Towards better clinical prediction models:

seven steps for development and an ABCD for validation

E.W. Steyerberg [1], Y. Vergouwe [1]

Department of Public Health, Erasmus MC - University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, the Netherlands

E.Steyerberg@ErasmusMC.nl

Abstract

Clinical prediction models provide risk estimates for the presence of disease (diagnosis) or an event in the future course of disease (prognosis) for individual patients. Although publications that present and evaluate such models are becoming more frequent, the methodology is often suboptimal. We propose that seven steps should be considered in developing prediction models: 1) consideration of the research question and initial data inspection; 2) coding of predictors; 3) model specification; 4) model estimation; 5) evaluation of model performance; 6) internal validation; and 7) model presentation. The validity of a prediction model is ideally assessed in fully independent data, where we propose four key measures to evaluate model performance: calibration-in-the-large, or the model intercept (A); calibration slope (B); discrimination, with a concordance statistic (C); and clinical usefulness, with decision curve analysis (D). As an application, we develop and validate prediction models for 30-day mortality in patients with an acute myocardial infarction. This illustrates the usefulness of the proposed framework to strengthen the methodological rigor and quality for prediction models in cardiovascular research.

Keywords: Prediction model; non-linearity; missing values; shrinkage; calibration; discrimination; clinical usefulness

Introduction

Prediction of the presence of disease (diagnosis) or an event in the future course of disease (prognosis) becomes more and more important in the current era of personalised medicine. Improvements in imaging, biomarkers and 'omics' research lead to many new predictors for diagnosis and prognosis.

Clinical prediction models may combine multiple predictors to provide insight into the relative effects of predictors in the model. For example, we may be interested in the independent prognostic value of inflammatory markers such as C-reactive protein for the clinical course and outcome of an acute coronary syndrome . Clinical prediction models may also provide absolute risk estimates for individual patients (2) (3). We focus here on the second role of such models, which are commonly developed with regression analysis techniques. Logistic regression analysis is most commonly used for the prediction of binary events, such as 30-day mortality. Cox regression is most common for time-to-event data, such as long-term mortality. We focus on prediction models for binary events, and indicate differences with time-to-event data where most relevant.

We note that rigorous development and validation of prediction models is important. Despite a substantial body of methodological literature and published guidance on how to perform prediction research, most models suffer from methodological shortcomings, or are at least reported poorly (4) (5) (6) (7). We propose a systematic approach to model development and validation, illustrated with prediction of 30-day mortality in patients suffering from an acute myocardial infarction. We compare the performance of a simple model (including age

only) to a more complex model (including age and other key predictors), which are developed either in a small or a large data set.

Case study: prediction of 30-day mortality in acute myocardial infarction

As an illustrative case study, we consider a cohort of 40,830 patients suffering from an acute myocardial infarction who were enrolled in the GUSTO-I trial. The data from this trial have been used for many analyses, including the development of a prediction model for 30-day mortality and various methodological studies . We consider the development of prediction models in patients enrolled in the US (N= 23,034, 1,565 deaths, and a small subset with N=259, 20 deaths). We validate these models in patients enrolled outside of the US (N=17,796, 1,286 deaths). The first model only includes age as a continuous, linear term in a logistic regression analysis, while a slightly more complex model includes age, Killip class, systolic blood pressure and heart rate. Programming code for the analyses is available at www.clinicalpredictionmodels.org with R software (R Foundation for Statistical Computing, Vienna, Austria, www.r-project.org). The original GUSTO-I model was based on 40,830 patients and included 14 predictors .

Seven steps to model development

We propose seven logically distinct steps in the development of prediction models with regression analysis. These steps are addressed below, with more detail provided elsewhere (16). A glossary is provided which summarizes definitions and characteristics of terms relevant to prediction model development and validation (Appendix).

1. *Problem definition and data inspection:* An important preliminary step is to carefully consider the prediction problem. Questions to be addressed include:

- What is the precise research question? In prediction research, we often are both interested in insight in what factors predict the endpoint, and in the pragmatic issue of providing estimates of the absolute risk for the endpoint, based on a combination of factors (17). Indeed, in the development of the GUSTO-I model, the title mentions 'Predictors of ..', suggesting a focus on insight in which prognostic factors predict 30-day mortality. The analysis however also describes the development of a multivariable prediction model, where a combination of factors predicts absolute risk with a logistic regression formula . For insight in the importance of predictors, effects are usually expressed on a relative scale, e.g. as an odds ratio (OR). For example, older age is associated with higher 30-day mortality, reflected in an OR of 2.3 per 10 years in GUSTO-I, with 95% confidence interval 2.2 – 2.4. In contrast, risk predictions are expressed as probabilities on an absolute scale between 0 and 100%. For example, predicted risks for 30-day mortality are 2% and 20% for 50 and 80-year-old patients, respectively. Uncertainty can be indicated with 95% confidence intervals. This is relevant for relative effects, but may confuse rather than help patients and clinicians when provided with absolute risk predictions (18).

- What is already known about the predictors? Literature review and close interactions between statisticians and clinical researchers are important to incorporate subject matter knowledge in the modeling process. The full GUSTO-I data set was of exceptional size, such that many modeling decision could reliably be guided by findings in the data. With smaller data sets, drawn from GUSTO-I, simulations showed that developed prediction models are unstable and produce too extreme predictions (11).

- How were patients selected? Patient data used for model development are commonly collected for a different purpose. The GUSTO-I trial was designed to study the therapeutic effects of streptokinase and tissue plasminogen activator. The inclusion criteria for the trial were relatively broad, which implies that

the GUSTO-I patients may be reasonably representative for the population of patients with an acute MI. We recognize that representativeness is usually a concern when RCT data are used for prognostic research, but this may be outbalanced by the superior quality of the data. For prediction of risk in current medical practice, the GUSTO-I data are likely outdated, since accrual in the trial was over 20 years ago.

Another issue is how we should deal with any treatment effects in prognostic analyses. The treatment effect may be of specific interest, and adjustment for baseline prognostic factors has several advantages in the estimation of a treatment effect that is applicable to individual patients (19) (20). If the focus is on absolute risk prediction, treatment effects have often been ignored, also since they are usually relatively small compared to the effects of prognostic factors. In our analyses of the GUSTO-I trial we study prognostic effects without consideration of the treatment.

For diagnostic prediction models, we note that these should be developed in subjects suspected of having a particular condition (3). The selection should mimic the clinical setting, e.g. we may consider patients with chest pain suspected of obstructive coronary artery disease for a diagnostic prediction model that estimates the presence of obstructive disease .

- Were the predictors reliably and completely measured? Many data sets are incomplete with respect to the values for some potential predictors. By default, patients with any missing value are excluded from statistical analyses (complete case analysis, or available case analysis). This is inefficient since available information of other predictors is lost. To solve this problem we may fill in best estimates for the missing values, exploiting the correlation between variables in the data set (both predictor, endpoint, and auxiliary variables such as calendar time and place) (22). Various statistical approaches are available to perform such imputation of missing values. Multiple imputation is a procedure to fill in missing values multiple times (typically at least 5 times) to appropriately address the randomness of the estimation

procedure. We recognize that imputation should be performed carefully, but is usually preferable to a complete case analysis (22).

In GUSTO-I, the collection of baseline data was prospective, since it was an RCT, and the definition of all predictors was carefully documented in the trial protocol. This makes us confident regarding the quality of the data. It also limited the occurrence of missing values, which were filled in with a single imputation procedure considering correlations between predictor variables .

- Is the endpoint of interest? Hard endpoints are usually preferred, such as 30-day mortality in GUSTO-I, which was also the primary endpoint in the trial. Another important issue is the frequency of the endpoint, which determines the effective sample size (rather than the total number of patients). The GUSTO-I data set had 2851 deaths, which allows for reliable prediction modeling, where a common rule of thumb is to require at least ten events per variable (EPV) (2) (12).

2. *Coding of predictors:* Categorical and continuous predictor variables can be coded in different ways. Categories with infrequent occurrence may for instance be collapsed with others. In GUSTO-I, location of the infarction might well be coded as anterior versus other, rather than as anterior, inferior, other, since a location other than anterior or inferior was infrequent (3% of the patients), and did not lead to a better model fit (p=0.30, likelihood ratio test for improving the logistic regression model).

Continuous predictors can often be modeled as a linear association in a regression model, at least as a starting point (23). The interpretation of the relative effect of a predictor requires attention for the scale of measurement. For example, the importance of age is easier interpreted when scaled per decade (OR 2.3, more than a doubling in odds per decade) than per year (OR 1.09, 9% higher odds per year older). Linear terms are obviously not appropriate for predictors with non-linear associations, such as a J-shape or U-shape. Such shapes can efficiently be modeled using restricted cubic splines (2) or fractional polynomials (23). These functions provide flexibility but use only few extra regression

coefficients. We emphasize that continuous predictors should not be dichotomized (categorization as below versus above a certain cut-off) in the model development phase, since valuable information is lost (24). In a later phase we may search for a user-friendly format of the prediction model and categorize some predictors (e.g. in 3, 4, or 5 categories), if the loss of predictive information is limited .

3. *Model specification:* Various strategies may be followed to choose predictors for inclusion in the prediction model. Stepwise selection methods are widely used to reduce a set of candidate predictors, but have many disadvantages. In particular when the numbers of events are low, the selection is instable, the estimated regression coefficients are too extreme, and the performance of the selected model is overestimated (2) (10) (16) (26). In the small sample of 259 patients, age and systolic blood pressure were statistically significant predictors, but Killip class (p=0.31) and heart rate (p=0.92) were not. In the total GUSTO-I data set, the sample size was large enough to rely on statistical testing for identification of predictors. All 14 predictors selected for the GUSTO-I model had p-values below 0.01 .

A related issue is how researchers should deal with assumptions in regression models, such as that the effect of one predictor is the same irrespective of the value of other predictors. This additivity assumption can be relaxed by including statistical interaction terms (2). In the full GUSTO-I data set, one such term was included in the prediction model (age*Killip class) . It appeared that with higher age, the prognostic effect of higher Killip class was less strong than for younger patients. For time to event data, the Cox model assumes proportionality of effects, which can be tested with interaction terms including time (2).

We note that iterative cycles of testing of the importance of predictors, assumptions in the model, and adaptation may lead to a model that fits the data well. But such a model may provide predictions that do not generalize to new subjects outside the data under study (“overfitting”) (2) (16). A simple, robust model may not fit the data perfectly, but should be preferred to an overly fine-tuned

model for the specific data under study. Similarly, predictor selection should usually consider clinical knowledge and previous studies rather than solely rely on statistical selection methods (2) (16).

4. *Model estimation:* Once a model is specified, regression coefficients need to be estimated. For logistic and Cox regression models, we commonly estimate coefficients with maximum likelihood (ML) methods. Some modern techniques have been developed that aim to limit overfitting of a model to the available data, such as statistical shrinkage techniques (27), penalized ML estimation (28), and the least absolute shrinkage and selection operator (LASSO) (11) (29). For the model estimated in 259 patients, we found a shrinkage factor of 0.82. The regression coefficients should be multiplied by that value to provide more reliable predictions for new patients.

5. *Model performance:* For a proposed model, researchers need to determine the quality. Several performance measures are commonly used, including measures for model calibration and discrimination. We discuss these and measures for clinical usefulness with model validation.

6. *Model validity:* It is important to separate internal and external validity. Internal validity of a prediction model refers to the validity of claims for the underlying population that the data originated from ('reproducibility') (30). Internal validation may especially address the stability of the selection of predictors, and the quality of predictions. Using a random sample for model development, and the remaining patients for validation ('split sample validation') is a common, but suboptimal form of internal validation (13). Better methods are cross-validation and bootstrap resampling, where samples are drawn with replacement from the development sample (2). In the small sample of 259 patients, bootstrapping indicated that the discriminative ability was expected to decrease from 0.82 to 0.78 in new patients. Such internal validation should always be attempted when developing a prediction model.

External validity refers to generalizability of claims to 'plausibly related' populations (30). External validation is commonly considered a stronger test for prediction models than internal validation, since it addresses transportability rather than reproducibility. External validity may be evaluated by studying patients who were more recently treated (temporal validation), from other hospitals (geographic validation), or treated in fully different settings (strong external validation) (30) (31).

7. *Model presentation:* As a final step we propose to consider is the presentation of a prediction model, such that it best addresses the clinical needs. Regression formulas can be used, such as the formula published with the GUSTO-I model . Many paper-based alternatives are available for easy applicability of a model, including score charts and nomograms (2). A recent trend is to present prediction models as web-based calculators, or as apps for mobile phones and tablets. In the future, prediction models may well be embedded in decision aids and in electronic patient records to support clinical decision-making.

An ABCD for model validation

Whatever the method used to develop a model, one could argue that validity is all that matters (7). We propose four key measures in the assessment of the validation of prediction models, related to calibration, discrimination and clinical usefulness (32) (33) (34) (35). The measures are illustrated by studying the external validity of the models developed in 259 or 23,034 patients enrolled in GUSTO-I in the US and tested in 17,796 GUSTO-I patients from outside the US.

A, Alpha: Calibration-in-the-large

Calibration refers to the agreement between observed endpoints and predictions (33). For example, if we predict a 5% risk that a patient will die within 30 days, the observed proportion should be approximately 5 deaths per 100 with such a prediction. Calibration can well be assessed graphically, in a plot with predictions on the x-axis and the observed endpoint on the y-axis (Figure 1). The observed values on the y-axis are 0 or 1 (dead/alive), while the predictions on the x-axis range between 0 and 100%. To visualise the agreement between the observed and predicted values, smoothing techniques can well be used to depict the association (36). We can also plot the observed proportions of death for groups of patients with similar predicted risk , for instance by deciles of predictions. Considering such groups with their deviations from the ideal line makes the plot a graphical illustration of the often used Hosmer-Lemeshow goodness-of-fit test. We do not recommend this test for assessment of calibration. It does not indicate the direction of any miscalibration and only provides a p-value for differences between observed and predicted endpoints per group of patients (commonly deciles) (33). Such

grouping is arbitrary and imprecise, and p-values depend on the combination of the extent of miscalibration and sample size. Rather, we emphasize the older recalibration idea as proposed by Cox in 1958 (37). Perfect predictions should be on the ideal line, described with an intercept alpha ('A') of 0 and slope beta ('B') of 1. The log odds of predictions are used as the predictor of the 0/1 outcome, or the log(hazard) for time to event outcomes (37) (38). Imperfect calibration can be characterized by deviations from these ideal values.

The intercept A relates to calibration-in-the-large, which compares the mean of all predicted risks with the mean observed risk. The parameter hence indicates the extent that predictions are systematically too low or too high. At model development observed incidence and mean predicted risk are equal for regression models, and assessment of calibration-in-the-large makes no sense. At external validation, calibration-in-the-large problems have often been found, for example for the Framingham model in multiple ethnic groups (39), or the Framingham variant developed by NICE for the UK (31). If we test the prediction model developed in 23,034 US patients outside of the US, we note a slightly higher mortality (A=0.07, p=0.38; equivalent to an odds ratio of non-US vs US of exp(0.07) = 1.07, right panel of Figure 1). Note that no intercept is calculated if a Cox model is used at external validation (35). Other types of models, such as Weibull regression models, can be used to assess calibration-in-the-large for survival models (40).

B, Beta: Calibration slope

The calibration slope B is often smaller than 1 if a model was developed in a relatively small data set. For the model based on 259 patients, the slope was 0.70 among the 17,796 non-US

patients, in line with what was expected at internal validation (shrinkage factor 0.82). For the model based on 23,034 patients, B was close to one at internal validation by cross-validation or bootstrapping , as well as at external validation in non-US patients (slope 0.99, p=0.55 for comparison to slope=1, Figure 1).

C, Concordance statistic: discrimination

Discrimination refers to the ability of the model to distinguish a patient with the endpoint (dead) from a patient without (alive). A better discriminating model has more spread between the predictions than a poorly discriminating model (34). Indeed, a model that predicts for all subjects the same predicted risk equal to the incidence shows perfect calibration, but is useless since it does not discriminate between patients. A validation plot showing a wide spread between predictions as a histogram, or between deciles of predicted risk, indicates a good discriminating model (Figure 1). Discriminative ability is commonly quantified with a concordance (*c*) statistic. For a binary endpoint, *c* is identical to the area under the receiver operating characteristic (ROC) curve, which plots the sensitivity (true positive rate) against 1 – specificity (false positive rate) for consecutive cut-offs for the predicted risk. For a time-to event endpoint, such as survival, the calculation of *c* may be affected by the amount of incomplete follow-up (censoring). We note that the *c* statistic is insensitive to systematic errors in calibration, and considers the rather artificial situation of classification in a pair of patients with and without the endpoint.

In GUSTO-I, the *c* statistic indicates the probability that among two patients, one dying before 30 days, and one surviving, the patient bound to die will have a higher predicted risk

than the surviving patient. The more complex prediction model had c statistics over 0.8 (0.813 [0.802-0.824] and 0.812 [0.800-0.824] at internal and external validation, respectively). This performance was much better than a model which only included age (0.75 at external validation, left panel in Figure 1). With development in only 259 patients, the apparent *c* statistic was 0.82, but 0.78 at internal validation, and 0.80 [0.79-0.82] at external validation. This illustrates that the availability of a smaller data set decreases model performance at external validation.

D: Decision curve analysis

Calibration and discrimination are important aspects of a prediction model, and consider the full range of predicted risks. However, these aspects do not assess clinical usefulness, i.e. the ability to make better decisions with a model than without (33). If a prediction model aims to guide treatment decisions, a cut-off is required to classify patients as either low risk (no treatment) or high risk (treatment is indicated). The cut-off is a decision threshold. At the threshold, the likelihood of benefit, e.g. reduced mortality as a result of thrombolytic therapy, exactly balances the likelihood of harm, e.g. bleeding risk and financial costs. A threshold value of e.g. 2% indicates that death of a non-treated patient is 98:2=49 times worse than the complications of a bleeding incident and costs of an unnecessarily treated patient. It is usually difficult to define a threshold since empirical evidence for the relative weight of benefits and harms is often lacking. Further, some patients may be prepared to take higher risk for a possible benefit than others. It is therefore advised to consider a range of thresholds when quantifying the clinical usefulness of a prediction model (41).

Once a threshold has been applied to classify patients as low versus high risk, sensitivity and specificity are often used as measures for usefulness. Finding an optimal balance between these is again possible with consideration of harms and benefits of treatment, in combination with the incidence of the endpoint. The sum of sensitivity and specificity can only be used as a naïve summary indicator of usefulness, since such a sum ignores the relative weight of true positives (considered in sensitivity) and false positives (considered in 1 – specificity) (42).

Recently proposed and more appropriate summary measures include the net benefit (NB) (41). This measure is consistent with the use of an optimal decision threshold to classify patients (43). The relative weight of harms and benefits is used to define the threshold, and is used to calculate a weighted sum of true minus false positive classifications (41).

For GUSTO-I, we first note that the 7% overall 30-day mortality implies a maximum NB of 7%. This is obtained if we use a threshold of 0%. All patients are then candidates for tPA treatment since we assume no harm of treatment. If we appreciate that treatment involves some harm, the optimal decision threshold is above 0%. We find that treatment decision making on the basis of risk predictions from a model would give a slightly higher net benefit than treating everyone. For example at a threshold of 2%, we have 1225 true positive classifications (candidates for treatment, and died), but also 11,192 false positive classifications (candidates for treatment, but survived) among the 17,796 non-US patients. The net benefit is calculated as (1225 – 2/98 * 11,192) / 17,796 = 5.6% (Figure 2). This is only slightly better than treating all, where NB = (1286 – 2/98 * 16,510) / 17,796 = 5.3%, since 1286 died and 16,510 survived overall. The difference is 0.3%. This implies that for every 1000 patients where we apply the prediction model, 3 extra true positives are identified without increasing the false

positive rate. The net benefit was higher for a higher threshold, such as 5% (19 per 1000 extra net true positives). Based on age only, the net benefit was virtually absent for a 2% decision threshold (0.04%, Figure 2), and in-between for a model based on only 259 patients (0.2%, Figure 2). Figure 2 illustrates that using more prognostic information than age increases the clinical usefulness of a model, and that a larger sample size leads to a better performing model. Documentation and software for decision-curve analysis is publicly available (www.decisioncurveanalysis.org), considering both binary and time-to-event end points.

Concluding remarks

We discussed seven steps to reliably develop a prediction model, and four measures that are important at model validation. There are many details to consider for each development and validation step, which are discussed in the methodological literature. Involvement of statistical experts is usually required to well develop or validate a prediction model. We provided an illustration on predicting 30-day mortality in patients with an acute myocardial infarction. The exceptional size of the US part of the GUSTO-I trial implied that overfitting was not relevant. Overfitting is a key problem in many prediction models. This is either because the number of events is small, as illustrated with the substudy in 259 patients, or because many potential predictors are studied, such as in genetic or other 'omics' analyses.

At model validation, calibration, discrimination and clinical usefulness should be considered. A validation graph such as Figure 1 (or a variant with a 'calibration belt' (44)) is an important summary tool. We furthermore recognize that a key question is nowadays how we can quantify an increase in predictive ability by new markers. For markers we may consider

changes in the *A, B, C* and *D* measures related to calibration, discrimination and clinical usefulness. Miscalibration of predictions is common at external validation (*A* different from 0, *B* < 1), but we would expect this to be similar when adding a marker to a model. Some performance measures, such as the Net Reclassification Improvement require well calibrated predictions to be meaningful (45) (46). Discrimination (*C*) only shows modest increases for most currently available markers in cardiology (47). Some researchers have blamed the *c* statistic as being insensitive. One might argue that we should merely accept that markers with statistically significant effects, but with a modest relative effect size, will not impact tremendously on identifying those with or without the endpoint (48). Cost-effectiveness analyses should eventually guide us on the question whether an increase in performance is important enough to measure an additional marker in clinical practice . Measures for clinical usefulness such as the net benefit (which may be shown in decision curves, *D*) are easy to calculate, increasingly used, and give a first impression of effectiveness in terms of potentially better patient outcomes (43). The net reclassification improvement (NRI) is very similar in behavior to measures for discrimination (50). We hence did not discuss this quite popular measure in detail here. A fierce debate is ongoing on the relevance of the NRI in the evaluation of the predictive value of markers (42) (46) (51).

We see a role for our proposed framework to support methodological researchers who develop and validate prediction models. More importantly, clinical researchers may use the framework to systematically and critically assess a publication where a prediction model is developed or validated. We anticipate that following the framework, admittedly with room for

refinements, will strengthen the methodological rigor and quality of prediction models in cardiovascular research.

References

1. Van de Werf F, Bax J, Betriu A, Blomstrom-Lundqvist C, Crea F, Falk V, Filippatos G, Fox K, Huber K, Kastrati A, Rosengren A, Steg PG, Tubaro M, Verheugt F, Weidinger F, Weis M, Guidelines ESCCfP. Management of acute myocardial infarction in patients presenting with persistent ST-segment elevation: the Task Force on the Management of ST-Segment Elevation Acute Myocardial Infarction of the European Society of Cardiology. *Eur Heart J* 2008; 29(23):2909-2945.
2. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York: Springer; 2001.
3. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *Bmj* 2009; 338:b375.
4. Mushkudiani NA, Hukkelhoven CW, Hernandez AV, Murray GD, Choi SC, Maas AI, Steyerberg EW. A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. *J Clin Epidemiol* 2008; 61(4):331-343.
5. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. *BMC Med* 2010; 8:20.
6. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, Altman DG, Moons KG. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012; 9(5):1-12.
7. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, Voysey M, Wharton R, Yu LM, Moons KG, Altman DG. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC medical research methodology* 2014; 14(1):40.
8. 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(6):1659-1668.
9. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the GUSTO database. *Stat Med* 1998; 17(21):2501-2508.
10. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol* 1999; 52(10):935-942.
11. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med* 2000; 19(8):1059-1079.
12. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making* 2001; 21(1):45-56.
13. 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. *J Clin Epidemiol* 2001; 54(8):774-781.
14. 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. *Stat Med* 2004; 23(16):2567-2586.
15. Steyerberg EW, Eijkemans MJ, Boersma E, Habbema JD. Equally valid models gave divergent predictions for mortality in acute myocardial infarction patients in a comparison of logistic regression models. *J Clin Epidemiol* 2005; 58(4):383-390.

16. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. New York: Springer; 2009.
17. Shmueli G. To Explain or to Predict? *Statistical Science* 2010; 25(3):289-310.
18. Kattan MW. Doc, what are my chances? A conversation about prognostic uncertainty. *Eur Urol* 2011; 59(2):224.
19. Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Control Clin Trials* 1998; 19(3):249-256.
20. Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics? *Am Heart J* 2000; 139(5):745-751.
21. Genders TS, Steyerberg EW, Alkadhi H, Leschka S, Desbiolles L, Nieman K, Galema TW, Meijboom WB, Mollet NR, de Feyter PJ, Cademartiri F, Maffei E, Dewey M, Zimmermann E, Laule M, Pugliese F, Barbagallo R, Sinitsyn V, Bogaert J, Goetschalckx K, Schoepf UJ, Rowe GW, Schuijf JD, Bax JJ, de Graaf FR, Knuuti J, Kajander S, van Mieghem CA, Meijs MF, Cramer MJ, Gopalan D, Feuchtner G, Friedrich G, Krestin GP, Hunink MG, Consortium CAD. A clinical prediction rule for the diagnosis of coronary artery disease: validation, updating, and extension. *Eur Heart J* 2011; 32(11):1316-1330.
22. Altman DG, Bland JM. Missing data. *BMJ* 2007; 334(7590):424.
23. Royston P, Sauerbrei W. *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Chichester, England ; Hoboken, NJ: John Wiley; 2008.
24. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006; 25(1):127-141.
25. Boersma E, Poldermans D, Bax JJ, Steyerberg EW, Thomson IR, Banga JD, van De Ven LL, van Urk H, Roelandt JR, Group DS. Predictors of cardiac events after major vascular surgery: Role of clinical characteristics, dobutamine echocardiography, and beta-blocker therapy. *JAMA* 2001; 285(14):1865-1873.
26. Derksen S, Keselman H. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *Br J Math Stat Psychol* 1992; 45:265-282.
27. van Houwelingen JC, Le Cessie S. Predictive value of statistical models. *Stat Med* 1990; 9(11):1303-1325.
28. Moons KG, Donders AR, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *J Clin Epidemiol* 2004; 57(12):1262-1270.
29. Tibshirani R. Regression and shrinkage via the Lasso. *J R Stat Soc, Ser B* 1996; 58:267-288.
30. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999; 130(6):515-524.
31. Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. *BMJ* 2012; 344:e4181.
32. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Validity of prognostic models: when is a model clinically useful? *Semin Urol Oncol* 2002; 20(2):96-107.
33. 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 (Cambridge, Mass* 2010; 21(1):128-138.
34. Royston P, Altman DG. Visualizing and assessing discrimination in the logistic regression model. *Stat Med* 2010; 29(24):2508-2520.
35. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC medical research methodology* 2013; 13:33.

36. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med* 2014; 33(3):517-535.
37. Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958; 45:562-565.
38. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. *Med Decis Making* 1993; 13(1):49-58.
39. 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.
40. van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat Med* 1995; 14(18):1999-2008.
41. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; 26(6):565-574.
42. Greenland S. The need for reorientation toward cost-effective prediction. *Stat Med* 2008; 27(2):199-206.
43. Localio AR, Goodman S. Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. *Ann Intern Med* 2012; 157(4):294-295.
44. Nattino G, Finazzi S, Bertolini G. A new calibration test and a reappraisal of the calibration belt for the assessment of prediction models based on dichotomous outcomes. *Stat Med* 2014.
45. Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011; 30(1):11-21.
46. Leening MJ, Vedder MM, Witteman JC, Pencina MJ, Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Ann Intern Med* 2014; 160(2):122-131.
47. Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *Jama* 2009; 302(21):2345-2352.
48. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004; 159(9):882-890.
49. Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS, Go AS, Harrell FE, Jr., Hong Y, Howard BV, Howard VJ, Hsue PY, Kramer CM, McConnell JP, Normand SL, O'Donnell CJ, Smith SC, Jr., Wilson PW. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation* 2009; 119(17):2408-2416.
50. Van Calster B, Vickers AJ, Pencina MJ, Baker SG, Timmerman D, Steyerberg EW. Evaluation of markers and risk prediction models: overview of relationships between NRI and decision-analytic measures. *Med Decis Making* 2013; 33(4):490-501.
51. Vickers AJ, Pepe M. Does the net reclassification improvement help us evaluate models and markers? *Ann Intern Med* 2014; 160(2):136-137.

Table 1 Illustration of seven steps for developing a prediction model with the GUSTO-I data (n=40,830)

Step	Specific issues	GUSTO-I model
1. Problem definition and data inspection	Aim: predictors or predictions?	Aim is both insights in which predictors are important and to provide individualized risk predictions
	Selection, predictor definitions and completeness, endpoint definition	Prospective data collection in a randomized trial with a hard endpoint (30-day mortality). Missing predictor values were imputed.
2. Coding of predictors	Continuous predictors	Extensive checks of transformations for continuous predictors
	Combining categorical predictors	Categories kept separate, e.g. for location of infarction
3. Model specification	Selection of main effects?	Stepwise selection; appropriate because of very large sample size
	Assessment of assumptions?	Additivity checked with interaction terms; one included (age*Killip class)
4. Model estimation	Shrinkage included?	Not necessary
5. Model performance	Appropriate measures used?	Calibration and discrimination, but no indicators of clinical usefulness
6. Model validation	Internal validation including model specification and estimation?	Bootstrap and 10 fold cross-validation
7. Model presentation	Format appropriate for audience	No; complex formula in appendix

Table 2. Overview of four measures (ABCD) for model performance.

Aspect	Measure	Visualisation	Characteristics
Calibration	**A**: alpha Calibration-in-the-large	Calibration plot	Intercept in plot; compares mean observed with mean predicted
	B: beta Calibration slope		Regression slope in plot; related to shrinkage of regression coefficients
Discrimination	**C** statistic	ROC curve	Interpretation for a pair of subjects with and without the endpoint
Clinical usefulness	**D**ecision curve analysis Net benefit	Decision curve	Net true positive classification rate by using a model over a range of thresholds

Appendix

Glossary of terms relevant to prediction model development and validation

Term	Definition	Characteristics
Continuous predictors		
Restricted cubic splines	Smooth functions with linear tails.	One to three extra regression coefficients are sufficient to adequately model most non-linear shapes.(2)
Fractional polynomials	A combination of 1 or 2 polynomial transformations such as $x^2 + x^{-1}$, with powers from the set (-3, -2, -1,-0.5,0,0.5,1,2,3).	An automatic search can find transformations to adequately model non-linear shapes, including reciprocal, logarithm, square root, square and cubic transformations.(23)
Missing values		
Complete case analysis	Analysis that considers patients with complete information on all predictors and the endpoint available	Performed by default in statistical packages, but inefficient.(22)
Multiple imputation	Fill in missing values multiple times to allow for analyses of completed data sets	Statistically efficient, but relies on assumptions of the missing value generating mechanism, and a correct imputation model.(22)
Reliable estimation		
Events per variable	The number of patients with the event of interest divided by the number of predictor variables considered.	Indicates effective sample size, which is lower than the total number of patients. The count of predictor variables should include all regression coefficients considered for all candidate predictors (not only the predictors that are finally selected for a prediction model).(2)
Stepwise selection	Procedure to eliminate candidate predictors that are not relevant to making predictions	Reduces the set of predictors. Many variants of stepwise selection are possible, all with disadvantages especially in small data sets.(16)
Shrinkage	Reduce regression coefficients towards zero, such that less extreme predictions are made.	Improves predictions from models, especially in small data sets. Specific methods include penalized estimation, such as the LASSO.(11)

Validation		
Internal validation	Assesses the validity of claims for the underlying population where the data originated from ('reproducibility')	Common methods are cross-validation and bootstrap resampling.(2)
External validation	Assesses the validity of claims for 'plausibly related' populations ('generalizability, or 'transportability').	Study patients who were more recently treated (temporal validation), from other hospitals (geographic validation), or treated in fully different settings (strong external validation).(30)
Evaluation of predictions		
Calibration	Agreement between observed and predicted risks	Calibration is usually near perfect at model development, and especially of interest at external validation. (16)
Discrimination	Ability to distinguish a patient with the endpoint from a patient without	Discrimination is a key aspect of model performance, but the concordance statistic ('C') refers to the artificial situation of a considering a pair of patients.(33)
Evaluation of classifications		
Decision threshold	Cut-off for a predicted risk to define a low and a high risk group	The optimal threshold is defined by the balance between the harm of a false-positive classification and the benefit of a true-positive classification.(41)
Sensitivity (specificity)	Probability of true-positive (true-negative) classification among those with (without) the end point	Traditional measures to quantify classification performance, conditional on knowing the endpoint.(33)
Net benefit	A weighted sum of true-positive and false-positive classifications	Novel measure to quantify classification performance, taking a decision-analytic perspective.(41)

Figure 1

Validation plots for clinical prediction models applied in 17,796 patients enrolled in GUSTO-I outside the US. The models contained the predictors age (left panel), or age plus Killip class, blood pressure, and heart rate (right panels, with n=259 or n=23,034 US patients for model development).

A: calibration-in-the-large, calculated as the logistic regression model intercept given that the calibration slope equals 1; B: calibration slope in a logistic regression model with the linear predictor as the sole predictor; C: *c* statistic indicating discriminative ability.

Triangles represent deciles of subjects grouped by similar predicted risk. The distribution of subjects is indicated with spikes at the bottom of the graph, stratified by endpoint (deaths above the x-axis, survivors below the x-axis).

Figure 2

Decision curves for the prediction models applied in 17,796 patients enrolled in GUSTO-I outside the US. Solid line: Assume no patients are treated, net benefit is zero (no true positive and no false positive classifications); Grey line: assume all patients are treated; Dotted lines: patients are treated if predictions exceed a threshold, with 30-day mortality risk predictions based on age only, or a prediction model with age, Killip class, blood pressure, and heart rate, developed in n=259 or n=23,034 US patients. The graph gives the expected net benefit per patient relative to no treatment in any patient ("Treat none"). The threshold defines the weight *w* for false-positive (FP, treat while patient did not die) versus true-positive (TP, treat a patient who died) classifications. For example, a threshold of 2% implies that FP classifications are valued at 2/98 of TP classifications, and *w* is 0.02 /(1 – 0.02) = 0.0204. The clinical usefulness of a prediction model can then be summarized as: NB = (TP - *w* FP)/N, where N is the total number of patients (41).