when the variation between differently treated patients is not completely controlled for (residual confounding, including confounding by indication). Specifically, a recent study concluded that documented surgical ineligibility is common and associated with significantly increased long-term mortality among CAD patients undergoing PCI, even after adjustment for known risk factors [52]. Further research is necessary to understand under which conditions observational data could be exploited for estimating individualized treatment benefit.

PART III APPLICATIONS

In PART III of this thesis, methods for development and validation of prediction models were applied to several case studies of guiding clinical and public health decisions (Box 22.3).

Box 22.3 Summary of applications

We developed and validated prediction models to guide decisions in:

1. **Cardiovascular medicine**
   a. on the optimal revascularization strategy for patients with complex coronary artery disease (chapters 12-14)
   b. on the duration of Dual Anti Platelet Therapy (DAPT) after coronary stent implementation (chapter 15)

2. **Oncology**
   a. on neoadjuvant chemoradiotherapy for esophageal cancer patients (chapter 16)
   b. on axillary lymph node dissection for sentinel lymph node positive breast cancer patients (chapters 17 and 18)

3. **Public health**
   a. on selective screening for sexually transmitted infections in the general population (chapters 19 and 20)
   b. on selective antenatal, intrapartum and postnatal care of mothers and neonates in the general population of low and middle income countries (chapter 21)

Revascularization in patient with complex coronary artery disease

We developed the SYNTAX Score II based on Cox regression analysis of 8 prognostic factors that were interacting with the treatment (chapter 12). The SYNTAX score II identifies patients for whom either CABG or PCI had a more favorable long-term outlook, and patients for whom long-term outlooks between CABG and PCI were much the same. Other studies combined the anatomical complexity – as assessed by the SYNTAX score – with cardiac-surgery-based risk scores to guide decision making between PCI and CABG [53, 54]. The SYNTAX score II further improved decision making between PCI and CABG by augmenting the anatomical SYNTAX score with anatomical and clinical variables.

The SYNTAX Score II showed good agreement between predictions and observed outcomes in both treatment cohorts (PCI and CABG) of a large Japanese all-comers patient registry (chapter 13). This analysis, being retrospective, could not assess the treatment recommendations based on SYNTAX Score II for the simple fact that the treatment decisions were likely made based on a combination of measured variables (as included in SYNTAX Score II) and unmeasured variables (e.g. bleeding risk and frailty).

The introduction of newer-generation everolimus-eluting stent (EES) – with proven improvements in both safety and efficacy – has prompted the design of the EXCEL trial [55]. The Excel trial provides the first randomized patient data that will be available to assess the validity of the SYNTAX Score II treatment recommendations. Based on the SYNTAX Score II we predicted future outcomes of patients that were enrolled in the Excel trial, as one of the first attempts to enable independent future validation (chapter 14).

Duration of Dual Anti Platelet Therapy after coronary stent implementation

The proposed PRECISE-DAPT score aids prediction of out-of-hospital bleeding risk in patients treated with dual antiplatelet therapy (DAPT) after stent implementation (chapter 15). Because DAPT reduces ischemic recurrences after coronary stenting, but increases bleeding risk, international guidelines suggest individualization of the antiplatelet treatment duration [56, 57]. Indeed, multiple randomized studies showed bleeding liability associated with a prolonged as compared to shortened DAPT duration regimens [58-60]. In our analysis we observed that the increase in clinically significant bleeding related to prolonged treatment duration with DAPT occurred almost exclusively in patients with a higher bleeding risk score at baseline (PRECISE-DAPT score ≥25), whereas the impact of a prolonged DAPT regimen on bleeding in patients at lower bleeding risk was marginal. Thus, the PRECISE-DAPT has the potential to support clinical decision-making for DAPT treatment duration.

Neoadjuvant chemoradiotherapy in esophageal cancer patients

We developed and validated a prediction model for survival of esophageal cancer patients after treatment with neoadjuvant chemoradiotherapy (nCRT) and surgery. The impact of nCRT on prognostic factors effects was determined by including treatment (surgery alone or nCRT plus surgery) interactions with each of the prognostic factors. Interestingly, there was no overlap in significant independent prognostic factors between the two treatment groups, except for pN-stage, confirming previous reports [61, 62]. This finding underlined the continued significance of pN-stage as an important prognostic factor in the era of multimodality treatment [61-64]. The final prediction model, based on clinical N-stage, pathological T-stage and pathological N-stage, had moderate discriminatory ability, strengthening the need for new prognostic and predictive factors to improve survival prediction in the era of multimodality treatment for esophageal and junctional cancer.

Personalized axillary treatment of breast cancer patients

Separate models were developed and validated for the presence (chapter 16) of additional axillary non-SLN involvement and for the extent of axillary nodal involvement (chapter 17) in breast cancer patients with a positive sentinel lymph node (SLN). The models showed good discriminative ability and adequate calibration over the complete range of predicted probabilities, at internal and external validation. Predicting the extent of nodal involvement is useful, because early stage breast cancer patients with limited nodal involvement (1-2 positive lymph nodes) are no longer necessarily subjected to a completion axillary lymph node dissection (ALND), based on the results of various trials [65-67]. Our model outperformed previously proposed prediction models developed to identify patients at risk of having at least 4 involved lymph nodes – in terms of discriminative ability and calibration at external validation [68-70].

Selective screening for sexually transmitted infections

We developed Chlamydia trachomatis (Ct) prediction models with data from a large Chlamydia screening project (chapter 19). A prediction model based on readily available registry data may serve as a simple tool for selective screening at the population level [71]. With detailed questionnaire information the prediction model performed better than a previously proposed Ct prediction model, and may guide STI clinicians and GPs whom to offer opportunistic Ct testing [72]. We encouraged studies to identify which STI prediction models or predictors perform well and meet the necessary quality standards, including internal and external validation in existing datasets (chapter 20).

Selective care of mothers and neonates in the low and middle income countries

We developed and validated prognostic models for neonatal mortality in the general population in low and middle income countries, with specific reference to South Asia, on the basis of risk factors known at (i) the start of pregnancy, (ii) the start of delivery, and (iii) five minutes post-partum (chapter 21). At the start of pregnancy, prediction of neonatal death was difficult, although infants born to women of lower socioeconomic position and to women with certain demographic characteristics (young or advanced age, very short birth interval, primigravida) were at a higher risk of neonatal death. Predictive ability improved at the start of delivery, where multiple pregnancy and a premature start of delivery were highly predictive of neonatal death. Predictive ability was high at five minutes post-partum, where prematurity, multiple birth, and a poor condition of the infant were strong predictors of death. The models can be used to inform population-based prevention and more narrowly targeted facility-based interventions [73, 74], community-based interventions [74, 75] and individualized interventions [76, 77].

Limitations

Each prediction model in PART III was limited by the specific population they were developed or validated on. Further validation and possibly recalibration to specific populations is important. The SYNTAX Score II has been validated in a separate study in external registry data (chapter 12). However, future validation studies of the SYNTAX score II should ideally be done in sufficiently powered, randomized, all-comers studies comparing CABG against PCI, in which selection bias would be minimized. The Excel trial will be a perfect opportunity to prospectively validate the SYNTAX Score II. Recalibration will probably be required in the PCI arm because of the use of better (drug-eluting) stents.

A general concern in prediction research is the quality of the data (measurement error). The accuracy of predictor values – especially when using questionnaire data (chapters 19 and 21) – may vary between studies and populations.

Some of the studies were limited by missing values. We consistently used advanced multiple imputation approaches to make optimal use of the available data, and confirmed the insensitivity to the imputation process by comparing with results of complete case analysis [78].

FUTURE RESEARCH

We encourage studies to apply and test the proposed framework for assessing model performance in clustered data, especially in small clusters (Box 22.1). Simulations and case studies should concentrate on the sensitivity of the *mbc* and the *c-mbc* for violation of the underlying assumptions. Special consideration is required for regression models with time-varying regression coefficients – violating the proportional hazards assumption – with the use of the truncated *mbc* (appendix 3 of chapter 5) as a potential solution [79, 80]. In the presence of competing risks, this truncated *mbc* may also be used to assess the discriminative ability of proportional subdistribution hazards regression models [81, 82]. Further research is needed to understand the applicability of the *mbc*, and the general applicability of the proposed framework, in the competing risk setting. Finally, implementation of the *mbc* and the *c-mbc* in a standardized R-package (www.r-project.org) may augment its applicability.

The proposed methodological recommendations for modeling heterogeneity of treatment effect (Box 22.2) need to be tested in multiple case studies, preferably randomized trials. The proposed treatment benefit c-statistic needs further study. Possibly the formulation of concordance as a pairs of pairs comparison is also useful for quantification of the incremental value of markers, where the classical c-statistic has been criticized [83-86].

We recommend validation, updating, clinical impact assessment, and possibly extension with new predictors for all of the prediction models proposed in PART III [87, 88]. For example in oncology, further research is needed to find new prognostic factors for improved survival prediction in the era of multimodality treatment of esophageal and junctional cancer. Risk predictions, in combination with imaging, may also play a role to select patients after neoadjuvant chemotherapy for watchful waiting rather than surgery. We encourage prospective evaluation of such treatment strategies.

OVERALL CONCLUSIONS

The overall conclusions with respect to the three primary research questions of this thesis are as follows.

Aim 1 - How to validate prediction models in clustered data?

We proposed a framework for assessing the performance of prediction models in clustered data (Box 22.1). The framework explores variation in overall model performance (step 1), variation in patient heterogeneity (step 2), variation in regression coefficients (step 3), and variation in model performance when clusters are of limited sample size (step 4).

Aim 2 - How to develop and validate prediction models for guiding treatment decisions?

When using prediction models to guide treatment decisions we recommend (Box 22.2): (1) to take carefully modeled treatment interactions into consideration when estimating absolute treatment benefit for individual patients; (2) to assess how well a model separates patients who benefit from those who do not with the newly developed treatment benefit c-statistic rather than with conventional performance metrics; (3) to estimate individualized cost-effectiveness based on estimates of long-term life expectancy at the same level of granularity as estimates of the short-term risk reduction.

Aim 3 – How to apply methods for development and validation of predictions models for guiding treatment decisions?

We showed that prediction models are potentially useful tools for decision making in cardiovascular medicine, oncology and public health. The new methods proposed in PART I and II of this thesis were used as building blocks for guiding treatment decisions. Specifically, the methods for validation of prediction models in clustered data were applied to the 8 trial populations that were used to develop a prediction model for bleeding after coronary stent implementation (PRECISE-DAPT score) and the 4 regions that were used to develop prediction models for neonatal mortality in low and middle income countries. Also, the methods for development and validation of models that predict treatment benefit were successfully applied, e.g. to guide decisions on the optimal revascularization strategy for individual patients (SYNTAX Score II).

REFERENCES

1. Osler W. The Principles and Practice of Medicine. D. Appleton and Company, 1893.
2. Justice AC, Covinsky KE, Berlin JA. Assessing the Generalizability of Prognostic Information. Annals of Internal Medicine 1999; 130: 515-524.
3. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000; 19: 453-473.
4. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009; 338: b605.
5. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. Journal of Clinical Epidemiology 2015; 68: 25-34.
6. Tugwell P, Knottnerus JA. Clinical prediction models are not being validated. Journal of Clinical Epidemiology 2015; 68: 1-2.
7. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. Journal of Clinical Epidemiology; 68: 279-289.
8. Steyerberg EW, Harrell Jr FE. Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology 2016; 69: 245-247.
9. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016; 353.
10. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 1986; 7: 177-188.
11. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: An update. Contemporary Clinical Trials 2007; 28: 105-114.
12. Van Oirbeek R, Lesaffre E. An application of Harrell's C-index to PH frailty models. Stat Med 2010; 29: 3160-3171.
13. Van Oirbeek R, Lesaffre E. Assessing the predictive ability of a multilevel binary regression model. Computational Statistics &amp; Data Analysis 2012; 56: 1966-1980.
14. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2009; 172: 137-159.
15. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011; 342: d549.
16. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. Biometrika 2005; 92: 965-970.
17. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol 2010; 172: 971-980.
18. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer: New York, 2009.
19. Legrand C, Duchateau L, Janssen P, Ducrocq V, Sylvester R. Validation of prognostic indices using the frailty model. Lifetime Data Analysis 2009; 15: 59-78.
20. Efron B. Biased versus unbiased estimation. Advances in Mathematics 1975; 16: 259-277.
21. Greenland S. Principles of multilevel modelling. Int J Epidemiol 2000; 29: 158-167.
22. Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press: Cambridge, 2007.
23. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004; 82: 661-687.

24. Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. Lancet 2005; 365: 256-265.
25. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA 2007; 298: 1209-1212.
26. Vickers AJ, Kent DM. The Lake Wobegon Effect: Why Most Patients Are at Below-Average Risk. Ann Intern Med 2015; 162: 866-867.
27. Kent DM, Hayward RA, Dahabreh IJ. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centered evidence. Int J Epidemiol 2016; In press.
28. Verweij PJ, Van Houwelingen HC. Penalized likelihood in Cox regression. Stat Med 1994; 13: 2427-2436.
29. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996; 15: 361-387.
30. Farooq V, van Klaveren D, Steyerberg EW, Meliga E, Vergouwe Y, Chieffo A, Kappetein AP, Colombo A, Holmes DR, Jr., Mack M, Feldman T, Morice MC, Stahle E, Onuma Y, Morel MA, Garcia-Garcia HM, van Es GA, Dawkins KD, Mohr FW, Serruys PW. Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score II. Lancet 2013; 381: 639-650.
31. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355: 1064-1069.
32. Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? Am Heart J 2006; 151: 257-264.
33. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine--reporting of subgroup analyses in clinical trials. N Engl J Med 2007; 357: 2189-2194.
34. Stein CR, Kaufman JS, Ford CA, Leone PA, Feldblum PJ, Miller WC. Screening young adults for prevalent chlamydial infection in community settings. Ann Epidemiol 2008; 18: 560-571.
35. Haukoos JS, Hopkins E, Bender B, Sasson C, Al-Tayyib AA, Thrun MW, Denver Emergency Department HIVTRC. Comparison of enhanced targeted rapid HIV screening using the Denver HIV risk score to nontargeted rapid HIV screening in the emergency department. Ann Emerg Med 2013; 61: 353-361.
36. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Med Res Methodol 2006; 6: 18.
37. Stone NJ, Robinson J, Lichtenstein AH, Merz CN, Blum CB, Eckel RH, Goldberg AC, Gordon D, Levy D, Lloyd-Jones DM, McBride P, Schwartz JS, Shero ST, Smith SC, Jr., Watson K, Wilson PW. 2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 2013.
38. Cholesterol Treatment Trialists C, Mihaylova B, Emberson J, Blackwell L, Keech A, Simes J, Barnes EH, Voysey M, Gray A, Collins R, Baigent C. The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. Lancet 2012; 380: 581-590.
39. Ridker PM, Cook NR. Statins: new American guidelines for prevention of cardiovascular disease. The Lancet 2013.
40. Ridker PM, Wilson PW. A trial-based approach to statin guidelines. JAMA 2013; 310: 1123-1124.
41. Tajik P, Oude Rengerink K, Mol BW, Bossuyt PM. SYNTAX score II. Lancet 2013; 381: 1899.

42. Farooq V, van Klaveren D, Steyerberg EW, Serruys PW. SYNTAX score II - Authors' reply. Lancet 2013; 381: 1899-1900.
43. Holland PW. Statistics and Causal Inference. Journal of the American Statistical Association 1986; 81: 945-960.
44. Rubin DB. Causal Inference Using Potential Outcomes. Journal of the American Statistical Association 2005; 100: 322-331.
45. Steyerberg EW, Vedder MM, Leening MJG, Postmus D, D'Agostino RB, Van Calster B, Pencina MJ. Graphical assessment of incremental value of novel markers in prediction models: From statistical to decision analytical perspectives. Biometrical Journal 2014: n/a-n/a.
46. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006; 26: 565-574.
47. Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, Luce BR, Studies ITFoGRP--M. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices--Modeling Studies. Value Health 2003; 6: 9-17.
48. Stevens W, Normand C. Optimisation versus certainty: understanding the issue of heterogeneity in economic evaluation. Soc Sci Med 2004; 58: 315-320.
49. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, Augustovski F, Briggs AH, Mauskopf J, Loder E, Force IHEEPG-CGRPT. Consolidated Health Economic Evaluation Reporting Standards (CHEERS)--explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. Value Health 2013; 16: 231-250.
50. Ioannidis JPA, Garber AM. Individualized Cost-Effectiveness Analysis. PLoS Med 2011; 8: e1001058.
51. Boersma H, van der Vlugt MJ, Arnold AE, Deckers JW, Simoons ML. Estimated gain in life expectancy. A simple tool to select optimal reperfusion treatment in individual patients with evolving myocardial infarction. Eur Heart J 1996; 17: 64-75.
52. Waldo SW, Secemsky EA, O'Brien C, Kennedy KF, Pomerantsev E, Sundt TM, McNulty EJ, Scirica BM, Yeh RW. Surgical Ineligibility and Mortality Among Patients With Unprotected Left Main or Multivessel Coronary Artery Disease Undergoing Percutaneous Coronary InterventionCLINICAL PERSPECTIVE. Circulation 2014; 130: 2295-2301.
53. Farooq V, Brugaletta S, Serruys PW. Contemporary and evolving risk scoring algorithms for percutaneous coronary intervention. Heart 2011; 97: 1902-1913.
54. Capodanno D, Miano M, Cincotta G, Caggegi A, Ruperto C, Bucalo R, Sanfilippo A, Capranzano P, Tamburino C. EuroSCORE refines the predictive ability of SYNTAX score in patients undergoing left main percutaneous coronary intervention. Am Heart J 2010; 159: 103-109.
55. Dangas GD, Serruys PW, Kereiakes DJ, Hermiller J, Rizvi A, Newman W, Sudhir K, Smith RS, Jr., Cao S, Theodoropoulos K, Cutlip DE, Lansky AJ, Stone GW. Meta-analysis of everolimus-eluting versus paclitaxel-eluting stents in coronary artery disease: final 3-year results of the SPIRIT clinical trials program (Clinical Evaluation of the Xience V Everolimus Eluting Coronary Stent System in the Treatment of Patients With De Novo Native Coronary Artery Lesions). JACC Cardiovasc Interv 2013; 6: 914-922.
56. Authors/Task Force m, Windecker S, Kolh P, Alfonso F, Collet JP, Cremer J, Falk V, Filippatos G, Hamm C, Head SJ, Juni P, Kappetein AP, Kastrati A, Knuuti J, Landmesser U, Laufer G, Neumann FJ, Richter DJ, Schauerte P, Sousa Uva M, Stefanini GG, Taggart DP, Torracca L, Valgimigli M, Wijns W, Witkowski A. 2014 ESC/EACTS Guidelines on myocardial revascularization: The Task Force on Myocardial Revascularization of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS)Developed with the special contribution of the European Association of Percutaneous Cardiovascular Interventions (EAPCI). Eur Heart J 2014; 35: 2541-2619.

57. Levine GN, Bates ER, Bittl JA, Brindis RG, Fihn SD, Fleisher LA, Granger CB, Lange RA, Mack MJ, Mauri L, Mehran R, Mukherjee D, Newby LK, O'Gara PT, Sabatine MS, Smith PK, Smith SC, Jr., Focused Update Writing G. 2016 ACC/AHA Guideline Focused Update on Duration of Dual Antiplatelet Therapy in Patients With Coronary Artery Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 2016.
58. Navarese EP, Andreotti F, Schulze V, Kolodziejczak M, Buffon A, Brouwer M, Costa F, Kowalewski M, Parati G, Lip GY, Kelm M, Valgimigli M. Optimal duration of dual antiplatelet therapy after percutaneous coronary intervention with drug eluting stents: meta-analysis of randomised controlled trials. BMJ 2015; 350: h1618.
59. Valgimigli M, Campo G, Monti M, Vranckx P, Percoco G, Tumscitz C, Castriota F, Colombo F, Tebaldi M, Fuca G, Kubbajeh M, Cangiano E, Minarelli M, Scalone A, Cavazza C, Frangione A, Borghesi M, Marchesini J, Parrinello G, Ferrari R, Prolonging Dual Antiplatelet Treatment After Grading Stent-Induced Intimal Hyperplasia Study I. Short- versus long-term duration of dual-antiplatelet therapy after coronary stenting: a randomized multicenter trial. Circulation 2012; 125: 2015-2026.
60. Mauri L, Kereiakes DJ, Yeh RW, Driscoll-Shempp P, Cutlip DE, Steg PG, Normand SL, Braunwald E, Wiviott SD, Cohen DJ, Holmes DR, Jr., Krucoff MW, Hermiller J, Dauerman HL, Simon DI, Kandzari DE, Garratt KN, Lee DP, Pow TK, Ver Lee P, Rinaldi MJ, Massaro JM, Investigators DS. Twelve or 30 months of dual antiplatelet therapy after drug-eluting stents. N Engl J Med 2014; 371: 2155-2166.
61. Gu Y, Swisher SG, Ajani JA, Correa AM, Hofstetter WL, Liao Z, Komaki RR, Rashid A, Hamilton SR, Wu TT. The number of lymph nodes with metastasis predicts survival in patients with esophageal or esophagogastric junction adenocarcinoma who receive preoperative chemoradiation. Cancer 2006; 106: 1017-1025.
62. Rizk NP, Venkatraman E, Bains MS, Park B, Flores R, Tang L, Ilson DH, Minsky BD, Rusch VW. American Joint Committee on Cancer staging system does not accurately predict survival in patients receiving multimodality therapy for esophageal adenocarcinoma. J Clin Oncol 2007; 25: 507-512.
63. Ajani JA, Correa AM, Swisher SG, Wu TT. For localized gastroesophageal cancer, you give chemoradiation before surgery, but then what happens? J Clin Oncol 2007; 25: 4315-4316.
64. Holscher AH, Drebber U, Schmidt H, Bollschweiler E. Prognostic classification of histopathologic response to neoadjuvant therapy in esophageal adenocarcinoma. Ann Surg 2014; 260: 779-784; discussion 784-775.
65. Galimberti V, Cole BF, Zurrida S, Viale G, Luini A, Veronesi P, Baratella P, Chifu C, Sargenti M, Intra M, Gentilini O, Mastropasqua MG, Mazzarol G, Massarut S, Garbay JR, Zgajnar J, Galatius H, Recalcati A, Littlejohn D, Bamert M, Colleoni M, Price KN, Regan MM, Goldhirsch A, Coates AS, Gelber RD, Veronesi U. Axillary dissection versus no axillary dissection in patients with sentinel-node micrometastases (IBCSG 23-01): a phase 3 randomised controlled trial. Lancet Oncol 2013; 14: 297-305.
66. Giuliano AE, McCall L, Beitsch P, Whitworth PW, Blumencranz P, Leitch AM, Saha S, Hunt KK, Morrow M, Ballman K. Locoregional recurrence after sentinel lymph node dissection with or without axillary dissection in patients with sentinel lymph node metastases: the American College of Surgeons Oncology Group Z0011 randomized trial. Ann Surg 2010; 252: 426-432; discussion 432-423.
67. Donker M, van Tienhoven G, Straver ME, Meijnen P, van de Velde CJ, Mansel RE, Cataliotti L, Westenberg AH, Klinkenbijl JH, Orzalesi L, Bouma WH, van der Mijle HC, Nieuwenhuijzen GA, Veltkamp SC, Slaets L, Duez NJ, de Graaf PW, van Dalen T, Marinelli A, Rijna H, Snoj M, Bundred NJ, Merkus JW, Belkacemi Y, Petignat P, Schinagl DA, Coens C, Messina CG, Bogaerts J, Rutgers EJ. Radiotherapy or surgery of the axilla after a positive sentinel node in breast cancer (EORTC

10981-22023 AMAROS): a randomised, multicentre, open-label, phase 3 non-inferiority trial. Lancet Oncol 2014; 15: 1303-1310.

68. Katz A, Smith BL, Golshan M, Niemierko A, Kobayashi W, Raad RA, Kelada A, Rizk L, Wong JS, Bellon JR, Gadd M, Specht M, Taghian AG. Nomogram for the prediction of having four or more involved nodes for sentinel lymph node-positive breast cancer. J Clin Oncol 2008; 26: 2093-2098.
69. Meretoja TJ, Audisio RA, Heikkila PS, Bori R, Sejben I, Regitnig P, Luschin-Ebengreuth G, Zgajnar J, Perhavec A, Gazic B, Lazar G, Takacs T, Kovari B, Saidan ZA, Nadeem RM, Castellano I, Sapino A, Bianchi S, Vezzosi V, Barranger E, Lousquy R, Arisio R, Foschini MP, Imoto S, Kamma H, Tvedskov TF, Jensen MB, Cserni G, Leidenius MH. International multicenter tool to predict the risk of four or more tumor-positive axillary lymph nodes in breast cancer patients with sentinel node macrometastases. Breast Cancer Res Treat 2013; 138: 817-827.
70. Chagpar AB, Scoggins CR, Martin RC, 2nd, Cook EF, McCurry T, Mizuguchi N, Paris KJ, Carlson DJ, Laidley AL, El-Eid SE, McGlothin TQ, McMasters KM. Predicting patients at low probability of requiring postmastectomy radiation therapy. Ann Surg Oncol 2007; 14: 670-677.
71. van den Broek IV, Brouwers EE, Gotz HM, van Bergen JE, Op de Coul EL, Fennema JS, Koekenbier RH, Pars LL, van Ravesteijn SM, Hoebe CJ. Systematic selection of screening participants by risk score in a Chlamydia screening programme is feasible and effective. Sex Transm Infect 2012; 88: 205-211.
72. Gotz HM, van Bergen JE, Veldhuijzen IK, Broer J, Hoebe CJ, Steyerberg EW, Coenen AJ, de Groot F, Verhooren MJ, van Schaik DT, Richardus JH. A prediction rule for selective screening of Chlamydia trachomatis infection. Sex Transm Infect 2005; 81: 24-30.
73. Lee AC, Katz J, Blencowe H, Cousens S, Kozuki N, Vogel JP, Adair L, Baqui AH, Bhutta ZA, Caulfield LE, Christian P, Clarke SE, Ezzati M, Fawzi W, Gonzalez R, Huybregts L, Kariuki S, Kolsteren P, Lusingu J, Marchant T, Merialdi M, Mongkolchati A, Mullany LC, Ndirangu J, Newell ML, Nien JK, Osrin D, Roberfroid D, Rosen HE, Sania A, Silveira MF, Tielsch J, Vaidya A, Willey BA, Lawn JE, Black RE, Group CS-PBW. National and regional estimates of term and preterm babies born small for gestational age in 138 low-income and middle-income countries in 2010. Lancet Glob Health 2013; 1: e26-36.
74. March of Dimes, PMNCH, Save the Children, WHO. Born Too Soon: The Global Action Report on Preterm Birth. In Born Too Soon: The Global Action Report on Preterm Birth, Editor (ed)^(eds). World Health Organization: City, 2012.
75. Houweling TA, Tripathy P, Nair N, Rath S, Gope R, Sinha R, Looman CW, Costello A, Prost A. The equity impact of participatory women's groups to reduce neonatal mortality in India: secondary analysis of a cluster-randomised trial. Int J Epidemiol 2013; 42: 520-532.
76. Wall SN, Lee AC, Niermeyer S, English M, Keenan WJ, Carlo W, Bhutta ZA, Bang A, Narayanan I, Ariawan I, Lawn JE. Neonatal resuscitation in low-resource settings: what, who, and how to overcome challenges to scale up? Int J Gynaecol Obstet 2009; 107 Suppl 1: S47-62, S63-44.
77. Lawn JE, Mwansa-Kambafwile J, Barros FC, Horta BL, Cousens S. 'Kangaroo mother care' to prevent neonatal deaths due to pre-term birth complications. Int J Epidemiol 2011; 40: 525-528.
78. Van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 2011; 45: 1-67.
79. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological) 1972; 34: 187-220.
80. Hastie T, Tibshirani R. Varying-Coefficient Models. Journal of the Royal Statistical Society. Series B (Methodological) 1993; 55: 757-796.
81. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association 1999; 94: 496-509.
82. Wolbers M, Blanche P, Koller MT, Witteman JC, Gerds TA. Concordance for prognostic models with competing risks. Biostatistics 2014; 15: 526-539.

83. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007; 115: 928-935.
84. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008; 27: 157-172; discussion 207-112.
85. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010; 21: 128-138.
86. Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med 2011; 30: 11-21.
87. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014; 35: 1925-1931.
88. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG, Group P. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013; 10: e1001381.

Summary

William Osler noted in 1893 that “If it were not for the great variability between individuals, medicine might as well be a science, not an art” [1]. In contrast, this thesis is based on the scientific paradigm that prediction models have the potential to guide medical decisions by exploiting identifiable heterogeneity across individual patients. Prediction research focuses on the development of well performing prediction models and on the assessment of their generalizability and applicability.

Recently, the interpretation of performance heterogeneity across clusters of patients, e.g. grouped by hospitals, has received particular attention [2-4]. In PART I we aimed to develop methods for measuring prediction model performance (validation) within clusters of patients and for quantifying heterogeneity of prediction model performance across patient clusters.

Treatments that demonstrate effect on average in randomized clinical trials help some patients but not others. A major focus of patient-centered outcomes research and personalized medicine is to identify this heterogeneity of treatment effect (HTE) such that treatment can be targeted to those who benefit, and avoided in those where it is useless or harmful [5-7]. In PART II we aimed to propose methods for development and validation of models that exploit heterogeneity of treatment effect across patients, with an extension to modeling individualized cost-effectiveness.

In PART III we applied methods for development and validation of prediction models to several case studies in cardiovascular medicine, oncology, and public health.

PART I RISK HETEROGENEITY IN CLUSTERED DATA

Model performance may differ substantially – beyond chance – between clusters of patients. We explored the variation in model performance by a random-effects meta-analysis of estimates of calibration (agreement between predictions and observations) and discrimination (separation of patients with good outcome from those with poor outcome) in **chapters 2 and 3**. We summarized the random-effects meta-analyses with a pooled estimate and a 95% prediction interval for each performance measure.

Discriminative ability is often measured by the concordance probability, i.e. the probability that a model predicts a higher risk for a patient with a poorer outcome. The concordance probability may vary across patient clusters because of differences in patient case-mix or differences in regression coefficient validity. We explained that model-based concordance probability estimates differ between patient clusters only due to differences in patient case-mix (**chapter 4**). We explored the variation in discriminative ability across clusters by the newly developed model-based concordance (*mbc*) in **chapter 5**. In simulations and case studies, the difference between the *mbc* at model development and the *mbc* at external validation well indicated the change in discriminative ability attributable to a difference in case-mix. We explored the impact of regression coefficient validity by the *mbc* of a recalibrated model (*c-mbc*) in **chapter 5**. The difference between the *c-mbc* and the *mbc* in external validation data expressed the change in discriminative ability due to the (in)correctness of the regression coefficients. Thanks to their censoring-robustness, the *mbc* and the *c-mbc* facilitated measurements of concordance that were not influenced by differences in censoring distributions between development and external validation settings.

Cluster-specific concordance probability estimates may be unstable in case of limited sample size, similar to extreme cluster-specific calibration estimates. We therefore fitted multilevel regression models to obtain stabilized estimates of cluster-level calibration (**chapters 6 and 7**) [3, 8]. We used these estimates to calculate the *c-mbc*, that is the expected concordance probability under the assumption that the random effect estimates of calibration are true (**chapter 7**). In simulations, the root mean squared error (rmse) of the *c-mbc* was lower than the rmse of cluster-specific concordance probability estimates.

We conclude with a framework for assessing the performance of prediction models in clustered data (**chapter 22**). The framework proposes to explore: (1) variation in overall model performance, by random-effects meta-analysis of cluster-specific performance estimates; (2) variation in discriminative ability due to case-mix differences, by the newly developed *mbc*; (3) variation in discriminative ability due to differences in regression coefficient validity, by the newly developed *c-mbc*; and (4) variation in model performance when clusters are of limited sample size, by multilevel regression analysis in combination with the *c-mbc*.

PART II HETEROGENEITY OF TREATMENT EFFECT

We analyzed treatment effect heterogeneity for patients with complex coronary artery disease in the SYNTAX trial. We compared modeling approaches to estimate the individual survival benefit of treatment with either coronary artery bypass graft surgery or percutaneous coronary intervention (**chapter 8**). Statistical treatment interactions with each of the predictors fitted much better to the data compared to the treatment interaction with predicted prognosis as a single factor. In **chapter 9**, we proposed to study the potentially differential treatment effect of statin treatment for prevention of cardiovascular disease. We suggested estimating individualized risk reductions in a re-analysis of statin trial data, by adding a statistical interaction between treatment and absolute risk predictions.

For assessing a model's decision making potential, the relevance of the conventional risk concordance-statistic (c-statistic) has been questioned [9]. Performance measures for treatment selection should assess how well a model discriminates patients who benefit from those who do not. We proposed a treatment benefit c-statistic, which measures a prediction model's ability to predict treatment benefit (**chapter 10**). Hereto we defined observed treatment benefit by the outcome difference in *pairs* of patients matched on predicted benefit but discordant for treatment assignment. The increase in the treatment benefit c-statistic when comparing models with and without treatment interactions indicated a major improvement in discriminative ability, in contrast with a negligible increase of the risk c-statistic.

The perceived benefit of a treatment depends on a valuation of the benefits and harms (including costs) of the treatment alternatives [10]. For individual patients, the trial-based short-term treatment benefit estimates are often extrapolated to lifetime treatment benefit estimates based on sex and age specific population life tables. In **chapter 11**, we studied the impact of the unrealistic underlying assumption of equal post-trial life expectancy for low and high risk patients of the same sex and age. Individualized cost-effectiveness estimates of aggressive thrombolysis for patients with an acute myocardial infarction were clearly biased when models of short-term risk reduction and life expectancy were of unequal granularity.

In conclusion, when using prediction models to guide treatment decisions we recommend: (1) to take carefully modeled treatment interactions into consideration when estimating absolute treatment benefit for individual patients; (2) to assess how well a model separates patients who benefit from those who do not with the newly developed treatment benefit c-statistic rather than with conventional performance metrics; (3) to estimate individualized cost-effectiveness based on estimates of long-term life expectancy at the same level of granularity as estimates of the short-term risk reduction (**chapter 22**).

PART III APPLICATIONS

We developed and validated a prediction model for survival of patients with complex coronary artery disease (SYNTAX Score II) after coronary artery bypass graft surgery or percutaneous coronary intervention (**chapter 12**). The SYNTAX Score II improved treatment decision making and showed good agreement between predictions and observed outcomes in both treatment cohorts of a large Japanese all-comers patient registry (**chapter 13**). Based on the SYNTAX Score II we predicted future outcomes of patients that were enrolled in the EXCEL trial, as one of the first attempts to enable independent future validation (**chapter 14**).

We developed and validated the PRECISE-DAPT score to aid prediction of out-of-hospital bleeding risk in patients treated with dual antiplatelet therapy (DAPT) after coronary stent implementation (**chapter 15**). In our analysis we observed that the increase in clinically significant bleeding related to prolonged treatment duration with DAPT occurred almost exclusively in patients with a high PRECISE-DAPT score. The impact of a prolonged DAPT regimen on bleeding in patients at lower bleeding risk was marginal.

We developed and validated a prediction model for survival of esophageal cancer patients after treatment with neoadjuvant chemoradiotherapy and surgery (**chapter 16**). This prediction model, based on tumor size (T-stage) and the extent of regional lymph node metastasis (N-stage), had moderate discriminatory ability, strengthening the need for new factors to improve survival prediction in the era of neoadjuvant chemoradiotherapy.

We developed and validated separate models for the presence of additional axillary involvement (**chapter 17**) and for the extent of axillary nodal involvement (**chapter 18**) in breast cancer patients with a positive sentinel lymph node. The models showed good discriminative ability and adequate calibration, at internal and external validation. Predicting the extent of nodal involvement is useful for early stage breast cancer patients. Patients with limited metastasis (1-2 positive lymph nodes) are not necessarily subjected to a complete axillary lymph node dissection.

We developed Chlamydia trachomatis (Ct) prediction models with data from a large Chlamydia screening project (**chapter 19**). A prediction model based on readily available registry data may serve as a simple tool for selective screening at the population level. With detailed questionnaire information the prediction model may guide STI clinicians and GPs whom to offer opportunistic Ct testing. We encouraged studies to identify which STI prediction models or predictors perform well and meet the necessary quality standards, including internal validation and external validation in existing datasets (**chapter 20**).

We developed and validated prediction models for neonatal mortality in the general population in low and middle income countries (**chapter 21**). We used risk factors known at (1) the start of pregnancy, (2) the start of delivery, and (3) five minutes post-partum. The models can be used to inform population-based prevention and more narrowly targeted facility-based interventions, community-based interventions and individualized interventions.

We conclude that prediction models are useful tools for decision making in cardiovascular medicine, oncology and public health. The methods proposed in PART I and II, for validation of prediction models in clustered data and for development and validation of models that predict treatment benefit respectively, were successfully applied.

REFERENCES

1. Osler W. The Principles and Practice of Medicine. D. Appleton and Company, 1893.
2. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015; 68: 279-289.
3. Steyerberg EW, Harrell Jr FE. Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology 2016; 69: 245-247.
4. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016; 353.
5. Rothwell PM. Can overall results of clinical trials be applied to all patients? Lancet 1995; 345: 1616-1619.
6. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004; 82: 661-687.
7. Vickers AJ, Kent DM. The Lake Wobegon Effect: Why Most Patients Are at Below-Average Risk. Ann Intern Med 2015; 162: 866-867.
8. Legrand C, Duchateau L, Janssen P, Ducrocq V, Sylvester R. Validation of prognostic indices using the frailty model. Lifetime Data Analysis 2009; 15: 59-78.
9. Tajik P, Oude Rengerink K, Mol BW, Bossuyt PM. SYNTAX score II. Lancet 2013; 381: 1899.
10. Ioannidis JPA, Garber AM. Individualized Cost-Effectiveness Analysis. PLoS Med 2011; 8: e1001058.

Samenvatting

William Osler stelde in 1893: “Als er niet zo veel variatie tussen individuen bestond, zou de geneeskunde net zo goed een wetenschap kunnen zijn, in plaats van een kunst” [1]. Daarentegen is dit proefschrift gebaseerd op het wetenschappelijk paradigma dat predictiemodellen het vermogen hebben om medische beslissingen te ondersteunen door het benutten van waarneembare heterogeniteit tussen individuele patiënten. Predictieonderzoek is gericht op het ontwikkelen van goed presterende predictiemodellen en op de beoordeling van hun generaliseerbaarheid en toepasbaarheid.

Recentelijk heeft de heterogeniteit van voorspellend vermogen over verschillende clusters van patiënten, bijvoorbeeld gegroepeerd binnen ziekenhuizen, specifieke aandacht gekregen [2-4]. In DEEL I beoogden we methoden te ontwikkelen voor het meten van het voorspellend vermogen (valideren) van een predictiemodel binnen clusters van patiënten en voor het kwantificeren van de heterogeniteit van het voorspellend vermogen over patiëntclusters.

Behandelingen die gemiddeld effectief zijn in gerandomiseerde klinische trials, zijn nuttig voor sommige, maar niet voor alle patiënten. Een belangrijk aandachtsgebied van patiëntgericht uitkomstonderzoek en meer op het individu gerichte geneeskunde is het identificeren van deze heterogeniteit van behandeleffect. Daarmee kan de behandeling specifiek worden gericht op patiënten die daar voordeel van ondervinden. En kan worden voorkomen dat behandelingen worden gegeven aan patiënten die daar geen voordeel of zelfs schade van ondervinden [5-7]. In DEEL II beoogden we methoden voor te stellen voor het ontwikkelen en valideren van modellen die heterogeniteit van behandeleffect benutten, met een uitbreiding naar het modelleren van kosteneffectiviteit voor individuele patiënten.

In DEEL III pasten we methoden toe voor het ontwikkelen en valideren van predictiemodellen in verschillende casestudies binnen de cardiovasculaire geneeskunde, de oncologie, en de maatschappelijke gezondheidszorg.

DEEL I HETEROGENITEIT VAN RISICO IN GECLUSTERDE DATA

Het voorspellend vermogen van een model verschilt substantieel – meer dan op basis van kans alleen – tussen clusters van patiënten. We onderzochten de variatie in voorspellend vermogen met een random-effects meta-analyse van schattingen van kalibratie (overeenstemming tussen voorspellingen en waarnemingen) en discriminatie (onderscheid tussen patiënten met een goede uitkomst en patiënten met een slechte uitkomst) in **hoofdstukken 2 en 3**. We vatten de random-effects meta-analyses samen met een gepoolde schatting en een 95% predictie-interval voor elke maat van voorspellend vermogen.

Discriminerend vermogen wordt vaak gemeten aan de hand van de concordantie, ofwel de kans dat het model een hoger risico voorspelt voor een patiënt met een slechtere uitkomst. De concordantie kan variëren tussen clusters patiënten als gevolg van verschillen in patiëntenmix of verschillen in validiteit van regressiecoëfficiënten. We zetten uiteen dat een modelgebaseerde schatting van de concordantie alleen verschilt tussen patiëntenclusters als gevolg van verschillen in patiëntenmix (**hoofdstuk 4**). We onderzochten de variatie in discriminerend vermogen tussen clusters met de nieuw ontwikkelde modelgebaseerde concordantie (model-based concordance; *mbc*) in **hoofdstuk 5**. In simulaties en casestudies gaven de verschillen tussen de *mbc* bij de ontwikkeling van een model en de *mbc* bij de externe validatie van een model een goede indicatie van de verandering in discriminerend vermogen als gevolg van verschillen in patiëntenmix. We onderzochten de invloed van de validiteit van de regressiecoëfficiënten met de *mbc* van een gekalibreerd model (calibrated *mbc*; c-*mbc*) in **hoofdstuk 5**. Het verschil tussen de *c-mbc* en de *mbc* in externe validatie data drukte de verandering in discriminerend vermogen als gevolg van de (on)juistheid van de regressiecoëfficiënten uit. Dankzij hun robuustheid voor censurering van waarnemingen, faciliteren de *mbc* en de *c-mbc* schattingen van de concordantie die niet worden beïnvloed door verschillen in censurering tussen de ontwikkelomgeving en de externe validatie-omgeving.

Cluster-specifieke schattingen van de concordantie kunnen instabiel zijn in het geval van beperkte steekproefomvang, vergelijkbaar met extreme cluster-specifieke schattingen van de kalibratie. Daarom gebruikten we hiërarchische regressieanalyse voor het verkrijgen van stabiele schattingen van de kalibratie op clusterniveau (**hoofdstukken 6 en 7**) [3, 8]. We gebruikten die schattingen om de *c-mbc* te berekenen, ofwel de verwachte concordantie onder de aanname dat de random-effect schattingen van de kalibratie waar zijn (**hoofdstuk 7**). De gemiddelde kwadratische fout van de *c-mbc* was lager in simulaties dan die van cluster-specifieke concordantieschattingen.

We leidden uit DEEL I een raamwerk af voor het beoordelen van het voorspellend vermogen van predictiemodellen in geclusterde data (**hoofdstuk 22**). Het raamwerk bevat de analyse van: (1) variatie in algeheel voorspellend vermogen, door middel van random-effects meta-analyse van cluster-specifieke schattingen van voorspellend vermogen; (2) variatie in discriminerend vermogen als gevolg van verschillen in patiëntenmix, door middel van de nieuw ontwikkelde *mbc*; (3) variatie in discriminerend vermogen als gevolg van verschillen in validiteit van regressiecoëfficiënten, door middel van de nieuw ontwikkelde *c-mbc*; en (4) variatie in voorspellend vermogen wanneer clusters een beperkte steekproefomvang hebben, door middel van hiërarchische regressieanalyse in combinatie met de *c-mbc*.

DEEL II HETEROGENITEIT VAN BEHANDELEFFECT

We analyseerden heterogeniteit van behandeleffect voor patiënten met een ernstige coronaire hartziekte in de SYNTAX trial. We vergeleken modelleerwijzen voor het schatten van het individuele overlevingsvoordeel van behandeling met een coronaire bypass operatie boven een percutane coronaire interventie (**hoofdstuk 8**). Statistische interacties tussen de behandeling en elk van de voorspellende factoren pasten veel beter bij de data vergeleken met een interactie tussen de behandeling en voorspelde prognose als een enkele factor. In **hoofdstuk 9** suggereereden we het potentieel verschillend behandeleffect van statines, ter preventie van hart- en vaatziekten, te bestuderen. We stelden voor de individuele risicoreductie te schatten in een hernieuwde analyse van statine trial data, door het toevoegen van een statistische interactie tussen behandeling en absolute risicopredicties.

De relevantie van de conventionele risico concordantie-statistiek (c-statistiek) – ter beoordeling van het potentieel van een model om beslissingen te ondersteunen – is ter discussie gesteld [9]. Een maat voor het vermogen om de juiste behandeling te selecteren, zou moeten kwantificeren hoe goed een model onderscheid maakt tussen patiënten voor wie de behandeling voordeel biedt en patiënten voor wie de behandeling schadelijk is. Wij stelden een behandelvoordeel c-statistiek voor, die het vermogen van een predictiemodel meet om behandelvoordeel te voorspellen (**hoofdstuk 10**). Hiertoe definieerden we waargenomen behandelvoordeel aan de hand van het verschil in uitkomsten van *paren* patiënten die gekoppeld zijn op basis van voorspeld voordeel, maar die gerandomiseerd zijn naar een verschillende behandeling. De toename in de behandelvoordeel c-statistiek wanneer we modellen met en zonder behandelinteracties vergeleken, indiceerde een substantiële verbetering in discriminerend vermogen, in tegenstelling tot een verwaarloosbare toename in de risico c-statistiek.

Het gepercipieerde voordeel van een behandeling hangt af van de waardering van de voordelen en nadelen (inclusief kosten) van behandelalternatieven [10]. Voor individuele patiënten wordt het korte-termijn voordeel van een behandeling dat gevonden wordt in een klinische trial meestal geëxtrapoleerd naar een levenslang behandelvoordeel op basis van geslachts- en leeftijdsspecifieke sterftetafels. In **hoofdstuk 11** bestudeerden we de impact van de irrealistische onderliggende aanname van gelijke levensverwachting na afloop van de trial voor laag- en hoog-risico patiënten van gelijk geslacht en gelijke leeftijd. Individuele kosteneffectiviteitsschattingen van agressieve trombolyse voor patiënten met een acuut hartinfarct bevatten structurele afwijkingen wanneer modellen van korte-termijn risicoreductie en levensverwachting verschilden van detailniveau.

Concluderend, wanneer predictiemodellen worden gebruikt ter ondersteuning van behandelbeslissingen, raden wij aan: (1) behoedzaam gemodelleerde behandelinteracties te overwegen voor het schatten van absoluut behandelvoordeel voor individuele patiënten; (2) te beoordelen hoe goed het model patiënten onderscheidt die voordeel ondervinden aan de hand van de nieuw ontwikkelde behandeleffect c-statistiek in plaats van met de conventionele maten van voorspellend vermogen; (3) individuele kosteneffectiviteit te baseren op schattingen van lang-termijn levensverwachting op hetzelfde detailniveau als de schattingen van de korte-termijn risicoreductie (**hoofdstuk 22**).

DEEL III TOEPASSINGEN

We ontwikkelden en valideerden een predictiemodel voor overleving van patiënten met een ernstige coronaire hartziekte (SYNTAX Score II) na coronaire bypass chirurgie of een percutane coronaire interventie (**hoofdstuk 12**). De SYNTAX Score II verbeterde de besluitvorming over de behandeling en liet een goede overeenkomst zien tussen voorspellingen en waargenomen uitkomsten in beide behandelcohorten van een Japanse niet geselecteerde patiëntenpopulatie (**hoofdstuk 13**). Op basis van de SYNTAX Score II voorspelden we toekomstige uitkomsten van patiënten die zijn geïncludeerd in de EXCEL trial, als een van de eerste pogingen om onafhankelijke toekomstige validatie mogelijk te maken (**hoofdstuk 14**).

We ontwikkelden en valideerden de PRECISE-DAPT score voor het voorspellen van bloedingsrisico na ontslag uit het ziekenhuis van patiënten die worden behandeld met duale antiplaatjes therapie (DAPT) na implementatie van coronaire stents (**hoofdstuk 15**). We observeerden in onze studie dat de toegenomen kans op klinisch significante bloedingen gerelateerd met een verlengde behandelingsduur van DAPT vrijwel geheel toe te wijzen was aan patiënten met een hoge PRECISE-DAPT score. De impact van een verlengd DAPT regiem op bloedingen in laag-risico patiënten was marginaal.

We ontwikkelden en valideerden een predictiemodel voor overleving van slokdarmkankerpatiënten na behandeling met neoadjuvante chemoradiotherapie en chirurgie (**hoofdstuk 16**). Dit predictiemodel, gebaseerd op de grootte van de tumor (T-stadium) en de uitbreiding naar regionale lymfeklieren (N-stadium), had middelmatig discriminerend vermogen en benadrukte de behoefte aan nieuwe factoren om overlevingsvoorspellingen te verbeteren in het tijdperk van neoadjuvante chemoradiotherapie.

We ontwikkelden en valideerden aparte modellen voor de aanwezigheid van resterende positieve okselklieren (**hoofdstuk 17**) en voor het aantal positieve okselklieren (**hoofdstuk 18**) in borstkankerpatiënten met een positieve schildwachtklier. De modellen hadden goed discriminerend vermogen en adequate kalibratie, zowel bij interne als bij externe validatie. Voorspellen van de mate van uitzaaiing in de okselklieren is nuttig voor patiënten met borstkanker in een vroeg stadium. Patiënten met beperkte uitzaaiing (1-2 positieve lymfklieren) hoeven niet meer noodzakelijkerwijs een complete okselklierdissectie te ondergaan.

We ontwikkelden Chlamydia trachomatis (Ct) predictiemodellen met data van een groot Chlamydia screening project (**hoofdstuk 19**). Een predictiemodel gebaseerd op eenvoudig verkrijgbare gegevens uit de basisregistratie kan dienen als een hulpmiddel voor selectieve screening op bevolkingsniveau. Aan de hand van gedetailleerde informatie uit questionnaires kan het predictiemodel huisartsen en SOA-artsen ondersteunen bij hun beslissing aan wie opportunistisch een Ct-test aan te bieden. We moedigden studies aan die identificeren welke SOA-predictiemodellen en -predictoren een goed voorspellend vermogen hebben en voldoen aan de noodzakelijke kwaliteitsstandaarden, inclusief interne validatie en externe validatie in bestaande datasets (**hoofdstuk 20**).

We ontwikkelden en valideerden predictiemodellen voor neonatale sterfte in de algemene bevolking van lage- en middeninkomenslanden (**hoofdstuk 21**). We gebruikten risicofactoren die bekend zijn (1) bij de start van de zwangerschap, (2) bij de start van de bevalling, en (3) vijf minuten postpartum. De modellen kunnen worden gebruikt voor de onderbouwing van preventie op bevolkingsniveau en van interventies toegespitste op het niveau van instellingen, gemeenschappen en individuen.

We concluderen dat predictiemodellen nuttige hulpmiddelen zijn voor het nemen van beslissingen in de cardiovasculaire geneeskunde, de oncologie en de maatschappelijke gezondheidszorg. De methoden die zijn voorgesteld in DEEL I en II werden succesvol toegepast, respectievelijk voor validatie van predictiemodellen in geclusterde data en voor ontwikkeling en validatie van modellen die behandeleffecten voorspellen.

REFERENTIES

1. Osler W. The Principles and Practice of Medicine. D. Appleton and Company, 1893.
2. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015; 68: 279-289.
3. Steyerberg EW, Harrell Jr FE. Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology 2016; 69: 245-247.
4. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016; 353.
5. Rothwell PM. Can overall results of clinical trials be applied to all patients? Lancet 1995; 345: 1616-1619.
6. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004; 82: 661-687.
7. Vickers AJ, Kent DM. The Lake Wobegon Effect: Why Most Patients Are at Below-Average Risk. Ann Intern Med 2015; 162: 866-867.
8. Legrand C, Duchateau L, Janssen P, Ducrocq V, Sylvester R. Validation of prognostic indices using the frailty model. Lifetime Data Analysis 2009; 15: 59-78.
9. Tajik P, Oude Rengerink K, Mol BW, Bossuyt PM. SYNTAX score II. Lancet 2013; 381: 1899.
10. Ioannidis JPA, Garber AM. Individualized Cost-Effectiveness Analysis. PLoS Med 2011; 8: e1001058.

Dankwoord

Acknowledgements

Yvonne Vergouwe Veel dank voor de kans die je me geboden hebt om mij om te scholen van actuaris naar onderzoeker in de medische besliskunde. En voor de gestructureerde en altijd waardevolle opmerkingen en aanvullingen op mijn concept artikelen en presentaties. Als dit proefschrift enigszins leesbaar is, dan is dat voor een aanzienlijk deel aan jou te danken!

Ewout Steyerberg Veel dank voor de ruimte die je me gegeven hebt om mij breed te ontwikkelen binnen de medische besliskunde. En voor de inspirerende snelheid waarmee je mijn vragen beantwoordde en concept artikelen van commentaar voorzag. Ik kijk met veel plezier terug op de kilometers die we gezamenlijk hardliepen in Boston, en natuurlijk in Bretton Woods!

David Kent Many thanks for going to a lot of trouble to offer me the opportunity for working in Boston. And for teaching me to write just a little less telegraphic. And last but not least, for taking me along with your daughter and friends to Cape Cod!

Professor Serruys Veel dank voor het vertrouwen dat u mij gegeven heeft in de samenwerking met uw fellow onderzoekers. De meest succesvolle hoofstukken in dit proefschrift zijn mede onder uw inspirerende begeleiding tot stand gekomen.

Mithat Gönen Many thanks for inspiring discussions on the most technical parts of my thesis which are founded on your earlier work. And for generously taking me to very pleasant lunches each time I visited you in NYC.

Vasim Farooq Many thanks for an honestly perfect collaboration that led to the most successful chapters of this thesis. You have opened up new doors in my career.

Carlos Campos Many thanks for a stimulating and friendly collaboration, which will hopefully once lead to visiting Pantanal!

Joel Shapiro Veel dank voor de inspirerende discussies op medisch gebied, maar vooral ook daar buiten. Ik zou onze vriendschappelijke samenwerking heel graag verder uitbouwen.

Daan Nieboer Veel dank voor de snelle, adequate antwoorden op het gebied van statistiek en R. En voor de gezelligheid op congressen!

Peter Austin, Francesco Costa, Marco Valgimigli, Ingrid van den Hoven, Hannelore Götz, Tanja Houweling Many thanks for trusting me in various collaborations which enabled me to become acquainted with a wide range of medical research.

Coby Pikaar, Bert van Klaveren Mam, pap, heel veel dank voor jullie onvoorwaardelijke liefde en grenzeloze vertrouwen. “Ga studeren wat je leuk vindt” heeft met dit proefschrift extra betekenis gekregen.

Andrea van Klaveren-Woltman Je hebt in 2015 ternauwernood een ongeluk overleefd, met zwaar letsel als gevolg. Toch bleef je op wonderbaarlijke wijze uitblinken in alle functies: als moeder, als partner en... als coach van een beginnend onderzoeker. Heel veel dank daarvoor. Ik hoop uit de grond van mijn hart – en ben vol vertrouwen – dat je weer optimaal je dromen zal kunnen najagen.

List of publications

1. Upshaw JN, Konstam MA, van Klaveren D, Noubary F, Huggins GS, Kent DM. Multistate Model to Predict Heart Failure Hospitalizations and All-Cause Mortality in Outpatients With Heart Failure With Reduced Ejection Fraction: Model Derivation and External Validation. **Circ Heart Fail.** 2016 Aug;9(8). pii: e003146.

2. Sotomi Y, Cavalcante R, van Klaveren D, Ahn JM, Lee CW, de Winter RJ, Wykrzykowska JJ, Onuma Y, Steyerberg EW, Park SJ, Serruys PW. Individual Long-Term Mortality Prediction Following Either Coronary Stenting or Bypass Surgery in Patients With Multivessel and/or Unprotected Left Main Disease: An External Validation of the SYNTAX Score II Model in the 1,480 Patients of the BEST and PRECOMBAT Randomized Controlled Trials. **JACC Cardiovasc Interv.** 2016 Aug 8;9(15):1564-72.

3. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. **J Clin Epidemiol.** 2016 Jun 2. pii: S0895-4356(16)30140-8.

4. van Klaveren D, Gönen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings. **Stat Med.** 2016 Oct 15;35(23):4136-52.

5. Shapiro J, Biermann K, van Klaveren D, Offerhaus GJ, Ten Kate FJ, Meijer SL,van Berge Henegouwen MI, Steyerberg EW, Wijnhoven BP, Lanschot JJ. Prognostic Value of Pretreatment Pathological Tumor Extent in Patients Treated With Neoadjuvant Chemoradiotherapy Plus Surgery for Esophageal or Junctional Cancer. **Ann Surg.** 2016 Jan 14. [Epub ahead of print]

6. Shapiro J, van Klaveren D, Lagarde SM, Toxopeus EL, van der Gaast A, Hulshof MC, Wijnhoven BP, van Berge Henegouwen MI, Steyerberg EW, van Lanschot JJ.Prediction of survival in patients with oesophageal or junctional cancer receiving neoadjuvant chemoradiotherapy and surgery. **Br J Surg.** 2016 Jul;103(8):1039-47

7. Wiggers JK, Groot Koerkamp B, Cieslak KP, Doussot A, van Klaveren D, Allen PJ, Besselink MG, Busch OR, D'Angelica MI, DeMatteo RP, Gouma DJ, Kingham TP, van Gulik TM, Jarnagin WR. Postoperative Mortality after Liver Resection for Perihilar Cholangiocarcinoma: Development of a Risk Score and Importance of Biliary Drainage of the Future Liver Remnant. **J Am Coll Surg.** 2016 Aug;223(2):321-331.e1.

8. van Klaveren D*, Götz HM*, Op de Coul EL, Steyerberg EW, Vergouwe Y. Prediction of Chlamydia trachomatis infection to facilitate selective screening on population and individual level: a cross-sectional study of a population-based screening programme. **Sex Transm Infect.** 2016 Sep;92(6):433-40.

9. van den Hoven I, van Klaveren D, Voogd AC, Vergouwe Y, Tjan-Heijnen V, Roumen RM. A Dutch Prediction Tool to Assess the Risk of Additional Axillary Non-Sentinel Lymph Node Involvement in Sentinel Node-Positive Breast Cancer Patients. **Clin Breast Cancer.** 2016 Apr;16(2):123-30.

10. Osnabrugge RL, Magnuson EA, Serruys PW, Campos CM, Wang K, van Klaveren D, Farooq V, Abdallah MS, Li H, Vilain KA, Steyerberg EW, Morice MC, Dawkins KD, Mohr FW, Kappetein AP, Cohen DJ; SYNTAX trial investigators. Cost-effectiveness of percutaneous coronary intervention versus bypass surgery from a Dutch perspective. **Heart.** 2015 Dec;101(24):1980-8.

11. Campos CM, Garcia-Garcia HM, van Klaveren D, Ishibashi Y, Cho YK, Valgimigli M, Räber L, Jonker H, Onuma Y, Farooq V, Garg S, Windecker S, Morel MA, Steyerberg EW, Serruys PW. Validity of SYNTAX score II for risk stratification of percutaneous coronary interventions: A patient-level pooled analysis of 5,433 patients enrolled in contemporary coronary stent trials. **Int J Cardiol.** 2015;187:111-5.

12. van Klaveren D, Vergouwe Y, Farooq V, Serruys PW, Steyerberg EW. Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions. **J Clin Epidemiol.** 2015 Nov;68(11):1366-74.

13. Zhang YJ, Iqbal J, van Klaveren D, Campos CM, Holmes DR, Kappetein AP, Morice MC, Banning AP, Grech ED, Bourantas CV, Onuma Y, Garcia-Garcia HM, Mack MJ, Colombo A, Mohr FW, Steyerberg EW, Serruys PW. Smoking is associated with adverse clinical outcomes in patients undergoing revascularization with PCI or CABG: the SYNTAX trial at 5-year follow-up. **J Am Coll Cardiol.** 2015 Mar 24;65(11):1107-15.

14. Polinder S, Haagsma JA, van Klaveren D, Steyerberg EW, van Beeck EF. Health-related quality of life after TBI: a systematic review of study design, instruments, measurement properties, and outcome. **Popul Health Metr.** 2015 Feb 17;13:4.

15. Campos CM*, van Klaveren D*, Farooq V, Simonton CA, Kappetein AP, Sabik JF 3rd, Steyerberg EW, Stone GW, Serruys PW; EXCEL Trial Investigators. Long-term forecasting and comparison of mortality in the Evaluation of the Xience Everolimus Eluting Stent vs. Coronary Artery Bypass Surgery for Effectiveness of Left Main Revascularization (EXCEL) trial: prospective validation of the SYNTAX Score II. **Eur Heart J.** 2015 May 21;36(20):1231-41.

16. Zhang YJ, Iqbal J, Campos CM, van Klaveren D, Bourantas CV, Dawkins KD, Banning AP, Escaned J, de Vries T, Morel MA, Farooq V, Onuma Y, Garcia-Garcia HM, Stone GW, Steyerberg EW, Mohr FW, Serruys PW. Prognostic value of site SYNTAX score and rationale for combining anatomic and clinical factors in decision making: insights from SYNTAX trial. **J Am Coll Cardiol.** 2014 Aug 5;64(5):423-32.

17. Campos CM, van Klaveren D, Iqbal J, Onuma Y, Zhang YJ, Garcia-Garcia HM, Morel MA, Farooq V, Shiomi H, Furukawa Y, Nakagawa Y, Kadota K, Lemos PA, Kimura T, Steyerberg EW, Serruys PW. Predictive Performance of SYNTAX Score II in Patients With Left Main and Multivessel Coronary Artery Disease-analysis of CREDO-Kyoto registry. **Circ J.** 2014;78(8):1942-9. Epub 2014 Jul 7.

18. Iqbal J, Vergouwe Y, Bourantas CV, van Klaveren D, Zhang YJ, Campos CM, García-García HM, Morel MA, Valgimigli M, Windecker S, Steyerberg EW, Serruys PW. Predicting 3-year mortality after percutaneous coronary intervention: updated logistic clinical SYNTAX score based on patient-level data from 7 contemporary stent trials. **JACC Cardiovasc Interv.** 2014 May;7(5):464-70.

19. Götz HM, van Klaveren D. Use of prediction rules in control of sexually transmitted infections: challenges and chances. **Sex Transm Dis.** 2014 May;41(5):331-2

20. van Klaveren D, Vergouwe Y, Steyerberg EW. Refining the American guidelines for prevention of cardiovascular disease. **Lancet.** 2014 Feb 15;383(9917):598.

21. van Klaveren D, Steyerberg EW, Vergouwe Y. Interpretation of concordance measures for clustered data. **Stat Med.** 2014 Feb 20;33(4):714-6.

22. van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. **BMC Med Res Methodol.** 2014 Jan 15;14:5.

23. Farooq V, van Klaveren D, Steyerberg EW, Serruys PW. SYNTAX score II - Authors' reply. **Lancet.** 2013 Jun 1;381(9881):1899-900.

24. Farooq V*, van Klaveren D*, Steyerberg EW, Meliga E, Vergouwe Y, Chieffo A, Kappetein AP, Colombo A, Holmes DR Jr, Mack M, Feldman T, Morice MC, Ståhle E, Onuma Y, Morel MA, Garcia-Garcia HM, van Es GA, Dawkins KD, Mohr FW, Serruys PW. Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score II. **Lancet.** 2013 Feb 23;381(9867):639-50.

25. van den Oord SC, Sijbrands EJ, ten Kate GL, van Klaveren D, van Domburg RT, van der Steen AF, Schinkel AF. Carotid intima-media thickness for cardiovascular risk assessment: systematic review and meta-analysis. **Atherosclerosis.** 2013 May;228(1):1-11.

* Equal contribution

PhD Portfolio

SUMMARY OF PhD TRAINING AND TEACHING

Name:	David van Klaveren
Erasmus MC Department:	Public Health
PhD period:	2012-2016
Promotor:	Prof. dr. E.W. Steyerberg
Copromotor:	Dr. Y. Vergouwe

1. PhD Training	**Year**	**Workload (days)**
General courses		
Statistical Issues in Drug Development. Stephen Senn	2014	2
An introduction to the joint modeling of longitudinal and survival outcomes, with applications in R. Dimitris Rizopoulos	2013	2
Regression modeling strategies. Frank Harrell	2012	1
Analysis of interval-censored survival data. Philip Hougaard	2012	1
Absolute risk prediction. Mitchell Gail and Ruth Pfeiffer	2012	1
Conferences		
Annual North American Meeting of the Society for Medical Decision Making, Saint Louis. Oral presentation: Stratified medicine and cost-effectiveness: strong influence of choices in modeling short-term, trial-based, mortality risk reduction and post-trial life expectancy.	2015	4
Annual Conference of the International Society for Clinical Biostatistics, Utrecht. Oral presentation: A new concordance measure for a competing risk regression model with proportional subdistribution hazards.	2015	4
Annual Meeting of the Society for Clinical Trials, Arlington.	2015	4
Biennial European conference of the Society for Medical Decision Making, Antwerp. Oral presentation: Assessing the influence of case-mix heterogeneity on the discriminative ability of a risk prediction model: the model-based concordance-index.	2014	3
Annual conference of the Netherlands Epidemiology Society, Leiden. Oral presentation: Assessing the influence of case-mix heterogeneity on the discriminative ability of a risk prediction model: the model-based concordance-index.	2014	3

	Year	Workload (days)
Conferences		
Annual Conference of the International Society for Clinical Biostatistics, Vienna, 2014. Poster presentation: A censoring-robust concordance measure for proportional hazards regression models in external validation data: the calibrated Gönen and Heller estimator.	2014	4
Annual Conference of the International Society for Clinical Biostatistics, Vienna. Poster presentation: Assessing the influence of case-mix heterogeneity on the discriminative ability of a risk prediction model: the model-based concordance-index.	2014	4
Annual Conference of the International Society for Clinical Biostatistics, Munich. Oral presentation: Using interactions in prediction modelling to account for heterogeneity in treatment effects.	2013	4
Annual conference of the Netherlands Epidemiology Society, Utrecht. Oral presentation: Using interactions in prediction modelling to account for heterogeneity in treatment effects.	2013	3
Annual conference of the Netherlands Epidemiology Society, Rotterdam. Oral presentation: Assessing discriminative ability in clustered data.	2012	3
Annual Conference of the International Society for Clinical Biostatistics, Bergen. Poster presentation. Assessing discriminative ability in clustered data.	2012	4
Spring meeting of the Eastern American Region of the International Biometric Society, Washington.	2012	4
Other presentations		
Annual Meeting of the European Association for Cardio-Thoracic Surgery, Amsterdam. Oral presentation: Making a risk model and testing predictability.	2015	1
Research meeting of the Public Health department: Prognosis, treatment, and individualized cost-effectiveness.	2014	1
Research meeting of the Public Health department: Guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients.	2014	1
Research meeting of the Public Health department: Assessing discriminative ability of prognostic models in clustered data.	2013	1

	Year	Workload (days)
Seminars and workshops		
Weekly Research meetings of the department of Public Health	2011-	10
2. Teaching		
Lecturing Clinical Epidemiology for the Netherlands Institute for Health Sciences (NIHES)	2016	6
Supervising practicals of the NIHES courses Advanced Analysis of Prognosis Studies and Clinical Epidemiology	2012-	3
Statistical consulting	2012-	5
3. Other activities		
Peer review of 9 papers for international journals	2013-	3
Organizing research meetings of the Medical Decision Making section	2012-	5
4. Awards		
Young Investigator Award of the Journal of Clinical Epidemiology	2015	
Best Debut Publication Award of the Erasmus MC Public Health Department	2013	

About the author

David van Klaveren was born on July 31, 1973 in Rotterdam, The Netherlands. In 1991 he graduated from secondary school (Christelijke Scholengemeenschap Walcheren, Middelburg) and started studying Technical Mathematics at Delft University of Technology. In 1997 he obtained his master's degree with a thesis on quantitative aspects of risk analysis.

After his graduation David established a Risk Analysis Section within Strukton, a large infrastructural building contractor in Maarssen. He switched to the insurance business in 1998 to work as an actuarial consultant for Posthuma Partners in The Hague. In 2003 he started working for the Dutch branch of the global insurance company Allianz in Rotterdam, first as a senior actuary, and later on as head of the Pricing Department. In 2009 he completed, *cum laude*, the Actuarial Sciences Master's program at the University of Amsterdam.

In 2011 David changed his career to become a researcher in medical decision making at the Department of Public Health of Erasmus MC in Rotterdam. In 2015 he received the Young Investigator Award from the Journal of Clinical Epidemiology. Since 2015 he is associate editor of BMC Medical Informatics and Decision Making. Before returning to Erasmus MC in September 2016, he worked for one year as a visiting scholar at the Predictive Analytics and Comparative Effectiveness Center of Tufts Medical Center in Boston, MA.

William Osler noted in 1893 that “If it were not for the great variability between individuals, medicine might as well be a science, not an art”.

In contrast, this thesis is based on the scientific paradigm that prediction models have the potential to guide medical decisions by exploiting identifiable heterogeneity across individual patients.

Prediction research focuses on the development of well performing prediction models and on the assessment of their generalizability and applicability. Several methods to measure prediction model performance across clusters of patients are proposed in **PART I** of this thesis. **PART II** contains novel methods for development and validation of models that incorporate heterogeneity of treatment effect across patients.
In **PART III**, methods for development and validation of prediction models are applied to several case studies in cardiovascular medicine, oncology, and public health.

Propositions belonging to the thesis

Heterogeneity in Prediction Research: Methods and applications

1. Heterogeneity of model performance across clusters of patients is more informative than overall model performance in pooled patient data. *(this thesis)*

2. Understanding heterogeneity of discriminative ability across clusters of patients requires understanding heterogeneity of both patient case-mix and overall regression coefficient validity. *(this thesis)*

3. Heterogeneity of absolute treatment effect across patients is underestimated if heterogeneity of baseline risk or heterogeneity of relative treatment effect is ignored. *(this thesis)*

4. Heterogeneity of cost-effectiveness across patients is overestimated when individualized cost-effectiveness estimates are based on long-term survival models that are less granular than short-term treatment benefit models. *(this thesis)*

5. Guiding clinical and public health decisions by valid prediction models has the potential to: save life years; reduce treatment harm; and avert costs. *(this thesis)*

6. After the data are used to determine the model, they will no longer be needed to estimate the predictive power.
   Edward Korn, Richard Simon (Statistics in Medicine 1990)

7. Certain deliberately induced biases can dramatically improve estimation.
   Bradley Efron (Advances in Mathematics 1975)

8. The modest benefit ascribed to many treatments in clinical trials can be misleading because modest average effects may reflect a mixture of substantial benefits for some, little benefit for many, and harm for a few.
   Richard Kravitz, Naihua Duan, Joel Braslow (The Milbank Quarterly 2004)

9. It is a mistake to conclude from the Fundamental Problem of Causal Inference that causal inference is impossible.
   Paul Holland (Journal of the American Statistical Association 1986)

10. Claimed research findings may often be simply accurate measures of the prevailing bias.
    John Ioannidis (PLOS Medicine 2005)

11. Scientists are people who know more and more about less and less, until they know everything about nothing.
    Konrad Lorenz

David van Klaveren
January 13, 2017