Evaluation of Biomarkers: Application in Urological Cancers

Moniek M. Vedder

Evaluation of Biomarkers: Application in Urological Cancers
Thesis, Erasmus MC, University Medical Center Rotterdam

ISBN: 978-94-6169-775-2
Cover Design: Matzwart, www.matzwartwerk.nl
Lay-out and printing: Optima Grafische Communicatie, Rotterdam

The research projects in this thesis were financially supported by the European Community's Seventh Framework program FP7/2007-2011 under grant agreement 201663 (UROMOL project) and the Center for Translational Molecular Medicine (CTMM) [The Prostate Cancer Molecular Medicine (PCMM) project]. This thesis was printed with financial support of the Department of Public Health of the Erasmus MC Rotterdam.

Evaluation of Biomarkers: Application in Urological Cancers

Evaluatie van biomarkers: toepassingen in urologische kankers

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
woensdag 13 januari 2016 om 11.30 uur

door

Moniek Martine Vedder

geboren te Rotterdam

Erasmus University Rotterdam

PROMOTIECOMMISSIE

Promotor: Prof.dr. E.W. Steyerberg

Overige leden: Prof.dr. K.G.M. Moons
Dr. M.J. Roobol-Bouts
Prof.dr. E.C. Zwarthoff

Copromotor: Dr. E.W. de Bekker-Grob

CONTENTS

PART I

Introduction

Chapter 1

General introduction

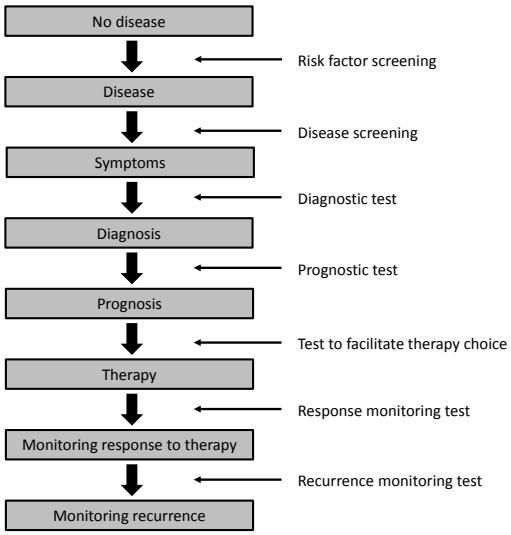
1.1 THE USE OF BIOMARKERS IN MEDICINE

Biomarkers are increasingly studied in medical research. A biomarker is a measurable characteristic of the biology of a particular health state. Biomarkers can for instance be measured in tissue, stool or bodily fluids such as blood or urine. It is an indicator for a variation from the normal values and could be used as an indicator for disease [1, 2]. Therefore, biomarkers and risk scores including these biomarkers can be used to study the probability of having a certain disease.

Some biomarkers used as predictors for deviation from a healthy state are simple to assess; body temperature for instance is a biomarker for fever. Other biomarkers require more comprehensive methods to determine. In genetics, a biomarker is a gene sequence that is either in the germline or acquired during life. Tumour related biomarkers often are produced by cancerous tissue. These biomarkers can be found using specially designed methods.

Different stages in the spectrum of a disease can be determined using different biomarkers; for example, in the diagnostic phase or monitoring phase after a patient has been treated (see figure 1.1). Biomarkers can be used as an early indicator in the case of risk factor screening. When a patient presents with symptoms, diagnostic biomarkers are sometimes available to diagnose the disease. The use of biomarkers in diagnosis and staging of cancer is increasingly popular [3, 4]. Such diagnostic tests may help in

Figure 1.1 Biomarkers in the spectrum of disease (adapted from: [6])



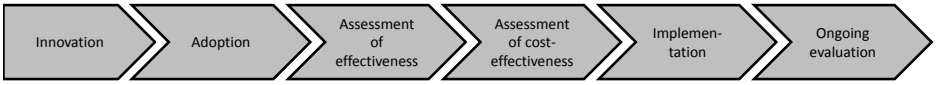
identifying which patients have cancer, and thus need treatment, and those who do not. When a biomarker is considered for use, the biomarker first has to be evaluated for its usefulness [5].

1.2 EVALUATION OF BIOMARKERS

There are several ways to evaluate new biomarkers. The type of method that should be used depends on the research question. In the early evaluation in the developmental phase of a new test the focus is on what biomarkers have sufficient potential and how cost-effective they are. In the later stage of evaluation we focus on the risk prediction of the biomarker using external validation of risk scores, and on cost-effectiveness in a clinical setting.

A systematic way to evaluate new technologies in medicine is Health Technology Assessment (HTA). HTA is a multidisciplinary field of policy analysis: it studies the medical but also social-economic implications of development and use of health technology. New biomarker tests are considered a new health technology, and the goal of an HTA analysis is to evaluate its consequences. It aids in answering several questions on effectiveness, comparative effectiveness and cost-effectiveness. Steps in the process of an HTA are shown in figure 1.2.

Figure 1.2 Steps in the process of Health Technology Assessment



A new biomarker should not be tested on its own, but within the body of knowledge that exists [7]. For certain types of cancer, this means including the biomarker in a previously developed risk prediction model or risk calculator. These prediction models most often consider binary events, which are already present in a patient (disease, i.e. diagnosis), or occur in the future (events, i.e. prognosis). These risk prediction models can aid clinicians in making a personalized decision for the patient in question [8, 9].

In this thesis, the focus is on the evaluation of biomarkers with an application in urological cancers. In evaluating the incremental predictive or diagnostic accuracy of a new biomarker, it is commonplace to use extended models and then test using a comparison of predictive values derived from the baseline model and from the model

extended with the new biomarker [10-12]. Different methods for this evaluation are available.

Evaluation methods of biomarker effectiveness

The classic measures of performance for a binary classification test are sensitivity and specificity. Sensitivity measures the proportion of positives which are correctly identified as such, i.e. positive test outcome when a disease is present. Specificity measures the proportion of negatives which are correctly identified, i.e. a negative test outcome when a disease is absent. Sensitivity and specificity form the base for several statistical methods for characterizing the degree of improved predictive power. These include the area under the curve (AUC) of a receiver operating characteristic curves with all possible cut-offs for predictions, the Net Reclassification Improvement (NRI), and decision curve analysis (DCA) over a plausible range of cut-off for predictions. A final step for the evaluation of biomarkers is cost-effectiveness analysis.

Cost-effectiveness of biomarkers

Cost-effectiveness analysis compares the relative costs and outcomes (effects) of two or more strategies. The cost-effectiveness is considered as a ratio between the incremental costs, usually monetary value, and the outcomes. In medicine, a popular health outcome considered is the Quality Adjusted Life Year (QALY). QALYs use utilities as effectiveness-measurement and are meant to capture both quality and quantity of life. The cost/QALY ratio is used as summary statistic. It is often evaluated in relation to a willingness-to-pay threshold, with sensitivity analyses to account for uncertainties and assumptions.

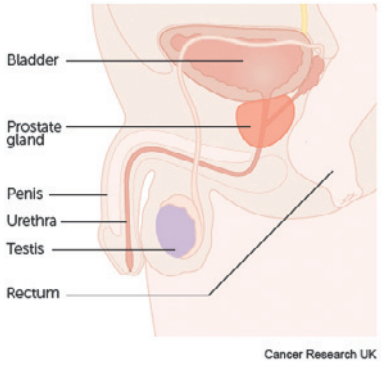
1.3 UROLOGICAL CANCERS

In this thesis, we study the evaluation of the use of new biomarkers for two urological cancers: prostate cancer and superficial bladder cancer. This is a field with active ongoing research, searching for the best biomarkers for diagnosis, to predict disease state and disease progression. The American Urological Association guidelines from 2007 and 2013 state that the development of biomarkers is highly desirable. Many biomarkers are available, however evidence for the effectiveness is poor.

Prostate cancer

Prostate cancer is the development of cancer in the prostate, a gland in the male reproductive system. Figure 1.3 shows the size and location of a healthy prostate gland.

Figure 1.3 The male reproductive system including the prostate gland (adapted from: [13])



Prostate cancer is the most common cancer in Europe for males with over 400,000 new cases diagnosed in 2012 [14]. It may initially cause no problems, and it is mainly found in men that have died from other causes [15]. Many cases can be safely followed with active surveillance or treatments that may include a combination of surgery (prostatectomy), radiation therapy, hormone therapy or chemotherapy.

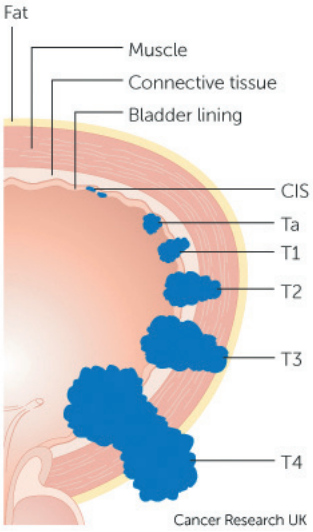
A well-known and commonly used biomarker in medicine is PSA, the Prostate Specific Antigen. PSA was first measured quantitatively in the blood in 1986 and PSA testing has been the mainstay for prostate cancer diagnosis in clinical practice since it was discovered [16]. PSA is not a unique indicator of prostate cancer, but may also detect prostatitis or benign prostatic hyperplasia. Thirty percent of patients with high PSA have prostate cancer diagnosed after biopsy. Hence, 70% of biopsies are potentially avoidable. New biomarkers can improve on the prediction of prostate cancer upon PSA alone and thus reduce the number of unnecessary biopsies.

Risk prediction models that predict the presence of prostate cancer include, besides PSA, digital rectal exam (DRE) and/or transrectal ultrasound (TRUS) assessed prostate volume, age, and the grade of prostate cancer. These factors together predict the chance of finding prostate cancer at biopsy. Prostate biopsies carry risks such as infection and/or hospitalization. PSA testing identifies many non-aggressive cancers. Uncertainty and fear cause patients and doctors to treat aggressively. Therefore, we need to better identify which patients need biopsies and treatments. New biomarkers that are considered for the prediction of prostate cancer include subforms of PSA, such as free PSA and the Prostate Health Index, PCA3 (an urine test for RNA that is elevated in prostate cancer), and 4k-panel (a serum RNA test) [17-21].

Bladder cancer

Bladder cancer is the most common malignancy of the urinary tract and a major health issue [14, 22]. Most patients with bladder cancer are diagnosed with non-muscle invasive bladder cancer (NMIBC: stage Ta or T1) [23] , shown in figure 1.4. After transurethral resection (TUR) of the tumour, recurrence of disease occurs in 30–60% of patients and, approximately, 10–15% develop progression to muscle-invasive disease in 5-year after diagnosis [24]. Therefore, regular cystoscopy is carried out for surveillances after TUR. This is a burden for patients. It is associated with urinary tract infections and other negative side effects, and potentially superfluous hospital visits.

Figure 1.4 Diagram showing the T stages of bladder cancer (adapted from: [25])



To better target surveillance, risk scores for recurrence and progression prediction have been developed. The best known are the European Organisation for Research and Treatment of Cancer (EORTC) [26] and the Spanish Urological Club for Oncological Treatment (CUETO) [27] risk scores. The use of independent validation samples is essential for measuring the predictive ability of these risk prediction scores.

1.4 AIM OF THIS THESIS

The aim of this thesis is to study methods to evaluate biomarkers and their application in two urological cancers: prostate or bladder cancer. The main study questions are:

1. What are controversies in performing and reporting the Net Reclassification Improvement (NRI) and how can the graphical assessment of incremental value of new biomarkers be approved?
2. Under what conditions is adding a new biomarker to PSA testing cost-effective for prostate cancer screening?
3. What is the added value in predictive ability of new biomarkers to existing risk prediction models for prostate cancer?
4. How well can recurrence and progression of bladder cancer be predicted with current models?

1.5 OUTLINE OF THIS THESIS

This thesis consists of six parts. Part I contains the general introduction. Part II (Chapters 2 and 3) focuses on the methods of biomarker evaluation. Chapter 2 assesses the computation, interpretation and controversies to the topic of the Net Reclassification Improvement (NRI). Chapter 3 describes the graphical assessment of incremental value of new biomarkers in prediction models. These chapters aim to answer study question 1.

Part III (Chapter 4) involves the early HTA of new biomarkers to answer study question 2. In this part, the cost-effectiveness of prostate cancer screening using a PSA test combined with a new biomarker is determined.

In part IV (Chapters 5 to 7) the focus is on late evaluation of risk prediction scores and new biomarkers. In this part of the thesis, the added value of percentage of free to PSA, PCA3, and a kallikrein panel to the ERSPC risk calculator for prostate cancer in prescreened men is determined. Furthermore, it describes a comparison of two prostate cancer risk calculators that include the Prostate Health Index. These chapters aim to answer study question 3. Finally, risk prediction scores for recurrence and progression of non-muscle invasive bladder cancer are validated in men and women with primary bladder cancer to answer study question 4.

Part V is a general discussion, which summarizes the main findings of this thesis and answers the research questions stated in the General introduction. In addition, practical implications are discussed with recommendations for future research.

The last part of this thesis contains a summary in both English and Dutch. It continues with a list of publications and acknowledgements and concludes with and the PhD portfolio of the Erasmus University Rotterdam and the curriculum vitae of the author.

REFERENCES

1. Biomarkers Definitions Working G. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001 Mar;69(3):89-95.
2. Strimbu K, Tavel JA. What are biomarkers? Curr Opin HIV AIDS. 2010 Nov;5(6):463-6.
3. Mishra A, Verma M. Cancer biomarkers: are we ready for the prime time? Cancers (Basel). 2010;2(1):190-208.
4. Henry NL, Hayes DF. Cancer biomarkers. Mol Oncol. 2012 Apr;6(2):140-6.
5. Steyerberg EW. Clinical Prediction Models. Gail M, Krickeberg K, Samet J, Tsiatis A, Wong W, editors: Springer; 2009.
6. Redekop K, Uyl-de Groot C. Diagnostiek en economische evaluatie. Van kosten tot effecten: Een handleiding voor economische evaluatiestudies in de gezondheidszorg: Elsevier gezondheidszorg; 2010.
7. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2015 Apr 18.
8. AlHilli MM, Mariani A, Bakkum-Gamez JN, Dowdy SC, Weaver AL, Peethambaram PP, *et al.* Risk-scoring models for individualized prediction of overall survival in low-grade and high-grade endometrial cancer. Gynecol Oncol. 2014 Jun;133(3):485-93.
9. Porta N, Calle ML, Malats N, Gomez G. A dynamic model for the risk of bladder cancer progression. Stat Med. 2012 Feb 10;31(3):287-300.
10. 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 Jan 30;27(2):157-72; discussion 207-12.
11. Pencina MJ, D'Agostino RB, Vasan RS. Statistical methods for assessment of added usefulness of new biomarkers. Clin Chem Lab Med. 2010 Dec;48(12):1703-11.
12. Pencina MJ, D'Agostino RB, Pencina KM, Janssens AC, Greenland P. Interpreting incremental value of markers added to risk prediction models. Am J Epidemiol. 2012 Sep 15;176(6): 473-81.
13. Cancer Research UK. The prostate. [updated 20-02-2014]; Available from: http://www.cancerresearchuk.org/about-cancer/type/prostate-cancer/about/the-prostate.
14. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JW, Comber H, *et al.* Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. Eur J Cancer. 2013 Apr;49(6):1374-403.
15. Bell KJ, Del Mar C, Wright G, Dickinson J, Glasziou P. Prevalence of incidental prostate cancer: A systematic review of autopsy studies. Int J Cancer. 2015 Mar 26.
16. Stamey TA, Yang N, Hay AR, McNeal JE, Freiha FS, Redwine E. Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate. N Engl J Med. 1987 Oct 8;317(15): 909-16.

17. Jansen FH, van Schaik RH, Kurstjens J, Horninger W, Klocker H, Bektic J, *et al.* Prostate-specific antigen (PSA) isoform p2PSA in combination with total PSA and free PSA improves diagnostic accuracy in prostate cancer detection. Eur Urol. 2010 Jun;57(6):921-7.

18. Bussemakers MJ, van Bokhoven A, Verhaegh GW, Smit FP, Karthaus HF, Schalken JA, *et al.* DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. Cancer Res. 1999 Dec 1;59(23):5975-9.

19. Hessels D, Schalken JA. The use of PCA3 in the diagnosis of prostate cancer. Nat Rev Urol. 2009 May;6(5):255-61.

20. Vickers AJ, Cronin AM, Aus G, Pihl CG, Becker C, Pettersson K, *et al.* A panel of kallikrein markers can reduce unnecessary biopsy for prostate cancer: data from the European Randomized Study of Prostate Cancer Screening in Goteborg, Sweden. BMC Med. 2008;6:19.

21. Gupta A, Roobol MJ, Savage CJ, Peltola M, Pettersson K, Scardino PT, *et al.* A four-kallikrein panel for the prediction of repeat prostate biopsy: data from the European Randomized Study of Prostate Cancer screening in Rotterdam, Netherlands. Br J Cancer. 2010 Aug 24; 103(5):708-14.

22. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, DM. P. GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10. Lyon, France: International Agency for Research on Cancer; [27-01-2013]; Available from: http://globocan.iarc.fr/.

23. Babjuk M, Burger M, Zigeuner R, Shariat SF, van Rhijn BW, Comperat E, *et al.* EAU guidelines on non-muscle-invasive urothelial carcinoma of the bladder: update 2013. Eur Urol. 2013 Oct;64(4):639-53.

24. Kirkali Z, Chan T, Manoharan M, Algaba F, Busch C, Cheng L, *et al.* Bladder cancer: epidemiology, staging and grading, and diagnosis. Urology. 2005 Dec;66(6 Suppl 1):4-34.

25. Cancer Research UK. Bladder cancer stage and grade. [updated 22-10-2013]; Available from: http://www.cancerresearchuk.org/about-cancer/type/bladder-cancer/treatment/bladder-cancer-stage-and-grade.

26. Sylvester RJ, van der Meijden AP, Oosterlinck W, Witjes JA, Bouffioux C, Denis L, *et al.* Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. Eur Urol. 2006 Mar;49(3):466-5; discussion 75-7.

27. Fernandez-Gomez J, Madero R, Solsona E, Unda M, Martinez-Pineiro L, Gonzalez M, *et al.* Predicting nonmuscle invasive bladder cancer recurrence and progression in patients treated with bacillus Calmette-Guerin: the CUETO scoring model. J Urol. 2009 Nov;182(5): 2195-203.

PART II

Methods of biomarker evaluation

Chapter 2

Net Reclassification Improvement: Computation, interpretation, and controversies: A literature review and clinician's guide

Leening MJ, Vedder MM, Witteman JC, Pencina MJ, Steyerberg EW

Ann Intern Med. 2014 Jan 21;160(2):122-31.

ABSTRACT

The net reclassification improvement (NRI) is an increasingly popular measure to evaluate improvements in risk predictions. In this chapter we first review 67 publications in high-impact general clinical journals that considered the NRI. We find incomplete reporting of methods for the NRI, incorrect calculation, and common misinterpretations. To aid improved applications of NRI, we elaborate on several aspects of the computation and interpretation in various settings. We discuss limitations and controversies, including the impact of miscalibration of prediction models, the use of continuous NRI and 'clinical' NRI, and the relation with decision-analytic measures. We propose a systematic approach towards presenting NRI analysis: detail and motivate the methods used for computation of the NRI; use clinically meaningful risk cut-offs for category-based NRI; always report the separate NRI components; address issues of calibration; and do not interpret the overall NRI as a percentage of the study population reclassified. Promising NRI findings need to be followed with decision-analytic or formal cost-effectiveness evaluations.

2.1 INTRODUCTION

Ever since the first introduction of the term *risk factor* little over 50 years ago in the Annals of Internal Medicine [1], a plethora of such factors has been identified. Risk factors have been incorporated into statistical models to predict future disease, to more adequately diagnose patients, or to predict outcomes after disease has been diagnosed. A substantial number of clinical guidelines have incorporated risk prediction models to aid clinicians in everyday decision making in various fields of medicine, including cardiology, oncology, and respiratory medicine [2-8].

Many markers, such as biomarkers, genetic factors, and imaging results, have been proposed to improve on these prediction models. In the last three decades, the most commonly used measure to quantify these improvements has been the change in *c*-statistic, also known as the area under the receiver operating characteristic curve (AUC). Limitations of the AUC have been emphasized, including the difficulty to interpret the usually small changes in this statistic and that the magnitude of improvement is related to the performance of the baseline model [9-12]. A more relevant criterion may be to assess whether the addition of the marker to an existing model will influence clinical practice [13], which is the case if the newly predicted risk crosses a clinically meaningful threshold for an individual. This has led to the introduction of the concept of risk reclassification [14]. It involves cross-tabulating categories of predicted risk for two prediction models to see how persons are classified differently using these two models, usually concerning one model with the new marker under study and one without. The subsequent changes in risk classification can be quantified by the net reclassification improvement (NRI) [15]. Risk reclassification analysis with the NRI have become extremely popular: over 1000 publications have thus far cited the 2008 article introducing the NRI [15]. However, reporting of the methods used is of heterogeneous quality [16] and misconceptions are common in interpreting the NRI [17].

In this chapter we aim to provide a systematic assessment of the reporting practices in analyses involving the NRI and address some controversies relating to its use and interpretation. We also propose recommendations on how to report and interpret the NRI [18].

2.2 OVERVIEW OF CURRENT REPORTING

Literature Search and Data Extraction

We systematically collected studies that computed NRI or discussed results from NRI analysis. Thereto we used the Thomson Reuters Web of Knowledge (version 5.9) to identify all publications that cited one of four methodological articles by Pencina *et al.* [15,19-21] or a methodological review on reclassification measures by Cook and Ridker [22]. The search was last updated April 23 2013, yielding 1250 unique citations (supplementary figure 2.1). We selected all 67 citations in the four general clinical journals with the highest impact factors (*N Engl J Med*, *Lancet*, *JAMA*, and *Ann Intern Med*) [22-88] for data extraction (supplementary tables 2.1 and 2.2). Our motivation was that these articles may be expected to have broad impact, and be used as examples for others.

Two evaluators (M.J.G.L. and M.M.V.) independently extracted data from the publications. Cases on which the evaluators disagreed were discussed to reach consensus with a third evaluator (E.W.S.). All publications were searched for calculations or results of NRI. If so, we checked which version of the NRI was used; the category-based NRI [15] or the continuous (i.e. category-free) NRI [20] (see table 2.1). Next, we reviewed all articles whether risk categories corresponding to diagnostic or treatment thresholds from clinical guidelines were used to evaluate the category-based NRI, or whether other categorization was justified. We determined what NRI components were reported: solely the overall NRI, or the event NRI and the nonevent NRI as well (see table 2.1). Moreover, we categorized studies reporting on estimates of the overall NRI into those reporting it as a unit-less statistic or as a percentage.

Table 2.1 Formulas and interpretation of the NRI

Category-based NRI	
Event NRI	= Pr(up|event) – Pr(down|event)
	= (number of events classified up – number of events classified down) ÷ number of events
	The net percentage of persons with the event of interest correctly classified upwards.
	The category-based event NRI can be interpreted as a percentage with a range of -100% to +100%. *
Nonevent NRI	= Pr(down|nonevent) – Pr(up|nonevent)
	= (number of nonevents classified down – number of nonevents classified up) ÷ number of nonevents
	The net percentage of persons without the event of interest correctly classified downwards.

Table 2.1 Formulas and interpretation of the NRI (continued)

Category-based NRI	
	The category-based nonevent NRI can be interpreted as a percentage with a range of -100% to +100%. *
Overall NRI	= [Pr(up|event) – Pr(down|event)] + [Pr(down|nonevent) – Pr(up|nonevent)]
	= event NRI + nonevent NRI
	The sum of the net percentages of correctly reclassified persons with and without the event of interest.
	Thereby, the category-based overall NRI is a statistic that is implicitly weighted for the event-rate and cannot be interpreted as a percentage.
	The theoretical range of the category-based overall NRI is -2 to +2.
Continuous NRI	
Event NRI	= Pr(higher|event) – Pr(lower|event)
	= (number of events with increased predicted risk – number of events with decreased predicted risk) ÷ number of events
	The net percentage of persons with the event of interest correctly assigned a higher predicted risk.
	The continuous event NRI can be interpreted as a percentage with a range of -100% to +100%. *
Nonevent NRI	= Pr(lower|nonevent) – Pr(higher|nonevent)
	= (number of nonevents with decreased predicted risk – number of nonevents with increased predicted risk) ÷ number of nonevents
	The net percentage of persons without the event of interest correctly assigned a lower predicted risk.
	The continuous nonevent NRI can be interpreted as a percentage with a range of -100% to +100%. *
Overall NRI	= [Pr(higher|event) – Pr(lower|event)] + [Pr(lower|nonevent) – Pr(higher|nonevent)]
	= event NRI + nonevent NRI
	The sum of the net percentages of persons with and without the event of interest correctly assigned a different predicted risk.
	Thereby, the continuous overall NRI is a statistic that is implicitly weighted for the event-rate and cannot be interpreted as a percentage.
	The theoretical range of the continuous overall NRI is -2 to +2.

* Negative percentages are to be interpreted as a worsening in risk classification (i.e. the number of incorrectly reclassified (non)events exceeds the number of correctly reclassified (non)events).
NRI = net reclassification improvement; Pr = probability.

Results

The predominant reason for citing one of the methodological articles was the computation of NRI estimates (n=39, table 2.2). In 2 (5%) articles only the continuous NRI was computed. In 5 articles the NRI was used to compare two different models instead of the nested addition of one or more new risk markers to a simpler model.

Table 2.2 Results from the literature review on reporting of the net reclassification improvement

Reporting of NRI feature	Studies, n (%)
Context for citing methodological article on NRI	
Claim to have calculated NRI	39 (58.2) *
Discuss NRI results from previous analysis	4 (6.0) *
Suggest alternative methods for quantifying predictive abilities	16 (23.9) *
Compute other (non-NRI) measures elaborated on in this chapter	8 (11.9) *
Risk categorization	
Only continuous (category-free) NRI computed	2 (5.1) †
Categorization for computing NRI detailed	34 (91.9) ‡
Categorization for computing NRI justified in text	10 (27.0) ‡
Reference given for NRI categorization	14 (37.8) ‡
Categorization for computing NRI correspond to diagnostic or therapeutic implications in clinical guidelines	4 (10.8) ‡
Time horizon and follow-up	
Predicted horizon detailed	30 (78.9) ||
Observed follow-up detailed (mean, median, or maximum)	37 (97.4) ||
Predicted time horizon longer than observed follow-up	7 (23.3) ¶
Components §	
Overall NRI	36 (92.3) †
Event NRI and nonevent NRI in text or tables	11 (28.2) †
Reclassification table for main findings	25 (67.6) ‡
Unit §	
Reported as a percentage	24 (66.7) **
Interpreted as a percentage or proportion	8 (22.2) **

* Out of all 67 publications included in the literature review.
† Out of 39 studies that calculated NRI.
‡ Out of 37 studies that calculated category-based NRI.
§ For more details on the components and units of the NRI refer to table 2.1.
|| Out of 38 prospective studies that calculated NRI.
¶ Out of 30 prospective studies that calculated NRI and detailed predicted horizon and follow-up.
** Out of all 36 studies that reported the overall NRI.
NRI = net reclassification improvement.

Of the 37 articles that computed category-based NRI results 34 (92%) detailed the cut-offs for the risk categories chosen. The number of risk categories defined to compute NRI varied between 2 and 6, with 3 being the most commonly used (supplementary table 2.1). These risk categories were justified in the text or by references in 22 (59%) instances and fully matched clinically meaningful categories with clear implications from guidelines in 4 (11%) instances (table 2.2). For outcomes other than atherosclerotic cardiovascular disease, the motivation for the risk categorization could not be retraced in 10 out of 12 instances. Another 8 studies on the prediction of various manifestations

of cardiovascular disease used cut-offs for the NRI that are subject of ongoing debate [28,60,70,89,90], for instance a 6% rather than a 10% 10-year risk cut-off for low risk of coronary heart disease. In 14 publications cut-offs for coronary risk stratification were applied to broader definitions of cardiovascular disease (supplementary table 2.1).

Among 38 prospective studies calculating the NRI, 30 (79%) clearly reported the horizon at which the risk predictions were evaluated. As for the horizon at which the risk predictions were evaluated, 30 out of 38 (79%) prospective studies calculating the NRI reported what time point the predicted risks were assessed. In 7 out of 30 (23%) instances where both predicted horizon and observed follow-up was detailed we could infer that the authors studied a predicted horizon beyond the observed follow-up time (table 2.2). We identified another 7 studies that used events occurring beyond the predicted horizon in the reclassification analysis.

Nearly all studies reported the overall NRI. Only 11 (28%) articles presented its components, the event NRI and nonevent NRI, in the results section. However, a total of 25 (68%) presented reclassification tables stratified for events and nonevents (table 2.2). This allowed for computing the event NRI and nonevent NRI by a knowledgeable reader. By combining the components presented in the text and the reclassification tables we identified 29 (74%) studies with information on event NRI and nonevent NRI presented for at least one reclassification analysis. Notably, one study claimed to have calculated the NRI, but no such results could be retraced. Another study only presented p-values, but no point estimates of the NRI.

Of all 36 studies presenting estimates of the overall NRI, 24 (67%) expressed the overall NRI as a percentage (table 2.2). A total of 8 (22%) articles in our review interpreted the overall NRI as a percentage or proportion of the entire study population that was correctly reclassified, or used similar wording, such as interpreting an overall NRI of 0.29 as *"[...], 29% of patients were correctly reclassified [...]"* [17,39].

2.3 NRI COMPUTATION, COMPONENTS, AND INTERPRETATION

Predicted Time Horizons and Follow-up

When prospective data are involved, such as cardiovascular events occurring during follow-up, it should be clear what time-horizon was used to calculate the predicted risks. Since virtually every prospective study has some loss to follow-up, it is important to adequately handle observations with incomplete follow-up in the analysis. In our

review, studies published shortly after the introduction of the NRI often did not report how incomplete follow-up was handled. Some studies classified censored observations as nonevents ('naive extrapolation') or excluded persons with incomplete follow-up. Better methods have been proposed to limit loss of useful information. They include either the Kaplan-Meier-based estimates of the expected number of events and nonevents ('prospective NRI') [20,78], or inverse-probability weighting [91]. Similarly, not every study has sufficient follow-up available for the predicted time horizons used in clinical guidelines (e.g. 10-year risk of coronary heart disease [89]). In our literature review authors made various attempts to overcome this problem, such as using Weibull extrapolation [48,53], adjusted the predicted risk cut-offs by the ratio of actual to desired follow-up [24], or extrapolating the observed rates on the Kaplan-Meier survival estimates to the predicted time horizon for presentation purposes [22].

Risk Categories

The NRI was introduced with the example of the added value of HDL-cholesterol to coronary risk prediction in the Framingham Heart Study [15]. Current clinical guidelines on primary prevention of cardiovascular disease recommend clear cut-offs on initiation of statin treatment [2,3,89,90]. These recommendations are supported by cost-effectiveness analyses. The NRI captures the change in a person's predicted risk that crosses one of such cut-offs and thus translates into a clinically meaningful change in treatment recommendations.

Our review of the literature confirms the findings by Tzoulaki and colleagues: selected risk cut-offs are generally poorly motivated and rarely correspond to therapeutic implications. Both shortcomings have been shown to yield significantly higher NRI estimates [16,81]. In some cases, the existing clinical cut-offs may result in limited reclassification. For example, when studying a population at a very low risk of cardiovascular disease, only a small number of participants would be considered high-risk and as a results few will cross the recommended risk thresholds after the addition of a new marker [92]. Using the existing cut-offs illustrates the limited utility of a new marker in real life application to such a low-risk population. Choosing *a priori* clinically meaningful cut-offs has been frequently emphasized [15,16,19,20,60,63,81,92-98]. Besides the reasons of clinical interpretation, the estimates of the NRI and its components increase with the number of categories [95,99]. Limiting analysis to clinically meaningful categories will forestall authors from presenting results from the cut-offs with the highest magnitude of NRI in their data. Moreover, consistent use of cut-offs enhances comparability of results on the same markers between studies, provided that the same outcome definition and time-horizons are used.

Although there is an abundance of risk prediction algorithms described in the medical literature, only a limited number of clinical guidelines outside the field of cardiology explicitly recommend risk thresholds for use in clinical practice. In the fields where meaningful cut-offs are lacking or evolving, various options have been suggested to overcome this problem. Each has its own caveats. First, in some cases classification thresholds exist for related outcomes. For example, 20% 10-year risk of 'hard coronary heart disease' corresponds to a 25% 10-year risk of 'total coronary heart disease' [100]. In these situations a conversion factor based on the ratios of event-rate can be used to translate cut-offs from one application to another, in this example a ratio of 1.25. Such conversion assumes that the associated clinical implications are similar for the different outcome definitions. This may not always be true. For example, the protective effect of statins on the occurrence of cardiovascular manifestations other than coronary heart disease, such as heart failure, may be less [101]. Similarly, conversion factors can be used to define risk cut-offs for different predicted time-horizons (e.g. 30-year versus 10-year risks [102]). In the absence of published conversion factors the data under study can be examined to define the relative occurrence of the outcomes. Second, some researchers have suggested defining risk categories based on the event-rate. A cut-off equal to the event-rate would be used for binary classification, and cut-offs at half the event-rate, the event-rate, and twice the event-rate when more than two categories are desired [99,103]. Such cut-offs, however, have no direct clinical interpretation. The appropriateness of risk cut-offs should be related to the anticipated use of the prediction model. As an example, risk thresholds for a model used to select patients with chest pain for early discharge from an emergency department will be set at a much lower level of myocardial infarction risk compared to thresholds for a model used to identify patients with chest pain that will benefit from early invasive coronary angiography. Third, the continuous NRI was introduced as an alternative in the absence of any categorization (table 2.1) [20]. However, the continuous NRI does not quantifying clinical impact of risk reclassification (see 'Limitations and Controversies' section below). The relation between cut-offs and the risk distribution in the data can be elegantly visualized in a reclassification graphs with superimposed cut-offs (e.g. supplementary figure 2.2).

Case-Control Studies

For reasons of costs and feasibility, the predictive value of new biomarkers is often studied in subsets of persons with events and nonevents from larger prospective studies, especially when the event-rates are low. The NRI can be used in both cohort studies and (nested) case-control studies [20]. In (nested) case-control studies the ratio of events (cases) to nonevents (controls) is determined by the researcher by selective oversampling of cases, which implies artificial weighing by the investigators [43]. This

should not lead to different estimates in magnitudes of the NRI when compared to results derived from a full cohort provided that the cases and controls are randomly selected [20,104]. However, difficulties arise when selected controls are not representative of the entire underlying subset of controls they were drawn from, such as is the case when matching on certain risk factors (even as simple as age and gender) is performed [104-106]. This can be overcome by weighing for the inverse of the sampling probability for cases and controls [101,104].

Components and Interpretation

Although clearly advised to do so by the original article introducing the NRI [15], we noticed that only a limited number of studies explicitly reported the components of the overall NRI. The components are easier to interpret than the combined number: when evaluating only a single cut-off, the event NRI equals the improvement in sensitivity and the nonevent NRI equals the improvement in specificity [15]. The NRI components then express the net percentages of persons with or without events correctly reclassified (table 2.1). Negative percentages for the components are to be interpreted as a net worsening in risk classification. The overall NRI is a sum of these two underlying components; as a result an identical point estimate of this statistic may have different interpretations depending on its components [62,93]. Large positive values of the event NRI indicate that the investigated marker aids in the detection of persons with the outcome of interest. This enables clinicians to initiate targeted treatment and thereby prevent events. On the other hand, an overall NRI driven by the nonevent NRI indicates the marker's property of correctly lowering risk estimates for nonevents, and is thus useful for reducing overtreatment. But such markers will have limited contribution to lowering the burden of disease. This illustrates the difficulty of interpreting the overall NRI without knowledge on the components [107]. Although tempting, the overall NRI cannot straightforwardly be interpreted as the *"[...] net percentage of persons correctly reclassified [...]"* [48] due to the implicit weighing by the event-rate: the overall NRI is a sum of two fractions with different denominators (i.e. the number of events and nonevents) [17]. Such misinterpretations may well have contributed to the popularity of the overall NRI, which therefore should not be presented as a percentage but as a unit-less statistic [17]. Moreover, the components of the overall NRI may be reasonably well interpretable, while their sum is less so due to the implicit weighing related to the event-rate (the costs of misclassification are assumed to be proportional to the odds of nonevents) (table 2.1) [108].

As with most summary statistics, the NRI should not be interpreted on its own but in the context of other complementary statistical measures. If a marker is not associated with the outcome or does not yield an increase in AUC, a positive NRI is not to be

expected [94]. In rare instances where this does occur random chance or differences in calibration between the two models are the most likely causes. Also, presenting reclassification tables (in tabular or graphical form) will aid in the broader interpretation of summarized reclassification statistics (e.g. supplementary table 2.3 and supplementary figure 2.2).

2.4 LIMITATIONS AND CONTROVERSIES

Miscalibration

Unlike rank-based statistics, such as the AUC, the NRI is affected by miscalibration of a model (i.e. the average predicted risk is not close to the event-rate) [108-110]. Systematic miscalibration does not occur when the performance of models is assessed on the same dataset that was used to develop them, but is often present when validating prediction models in other populations. Well-recognized examples of this phenomenon include applications of the Framingham cardiovascular risk models to European populations [111-113]. When performing a head-to-head comparison between a Framingham function (using the published coefficients and baseline hazard) and a new risk function developed on the data under study, one might find NRI that favors the new model and no difference in the AUCs [114,115]. This can be avoided by deriving both the reference model and the model including the marker under investigation from the same set of data that is used to compute the NRIs, or by recalibrating both models in case of independent validation [116].

The traditional Hosmer-Lemeshow goodness-of-fit test is strongly depending on sample size of the study [117]. Therefore, calibration might better be assessed graphically in a plot with predicted risks on the horizontal axis and observed event-rates on the vertical axis (e.g. [54]). For perfectly calibrated models, the plot forms a diagonal line where the observed event-rates equal the predicted risks. Such graphs can show systematic under- or overestimation, as well as issues of overfitting, (which can be quantified using the calibration intercept and slope [118]).

Classification or Reclassification?

Some researchers have argued that before addressing the issue of reclassification, one should first focus on risk classification and examine the margins of a reclassification table [43]. Accordingly, examining reclassification is useful only to the extent to which it quantifies change in the size of these margins. This might be of particular relevance in head-to-head comparisons of non-nested models with substantial reclassification (i.e. the two models have low correlation). In this case it is of greater interest to know

how many persons are classified in the clinically relevant subgroups, rather than the exact reclassification within the inner cells of the table [93,96,54]. So when choosing between two competing models for clinical practice, the main question is which one leads to better classification (which relates to both discrimination and calibration of the models). On the other hand, when the focus is primarily on the future potential of a new marker, the improvements in discrimination and subsequent risk reclassification that it can induce are of main interest.

Continuous NRI

The continuous NRI was originally proposed to overcome the problem of selecting categories in applications where they do not naturally exist [20]. It does not require any risk categorization and considers all changes in predicted risk for all events and nonevents. This has several consequences. First, most changes in predicted risk do not translate into changes in clinical management (e.g. a middle-aged woman whose 10-year predicted coronary risk doubles from 1% to 2% will likely not be treated differently) [92,119]. Therefore the interpretation of the continuous NRI is quite different from the category-based NRI (table 2.1) [11]. Second, when considering the addition of a normally distributed marker, the continuous NRI is much less affected by the performance of the baseline model and can therefore be seen as a rescaling of the measures of association (e.g. an odds ratio of 1.65 per standard deviation corresponds to a continuous NRI of 0.395) [11,21]. As a consequence the continuous NRI is often positive for relatively weak markers [11]. Moreover, it is strongly affected by miscalibration, especially in the setting of external validation [110].

As such, continuous NRI is less suitable for head-to-head comparisons of competing models unless these models have been developed on the same data or are correctly calibrated. The most appealing application of the continuous NRI comes in quantifying the impact of an added predictor in settings where the distributions of other risk factors may not be representative of the population [120]. For example, when the same marker for coronary risk prediction is evaluated in two different populations, one with wide and one with narrow age ranges, the conclusions about its usefulness might be different if based on the increment in AUC [12]. The continuous NRI, however, would give a consistent message. The continuous NRI is hence marker-descriptive rather than model-descriptive. Furthermore, its magnitude should be assessed on its own scale [11] and should not be compared with the category-based version.

'Clinical NRI'

Reclassification measures, including the NRI, can be used to evaluate markers in specific subgroups of the study population defined by the reference model. Specifi-

cally, the added value of new risk markers may be of greater importance in persons with a risk categorization that has more uncertainty regarding the clinical implications (e.g. persons at intermediate risk of coronary heart disease [33,48,62,72,73,86]). This 'clinical NRI' [121], however, has been found to be biased because it does not take into account incorrect reclassification from other risk categories into the intermediate risk category [62]. Adding randomly generated noninformative markers to existing prediction models leads to positive clinical NRIs more frequently than expected on the basis of chance [99,122]. A method for correcting this systematic overestimation has been published [122].

Decision-Analytic Measures

The overall NRI implicitly weights for the event-rate (p), with $1/p$ and $1/(1-p)$ serving as costs for false negatives (events classified downwards) and false positives (nonevents classified upwards) [108,123]. However, a different weighing of false positives and false negatives is often clinically more appropriate [98]. This can readily be incorporated in a weighted version of the NRI if the event NRI and nonevent NRI are presented separately, or when a reclassification table is provided [20,124]. In its broadest form the weighted NRI can be interpreted as the average savings (for instance expressed in dollars or QALYs) per person resulting from using the new model over the old one [20].

The weighted NRI is a decision-analytic measure and is mathematically a transformation of changes in net benefit and relative utility [124]. These measures use the harm-to-benefit ratio to define a single optimal decision threshold for binary classification as high risk versus low risk [125]. The harm-to-benefit ratio also defines the weights of true positive and false positive classifications to calculate a single summary measure [124-126]. The use of such decision-analytic measures is, however, limited by the fact that weights for harms and benefits are not firmly established in most fields of medicine [126], although a range of decision thresholds can be considered in a sensitivity analysis with visualization in a 'decision curve' [127].

The non-weighted category-based NRI analysis is regarded as an early stage analysis in the evaluation of new markers or prediction models. For assessment of the potential clinical utility of promising markers, decision-analytic approaches are needed in the next step, beyond the NRI analyses, but prior to a full formal cost-effectiveness analysis that incorporates changes in both costs and clinical outcomes in more detail [13].

2.5 RECOMMENDATIONS

In our literature review we encountered several common flaws in the presentation and interpretation of the NRI and insufficient documentation of the computational methods. Motivated by our observations we propose a number of recommendations for clinical research [18] (table 2.3).

Table 2.3 Recommendations for reporting the net reclassification improvement

Methods	
Type of NRI	Specify the type of NRI computed in the methods section of the manuscript; category-based and/or continuous NRI.
Follow-up	Specify the horizon of risk prediction if NRI was computed for prognostic evaluations (e.g. 10-year risk).
	Describe how censored observations (e.g. persons lost to follow-up before the specified horizon) were handled.
	Use the event status at the predicted time horizon and ignore events occurring beyond the predicted time horizon (e.g. when predicting 10-year risk of CHD consider participants with a myocardial infarction occurring after 10 years of follow-up as nonevents).
Cut-offs	For category-based NRI ideally the categorization should have clear consequences in clinical practice.
	Where possible, give reference to formal clinical guidelines used to define the risk categories for the computation of the NRI.
	If alternative cut-offs were used, clearly motivate the chosen cut-offs.
Results	
Components	Report the NRI for events and nonevents separately.
	Reclassification tables stratified for persons with and without the event of interest are informative beyond the NRI (e.g. supplementary table 2.3).
Unit of NRI	The event and nonevent NRI components can be presented as a percentage, however the overall NRI has no unit and should therefore not be presented as a percentage (also see table 2.1).
Calibration	Provide information on the calibration for the models under comparison.
Discussion	
Interpretation	The components of the overall NRI can be interpreted as a net percentage of the number of persons with or without events.
	The overall NRI, however, should not be interpreted as a net percentage of the study population correctly reclassified.
Comparisons	Do not draw strong comparative conclusions based on direct comparisons of NRIs obtained in different populations, using different outcomes, or using different cut-offs.

CHD = coronary heart disease; NRI = net reclassification improvement.

It is essential to clearly define which type of NRI is used as their applicability and relevance vary substantially. What type of NRI and what cut-points are the most appropriate depends on a number of factors as discussed in this review. We recommend separate reporting of the NRI for events and nonevents in all circumstances. Also, the sum of the NRI components should not be interpreted as a percentage. If authors choose to present the category-based NRI, the implied costs of misclassification by the event-rate should be discussed. The cut-offs selected for the NRI analyses should preferably match risk thresholds that have clear clinical implications or can be well motivated on clinical grounds. In general, the category-based NRI is directly applicable in settings where meaningful risk categories exist and models are well-calibrated. If either of these two conditions is not satisfied, one needs to carefully determine what information NRI offers and whether it retains meaningful interpretation. Using cut-offs that have no direct clinical meaning impedes the interpretation of the categorical NRI. Several methods have been proposed to define cut-points in situations where meaningful thresholds do not exist, but each has its own caveats. Presenting graphical displays similar to a decision curve [127] for a range of cut-offs could be considered as an alternative. There are few settings in which the continuous NRI can be recommended. These include instances where the primary focus is on the strength of the marker rather than model performance. Authors must be careful not to overinterpret the magnitude of the continuous NRI, which is usually much larger than in the case of the category-based NRI, and ascertain that the models are well-calibrated. Finally, for mathematical reasons, we recommend against calculating p-values for any of the forms of the NRI when evaluating the contribution of a new marker [128,129]. Instead, after a marker has been shown to be statistically significantly associated with the outcome, only confidence intervals for the NRI should be presented.

Our recommendations are meant to improve completeness, transparency, and clinical relevance of research involving risk reclassification. However, since the scientific debate on the NRI and related performance measures is still ongoing, our recommendations may very well be subjective to advances or additions in the future.

REFERENCES

1. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J, 3rd. Factors of risk in the development of coronary heart disease - six-year follow-up experience. The Framingham Study. Ann Intern Med. 1961;55:33-50.
2. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). JAMA. 2001;285(19):2486-97.
3. Perk J, De Backer G, Gohlke H, Graham I, Reiner Z, Verschuren WMM, *et al.* European Guidelines on cardiovascular disease prevention in clinical practice (version 2012). The Fifth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of nine societies and by invited experts). Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). Eur Heart J. 2012; 33(13):1635-701.
4. Hamm CW, Bassand JP, Agewall S, Bax JJ, Boersma E, Bueno H, *et al.* ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: The Task Force for the management of acute coronary syndromes (ACS) in patients presenting without persistent ST-segment elevation of the European Society of Cardiology (ESC). Eur Heart J. 2011;32(23):2999-3054.
5. Visvanathan K, Chlebowski RT, Hurley P, Col NF, Ropka M, Collyar D, *et al.* American Society of Clinical Oncology clinical practice guideline update on the use of pharmacologic interventions including tamoxifen, raloxifene, and aromatase inhibition for breast cancer risk reduction. J Clin Oncol. 2009;27(19):3235-58.
6. Munshi NC, Anderson KC, Bergsagel PL, Shaughnessy J, Palumbo A, Durie B, *et al.* Consensus recommendations for risk stratification in multiple myeloma: report of the International Myeloma Workshop Consensus Panel 2. Blood. 2011;117(18):4696-700.
7. Worth LJ, Lingaratnam S, Taylor A, Hayward AM, Morrissey S, Cooney J, *et al.* Use of risk stratification to guide ambulatory management of neutropenic fever. Australian Consensus Guidelines 2011 Steering Committee. Intern Med J. 2011;41(1b):82-9.
8. Bates SM, Jaeschke R, Stevens SM, Goodacre S, Wells PS, Stevenson MD, *et al.* Diagnosis of DVT: antithrombotic therapy and prevention of thrombosis, 9th ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest. 2012;141(2 Suppl): e351S-418S.
9. 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):882-90.
10. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007;115(7):928-35.

11. Pencina MJ, D'Agostino RB, Pencina KM, Janssens ACJW, Greenland P. Interpreting incremental value of markers added to risk prediction models. Am J Epidemiol. 2012;176(6): 473-81.

12. Austin PC, Steyerberg EW. Predictive accuracy of risk factors and markers: a simulation study of the effect of novel markers on different performance measures for logistic regression models. Stat Med. 2013;32(4):661-72.

13. Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MSV, *et al.* Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation. 2009;119(17):2408-16.

14. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. Ann Intern Med. 2006;145(1):21-9.

15. 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(2):157-72.

16. Tzoulaki I, Liberopoulos G, Ioannidis JPA. Use of reclassification for assessment of improved prediction: an empirical evaluation. Int J Epidemiol. 2011;40(4):1094-105.

17. Leening MJG, Steyerberg EW. Fibrosis and mortality in patients with dilated cardiomyopathy. JAMA. 2013;309(24):2547-8.

18. Lilford RJ, Richardson A, Stevens A, Fitzpatrick R, Edwards S, Rock F, *et al.* Issues in methodological research: perspectives from researchers and commissioners. Health Technol Assess. 2001;5(8):1-57.

19. Pencina MJ, D'Agostino RB, Vasan RS. Statistical methods for assessment of added usefulness of new biomarkers. Clin Chem Lab Med. 2010;48(12):1703-11.

20. 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.

21. Pencina MJ, D'Agostino RB, Sr., Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. Stat Med. 2012;31(2):101-13.

22. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. Ann Intern Med. 2009;150(11):795-W143.

23. Adabag AS, Therneau TM, Gersh BJ, Weston SA, Roger VL. Sudden death after myocardial infarction. JAMA. 2008;300(17):2022-2029.

24. Auer R, Bauer DC, Marques-Vidal P, Butler J, Min LJ, Cornuz J, *et al.* Association of major and minor ECG abnormalities with coronary heart disease events. JAMA. 2012;307(14): 1497-1505.

25. Breteler MMB. Mapping out biomarkers for Alzheimer disease. JAMA. 2011;305(3):304-305.

26. Buckley DI, Fu R, Freeman M, Rogers K, Helfand M. C-reactive protein as a risk factor for coronary heart disease: a systematic review and meta-analyses for the US Preventive Services Task Force. Ann Intern Med. 2009;151(7):483-W161.

27. Chou R, Arora B, Dana T, Fu R, Walker M, Humphrey L. Screening asymptomatic adults with resting or exercise electrocardiography: a review of the evidence for the US Preventive Services Task Force. Ann Intern Med. 2011;155(6):375-U74.

28. Cook NR. Biomarkers for prediction of cardiovascular events. JAMA. 2009;302(19):2089.

29. Cornelis MC, Qi L, Zhang C, Kraft P, Manson J, Cai T, *et al.* Joint effects of common genetic variants on the risk for type 2 diabetes in U. S. men and women of European ancestry. Ann Intern Med. 2009;150(8):541-W98.

30. de Boer IH, Levin G, Robinson-Cohen C, Biggs ML, Hoofnagle AN, Siscovick DS, *et al.* Serum 25-hydroxyvitamin D concentration and risk for major clinical disease events in a community-based population of older adults: a cohort study. Ann Intern Med. 2012;156(9): 627-U62.

31. de Lemos JA, Drazner MH, Omland T, Ayers CR, Khera A, Rohatgi A, *et al.* Association of troponin T detected with a highly sensitive assay and cardiac structure and mortality risk in the general population. JAMA. 2010;304(22):2503-2512.

32. deFilippi CR, de Lemos JA, Christenson RH, Gottdiener JS, Kop WJ, Zhan M, *et al.* Association of serial measures of cardiac troponin T using a sensitive assay with incident heart failure and cardiovascular mortality in older adults. JAMA. 2010;304(22):2494-2502.

33. den Ruijter HM, Peters SAE, Anderson TJ, Britton AR, Dekker JM, Eijkemans MJ, *et al.* Common carotid intima-media thickness measurements in cardiovascular risk prediction: a meta-analysis. JAMA. 2012;308(8):796-803.

34. Devereaux PJ, Chan MTV, Alonso-Coello P, Walsh M, Berwanger O, Villar JC, *et al.* Association between postoperative troponin levels and 30-day mortality among patients undergoing noncardiac surgery. JAMA. 2012;307(21):2295-2304.

35. Di Angelantonio E, Gao P, Pennells L, Kaptoge S, Caslake M, Thompson A, *et al.* Lipid-related markers and cardiovascular disease prediction. JAMA. 2012;307(23):2499-2506.

36. Eddy DM, Adler J, Patterson B, Lucas D, Smith KA, Morris M. Individualized guidelines: the potential for increasing quality and reducing costs. Ann Intern Med. 2011;154(9):627-U139.

37. 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(9867):639-650.

38. Fonarow GC, Pan W, Saver JL, Smith EE, Reeves MJ, Broderick JP, *et al.* Comparison of 30-day mortality models for profiling hospital performance in acute ischemic stroke with vs without adjustment for stroke severity. JAMA. 2012;308(3):257-264.

39. Gulati A, Jabbour A, Ismail TF, Guha K, Khwaja J, Raza S, *et al.* Association of fibrosis with mortality and sudden cardiac death in patients with nonischemic dilated cardiomyopathy. JAMA. 2013;309(9):896-908.

40. Helfand M, Buckley DI, Freeman M, Fu R, Rogers K, Fleming C, *et al.* Emerging risk factors for coronary heart disease: a summary of systematic reviews conducted for the US Preventive Services Task Force. Ann Intern Med. 2009;151(7):496-W164.

41. Hingorani AD, Psaty BM. Primary prevention of cardiovascular disease. Time to get more or less personal? JAMA. 2009;302(19):2144-2145.

42. Hlatky MA. Framework for evaluating novel risk markers. Ann Intern Med. 2012;156(6): 468-469.

43. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. Ann Intern Med. 2008;149(10):751-W162.

44. Janssens ACJW, Ioannidis JPA, van Duijn CM, Little J, Khoury MJ, for the GRIPS Group. Strengthening the reporting of genetic risk prediction studies: the GRIPS statement. Ann Intern Med. 2011;154(6):421-W141.

45. Kaptoge S, Di Angelantonio E, Lowe G, Pepys MB, Thompson SG, Collins R, *et al.* C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. Lancet. 2010;375(9709):132-140.

46. Kaptoge S, Di Angelantonio E, Pennells L, Wood AM, White IR, Gao P, *et al.* C-reactive protein, fibrinogen, and cardiovascular disease prediction. N Engl J Med. 2012;367(14):1310-1320.

47. Kathiresan S, Melander O, Anevski D, Guiducci C, Burtt NP, Roos C, *et al.* Polymorphisms associated with cholesterol and risk of cardiovascular events. N Engl J Med. 2008;358(12): 1240-1249.

48. Kavousi M, Elias-Smale SE, Rutten JHW, Leening MJG, Vliegenthart R, Verwoert GC, *et al.* Evaluation of newer risk markers for coronary heart disease risk classification: a cohort study. Ann Intern Med. 2012;156(6):438-U88.

49. Keller T, Zeller T, Ojeda F, Tzikas S, Lillpopp L, Sinning C, *et al.* Serial changes in highly sensitive troponin I assay and early diagnosis of myocardial infarction. JAMA. 2011;306(24): 2684-2693.

50. Kengne AP, Echouffo-Tcheugui JB, Sobngwi E. Coronary artery calcium for guiding statin treatment. Lancet. 2012;379(9813):312.

51. Khera AV, Cuchel M, de la Llera-Moya M, Rodrigues A, Burke MF, Jafri K, *et al.* Cholesterol efflux capacity, high-density lipoprotein function, and atherosclerosis. N Engl J Med. 2011; 364(2):127-135.

52. Kim WR, Biggins SW, Kremers WK, Wiesner RH, Kamath PS, Benson JT, *et al.* Hyponatremia and mortality among patients on the liver-transplant waiting list. N Engl J Med. 2008; 359(10):1018-1026.

53. Kivimäki M, Batty GD, Hamer M, Ferrie JE, Vahtera J, Virtanen M, *et al.* Using additional information on working hours to predict coronary heart disease. Ann Intern Med. 2011; 154(7):457-W153.

54. Koller MT, Leening MJG, Wolbers M, Steyerberg EW, Hunink MGM, Schoop R, *et al.* Development and validation of a coronary risk prediction model for older U.S. and European persons in the Cardiovascular Health Study and the Rotterdam Study. Ann Intern Med. 2012;157(6):389-97.

55. Lubitz SA, Yin X, Fontes JD, Magnani JW, Rienstra M, Pai M, *et al.* Association between familial atrial fibrillation and risk of new-onset atrial fibrillation. JAMA. 2010;304(20):2263-2269.

56. Lyssenko V, Jonsson A, Almgren P, Pulizzi N, Isomaa B, Tuomi T, *et al.* Clinical risk factors, DNA variants, and the development of type 2 diabetes. N Engl J Med. 2008;359(21):2220-2232.

57. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010;363(2):166-176.

58. Martinez ME, Thompson P, Messer K, Ashbeck EL, Lieberman DA, Baron JA, *et al.* One-year risk for advanced colorectal neoplasia: US versus UK risk-stratification guidelines. Ann Intern Med. 2012;157(12):856-U192.

59. Matsushita K, Mahmoodi BK, Woodward M, Emberson JR, Jafar TH, Jee SH, *et al.* Comparison of risk prediction using the CKD-EPI equation and the MDRD study equation for estimated glomerular filtration rate. JAMA. 2012;307(18):1941-1951.

60. McEvoy JW. Coronary artery calcium score and cardiovascular event prediction. JAMA. 2010;304(7):741-742.

61. Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, *et al.* Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med. 2008; 359(21):2208-2219.

62. Melander O, Newton-Cheh C, Almgren P, Hedblad B, Berglund G, Engstrom G, *et al.* Novel and conventional biomarkers for prediction of incident cardiovascular events in the community. JAMA. 2009;302(1):49-57.

63. Melander O, Newton-Cheh C, Wang TJ. Biomarkers for prediction of cardiovascular events - reply. JAMA. 2009;302(19):2090.

64. Omland T, de Lemos JA, Sabatine MS, Christophi CA, Rice MM, Jablonski KA, *et al.* A sensitive cardiac troponin T assay in stable coronary artery disease. N Engl J Med. 2009;361(26): 2538-2547.

65. Palomaki GE, Melillo S, Bradley LA. Association between 9p21 genomic markers and heart disease: a meta-analysis. JAMA. 2010;303(7):648-656.

66. Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, Ridker PM. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. Ann Intern Med. 2009;150(2):65-72.

67. Paynter NP, Chasman DI, Pare G, Buring JE, Cook NR, Miletich JP, *et al.* Association between a literature-based genetic risk score and cardiovascular events in women. JAMA. 2010; 303(7):631-637.

68. Peralta CA, Shlipak MG, Judd S, Cushman M, McClellan W, Zakai NA, *et al.* Detection of chronic kidney disease with creatinine, cystatin C, and urine albumin-to-creatinine ratio and association with progression to end-stage renal disease and mortality. JAMA. 2011; 305(15):1545-1552.

69. Pischon T, Boeing H, Hoffmann K, Bergmann M, Schulze MB, Overvad K, *et al.* General and abdominal adiposity and risk of death in Europe. N Engl J Med. 2008;359(20):2105-2120.

70. Pletcher MJ, Tice JA, Pignone M. Modeling cardiovascular disease prevention. JAMA. 2010; 303(9):835.

71. Polak JF, Pencina MJ, Pencina KM, O'Donnell CJ, Wolf PA, D'Agostino RB, Sr. Carotid-wall intima-media thickness and cardiovascular events. N Engl J Med. 2011;365(3):213-221.

72. Polonsky TS, McClelland RL, Jorgensen NW, Bild DE, Burke GL, Guerci AD, *et al.* Coronary artery calcium score and risk classification for coronary heart disease prediction. JAMA. 2010;303(16):1610-1616.

73. Ripatti S, Tikkanen E, Orho-Melander M, Havulinna AS, Silander K, Sharma A, *et al.* A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. Lancet. 2010;376(9750):1393-1400.

74. Rosenberg S, Elashoff MR, Beineke P, Daniels SE, Wingrove JA, Tingley WG, *et al.* Multicenter validation of the diagnostic accuracy of a blood-based gene expression test for assessing obstructive coronary artery disease in nondiabetic patients. Ann Intern Med. 2010;153(7): 425-W158.

75. Schelbert EB, Cao JJ, Sigurdsson S, Aspelund T, Kellman P, Aletras AH, *et al.* Prevalence and prognosis of unrecognized myocardial infarction determined by cardiac magnetic resonance in older adults. JAMA. 2012;308(9):890-897.

76. Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB, Sr., *et al.* Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. Lancet. 2009;373(9665):739-745.

77. Selvin E, Steffes MW, Zhu H, Matsushita K, Wagenknecht L, Pankow J, *et al.* Glycated hemoglobin, diabetes, and cardiovascular risk in nondiabetic adults. N Engl J Med. 2010;362(9): 800-811.

78. Steyerberg EW, Pencina MJ. Reclassification calculations for persons with incomplete follow-up. Ann Intern Med. 2010;152(3):195-196.

79. Tammemägi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, *et al.* Selection criteria for lung-cancer screening. N Engl J Med. 2013;368(8):728-736.

80. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, *et al.* A predictive model for progression of chronic kidney disease to kidney failure. JAMA. 2011;305(15): 1553-1559.

81. Tzoulaki I, Liberopoulos G, Ioannidis JPA. Assessment of claims of improved prediction beyond the Framingham risk score. JAMA. 2009;302(21):2345-2352.

82. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, *et al.* Performance of common genetic variants in breast-cancer risk models. N Engl J Med. 2010; 362(11):986-993.

83. Wilson PWF. Challenges to improve coronary heart disease risk assessment. JAMA. 2009; 302(21):2369-2370.

84. Wormser D, Kaptoge S, Di Angelantonio E, Wood AM, Pennells L, Thompson A, *et al.* Separate and combined associations of body-mass index and abdominal adiposity with cardiovascular disease: collaborative analysis of 58 prospective studies. Lancet. 2011;377(9771): 1085-1095.

85. Wormser D, Di Angelantonio E, Sattar N, Collins R, Thompson S, Danesh J, *et al.* Body-mass index, abdominal adiposity, and cardiovascular risk - reply. Lancet. 2011;378(9787):228.

86. Yeboah J, McClelland RL, Polonsky TS, Burke GL, Sibley CT, O'Leary D, *et al.* Comparison of novel risk markers for improvement in cardiovascular risk assessment in intermediate-risk individuals. JAMA. 2012;308(8):788-795.

87. Zethelius B, Berglund L, Sundström J, Ingelsson E, Basu S, Larsson A, *et al.* Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. N Engl J Med. 2008;358(20):2107-2116.

88. Zoungas S, Patel A, Chalmers J, de Galan BE, Li Q, Billot L, *et al.* Severe hypoglycemia and risks of vascular events and death. N Engl J Med. 2010;363(15):1410-1418.

89. Greenland P, Alpert JS, Beller GA, Benjamin EJ, Budoff MJ, Fayad ZA, *et al.* 2010 ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation. 2010;122(25):e584-636.

90. Mosca L, Benjamin EJ, Berra K, Bezanson JL, Dolor RJ, Lloyd-Jones DM, *et al.* Effectiveness-based guidelines for the prevention of cardiovascular disease in women--2011 update: a guideline from the American Heart Association. Circulation. 2011;123(11):1243-62.

91. Cai T, Tian L, Lloyd-Jones DM. Comparing costs associated with risk stratification rules for t-year survival. Biostatistics. 2011;12(4):597-609.

92. Leening MJG, Cook NR. Net reclassification improvement: a link between statistics and clinical practice. Eur J Epidemiol. 2013;28(1):21-3.

93. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. Am J Epidemiol. 2011;173(11):1327-35.

94. Mihaescu R, van Zitteren M, van Hoek M, Sijbrands EJ, Uitterlinden AG, Witteman JCM, *et al.* Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. Am J Epidemiol. 2010;172(3): 353-61.

95. Muhlenbruch K, Heraclides A, Steyerberg EW, Joost HG, Boeing H, Schulze MB. Assessing improvement in disease prediction using net reclassification improvement: impact of risk cut-offs and number of risk categories. Eur J Epidemiol. 2013;28(1):25-33.

96. Pepe MS, Janes H. Commentary: Reporting standards are needed for evaluations of risk reclassification. Int J Epidemiol. 2011;40(4):1106-8.

97. McGeechan K, Macaskill P, Irwig L, Liew G, Wong TY. Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. Arch Intern Med. 2008;168(21): 2304-10.

98. Greenland S. The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina *et al.* Stat Med. 2008;27(2):199-206.

99. Cook NR, Paynter NP. Performance of reclassification statistics in comparing risk prediction models. Biom J. 2011;53(2):237-58.

100. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) Final Report. Circulation. 2002;106(25):3143-421.

101. Cook NR, Paynter NP, Eaton CB, Manson JE, Martin LW, Robinson JG, *et al.* Comparison of the Framingham and Reynolds Risk Scores for global cardiovascular risk prediction in the multiethnic Women's Health Initiative. Circulation. 2012;125(14):1748-56, S1-11.

102. Pencina MJ, D'Agostino RB, Sr., Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the framingham heart study. Circulation. 2009;119(24): 3078-84.

103. Takahara M, Katakami N, Kaneto H, Shimomura I. Risk categorization for calculating net reclassification improvement. Eur J Epidemiol. 2013;28(7):607-9.

104. Ganna A, Reilly M, de Faire U, Pedersen N, Magnusson P, Ingelsson E. Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. Am J Epidemiol. 2012;175(7):715-724.

105. Pepe MS, Fan J, Seymour CW, Li C, Huang Y, Feng Z. Biases introduced by choosing controls to match risk factors of cases in biomarker research. Clin Chem. 2012;58(8):1242-51.

106. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology - 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

107. Kavousi M, Leening MJG, Witteman JCM. Markers for prediction of cardiovascular disease risk. JAMA. 2012;308(24):2561.

108. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Comments on 'Integrated discrimination and net reclassification improvements—Practical advice'. Stat Med. 2008; 27(2):207-12.

109. Pepe MS, Feng Z, Gu JW. Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina *et al.* Stat Med. 2008;27(2):173-81.

110. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. Stat Med. 2013:Epub ahead of print.

111. Brindle P, Emberson J, Lampe F, Walker M, Whincup P, Fahey T, *et al.* Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study. BMJ. 2003; 327(7426):1267.

112. Hense HW, Schulte H, Lowel H, Assmann G, Keil U. Framingham risk function overestimates risk of coronary heart disease in men and women from Germany--results from the MONICA Augsburg and the PROCAM cohorts. Eur Heart J. 2003;24(10):937-45.

113. Koller MT, Steyerberg EW, Wolbers M, Stijnen T, Bucher HC, Hunink MGM, *et al.* Validity of the Framingham point scores in the elderly: results from the Rotterdam study. Am Heart J. 2007;154(1):87-93.

114. Merry AHH, Boer JMA, Schouten LJ, Ambergen T, Steyerberg EW, Feskens EJM, *et al.* Risk prediction of incident coronary heart disease in the Netherlands: re-estimation and improvement of the SCORE risk function. Eur J Prev Cardiol. 2012;19(4):840-8.

115. Siontis GCM, Tzoulaki I, Siontis KC, Ioannidis JPA. Comparisons of established risk prediction models for cardiovascular disease: systematic review. BMJ. 2012;344:e3318.

116. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2009.

117. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. Crit Care Med. 2007;35(9):2052-6.

118. Cox DR. Two further applications of a model for binary regression. Biometrika. 1958;45(3-4):562-565.

119. Cook NR, Paynter NP. Comments on 'Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers' by M. J. Pencina, R. B. D'Agostino, Sr. and E. W. Steyerberg. Stat Med. 2012;31(1):93-5; author reply 96-7.

120. Vergouwe Y, Moons KGM, 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):971-80.

121. Cook NR. Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina *et al.* Stat Med. 2008; 27(2):191-5.

122. Paynter NP, Cook NR. A bias-corrected net reclassification improvement for clinical subgroups. Med Decis Making. 2013;33(2):154-62.

123. Vickers AJ, Elkin EB, Steyerberg EW. Net reclassification improvement and decision theory. Stat Med. 2009;28(3):525-6; author reply 526-8.

124. 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.

125. Localio AR, Goodman S. Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. Ann Intern Med. 2012;157(4):294-5.

126. 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(1):128-38.

127. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-74.

128. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. BMC Med Res Methodol. 2011;11:13.

129. Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. Stat Med. 2013;32(9):1467-82.

SUPPLEMENTARY TABLES AND FIGURES

Supplementary table 2.1 List and main characteristics of the 67 articles

First author	Year	Article type	Marker / Comparison	Outcome of interest / Topic	Cut-offs used for NRI
Adabag, A.S. (23)	2008	original article	-	sudden death after MI	NA *
Auer, R. (24)	2012	original article	ECG abnormalities	CHD	7.5% and 15% at 7.5 years
Breteler, M.M.B. (25)	2011	editorial	-	dementia	NA *
Buckley, D.I. (26)	2009	meta-analysis	CRP	CHD	NA *
Chou, R. (27)	2011	review	resting or exercise ECG	CVD	NA *
Cook, N.R. (22)	2009	methods	CRP	CVD	5%, 10%, and 20% at 10 years †
Cook, N.R. (28)	2009	letter	CRP	CVD	NA *
Cornelis, M.C. (29)	2009	original article	genetic risk score	type 2 DM	NA *
de Boer, I.H. (30)	2012	original article	25(-OH) vitamin D	composite of hip fracture, MI, cancer, and death	50 nmol/L vs. season specific using a 10 year horizon
de Lemos, J.A. (31)	2010	original article	troponin T	cardiac structure and mortality	NA *
deFilippi, C.R. (32)	2010	original article	troponin T	heart failure and CVD mortality	10% and 20% at 10 years
den Ruijter, H.M.	2012	original article	cIMT	MI and stroke	5%, (10%), and 20% at 10 years
Devereaux, P.J. (34)	2012	original article	troponin T	mortality after noncardiac surgery	1%, 5%, and 10% at 30 days
Di Angelantonio, E. (35)	2012	original article	cholesterol, apolipoprotein, and Lp(a)	CVD	10% and 20% at 10 years
Eddy, D.M. (36)	2011	original article	hypertension guidelines	MI and stroke	not specified
Farooq, V. (37)	2013	original article	coronary revascularization strategies	mortality	NA *
Fonarow, G.C. (38)	2012	original article	NIH stroke scale	stroke fatality	not specified using a 30 days horizon
Gulati, A. (39)	2013	original article	myocardial fibrosis	mortality major arrhythmia	5%, 10%, and 20% at 5 years 15% at 5 years

Supplementary table 2.1 List and main characteristics of the 67 articles (continued)

First author	Year	Article type	Marker / Comparison	Outcome of interest / Topic	Cut-offs used for NRI
Helfand, M. (40)	2009	review	CRP, CAC score, Lp(a), homocysteine, leukocyte count, fasting glucose periodontal disease, ABI, and cIMT	CHD	NA *
Hingorani, A.D. (41)	2009	commentary	-	CVD	NA *
Hlatky, M.A. (42)	2012	editorial	-	CHD	NA *
Janes, H. (43)	2008	methods	-	risk stratification tables	NA *
Janssens, A.C.J.W. (44)	2011	methods	-	GRIPS statement	NA *
Kaptoge, S. (45)	2010	original article	CRP	CHD, stroke, and mortality	NA *
Kaptoge, S. (46)	2012	original article	CRP and fibrinogen	CVD	10% and 20% at 10 years
Kathiresan, S. (47)	2008	original article	genetic risk score	CVD	10% and 20% at 10 years
Kavousi, M. (48)	2012	original article	NT-proBNP, vWF, fibrinogen, CKD, leukocyte count, CRP, homocysteine, uric acid, CAC score, cIMT, PAD, and PWV	CHD	10% and 20% at 10 years †
Keller, T. (49)	2011	original article	serial changes in troponin I	MI	NA *
Kengne, A.P. (50)	2012	letter	CRP, CAC score	CVD	NA *
Khera, A.V. (51)	2011	original article	cholesterol efflux capacity	obstructive CAD	NA *
Kim, W.R. (52)	2008	original article	hyponatremia	mortality in ESLD	NA *
Kivimäki, M. (53)	2011	original article	working hours	CHD	5% and 10% at 10 years
Koller, M.T. (54)	2012	original article	BMI, CRP, cIMT, ABI, and ECG-LVH	CHD	10% and 20% at 10 years
Lubitz, S.A. (55)	2010	original article	familial atrial fibrillation	atrial fibrillation	5% and 10% at 8 years
Lyssenko, V. (56)	2008	original article	genetic polymorphisms	type 2 DM	10% and 20% at an unspecified horizon
Manolio, T.A. (57)	2010	review	-	genetic risk prediction	NA *

Supplementary table 2.1 List and main characteristics of the 67 articles (continued)

First author	Year	Article type	Marker / Comparison	Outcome of interest / Topic	Cut-offs used for NRI
Martinez, M.E. (58)	2012	original article	U.K. and U.S. guidelines	advanced colorectal dysplasia	number, type, and size of adenomas using a 1 year horizon
Matsushita, K. (59)	2012	original article	CKD-EPI and MDRD equations	mortality and ESRD	eGFR of 90, 60, 45, 30, and 15 mL/min/1.73m^2 using an unspecified horizon
McEvoy, J.W. (60)	2010	letter	CAC score	CHD	NA *
Meigs, J.B. (61)	2008	original article	genetic risk score	type 2 DM	2% and 8% at 8 to 10 years
Melander, O. (62)	2009	original article	CRP, cystatin C, Lp-PLA2, MR-proADM, MR-proANP, and NT-proBNP	CHD and CVD	6%, 10%, and 20% at 10 years
Melander, O. (63)	2009	letter reply	CRP	CVD	NA *
Omland, T. (64)	2009	original article	troponin T	CVD mortality, heart failure, and MI	NA *
Palomaki, G.E. (65)	2010	meta-analysis	chromosome 9p21 polymorphisms	CHD	5%, 10%, and 20% at 10 years †
Paynter, N.P. (66) ‡	2009	original article	chromosome 9p21.3 polymorphisms	CVD	5%, 10%, and 20% at 10 years †
Paynter, N.P. (67)	2010	original article	genetic risk score	CVD	5%, 10%, and 20% at 10 years
Peralta, C.A. (68)	2011	original article	creatinine, cystatin C, and urine albumin-to-creatinine ratio	mortality and ESRD	continuous NRI only at an unspecified horizon
Pischon, T. (69)	2008	original article	BMI and abdominal adiposity	mortality	2.5%, 5%, and 7.5% at 5 years
Pletcher, M.J. (70)	2010	letter	-	CVD	NA *
Polak, J.F. (71)	2011	original article	cIMT	CVD	6% and 20% at 10 years †
Polonsky, T.S. (72)	2010	original article	CAC score	CHD	3% and 10% at 5 years
Ripatti, S. (73)	2010	original article	genetic risk score	CHD	5%, 10%, and 20% at 10 years
Rosenberg, S. (74)	2010	original article	gene expression test	presence of obstructive CAD	20 and 50%

Supplementary table 2.1 List and main characteristics of the 67 articles (continued)

First author	Year	Article type	Marker / Comparison	Outcome of interest / Topic	Cut-offs used for NRI
Schelbert, E.B. (75)	2012	original article	unrecognized MI	mortality	continuous NRI only at an unspecified horizon
Schnabel, R.B. (76)	2009	original article	echocardiographic measurements	atrial fibrillation	5% and 15% at 10 years
Selvin, E. (77)	2010	original article	glycated hemoglobin	type 2 DM, CHD, and mortality	5%, 10%, and 20% at 10 years
Steyerberg, E.W. (78)	2010	letter	CRP	CVD	5%, 10%, and 20% at 10 years †
Tammemägi, M.C. (79)	2013	original article	smoking intensity and history of cancer	lung-cancer	1% and 2% at 6 years
Tangri, N. (80)	2011	original article	calcium, phosphate, bicarbonate, and albumin	CKD	not specified
Tzoulaki, I. (81)	2009	review	86 different predictors	CHD	NA *
Wacholder, S. (82)	2010	original article	genetic polymorphisms	breast cancer	NA *
Wilson, P.W.F. (83)	2009	editorial	-	CHD	NA *
Wormser, D. (84)	2011	original article	BMI and abdominal adiposity	CHD and stroke	5%, 10%, and 20% at 10 years
Wormser, D. (85)	2011	letter reply	BMI and abdominal adiposity	CHD	NA *
Yeboah, J. (86)	2012	original article	cIMT, CAC score, brachial FMD, ABI, CRP, and family history	CHD and CVD	5% and 20% at 10 years †
Zethelius, B. (87)	2008	original article	troponin I, NT-proBNP, cystatin C, and CRP	CVD mortality	6% and 20% at an unspecified horizon
Zoungas, S. (88)	2010	original article	severe hypoglycemia	CVD	NA *

* Not applicable: NRI was not calculated.

† Using observations from a follow-up period shorter than the predicted time horizon.

‡ Identified through hand search with erroneous citation linkage to a methodological article on NRI.

ABI = ankle-brachial index; BMI = body mass index; CAC = coronary artery calcium; CAD = coronary artery disease; CHD = coronary heart disease; cIMT = carotid intima-media thickness; CKD = chronic kidney disease; CKD-EPI = Chronic Kidney Disease Epidemiology Collaboration; CRP = C-reactive protein; CVD = cardiovascular disease; DM = diabetes mellitus; ECG = electrocardiography; ECG-LVH = electrocardiographic left ventricular hypertrophy; eGFR = estimated glomerular filtration rate; ESLD = end-stage liver disease; ESRD = end-stage renal disease; FMD = flow-mediated dilation; Lp(a) = lipoprotein(a); Lp-PLA2 Lipoprotein-associated phospholipase 2; MI = myocardial infarction; MDRD = Modification of Diet in Renal Disease; MR-proADM = midregional proadrenomedullin; MR-proANP = midregional proatrial natriuretic peptide; NA = not applicable; NRI = net reclassification improvement; NT-proBNP = N-terminal pro-B-type natriuretic peptide; PAD = peripheral arterial disease; PWV = pulse wave velocity; SBP = systolic blood pressure; SCD = sudden cardiac death; vWF = von Willebrand factor.

Supplementary table 2.2 Summary characteristics of the 67 articles

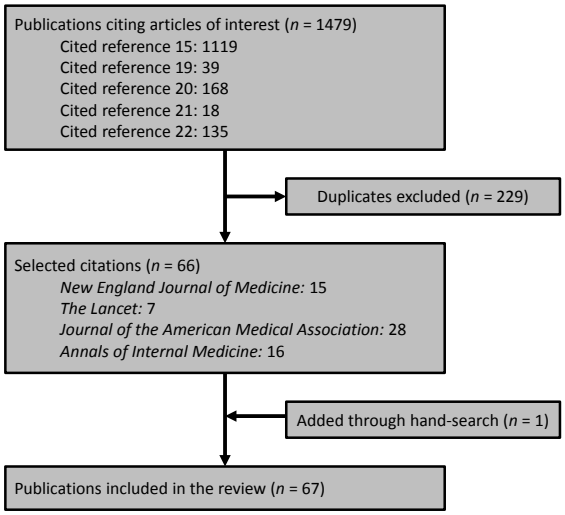
Study characteristic	Studies, n (%)
Journal	
New England Journal of Medicine	15 (22.4)
Lancet	7 (10.4)
Journal of the American Medical Association	28 (41.8)
Annals of Internal Medicine	17 (25.4)
Year of print publication	
2008	8 (11.9)
2009	13 (19.4)
2010	16 (23.9)
2011	12 (17.9)
2012	15 (22.4)
2013	3 (4.5)
Citing methodologic article	
Pencina *et al.* 2008 (15)	56 (83.6)
Pencina *et al.* 2010 (19)	3 (4.5)
Pencina *et al.* 2011 (20)	9 (13.4)
Pencina *et al.* 2012 (21)	0 (0.0)
Cook and Ridker 2009 (22)	11 (16.4)
Country (of address for correspondence)	
Australia	1 (1.5)
Canada	2 (3.0)
Finland	1 (1.5)
Germany	2 (3.0)
Greece	1 (1.5)
The Netherlands	6 (9.0)
Norway	1 (1.5)
South Africa	1 (1.6)
Sweden	4 (6.0)
Switzerland	1 (1.5)
U.S.	39 (58.2)
U.K.	8 (11.9)

Supplementary table 2.3 Example of a risk reclassification table stratified by event status

Model containing only Framingham risk score variables	Model containing Framingham risk score variables and coronary artery calcification score						
	<10%		**10%-20%**		**>20%**		**Total**
<10%							
Persons with event, *n*	71		50		4		125
Persons without event, *n*	2015		315		16		2346
Total persons, *n*	2086		365		20		2471
Observed risk (95% CI)	3%	(2%-5%)	14%	(10%-19%)	21%	(6%-60%)	
10%-20%							
Persons with event, *n*	19		75		55		149
Persons without event, *n*	262		364		144		770
Total persons, *n*	281		439		199		919
Observed risk (95% CI)	7%	(4%-12%)	17%	(13%-22%)	28%	(20%-37%)	
>20%							
Persons with event, *n*	0		9		62		71
Persons without event, *n*	17		60		140		217
Total persons, *n*	17		69		202		288
Observed risk (95% CI)	0%	(0%-0%)	13%	(6%-27%)	31%	(23%-40%)	
Total							
Persons with event, *n*	90		134		121		345
Persons without event, *n*	2294		739		300		3333
Total persons, *n*	2384		873		421		3678

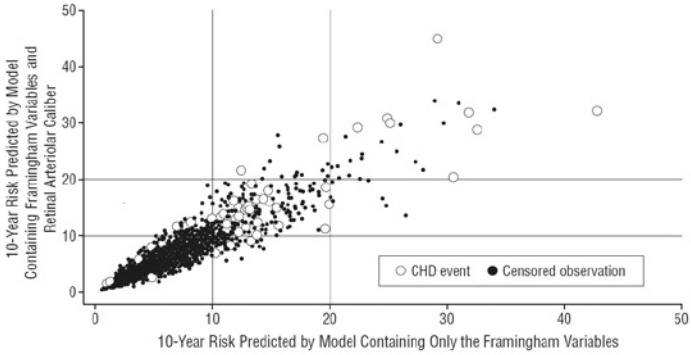
Risk reclassification for 10-year risk of incident coronary heart disease in participants from the Rotterdam Study predicted by a model containing only the Framingham risk score variables against risk predicted by a model containing Framingham risk score variables and coronary artery calcification score. The numbers of persons in the cells are rounded due to the use of Kaplan-Meier estimates for persons with incomplete follow-up (original publication:(48)). NA = not applicable; CI = confidence interval.

Supplementary figure 2.1 Flowchart for selection of the included articles



The search was last updated April 23, 2013

Supplementary figure 2.1 Example of a reclassification graph with superimposed cut-points of predicted risk



10-year risk of incident coronary heart disease in women from the Atherosclerosis Risk In Communities (ARIC) Study, predicted by a model containing only the Framingham risk score variables (horizontal axis) against risk predicted by a model containing Framingham risk score variables and retinal arteriolar caliber (vertical axis) (original publication: (97). Lines at 10% and 20% predicted 10-year risk are superimposed to show reclassification over clinically relevant cut-points (2,89) and thereby create a visual representation of a reclassification table (e.g. supplementary table 2.3). It can be noted that the vast majority of women in this study have a low (<10%) predicted risk of coronary heart disease, both with the Framingham variables and using the model extended with retinal arteriolar caliber. The graph also shows that only a limited amount of women is reclassified over the cut-points (i.e. only a small proportion of dots lies in the off-diagonal cells of the graph). CHD = coronary heart disease.

Chapter 3

Graphical assessment of incremental value of novel markers in prediction models: From statistical to decision analytical perspectives

Steyerberg EW, Vedder MM, Leening MJ, Postmus D, D'Agostino RB, Sr., Van Calster B, Pencina MJ

Biom J. 2015 Jul;57(4):556-70.

ABSTRACT

Background: New markers may improve prediction of diagnostic and prognostic outcomes. We aimed to review various options for graphical display and summary measures to assess the predictive value of markers over standard, readily available predictors.

Methods: We illustrated various approaches using previously published data on 3264 participants from the Framingham Heart Study, where 183 developed coronary heart disease (10-year risk 5.6%). We considered performance measures for the incremental value of adding HDL cholesterol to a prediction model.

Results: An initial assessment may consider statistical significance (HR=0.65, 95% confidence interval 0.53 to 0.80; Likelihood Ratio p<0.001), and distributions of predicted risks (densities or box plots) with various summary measures. A range of decision thresholds is considered in predictiveness and receiver operating characteristic curves, where the area under the curve (AUC) increased from 0.762 to 0.774 by adding HDL. We can furthermore focus on reclassification of participants with and without an event in a reclassification graph, with the continuous Net Reclassification Improvement (NRI) as a summary measure.

When we focus on one particular decision threshold, the changes in sensitivity and specificity are central. We propose a Net Reclassification Risk graph, which allows us to focus on the number of reclassified persons and their event rates. Summary measures include the binary AUC, the two-category NRI, and decision analytic variants such as the Net Benefit (NB).

Conclusions: Various graphs and summary measures can be used to assess the incremental predictive value of a marker. Important insights for impact on decision making are provided by a simple graph for the Net Reclassification Risk.

3.1 INTRODUCTION

Risk prediction models, or clinical prediction models, have been successfully developed to aid clinicians in personalized decision making in all major fields of modern medicine, including cardiovascular disease, diabetes, trauma, and cancer (Steyerberg *et al.*, 2013). Prediction models most often consider binary events, which are already present in a patient (disease, i.e. diagnosis), or occur in the future (events, i.e. prognosis). Probabilistic models for such outcomes may be evaluated with various performance measures, commonly related to discrimination (i.e. the ability to distinguish subjects with the event of interest from those without) and calibration (i.e. the agreement between predicted and observed probabilities of the event) (Harrell, 2001) (Steyerberg, 2009).

Nowadays, specific interest focuses on ways in which risk prediction can be improved using novel markers (Pencina *et al.*, 2008) (Pencina *et al.*, 2010) (Pencina *et al.*, 2011) (Pencina *et al.*, 2012b). Markers may include information from simple demographics or blood markers to genomics, proteomics, and advanced imaging techniques. Such markers hold the promise of bringing personalized medicine closer. An important question is how to evaluate the usefulness of a new marker in making better decisions in clinical practice, such as better targeting of statin therapy to those at increased risk of cardiovascular disease (Ridker *et al.*, 2007). Various measures have been proposed to quantify usefulness, including the Net Reclassification Improvement (NRI) (Pencina *et al.*, 2008), and decision analytic variants such as the Net Benefit (NB) (Vickers & Elkin, 2006) and Relative Utility (RU) (Baker *et al.*, 2009). The latter measures have important theoretical advantages (Van Calster *et al.*, 2013). They are directly translate to patient-centered outcomes, but are more difficult to communicate to a clinical audience (Localio & Goodman, 2012) (Leening & Steyerberg, 2013). Graphical displays may be of assistance in assessing usefulness of a marker for predictive purposes, since graphs allow for direct pattern recognition in addition to a table look-up function (Cleveland, 1985). We illustrate various graphical possibilities and summary measures using previously published data from the Framingham Heart Study, where we focus on adding high-density lipoprotein (HDL) cholesterol as a marker to improve predictions of 10-year risk of coronary heart disease (CHD). After describing the data and the statistical analysis, we consider performance measures related to continuous predictions and to dichotomizations to provide a classification as low versus high risk.

3.2 DATA

We provide an illustration with data published previously (Pencina *et al.*, 2008), relating to 3264 participants from the Framingham Heart Study aged 30–74. Participants with prevalent cardiovascular disease and missing standard risk factors were excluded. (Pencina *et al.*, 2008). Participants were followed for 10 years for the development of a first CHD event (including myocardial infarction, angina pectoris, coronary insufficiency, or CHD death). A total of 183 individuals developed CHD (5.6% 10-year cumulative incidence).

3.3 ANALYSIS

We focus on the improvement in model performance due to the addition of HDL cholesterol to a model that already contains sex, diabetes and smoking as dichotomous predictors and age, systolic blood pressure, and total cholesterol as continuous predictors. Adding HDL cholesterol as a continuous predictor to a Cox regression model was highly significant (HR=0.65 per SD increase, 95% confidence interval 0.53 to 0.80, p-value <0.001). We compare 2 sets of predicted probabilities of the 10-year CHD risk: one set of predictions based on a model without and one set of predictions based on a model with HDL cholesterol included. We consider the full risk distributions as well as categorization by the clinically motivated threshold of 20% 10-year CHD risk. (Greenland *et al.*, 2010). Since we focus on conceptual issues for binary classification, we simply stratify subjects as those with events versus those without events after 10-year follow-up. We recognize that survival data may require more sophisticated approaches to address censoring of incomplete observations, such as the use of Kaplan-Meier or Cox regression analysis to estimate expected events rather than naively use observed events (Steyerberg & Pencina, 2010) (Pencina *et al.*, 2011).

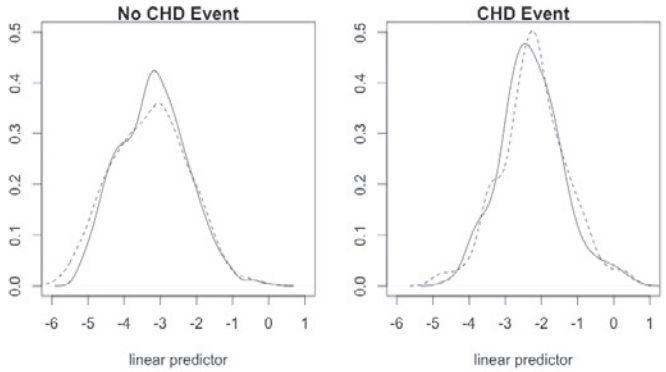
3.4 PERFORMANCE MEASURES RELATED TO CONTINUOUS PREDICTIONS

Densities and box plots with their summary measures

A better prediction model will provide predicted risks closer to the observed outcome: higher predicted risks for those with an event, and lower predicted risks for those without an event. Potential graphical illustrations include a density plot (figure 3.1) or a box plot (figure 3.2), where we can compare the predicted risks from models with and without the marker (Royston & Altman, 2010). In both plots, we may choose to show predicted values transformed to the linear predictor scale (log hazard for survival, log

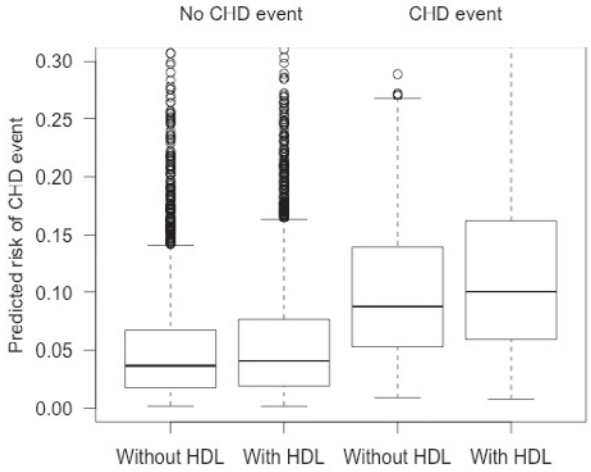
odds for binary predictions), or at the absolute risk scale (as a percentage). In both plots, we hope to see more separation in predicted risks, i.e. lower risk predictions for those without and higher risk predictions for those with events.

Figure 3.1 Density plots of logodds of predicted 10-year CHD risks ('linear predictor') for non-events and events in 3264 participants from the Framingham Heart Study.



Solid line: distribution of linear predictor without HDL cholesterol; dotted line: distribution of linear predictor with HDL cholesterol.

Figure 3.2 Box plots of predicted 10-year CHD risks for patients without and with a CHD event in 3264 participants from the Framingham Heart Study for the model without HDL and with HDL.



A density plot may most naturally use the linear predictor scale, since a reasonably Normal distribution is commonly noted at this scale (figure 3.1). A density plot for the linear predictors relates to summary measures that consider the log-based distances of predictions to observed outcomes. One such summary measure is the explained variation as defined by Nagelkerke, where the log-likelihood of a model is scaled between 0 and 100%:

$$R^2 = (1 - \exp(-LR/n)) / (1 - \exp(- -2LL_0 /n)),$$

where LR is the Likelihood Ratio statistic of the model, $-2LL_0$ is the -2 log likelihood of the Null model without any covariates, and n is the number of subjects (Nagelkerke, 1991). In our example, Nagelkerke's R^2 increased from 13.1% to 14.7% or +1.6%.

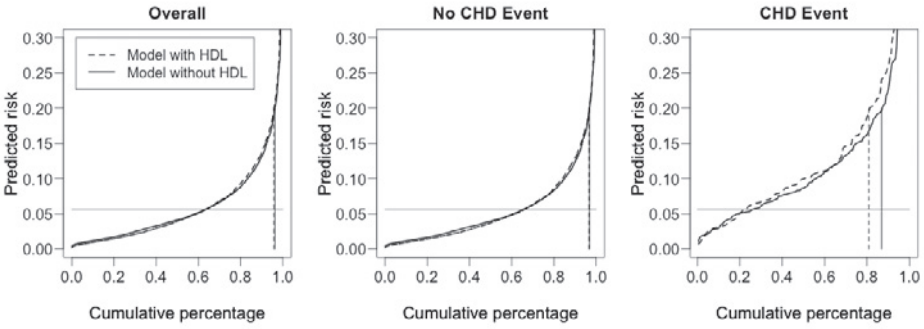
For box plots, the use of absolute risks has been proposed before. Box plots are attractive as a visual companion to measures such as the discrimination slope, which is defined as the difference in mean predicted risks for those with and without the event. The discrimination slope increased from 6.29% to 7.14% (+0.85%), which is identical to another relatively novel measure: the Integrated Discrimination Index (IDI, +0.0085) (Pencina *et al.*, 2008). The difference in Pearson's R<sup>2 (defined simply by comparing predicted risks to observed outcomes) was +1.9% and hence larger than the IDI, although these estimates should be asymptotically equivalent (Tjur, 2009).

Predictiveness curves

The absolute risk can also be shown as a cumulative distribution in a predictiveness curve, i.e. based on the ordering of risk from lowest to highest values (Pepe *et al.*, 2008). Better predictiveness is indicated by more spread in predictions, with many low predictions and a steep increase in the cumulative distribution of predictions at 1 minus the event rate. The event rate is useful as a reference line, with non-predictive model predictions being close to that line. In figure 3.3, we note that the distribution of the risks based on the model with HDL is only slightly more extreme than the distribution based on the model without HDL, consistent with the minor increase in R<sup>2 statistics as noted above. The performance of the models in the population is illustrated with a 20% risk threshold, where we note that both models would classify approximately 96% as low risk, and 4% as high risk.

As for figure 3.1 and figure 3.2, we can stratify the predictiveness curve by event status. The specificity is related to the cumulative distribution of risks for those with no CHD event. Specificities were very similar in figure 3.3, but the sensitivity of predictions from a model with HDL was higher than from a model without, consistent with figure 3.1 and figure 3.2.

Figure 3.3 Predictiveness curves for CHD events in 3264 participants from the Framingham Heart Study.

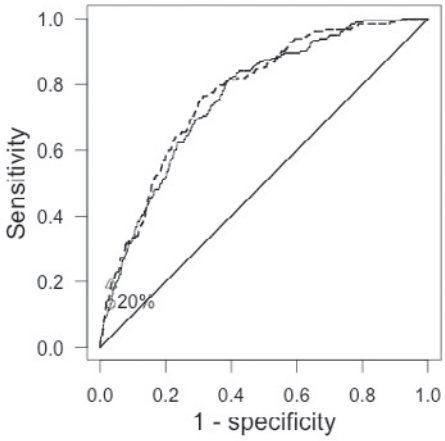


The horizontal lines indicate the event rate (5.6%). The vertical lines indicate the 20% decision threshold, which leads to similar specificity (96.82% and 96.66%) but higher sensitivity (13.1% and 19.1%) for the models without and with HDL cholesterol respectively.

ROC curves and area under the curve

The receiver operating characteristic (ROC) curve shows the relation between sensitivity (or the true positive rate, i.e. the fraction with a predicted probability above a cut-off among those developing CHD), versus 1 minus specificity (or the false positive rate, i.e. the fraction with a predicted probability above a cut-off among those not developing CHD, figure 3.4). The sensitivity and specificity pairs are calculated for consecutive cut-offs for the predicted probabilities of the 10-year CHD risk. The area under the ROC curve (AUC) is the most popular metric to quantify discriminative ability, i.e. the ability to distinguish those who will develop the event of interest from those who will not. The AUC is the probability that given two subjects (one with CHD and one without CHD), the model will assign a higher probability of CHD to the former. The AUC is a rank order statistic, and when estimated non-parametrically, related to the Mann-Whitney U statistic: AUC = U / (n1 * n2), where n1 and n2 are the numbers of persons with and without events (for which the product equals the number of possible pairs) (Hanley & McNeil, 1982). The AUC [95% confidence interval] for the model without vs with HDL was 0.762 [0.730 – 0.794] vs 0.774 [0.742 – 0.806]. We note that plotting ROC curves makes sense only when clinically relevant thresholds are indicated, at which sensitivity and specificity can be determined from the graph. The link between threshold and sensitivity/specificity is immediately clear in the predictiveness curve with stratification by event status (figure 3.3, middle and right panels). Furthermore, it should be noted that delta AUC depends strongly on the value of the reference model without the marker (Pencina *et al.*, 2012a) (Austin & Steyerberg, 2013).

Figure 3.4 Receiver operating characteristic (ROC) curves.



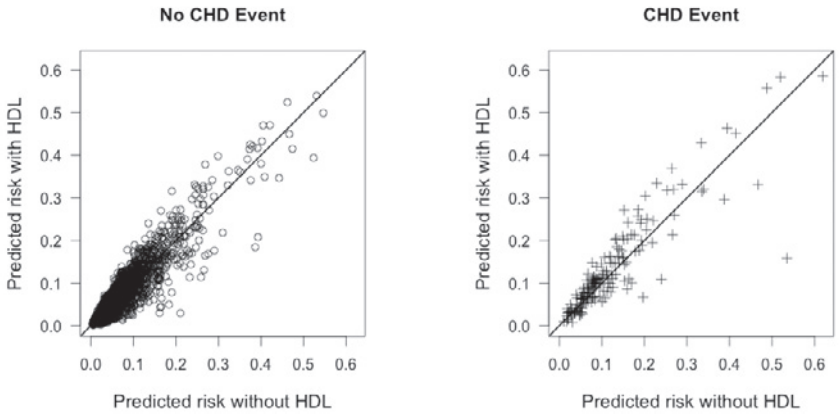
Solid lines: based on 10-year CHD predictions without HDL cholesterol; dotted lines: based on 10-year CHD predictions with HDL cholesterol in 3264 participants from the Framingham Heart Study. The 20% decision threshold is indicated with a triangle and a circle for the model with and without HDL cholesterol respectively.

Reclassification plots and the Net Reclassification Improvement

A simple and informative graphical summary can be obtained by plotting the predicted probabilities based on the model with versus that without the new marker(s) with different symbols denoting subjects with and without events (figure 3.5) (McGeechan *et al.*, 2008). If the new marker offers no improvement, the points will scatter around the diagonal line. If the new model is useful, the predicted probabilities for events will be larger using the new model and hence points denoting events will lie above the diagonal line. Similarly, predicted probabilities for non-events will be smaller using the new model and hence points denoting non-events will lie below the diagonal line. The extent to which a clear separation can be seen helps determine the degree of model improvement. A limitation is that for large data sets, the graphical impression may be difficult to notice with many overlapping points (McGeechan *et al.*, 2008).

This graphical presentation can be summarized using the continuous net reclassification improvement (NRI (>0)) (Pencina *et al.*, 2011). In figure 3.5, 62.3% among those with events had a predicted risk with HDL that was higher than the predicted risk without HDL, i.e. a better risk estimate. Conversely, 37.7% had a predicted risk with HDL that was lower than the predicted risk without HDL. The net proportion with better risk estimates (or event NRI (>0)) hence was 24.6%. Similarly, we note that 52.8% of those without an event had a lower prediction with HDL than without, while 47.2% had

Figure 3.5 Reclassification graphs for the addition of HDL cholesterol to 10-year CHD predictions in 3264 participants from the Framingham Heart Study.



Classification with a 20% risk cut-off is indicated with dotted lines.

a higher prediction (non-event NRI (>0)=5.5%). The total continuous NRI (>0) equals the sum of the two components above, 0.246+0.055 = 0.301.

It has been noted that NRI (>0) can be viewed as a measure of effect size of the new predictor in the context of a risk prediction model rather than a difference in performance of two models (Pencina *et al.*, 2012a). A feature of NRI (>0) is that it is only weakly related to the performance of the reference model. Under normality, in case of an independent predictor added to the risk prediction model, it is the same no matter how good or bad the reference model is (Pencina *et al.*, 2012a). Finally, it is important to note that the NRI (>0) considers any change in predicted risk, irrespective of the magnitude. The size of the risk difference is considered in measures such as the IDI (Pencina *et al.*, 2008). An overview of the discussed performance measures and options for display is shown in table 3.1.

3.5 PERFORMANCE MEASURES RELATED TO BINARY CLASSIFICATION

Sensitivity and specificity

When we focus on one particular decision threshold, such as 20% risk of CHD, the changes in sensitivity and specificity are central to quantify the incremental value of a marker (Van Calster *et al.*, 2013). As a first option for graphical display, we indicated the 20% risk threshold in the predictiveness and ROC curves (figure 3.3 and figure 3.4). At the 20% threshold, the sensitivity increased from 13.1% to 19.1%, and specificity

Table 3.1 Overview of performance measures with display options and summary measures of performance for predicting 10-year CHD risks in 3264 participants from the Framingham Heart Study. We first consider continuous predictions, followed by dichotomized classifications. Performance relates to the difference between a prediction model with and without HDL.

Performance measure	Display option	Performance	Comments
Continuous predictions			
Δ log-likelihood ΔNagelkerke R^2 Δlp events Δlp non-events	Fig 3.1: density plot	+19 +1.6% +0.058 - 0.084	We note a statistically significant improvement in model fit; 1.6% more variability is explained at the log-likelihood scale
Δpredicted risk events Δpredicted risk non-events Δdiscrimination slope (= IDI) ΔPearson R^2	Fig 3.2: Box plot Fig 3.3: Predictiveness curves	+0.81% –0.04% +0.85% +1.9%	Those with events receive higher predicted risks and may hence better be identified; 1.9% more variability is explained at a squared distance scale for y vs ŷ.
ΔAUC	Fig 3.4: ROC curve	0.012	The probability of correctly identifying who will develop a CHD event increases by 1.2% among a random pair of participants where 1 has an event and 1 has no event.
continuous NRI events continuous NRI non-events continuous NRI	Fig 3.5: Reclassification graph	24.6% 5.5% 0.301	A net 24.6% of those with events receive higher predicted risks, and a net 5.5% of non-events receive lower predicted risks. Their sum is 0.301.
Dichotomized classification			
ΔAUC	Fig 3.4: ROC curve	0.029	The probability of correctly identifying who will develop a CHD event increases by 2.9% among a random pair of participants where 1 has an event and 1 has no event, if a decision threshold of 20% 10-year CHD risk is applied.
NRI events NRI non-events NRI	Table 3.2 and Table 3.3: reclassification tables Fig 3.3: Predictiveness curve Fig 3.6: Net Reclassification Risk graph	6.0% -0.2% 0.058	A net 6.0% increase in high risk classifications for those with events, with a minor net decrease (-0.2%) in low risk classifications for those without events. The sum is the NRI (0.058), which is twice the increase in AUC for this binary classification.
Net Benefit Test threshold	Table 3.2 and Table 3.3: reclassification table Fig 3.7: Decision curve as a sensitivity analysis.	3/1000 1/335	The net number of true positive classifications increases by 3 in 1000 where HDL cholesterol is measured, on a scale from 0 to 5.6% (event rate). This implies that 335 measurements of HDL cholesterol have to be done to identify one more event as high risk without decreasing low risk classifications for those without events.

decreased from 96.8% to 96.7% with the addition of HDL cholesterol to the model. The increase in sensitivity and decrease in specificity is in line with theoretical expectations, since the decision threshold was higher than the event rate (Van Calster *et al.*, 2014).

Net Reclassification Risk graph

For a direct understanding of the clinical usefulness of a marker at a specific decision threshold, we propose a "Net Reclassification Risk" graph. This graph allows us to focus on the number of reclassified individuals and their observed event rates. The components for this graph are identical to what is used in a reclassification table. In a reclassification table (table 3.2), we stratify by event status to calculate net improvements for those with and without events. Alternatively, we may stratify by reclassification group to calculate numbers of persons and absolute risks (a net reclassification risk table, table 3.3). Here, the two most relevant groups are those reclassified from low to high and from high to low risk (low: <20% risk, high: >= 20% risk). The graphical display as in figure 3.6 allows for a straightforward interpretation: a larger fraction reclassified is better (width of the bars) and more separation of absolute risks is better (vertical spread). The area of each bar is proportional to the number of events in a reclassification group. The change in sensitivity is proportional to the difference between areas for the L→H and H→L groups in figure 3.6. A marker that can better identify events hence has a larger difference in areas. From table 3.3 and figure 3.6 we learn that 29 persons were reclassified from high to low risk (H→L), with a 10% event rate, and 45 reclassified from low to high risk (L→H), with a 31% event rate. We hence identify 45*0.31 – 29*0.10 = 11 more events. These 11 events account for a sensitivity increase of 11/183 (6.0%). The increase is partly achieved by defining a larger number of persons as high risk (45 vs 29 persons), which causes some overtreatment for those without events (more false-positives). The number of false-positives can be calculated from the probabilities

Table 3.2 Reclassification table with 2 categories based on a >20% 10-year CHD risk threshold to define low and high risk

		Model with HDL cholesterol	
Without event: n=3081		Low	High
Model without HDL cholesterol	Low	2952	31
	High	26	72
With event: n=183		Low	High
Model without HDL cholesterol	Low	145	14
	High	3	21

Table 3.3 Risk table for reclassified persons with the addition of HDL cholesterol, based on a >20% 10-year CHD risk threshold to define low and high risk.

Reclassification	Events	Non-events	N	Risk [95% CI]
Low → low	145	2952	3097	4.7% [4.0-5.5%]
High → Low	3	26	29	10.3% [2.7-28%]
Low → High	14	31	45	31.1% [19-47%]
High → High	21	72	93	22.6% [15-33%]
Total	183	3081	3264	5.6%

95% CI: 95% confidence interval
NRI, ΔNB, and ΔRU calculations from table 3.2 and table 3.3:
NRI events = (14-3)/ 183 = 6.0%
NRI non-events = (26 – 31) /3081 = –0.2%
NRI = [(14-3)/183]+[(26-31)/3081] = 6.0% + -0.2% = 0.058
ΔNB = (11 - 0.25*5) / 3264 = 0.003
ΔRU= (11 - 0.25*5) / 183 = 0.053

of a non-event, which is the complement of the event risk: 45*(1 - 0.31) – 29*(1 - 0.10) = 5 additional false-positives. This loss in specificity (5/3081, –0.2%) might be shown similarly in a Net Reclassification Risk graph, either by stacked bars, or by plotting the non-event rate at the y-axis (results not shown).

Summary measures for binary classification

The AUC for a binary ROC curve equals (sensitivity + specificity)/2. It increased from 0.550 to 0.579 (+0.029). In the case of a single cut-off, the event NRI and non-event NRI are the changes in sensitivity and specificity. Hence,

$$NRI = 2*\Delta AUC.$$ (Pencina *et al.*, 2008)

In our case study, NRI = 0.058 [=6.0% – 0.2%], and ΔAUC = 0.029 [=0.579 – 0.550]. ΔAUC and NRI weight changes in sensitivity and specificity equally. If the event rate is below 50%, this implies that a change in classification for an individual with an event is weighted relatively heavier than for an individual without an event. For event rates over 50%, specificity changes are weighted as relatively more important. Specifically, ΔAUC and NRI weight the numbers of extra TP and TN classifications by the odds of the event rate (Van Calster *et al.*, 2013).

Weighted sums of sensitivity and specificity

A weighted variant of summing sensitivity and specificity was already proposed by Peirce in 1884. (Peirce, 1884). It was recently reintroduced as the Net Benefit (NB). (Vickers & Elkin, 2006). For changes in the number of true positives (TP) and false positives (FP), improvement in NB is defined as:

Figure 3.6 Net Reclassification Risk graph for the addition of HDL cholesterol to 10-year CHD predictions in 3264 participants from the Framingham Heart Study.

L: Low risk classification; H: High risk classification, based on a 20% risk cut-off. Reclassified patients are in the groups H→L and L→H (Low risk reclassified as high risk, and high risk reclassified as low risk respectively). Uncertainty is indicated by 95% confidence intervals.

$$\Delta NB = (\Delta TP - w * \Delta FP) / \text{number of subjects}$$

where the weight w is the odds of the decision threshold. For example, a 20% CHD risk threshold means w = 0.20 / (1 – 0.20) = 0.25. The decision threshold of 20% hence implies that we weight a FP classification 0.25 times as important as a TP classification, or that 1 more TP classification is worth 4 more FP classifications. In table 3.3 and figure 3.6 we note 11 more TP at the price of 5 more FP classifications. The burden of overtreatment of those without events is explicitly weighted in the NB calculation by the odds of the decision threshold 0.20 / (1 - 0.20) = 0.25. This leads to a penalty for overtreatment of 5*0.25=1.25. The ΔNB hence is (11 - 1.25) / 3264 = 0.3%, equivalent to potentially identifying and treating an additional 3 events per 1000 persons screened without extra overtreatment after the addition of HDL cholesterol to the CHD prediction model.

The link between a decision threshold and the relative weight of TP vs FP classifications has a strong foundation in decision theory (Pauker & Kassirer, 1980). It is also used in other recently proposed weighted summary measures such as the change in relative utility (ΔRU) (Baker, 2009) and the weighted NRI (wNRI) (Pencina *et al.*, 2011). In our example, the decision threshold of 20% is higher than the event rate of 5.6%, and ΔRU is defined as:

$$\Delta RU = (\Delta TP - w * \Delta FP) / \text{number of events},$$

with weight w for the odds of the decision threshold. In our example ΔRU = (11 – 0.25 * 5) / 183 = 0.053. ΔRU focuses on the improvement of using the prediction model to

assign treatment over the baseline strategy with highest NB, either treat all or treat none (Baker, 2009). If the decision threshold is higher than the event rate, as in our example, this means that RU compares to treat none. Furthermore, the ΔRU divides the obtained improvement by the maximal improvement, which is the situation where treatment would be assigned to all individuals that develop the outcome of interest and to none without. The relation between ΔRU and ΔNB is: ΔRU = ΔNB / event rate, in this situation of threshold > event rate (Van Calster *et al.*, 2013). ΔNB considers the utility of the addition of a novel marker on an absolute scale, while ΔRU considers this utility relative to the maximum as defined by the event rate.

Finally, the weighted version of NRI only differs from ΔNB by a scaling factor: ΔNB is usually expressed in units of savings that result from a correct classification, whereas wNRI uses the savings in whatever unit is selected.

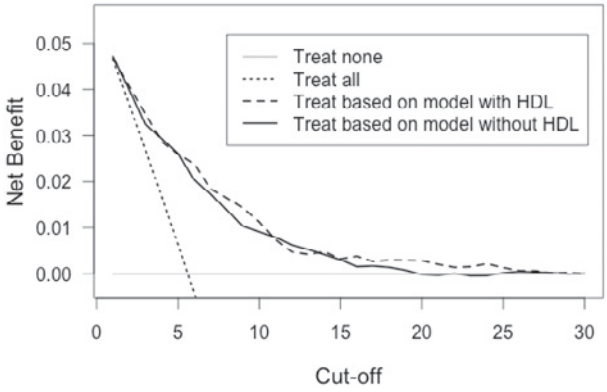
In our example, a 20% CHD risk decision threshold leads to ΔNB =0.3%, which can be interpreted as that prediction with HDL cholesterol increases the fraction of true positives identified in the population by 3 per 1000, without a change in false positives. ΔRU = 5.3%, which indicates a 5.3% relative gain in utility compared to the alternative of treating none. If we would assume that correct identification of a patient at risk of a CVD event results in savings of $100,000, whereas avoiding unnecessary treatment with statins saves $25,000, we obtain wNRI = $300 of average savings per person (0.3%*$100,00).

All decision analytic measures can be used to calculate the "test tradeoff", which indicates in how many persons the marker needs to be measured for a net increase in one true positive classification (i.e. identifying one additional person with the event as high risk who thereby qualifies for treatment) (Baker *et al.*, 2012). This test tradeoff has similarities to the well-known concept of 'number needed to treat' in trials and efficacy research. The test tradeoff is defined as 1 / ΔNB. Hence, the test tradeoff for measuring HDL cholesterol was 1 in 335.

Sensitivity to choice of cut-off

In the weighted variants of summing sensitivity and specificity, the decision threshold is essential, which is defined by the harm to benefit ratio. As a sensitivity analysis, it is recommended to consider a range of possible thresholds. This can be displayed in a decision curve (Vickers & Elkin, 2006). We note that the model with HDL cholesterol has a small but consistently higher NB for most thresholds in the clinically most relevant range from 5% to 25% 10-year risks (figure 3.7).

Figure 3.7 Decision curves comparing Net Benefit of four alternative strategies: treat all, treat none, and treatment decisions based on the predicted 10-year CHD risk (from models with or without HDL cholesterol) according to risk cut-offs between 0 and 30%.



3.6 DISCUSSION

In this review, we considered various graphs and summary measure to assess the incremental value of a novel marker in predicting presence of disease (diagnosis) or the occurrence of an event over time (prognosis). We first examined continuous predictions, which is in line with the usual statistical approach to develop a model with and without the marker under study. Several informative graphs can be created (figures 3.1 – 3.5), with appealing summary measures such as the increase in explained variability and increase in AUC, or the continuous NRI. Second, we examined binary classifications, which can well be summarized in a reclassification table (table 3.2). We propose an alternative design of such a table ("Net reclassification table", table 3.3) with a graph ("Net Reclassification Risk graph", figure 3.6) to readily assess the incremental value of a marker for improving clinical risk classification. This presentation draws our attention to the number of reclassified persons, i.e. from low to high and from high to low risk. This focus was also central in Cook's early proposals for assessing reclassification (Cook *et al.*, 2006) (Cook, 2007). Second, our attention is drawn to the absolute risk estimates, which naturally should be higher for those reclassified as at high risk then for those reclassified as low risk. The difference in areas of the low→high and high→low bars indicates how many more subjects with events can be identified by the marker. This increase in sensitivity is a key attribute in many decision problems. The harms for false positive classifications, which relate to the loss of specificity, are typically less than the benefits for true positive classifications. We recognize that more research is necessary on the properties of the Net Reclassification Risk graph and its potential applications.

An obvious limitation is that the graph cannot readily consider multiple categories, since a 3 category classification would already lead to 9 bars.

Interpretation of incremental value

A limitation of all our assessments is that it is difficult to define what increment in performance is 'important', 'substantial', or 'meaningful'. For statistical significance, the limit of 0.05 is widely accepted. A lower p-value is affected by the combination of effect size and sample size, and should hence not be sufficient to claim that a marker is important for better prediction. Effect size criteria are preferable, where a value of Cohen's D of 0.5 is widely accepted as reflecting a medium effect (Cohen, 1988). Some epidemiologists may also like estimates of relative risk per standard deviation increase in value of a marker. Pencina *et al.* (Pencina *et al.*, 2012a) found that a medium effect size or OR of 1.65 per SD corresponds to a continuous NRI of 0.395. The increase in AUC depends on the AUC of the baseline model, e.g. Cohen's D of 0.5 corresponds to +0.05 for a baseline model with AUC of 0.65, whereas only to +0.02 for a baseline model with AUC of 0.80 (Pencina *et al.*, 2012a). A third option is to consider cost-effectiveness criteria, where willingness to pay thresholds have been proposed in terms of euros (or dollars) per quality adjusted life-year (QALY) gained, for example $20,000 per QALY (Chapman *et al.*, 2000) (Postmus *et al.*, 2012). Performing a full cost-effectiveness analysis may be too much effort for every marker studied for incremental value. An intermediate solution is to consider measures such as the ΔNB, ΔRU, wNRI, or a reciprocal summary measure, the test tradeoff. In our example, the test tradeoff for measuring HDL cholesterol was 1 in 335. Whether this number is valued relatively low or relatively high depends on the context. For example, if we consider starting statin treatment for those classified as high risk, and assume a relative risk of 0.73 for the reduction of CHD events (Taylor *et al.*, 2013), we know that we can prevent 27% of the net reclassified CHD events. This absolute benefit should be weighted against the disadvantages of the marker measurement, such as financial costs, the burden or discomfort, and potential risks to the patient. The harms of overtreatment are already incorporated in the NB calculation, where the treatment threshold reflects the harm to benefit ratio (in our example 1:4, threshold 20% 10-year CHD risk).

Estimation issues

We focused on the apparent performance of models with and without markers. Especially in small data sets, it is well known that effect estimates for markers may be exaggerated. Such overfit leads to overoptimistic estimates of performance. Several internal validation techniques can be used to correct for optimism in performance, including cross-validation and bootstrap resampling (Steyerberg *et al.*, 2001). Internally validated performance estimates can be derived for all measures as presented. Ideally,

performance is determined on fully independent validation data. Many differences may exist between the independent validation data and the development setting, which makes that we may often be testing transportability of prediction models in time or place rather than validating marker performance in the more narrow sense of assessing reproducibility (Justice *et al.*, 1999). It is imperative that the NRI for quantifying added value of a marker should only be calculated after recalibrating or refitting model predictions (Hilden & Gerds, 2013) (Leening *et al.*, 2014a). Such recalibration places the predicted risks with and without the marker at an equal and fair level with respect to calibration. The NRI then indicates the improvement in classification attributable to the marker, but conditional on adequate calibration. An alternative approach is to evaluate reclassification in the context of better clinical decision making. We may then take predicted risks from a model with and without a marker literally, i.e. without recalibration, and preferably use performance measures that are consistent with a decision making framework, such as the ΔNB, ΔRU, or wNRI (Van Calster *et al.*, 2013).

A final limitation is that we focused on prediction of a binary endpoint, without considering the time-to-event nature of the data in our example. In several graphs and tables we conditioned on the event status, which is a simplification that is especially problematic if censoring occurs in many subjects before the end of follow-up. The Net Reclassification Risk graph can readily be made with Kaplan-Meier estimates rather than observed event rates (Steyerberg & Pencina, 2010), and definitions for survival data are available for many performance measures, such as Nagelkerke's R^2, Harrell's or Uno's c statistic (equivalent to AUC for binary endpoints) (Uno *et al.*, 2011), the IDI and NRI (Pencina *et al.*, 2011), and the ΔNB (Vickers *et al.*, 2008).

In conclusion, we recommend a clear distinction in the type of research question that is addressed in relation to the incremental value of a marker: better prediction or better classification. From a prediction perspective, continuous predictions may be considered with various graphs and summary measures. We are in favor of reclassification plots, but due to the lack of informativeness we recommend against publishing of ROC curves unless clinically motivated thresholds are indicated. For summary measures, we are in favor of measures that indicate the explained variability (increase in Nagelkerke's or Pearson's R^2, the IDI), increase in discrimination (ΔAUC), or effect size (continuous NRI). If we move to the evaluation of a classification, the category-based NRI may be used, with several important caveats (Leening *et al.*, 2014b). The Net Reclassification Risk graph may allow for a direct visual impression in the case of binary classification. We can then readily judge two important aspects of improved classification: what proportion of subjects are reclassified, and what is the difference in observed event rates

between reclassified groups. When classification leads to decisions, decision-analytic summary measures such as ΔNB, ΔRU and wNRI are adequate.

REFERENCES

Austin, P.C. & Steyerberg, E.W. (2013). Predictive accuracy of risk factors and markers: a simulation study of the effect of novel markers on different performance measures for logistic regression models. Stat Med, 32, 661-72.

Baker, S.G. (2009). Putting risk prediction in perspective: relative utility curves. J Natl Cancer Inst, 101, 1538-42.

Baker, S.G., Cook, N.R., Vickers, A. & Kramer, B.S. (2009). Using relative utility curves to evaluate risk prediction. J R Stat Soc Ser A Stat Soc, 172, 729-748.

Baker, S.G., Van Calster, B. & Steyerberg, E.W. (2012). Evaluating a new marker for risk prediction using the test tradeoff: an update. Int J Biostat, 8.

Chapman, R.H., Stone, P.W., Sandberg, E.A., Bell, C. & Neumann, P.J. (2000). A comprehensive league table of cost-utility ratios and a sub-table of "panel-worthy" studies. Med Decis Making, 20, 451-67.

Cleveland, W.S. (1985). The elements of graphing data. Wadsworth Advanced Books and Software: Monterey.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. New Jersey: Lawrence Erlbaum Associates.

Cook, N.R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation, 115, 928-35.

Cook, N.R., Buring, J.E. & Ridker, P.M. (2006). The effect of including C-reactive protein in cardiovascular risk prediction models for women. Ann Intern Med, 145, 21-9.

Greenland, P., Alpert, J.S., Beller, G.A., Benjamin, E.J., Budoff, M.J., Fayad, Z.A., Foster, E., Hlatky, M.A., Hodgson, J.M., Kushner, F.G., Lauer, M.S., Shaw, L.J., Smith, S.C., Jr., Taylor, A.J., Weintraub, W.S., Wenger, N.K., Jacobs, A.K. & American College of Cardiology Foundation/American Heart Association Task Force on Practice, G. (2010). 2010 ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation, 122, e584-636.

Hanley, J.A. & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143, 29-36.

Harrell, F.E. (2001). Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer: New York.

Hilden, J. & Gerds, T.A. (2013). A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. Stat Med.

Justice, A.C., Covinsky, K.E. & Berlin, J.A. (1999). Assessing the generalizability of prognostic information. Ann Intern Med, 130, 515-24.

Leening, M.J. & Steyerberg, E.W. (2013). Fibrosis and mortality in patients with dilated cardiomyopathy. JAMA, 309, 2547-8.

Leening, M.J., Steyerberg, E.W., Van Calster, B., D'Agostino, R.B. & Pencina, M.J. (2014a). Net reclassification improvement and integrated discrimination improvement require calibrated models: relevance from a marker and model perspective. Stat Med, in press.

Leening, M.J., Vedder, M.M., Witteman, J.C., Pencina, M.J. & Steyerberg, E.W. (2014b). Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. Ann Intern Med, 160, 122-31.

Localio, A.R. & Goodman, S. (2012). Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. Ann Intern Med, 157, 294-5.

McGeechan, K., Macaskill, P., Irwig, L., Liew, G. & Wong, T.Y. (2008). Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. Arch Intern Med, 168, 2304-10.

Nagelkerke, N.J. (1991). A note on a general definition of the coefficient of determination. Biometrika, 691-692.

Pauker, S.G. & Kassirer, J.P. (1980). The threshold approach to clinical decision making. N Engl J Med, 302, 1109-17.

Peirce, C.S. (1884). The numerical measure of the success of predictions. Science, 4, 453-454.

Pencina, M.J., D'Agostino, R.B., Pencina, K.M., Janssens, A.C. & Greenland, P. (2012a). Interpreting incremental value of markers added to risk prediction models. Am J Epidemiol, 176, 473-81.

Pencina, M.J., D'Agostino, R.B., Sr., D'Agostino, R.B., Jr. & Vasan, R.S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med, 27, 157-72; discussion 207-12.

Pencina, M.J., D'Agostino, R.B., Sr. & Demler, O.V. (2012b). Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. Stat Med, 31, 101-13.

Pencina, M.J., D'Agostino, R.B., Sr. & Steyerberg, E.W. (2011). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med, 30, 11-21.

Pencina, M.J., D'Agostino, R.B. & Vasan, R.S. (2010). Statistical methods for assessment of added usefulness of new biomarkers. Clin Chem Lab Med, 48, 1703-11.

Pepe, M.S., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I.M. & Zheng, Y. (2008). Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol, 167, 362-8.

Postmus, D., de Graaf, G., Hillege, H.L., Steyerberg, E.W. & Buskens, E. (2012). A method for the early health technology assessment of novel biomarker measurement in primary prevention programs. Stat Med, 31, 2733-44.

Ridker, P.M., Buring, J.E., Rifai, N. & Cook, N.R. (2007). Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. JAMA, 297, 611-9.

Royston, P. & Altman, D.G. (2010). Visualizing and assessing discrimination in the logistic regression model. Stat Med, 29, 2508-20.

Steyerberg, E.W. (2009). Clinical prediction models: a practical approach to development, validation, and updating. Springer: New York.

Steyerberg, E.W., Harrell, F.E., Jr., Borsboom, G.J., Eijkemans, M.J., Vergouwe, Y. & Habbema, J.D. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol, 54, 774-81.

Steyerberg, E.W., Moons, K.G., van der Windt, D.A., Hayden, J.A., Perel, P., Schroter, S., Riley, R.D., Hemingway, H., Altman, D.G. & Group, P. (2013). Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med, 10, e1001381.

Steyerberg, E.W. & Pencina, M.J. (2010). Reclassification calculations for persons with incomplete follow-up. Ann Intern Med, 152, 195-7.

Taylor, F., Huffman, M.D., Macedo, A.F., Moore, T.H., Burke, M., Davey Smith, G., Ward, K. & Ebrahim, S. (2013). Statins for the primary prevention of cardiovascular disease. Cochrane Database Syst Rev, 1, CD004816.

Tjur, T. (2009). Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination. The American Statistician, 63, 366-372.

Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B. & Wei, L.J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med, 30, 1105-17.

Van Calster, B., Steyerberg, E.W., D'Agostino, R.B., Sr. & Pencina, M.J. (2014). Sensitivity and specificity can change in opposite directions when new predictive markers are added to risk models. Med Decis Making, 34, 513-22.

Van Calster, B., Vickers, A.J., Pencina, M.J., Baker, S.G., Timmerman, D. & Steyerberg, E.W. (2013). Evaluation of markers and risk prediction models: overview of relationships between NRI and decision-analytic measures. Med Decis Making, 33, 490-501.

Vickers, A.J., Cronin, A.M., Elkin, E.B. & Gonen, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak, 8, 53.

Vickers, A.J. & Elkin, E.B. (2006). Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making, 26, 565-74.

PART III

Early HTA of a new biomarker

Chapter 4

Cost-effectiveness of prostate cancer screening using a PSA test combined with a novel biomarker

Vedder MM, de Bekker-Grob EW, de Koning HJ, Heijnsdijk EAM, Steyerberg EW

Submitted

ABSTRACT

Objectives: To investigate under what conditions adding a novel biomarker to Prostate Specific Antigen (PSA) testing is cost-effective for prostate cancer (PC) screening.

Patients and methods: We adapted the MISCAN-Prostate microsimulation-model, which includes data from the European Randomised study of Screening for Prostate Cancer (ERSPC) trial to estimate costs and Quality Adjusted Life Years (QALYs) for different PC screening strategies in men aged 55-69 years. The reference strategy was annual screening with PSA alone, with a relatively low Positive Predictive Value (PPV) of 23%. We varied costs and PPVs of a strategy that added the novel biomarker to PSA testing to determine the price thresholds for the novel biomarker to be cost-effective.

Results and limitation: Assuming a threshold of €20,000/QALY, annual screening with PSA combined with a novel biomarker was cost-effective at PPVs of 35% and 100%, if the costs of the novel biomarker were very low (€6 and €13, respectively). Assuming a threshold of €50,000/QALY, four-yearly screening and 100% PPV, the highest price threshold was €49. For testing with 100% PPV *after* initial PSA >3.0ng/ml, the threshold was substantially higher, at €255 for four-yearly screening and a threshold of €50.000/QALY. These price thresholds are uncertain due to modelling assumptions, such as that no additional burden was caused by the biomarker measurement.

Conclusion: PSA combined with a novel biomarker will only be a cost-effective alternative to screening with PSA alone if costs are very low, or selectively in those with high PSA.

4.1 INTRODUCTION

Prostate cancer (PC) is the most common cancer in Europe for males with over 400,000 new cases diagnosed in 2012 [1]. Prostate Specific Antigen (PSA) testing is the cornerstone for detecting PC. The European Randomised Study of Screening for Prostate Cancer (ERSPC) trial proved that population-based screening led to a reduction in PC specific mortality [2]. This large randomised controlled trial included 162,243 men from the Netherlands, Sweden, Finland, Belgium, France, Spain, Italy and Switzerland. Additional modelling showed that 73 life years and 56 quality adjusted life years (QALYs) per 1,000 men can be saved by annual screening in men aged 55-69 years [3].

Prostate cancer screening can be cost-effective for some strategies [4]. Screening with PSA for PC remains controversial because PSA testing has a limited specificity for early detection [5]. Many biopsies are done in men without cancer, and cancers are found that would not have been diagnosed in the absence of screening (overdetection) [2, 3, 6]. Novel biomarkers are needed that improve the accuracy of finding clinically significant PC at biopsy, and hence avoid unnecessary biopsies [7].

It is not known how accurate, and at what cost, such a biomarker would need to be to be considered a cost-effective addition to screening with the PSA test alone. To guide the future development of cost-effective biomarkers, it is important to obtain knowledge of the required performance characteristics.

We aimed to investigate under what conditions adding a novel biomarker to PSA testing is cost-effective for PC screening in men aged 55-69 years. We hereto used a microsimulation model based on the results of the ERSPC trial.

4.2 PATIENTS AND METHODS

Model description

The microsimulation screening analysis (MISCAN) prostate cancer model [3, 8, 9] was adapted to assess costs and effects for the current study. The MISCAN prostate cancer model simulates a large study population of men with individual life histories. Some men will develop PC, and some will die from the disease. A life history is defined by a sequence of states and the time spent in these states (dwelling time), generated by a semi-Markov process. From each state, a next state is generated with transition probabilities and dwelling times determined by the present state and age of the subjects (supplementary figure 4.1).

The cancers were divided into clinically diagnosed cancers that were found in between screening rounds or in men not screened, relevant screen-detected cancers and over-detected cancers that would not have become clinically diagnosed during a man's life ('overdiagnoses'). Negative biopsies after an initial PSA test with PSA>3.0ng/ml were defined as negative (potentially avoidable) biopsies. PCs were characterised according to their clinical stage, grade, and metastatic stage. Death from other causes was generated independently using standard life tables from the Netherlands. This simulation resulted in an age-specific and time-specific output of disease incidence and mortality for 1,000,000 men. All men then underwent simulated screening; this intervention changed some of the life histories. These changes in life histories constituted the effects of the intervention and were represented by the number of PCs, treatments, deaths, and QALYs induced or prevented by the intervention. The costs and quality-of-life outcomes of the intervention were determined from the total numbers of events.

Model input

The input variables for the MISCAN prostate cancer model included costs and utilities for various health states, and transition probabilities between states. Costs included the direct screening costs and medical costs for diagnosis and treatment. Costs and utility estimates both were based on previous publications [3, 6, 10] (see supplementary table 4.1). E.g., per biopsy there was a utility loss of 0.10, for a duration of three weeks, equivalent to a QALY loss of 0.006 per biopsy. The transition probabilities were based on the ERSPC study [11] (supplementary figure 4.1). In the ERSPC participating men underwent randomization to undergo PSA testing or not. Most centres used a PSA cut-off value of 3.0 ng per millilitre as an indication for biopsy, whereas others used a cut-off of 4.0 ng/ml, with additional tests for values between 2.5 and 4.0 ng/ml. The screening interval was 4 years, with the exception of Sweden, where it was 2 years. PC diagnosis was based on sextant biopsy outcome. Treatment was performed according to local policies and guidelines, independent of the study group [12].

Analyses

We compared screening with PSA combined with a novel biomarker to screening with PSA only in different screening strategies. Screening with a PSA test was assumed to consist of a single screening test with a sensitivity that depended on the stage of the cancer, ranging from 0.82 to 0.98 [8]. Screening with the PSA test combined with the novel biomarker was assumed to have equal sensitivity, but improved specificity. Hence, the Positive Predictive Value (PPV) was improved, with PPV defined as probability of having biopsy confirmed PC given a positive test result.

For this study, we assumed that the novel biomarker test was based on blood samples taken for PSA testing, so no separate sampling in subjects was necessary. We assumed that the novel biomarker test was performed simultaneously with the PSA test, thereby improving on specificity of the combined test and thereby decrease the number of biopsies needed to find the same number of cancers. Additionally, the decision whether or not to do a biopsy is based on the combined outcome of the PSA and the novel biomarker test; only when both tests were positive, a biopsy is carried out. While the PPV of the screening test increases by adding a biomarker, the overdetection of PC was assumed to remain the same, i.e. finding a cancer that would not have been clinically diagnosed without screening.

We used a previously published reference strategy [3] as our base-case, i.e. annual PSA screening in men aged 55-69 years and 80% attendance. We focused on two parameters that may vary and affect how adding a novel biomarker to PSA testing for PC compares to traditional screening with PSA only testing: the PPV and the novel biomarker test unit costs. Demography and age distributions of disease stages were based on European data and would equally affect the compared outcomes of the investigated screening strategies; they were therefore not varied. We assumed that PSA testing with a test threshold of 3.0ng/ml had a PPV of 22.7% for screen-detected cancers [13] and a PPV of 35.8% for cancers that were found in between screening rounds [14]. An increased PPV by adding a biomarker reduces the number of biopsies that is needed to find the same number of cancers. We studied PPVs of 25%, 35%, 50% and 100% for a screening strategy including a novel biomarker. A range from 25 to 100% PPV was chosen to cover the broad range of possible PPV values. All costs and utilities were discounted at an annual rate of 3% [15].

Threshold values for costs were determined to investigate when adding a novel biomarker would be cost-effective to add to PSA based PC screening. Cost-effectiveness thresholds of €20,000 and €50,000 per QALY gained were used in the light of earlier policy decisions in the Netherlands [16] and other Western European countries [17].

In a sensitivity analysis, we considered screening every four years rather than every year. This screening interval represents actual screening strategies predominantly used in the ERSPC study. Moreover, discount rates of 0% and 5% were used rather than 3%. Additionally, we analysed a strategy where the novel biomarker test was only performed *after* a PSA test >3.0ng/ml.

4.3 RESULTS

Reducing unnecessary biopsies

The reference strategy of annual screening with PSA alone led to 290 biopsies to diagnose 74 cancers per 1,000 men (table 4.1). When the combined test PPV increased from 22.7% in the reference strategy to 35%, the number of biopsies needed to find 74 cancers was reduced from 290 to 210. In the best-case scenario, the combination of PSA and the novel biomarker was assumed to have PPV of 100%. The number of biopsies would then decline from 290 to 113, a reduction with a factor 2.6. With the best-case novel biomarker, there were still 39 unnecessary biopsies carried out that would not have been initiated without screening, i.e. for men dying of other causes before the PC would become clinically significant. This is a factor of 5.5 (216/39) less biopsies that result in overdiagnosis compared to the reference strategy.

Table 4.1 Lifetime numbers and costs per 1,000 men for various health states expected with annual prostate cancer screening in men aged 55-69 years (costs and effects discounted at 3%)

Effects (N)	PSA test[1]	PSA + novel biomarker test[3]			
	PPV[2]: 22.7%	PPV: 25%	PPV: 35%	PPV: 50%	PPV: 100%
Screening attendance	3952	3952	3952	3952	3952
Biopsies	290	269	210	165	113
Diagnosis					
Relevant cancers	74	74	74	74	74
Overdiagnoses[6]	39	39	39	39	39
Negative (potentially avoidable) biopsies	177	156	97	52	0
Active surveillance	21	21	21	21	21
Radiation therapy (RT)	22	22	22	22	22
Radical prostatectomy (RP)	34	34	34	34	34
Terminal illness	9	9	9	9	9
Palliative therapy	11	11	11	11	11
Total costs prostate cancer screening and treatment (k€)					
PSA testing only	1,029				
Assuming additional novel biomarker costs of €0		1,025	1,015	1,008	999
Assuming additional novel biomarker costs of €50		1,223	1,213	1,205	1,196
Assuming additional novel biomarker costs of €100		1,421	1,411	1,403	1,394
QALYs gained compared to no screening	8.34	8.46	8.80	9.06	9.36
QALYs gained compared to PSA screening		+0.12	+0.47	+0.72	+1.03

[1] Strategy I: prostate cancer (PC) screening with a Prostate Specific Antigen (PSA) test alone
[2] PPV: positive predictive value
[3] Strategy II: PC screening with PSA and an additional biomarker, with a combined PPV of 25% (a), 35% (b), 50% (c), or 100% (d).
[6] The definition of an overdiagnosis is an indolent prostate cancer, which would not have been found during the life time without screening

For screening every 4 years, the number of biopsies was 232 for diagnosing 63 cancers per 1,000 men (table 4.2) in the PSA only strategy. At a PPV of 35%, the number of biopsies needed to find these cancers was 179; 63 unnecessary biopsies were avoided. Assuming a best-case novel biomarker (PPV of 100%), the number of biopsies declined from 232 to 116 biopsies. The number of unnecessary biopsies would decline from 169 to 53, a reduction by a factor 3.2 (169/53) (table 4.2).[1] In a strategy were the novel biomarker test was performed only after a PSA test >3.0ng/ml, the number of avoided

Table 4.2 Lifetime numbers and costs per 1,000 men for various health states expected with screening every 4 years in men aged 55-69 years (costs and effects discounted at 3%)

Effects (N)	PSA test[1]	PSA + novel biomarker test[3]			
	PPV[2]: 22.7%	PPV: 25%	PPV: 35%	PPV: 50%	PPV: 100%
Screening attendance	1091	1091	1091	1091	1091
Biopsies	232	218	179	150	116
Diagnosis					
Relevant cancers	63	63	63	63	63
Overdiagnoses[6]	53	53	53	53	53
Negative (potentially avoidable) biopsies	116	102	63	34	0
Active surveillance	16	16	16	16	16
Radiation therapy (RT)	20	20	20	20	20
Radical prostatectomy (RP)	27	27	27	27	27
Terminal illness	10	10	10	10	10
Palliative therapy	12	12	12	12	12
Total costs prostate cancer screening and treatment (k€)					
PSA testing only	848				
Assuming additional novel biomarker costs of €0		846	839	834	829
Assuming additional novel biomarker costs of €50		900	894	889	883
Assuming additional novel biomarker costs of €100		955	948	943	938
QALYs gained compared to no screening	6.31	6.39	6.61	6.78	6.98
QALYs gained compared to PSA screening		+0.08	+0.30	+0.47	+0.67

[1] Strategy I: prostate cancer (PC) screening with a Prostate Specific Antigen (PSA) test alone
[2] PPV: positive predictive value
[3] Strategy II: PC screening with PSA and an additional biomarker, with a combined PPV of 25% (a), 35% (b), 50% (c), or 100% (d).
[6] The definition of an overdiagnosis is an indolent prostate cancer, which would not have been found during the life time without screening

1 We note that the 53 cancers that would not have become clinically diagnosed during a person's life (overdiagnoses) will still remain because of biopsy referral based on PSA.

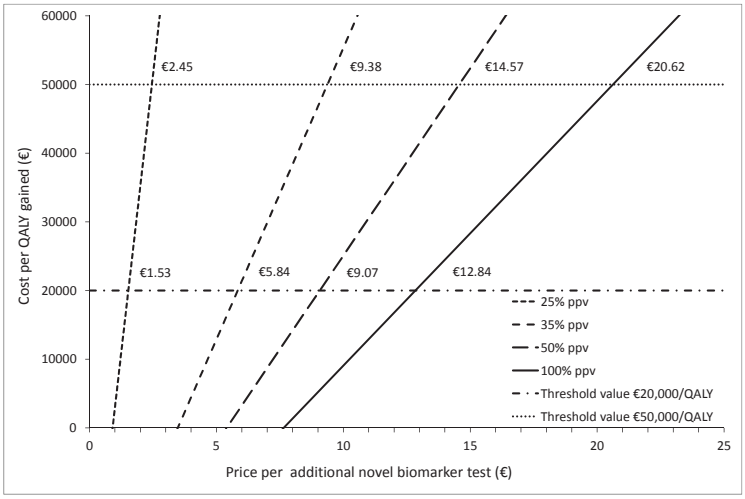
biopsies and QALYs saved were equal to these numbers in a test all scenario. However, because less novel biomarker tests were carried out, total costs for screening with this strategy were substantially reduced.

Price thresholds

With discounting at 3%, screening with PSA combined with a novel biomarker was cost-effective at a PPV of 35%, if the novel biomarker costs did not exceed €5.84 or €9.38, assuming a threshold of €20.000/QALY and €50.000/QALY, respectively (figure 4.1). With a PPV of 100% (perfect specificity), screening with PSA combined with a novel biomarker was cost-effective if the costs of the biomarker test did not exceed €13 or €21, assuming a threshold of €20.000/QALY and €50.000/QALY, respectively. Numbers were quite similar at discount rates of 0% or 5% (supplementary figure 4.2).

For screening every 4 year, the price for the biomarker might be higher. Maximum costs were €30 and €49 for thresholds of €20,000/QALY and €50,000/QALY, respectively, at

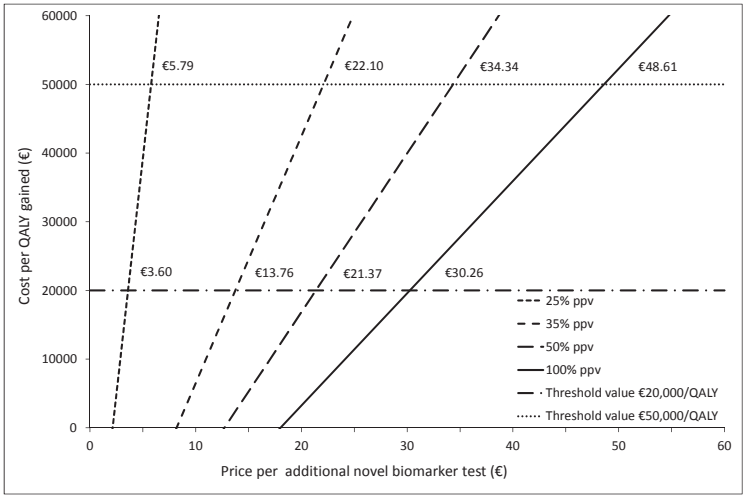
Figure 4.1 Price threshold analysis for costs of an additional novel biomarker (annual screening, costs and effects discounted at 3%)



PPV: positive predictive value of PSA screening combined with the novel biomarker test; QALY = Quality Adjusted Life Year

Intersections with the horizontal dashed lines (i.e., the acceptability thresholds) represent the threshold price per novel biomarker test at which PSA combined with a novel biomarker will be a cost-effective alternative for PSA screening alone assuming an acceptability threshold of €20,000/QALY or €50,000/QALY, with 3% discounting

Figure 4.2 Price threshold analyses for costs of an additional novel biomarker (screening every 4 years, costs and effects discounted at 3%)



PPV: positive predictive value of PSA screening combined with the novel biomarker test; QALY = Quality Adjusted Life Year
Intersections with the horizontal dashed lines (i.e., the acceptability thresholds) represent the threshold price per novel biomarker test at which PSA combined with a novel biomarker will be a cost-effective alternative for PSA screening alone assuming an acceptability threshold of €20,000/QALY or €50,000/QALY, with 3% discounting

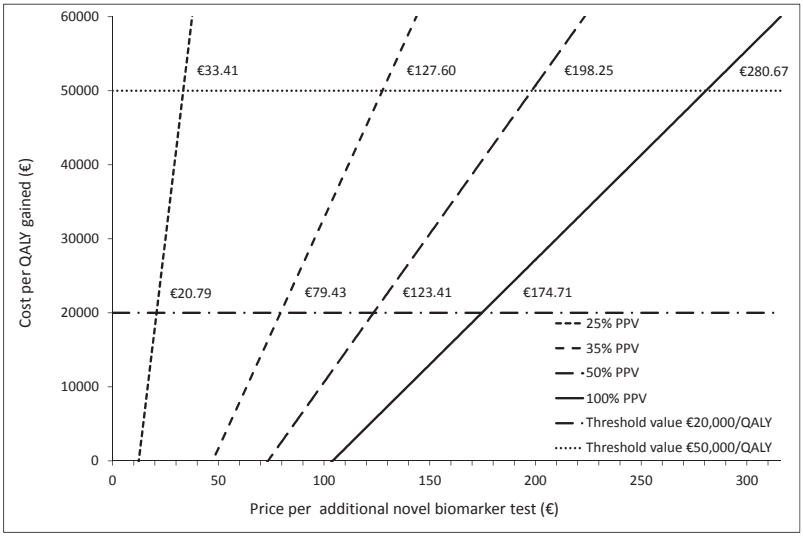
a discount rate of 3% and perfect PPV of 100% (figure 4.2). The discounting rate again had no meaningful impact on the maximum price (supplementary figure 4.3).

When the novel biomarker test is performed after an initial PSA test >3.0ng/ml, the price of the test may be substantially higher. For annual screening, this was €175 and €281 for thresholds of €20,000/QALY and €50,000/QALY, respectively, at a discount rate of 3% and perfect PPV of 100% (figure 4.3). With a strategy of screening every four years, this was €158 and €255 respectively (figure 4.4).

4.4 DISCUSSION

We found that adding a biomarker to PSA-based screening could potentially lead to a substantial reduction of the number of unnecessary biopsies for PC. Costs of a novel biomarker should however remain below €50 when used for all screening participants, even if the screening test based on the combination of PSA and the biomarker would have perfect specificity, and hence perfect PPV. Utilization of the novel biomarker test

Figure 4.3 Price threshold analysis for costs of an additional novel biomarker after an initial PSA test > 3.0 ng/ml (annual screening, costs and effects discounted at 3%)

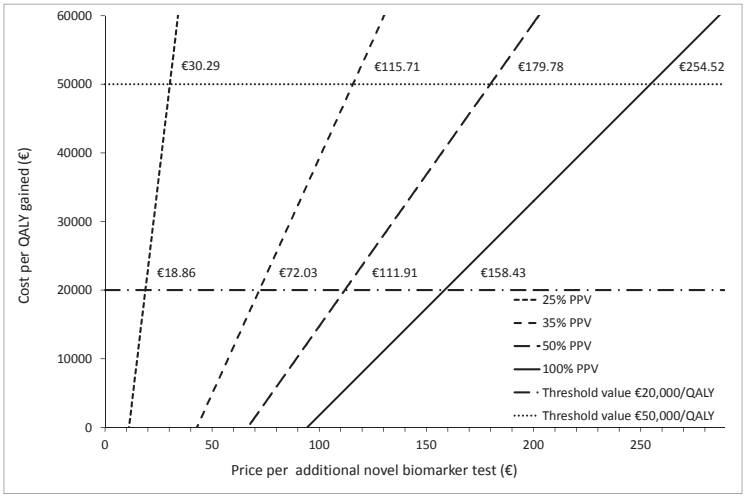


PPV: positive predictive value of PSA screening combined with the novel biomarker test; QALY = Quality Adjusted Life Year
Intersections with the horizontal dashed lines (i.e., the acceptability thresholds) represent the threshold price per novel biomarker test at which PSA combined with a novel biomarker will be a cost-effective alternative for PSA screening alone assuming an acceptability threshold of €20,000/QALY or €50,000/QALY, with 3% discounting

in men with an initial PSA >3.0ng/ml provides a better financial scope. Increasing specificity and PPV of PC testing from 22.7% to 100% implies that 5.5 times less biopsies are needed to find the same number of cancers. With screening every four years, this is a factor of 3.2. A range up to 100% PPV was chosen, although a novel biomarker test with 100% PVV is presumably not realistic.

The cost-effectiveness of new prostate cancer specific markers has been considered in a limited number of studies specifically for a screening setting [18]. A recent study investigated the Prostate Health Index (PHI), which combines PSA, free PSA, and a PSA precursor form [−2]proPSA. It was developed for use in the diagnosis of prostate cancer, and is more expensive than PSA testing alone. PHI was found to be cost-effective when used as an aid in distinguishing prostate cancer from benign prostatic conditions for men with a borderline PSA test result (e.g. PSA 2-10 ng/mL or 4-10 ng/mL) [18] compared to PC screening with repeated PSA testing for all men in the US.

Figure 4.4 Price threshold analyses for costs of an additional novel biomarker after an initial PSA test > 3.0 ng/ml (screening every 4 years, costs and effects discounted at 3%)



PPV: positive predictive value of PSA screening combined with the novel biomarker test; QALY = Quality Adjusted Life Year
Intersections with the horizontal dashed lines (i.e., the acceptability thresholds) represent the threshold price per novel biomarker test at which PSA combined with a novel biomarker will be a cost-effective alternative for PSA screening alone assuming an acceptability threshold of €20,000/QALY or €50,000/QALY, with 3% discounting

In our analysis, we first assumed that the novel biomarker test was done in all men and not opportunistically in men with borderline PSA values. Opportunistic testing based on PSA results would mean fewer novel biomarker tests to be carried out, and thus higher maximum costs per test may be acceptable. Our subsequent analyses assumed that a novel biomarker test was measured only after an initial PSA >3.0ng/ml and confirmed this notion. Further research is recommended to examine the impact of adding a novel biomarker in case of borderline PSA test results.

Studies on the added value of other recently described biomarkers have been contradictory. Prostate cancer antigen 3 (PCA3) and v-ets erythroblastosis virus E26 oncogene homolog (TMPRSS2-ERG) gene fusions are promising PC specific biomarkers that can be measured in urine [19, 20]. Other serum-based markers include percentage of free to total PSA (%fPSA), a kallikrein panel (4k-panel), and circulating tumour cells [7]. While these markers may have independent predictive value in addition to PSA, a new test based on these biomarkers is not likely to be cost-effective when used simultaneously with a PSA test due to costs exceeding the maximum threshold costs as presented in

this chapter. A resourceful way of using a novel biomarker test, for example in a range of borderline PSA test results, is then needed.

The reduction in unnecessary biopsies increases the QALYs saved with PC screening. Because a single biomarker with such good performance characteristics is not currently available, the results of the performed analyses represent thresholds for maximum costs of any novel biomarker with a given PPV. The results of this study can be used to judge the results of pilot studies of novel biomarkers [21, 22]; if those results suggest the biomarker has limited PPV due to low specificity, the biomarker would likely not be cost-effective [23, 24].

A strong point of this study is the systematic analysis based on the MISCAN prostate cancer model, which accurately simulates screening in a European situation. There are however some limitations. For the novel biomarker test, we only assume additional financial costs and no additional disadvantages. This is reasonable to assume when the novel biomarker test is blood-based and the novel biomarker test is performed within the same blood sample and in a short time frame. An extra blood or urine test would increase personnel and material costs, and performing an additional blood test may also increase the waiting time before a diagnosis is made. This is especially important to take into account when a novel biomarker test would be opportunistic, i.e. based on PSA results. Such disadvantages of additional testing imply less QALYs saved and a lower threshold for costs of the novel biomarker. Second, diagnostic tests for prostate cancer are nowadays increasingly based on magnetic resonance imaging (MRI). However, MRI targeted biopsy implemented in a prostate cancer screening setting is still in its infancy. Besides, adding MRI to PSA-based screening for prostate cancer is associated with many more assumptions in terms of personnel and material costs and QALY gain for patients.

Furthermore, there is a potential risk of misclassification due to the fact that ERSPC is based on sextant biopsy outcome. Using sextant biopsy for repeat screening has been studied before and the rate of deaths due to PC in men with an initial negative biopsy of 0.03% compared favourably to the 0.35% rate of overall PC mortality [25].

Although screening with a novel biomarker potentially decreases the number of biopsies needed drastically, this does not increase PC diagnoses due to the high sensitivity of the PSA test in our simulation model. Treatment costs after diagnosis remain unchanged in all strategies. Overtreatment of diagnosed PC will thus remain an important negative consequence of screening for PC.

In conclusion, a novel biomarker with high PPV may lead to a substantial reduction in unnecessary biopsies, thereby slightly increasing the QALYs saved by PC screening. Such a well-performing novel biomarker will only be cost-effective as an addition to screening with PSA alone under the condition of relatively low costs, or selective use of the novel biomarker test.

REFERENCES

1. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JW, Comber H, *et al.* Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. Eur J Cancer. 2013 Apr;49(6):1374-403.
2. Schroder FH, Hugosson J, Roobol MJ, Tammela TL, Zappa M, Nelen V, *et al.* Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. Lancet. 2014 Aug 6.
3. Heijnsdijk EA, Wever EM, Auvinen A, Hugosson J, Ciatto S, Nelen V, *et al.* Quality-of-life effects of prostate-specific antigen screening. N Engl J Med. 2012 Aug 16;367(7):595-605.
4. Heijnsdijk EA, de Carvalho TM, Auvinen A, Zappa M, Nelen V, Kwiatkowski M, *et al.* Cost-effectiveness of prostate cancer screening: a simulation study based on ERSPC data. J Natl Cancer Inst. 2015 Jan;107(1):366.
5. Schroder FH, Carter HB, Wolters T, van den Bergh RC, Gosselaar C, Bangma CH, *et al.* Early detection of prostate cancer in 2007. Part 1: PSA and PSA kinetics. Eur Urol. 2008 Mar; 53(3):468-77.
6. Draisma G, Boer R, Otto SJ, van der Cruijsen IW, Damhuis RAM, Schröder FH, *et al.* Lead times and overdetection due to prostate-specific antigen screening: Estimates from the European randomized study of screening for prostate cancer. Journal of the National Cancer Institute. 2003;95(12):868-78.
7. Febbo PG, Ladanyi M, Aldape KD, De Marzo AM, Hammond ME, Hayes DF, *et al.* NCCN Task Force report: Evaluating the clinical utility of tumor markers in oncology. J Natl Compr Canc Netw. 2011 Nov;9 Suppl 5:S1-32; quiz S3.
8. Draisma G, De Koning HJ. MISCAN: estimating lead-time and over-detection by simulation. BJU Int. 2003 Dec;92 Suppl 2:106-11.
9. Draisma G, Etzioni R, Tsodikov A, Mariotto A, Wever E, Gulati R, *et al.* Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context. J Natl Cancer Inst. 2009 Mar 18;101(6):374-83.
10. Heijnsdijk EA, der Kinderen A, Wever EM, Draisma G, Roobol MJ, de Koning HJ. Overdetection, overtreatment and costs in prostate-specific antigen screening for prostate cancer. Br J Cancer. 2009 Dec 1;101(11):1833-8.
11. Roobol MJ, Schroder FH. European Randomized Study of Screening for Prostate Cancer: achievements and presentation. BJU Int. 2003 Dec;92 Suppl 2:117-22.
12. Wolters T, Roobol MJ, Steyerberg EW, van den Bergh RC, Bangma CH, Hugosson J, *et al.* The effect of study arm on prostate cancer treatment in the large screening trial ERSPC. Int J Cancer. 2010 May 15;126(10):2387-93.
13. Postma R, Schroder FH, van Leenders GJ, Hoedemaeker RF, Vis AN, Roobol MJ, *et al.* Cancer detection and cancer characteristics in the European Randomized Study of Screening for Prostate Cancer (ERSPC)--Section Rotterdam. A comparison of two rounds of screening. Eur Urol. 2007 Jul;52(1):89-97.

14. Otto SJ, van der Cruijsen IW, Liem MK, Korfage IJ, Lous JJ, Schroder FH, *et al.* Effective PSA contamination in the Rotterdam section of the European Randomized Study of Screening for Prostate Cancer. Int J Cancer. 2003 Jun 20;105(3):394-9.

15. Siegel JE, Weinstein MC, Russell LB, Gold MR. Recommendations for reporting cost-effectiveness analyses. Panel on Cost-Effectiveness in Health and Medicine. JAMA. 1996 Oct 23-30;276(16):1339-41.

16. van den Akker-van Marle ME, van Ballegooijen M, van Oortmarssen GJ, Boer R, Habbema JD. Cost-effectiveness of cervical cancer screening: comparison of screening policies. J Natl Cancer Inst. 2002 Feb 6;94(3):193-204.

17. Eichler HG, Kong SX, Gerth WC, Mavros P, Jonsson B. Use of cost-effectiveness analysis in health-care resource allocation decision-making: how are cost-effectiveness thresholds expected to emerge? Value Health. 2004 Sep-Oct;7(5):518-28.

18. Nichol MB, Wu J, Huang J, Denham D, Frencher SK, Jacobsen SJ. Cost-effectiveness of Prostate Health Index for prostate cancer detection. BJU Int. 2012 Aug;110(3):353-62.

19. Leyten GH, Hessels D, Jannink SA, Smit FP, de Jong H, Cornel EB, *et al.* Prospective multicentre evaluation of PCA3 and TMPRSS2-ERG gene fusions as diagnostic and prognostic urinary biomarkers for prostate cancer. Eur Urol. 2014 Mar;65(3):534-42.

20. Vedder MM, de Bekker-Grob EW, Lilja HG, Vickers AJ, van Leenders GJ, Steyerberg EW, *et al.* The Added Value of Percentage of Free to Total Prostate-specific Antigen, PCA3, and a Kallikrein Panel to the ERSPC Risk Calculator for Prostate Cancer in Prescreened Men. Eur Urol. 2014 Aug 25.

21. Rutigliano MJ. Cost effectiveness analysis: a review. Neurosurgery. 1995 Sep;37(3):436-43; discussion 43-4.

22. Postmus D, de Graaf G, Hillege HL, Steyerberg EW, Buskens E. A method for the early health technology assessment of novel biomarker measurement in primary prevention programs. Stat Med. 2012 Oct 15;31(23):2733-44.

23. Oesterling JE. Using prostate-specific antigen to eliminate unnecessary diagnostic tests: significant worldwide economic implications. Urology. 1995 Sep;46(3 Suppl A):26-33.

24. Perlis RH. Translating biomarkers to clinical practice. Mol Psychiatry. 2011 Nov;16(11): 1076-87.

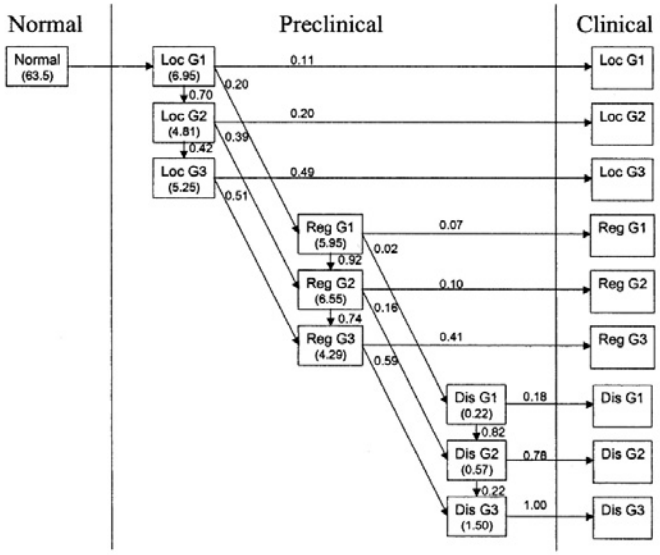
25. Schroder FH, van den Bergh RC, Wolters T, van Leeuwen PJ, Bangma CH, van der Kwast TH, *et al.* Eleven-year outcome of patients with prostate cancers diagnosed during screening after initial negative sextant biopsies. Eur Urol. 2010 Feb;57(2):256-66.

Supplementary table 4.1 MISCAN input data on: transition probabilities for prostate cancer stages and grades; Utility estimates and durations for various health states; and direct medical costs for prostate cancer screening diagnosis and primary therapy

State	Costs	Time period	Utility	Duration	Source
Screening attendance	**€24**		**0.99**	**1 week**	**(1,2)**
Invitation	€2				Estimation
Blood sample taking	€9.5				NZA
PSA determination	€12.5				NZA
Diagnosis	**€170**				
Biopsy	€92		0.90	3 weeks	NZA, (2)
PA research	€33				NZA
GP consulting	€45				20 min (tariff per hour €135.5)
Positive diagnosis			0.80	1 month	(3)
Primary therapy					
Staging	€200				Estimation
Radical prostatectomy	€11,800	At 2 mo after procedure	0.67	2 months	(4, 37, 38)
		At >2 mo to 1 yr after procedure	0.77	10 months	(5, 6)
Radiotherapy	€14,178	At 2 mo after procedure	0.73	2 months	(4, 7, 8)
		At >2 mo to 1 yr after procedure	0.78	10 months	(6, 9)
Active surveillance	€1,588	0.97	7 year	(10-13)	(10-13)
19 PSA tests	*€418*				
10 DRE	*€490*				Estimation €26 per test and 10 min.
Four biopsies	*€680*				
Follow-up	*€150*				
Post recovery period	**NA**		**0.95**	**9 year**	(4, 6)
Palliative therapy	**€12,276**		**0.60**	**30 months**	(14-19)
Terminal illness	**NA**		**0.40**	**6 months**	(14, 16, 17)

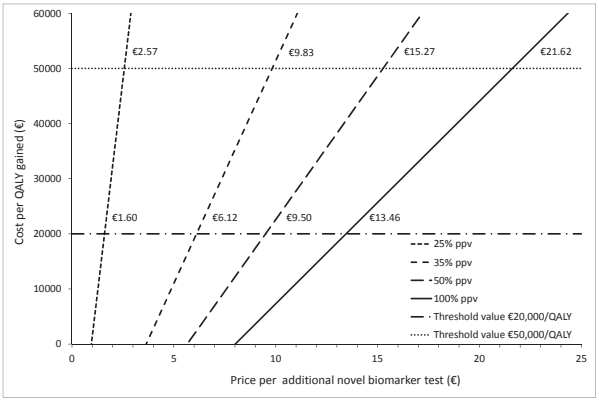
NZA = Dutch Healthcare Authority, data previously published in (20, 21).

Supplementary figure 4.1 The MISCAN model (adapted from [22])



The MISCAN model of the history of prostate cancer up to clinical diagnosis. The model distinguishes between cancer tumour–node–metastasis (TNM) stages (normal, localized [Loc: T1/2, N0/X, M0/X], regional [Reg: T3/4 or N+ and M0/X], and distant [Dis: any TN, M1]) and between differentiation grades (G1: Gleason score <7; G2: Gleason score = 7; and G3: Gleason score >7). Screening may detect cancer in one of the preclinical stages. The course of events may be interrupted by death from other causes. Key parameters of the basic model, fitted to all data, are indicated in the diagram. Transition probabilities are indicated next to the arrows; mean dwelling times in years are indicated in parentheses. Other parameters are cumulative incidence (the probability of ever getting prostate cancer) (0.19); Weibull shape parameters for dwelling times in the normal (10.7), localized (5.3), and regional or distant stages (5.0); sensitivities of the screening test for localized (0.64), regional (0.91), and distant stages (0.97); and contamination (30 tests per 1000 man-years). Time of death by causes not related to prostate cancer was obtained from the standard male life table (Statistics Netherlands, 1991–1995).

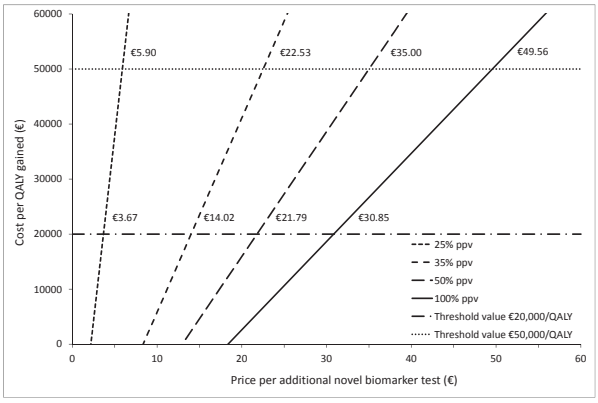
Supplementary figure 4.2 Impact of all-in price per additional novel biomarker on the cost-effectiveness of PSA screening combined with a novel biomarker versus screening with PSA alone (annual screening, costs and effects discounted at 5% from 2008)



PPV: positive predictive value of PSA screening combined with the novel biomarker test; QALY = Quality Adjusted Life Year

Intersections with the horizontal dashed lines (i.e., the acceptability thresholds) represent the threshold price per novel biomarker test at which PSA combined with a novel biomarker will be a cost-effective alternative for PSA screening alone assuming an acceptability threshold of €20,000/QALY or €50,000/QALY, with 5% discounting

Supplementary figure 4.3 Impact of all-in price per additional novel biomarker on the cost-effectiveness of PSA screening combined with a novel biomarker versus screening with PSA alone (screening every 4 years, costs and effects discounted at 5% from 2008)



PPV: positive predictive value of PSA screening combined with the novel biomarker test; QALY = Quality Adjusted Life Year

Intersections with the horizontal dashed lines (i.e., the acceptability thresholds) represent the threshold price per novel biomarker test at which PSA combined with a novel biomarker will be a cost-effective alternative for PSA screening alone assuming an acceptability threshold of €20,000/QALY or €50,000/QALY, with 5% discounting

REFERENCES

1. Essink-Bot ML, de Koning HJ, Nijs HG, Kirkels WJ, van der Maas PJ, Schroder FH. Short-term effects of population-based screening for prostate cancer on health-related quality of life. J Natl Cancer Inst. 1998 Jun 17;90(12):925-31.
2. de Haes JC, de Koning HJ, van Oortmarssen GJ, van Agt HM, de Bruyn AE, van Der Maas PJ. The impact of a breast cancer screening programme on quality-adjusted life-years. Int J Cancer. 1991 Oct 21;49(4):538-44.
3. Korfage IJ, de Koning HJ, Roobol M, Schroder FH, Essink-Bot ML. Prostate cancer diagnosis: the impact on patients' mental health. Eur J Cancer. 2006 Jan;42(2):165-70.
4. Stewart ST, Lenert L, Bhatnagar V, Kaplan RM. Utilities for prostate cancer health states in men aged 60 and older. Med Care. 2005 Apr;43(4):347-55.
5. Calvert NW, Morgan AB, Catto JW, Hamdy FC, Akehurst RL, Mouncey P, *et al.* Effectiveness and cost-effectiveness of prognostic markers in prostate cancer. Br J Cancer. 2003 Jan 13; 88(1):31-5.
6. Sanda MG, Dunn RL, Michalski J, Sandler HM, Northouse L, Hembroff L, *et al.* Quality of life and satisfaction with outcome among prostate-cancer survivors. N Engl J Med. 2008 Mar 20;358(12):1250-61.
7. Malmberg I, Persson U, Ask A, Tennvall J, Abrahamsson PA. Painful bone metastases in hormone-refractory prostate cancer: economic costs of strontium-89 and/or external radiotherapy. Urology. 1997 Nov;50(5):747-53.
8. Perez CA, Michalski J, Ballard S, Drzymala R, Kobeissi BJ, Lockett MA, *et al.* Cost benefit of emerging technology in localized carcinoma of the prostate. Int J Radiat Oncol Biol Phys. 1997 Nov 1;39(4):875-83.
9. Konski A, Sherman E, Krahn M, Bremner K, Beck JR, Watkins-Bruner D, *et al.* Economic analysis of a phase III clinical trial evaluating the addition of total androgen suppression to radiation versus radiation alone for locally advanced prostate cancer (Radiation Therapy Oncology Group protocol 86-10). Int J Radiat Oncol Biol Phys. 2005 Nov 1;63(3):788-94.
10. Bennett CL, Matchar D, McCrory D, McLeod DG, Crawford ED, Hillner BE. Cost-effective models for flutamide for prostate carcinoma patients: are they helpful to policy makers? Cancer. 1996 May 1;77(9):1854-61.
11. Zeliadt SB, Etzioni RD, Penson DF, Thompson IM, Ramsey SD. Lifetime implications and cost-effectiveness of using finasteride to prevent prostate cancer. Am J Med. 2005 Aug; 118(8):850-7.
12. Cooperberg MR, Carroll PR, Klotz L. Active surveillance for prostate cancer: progress and promise. J Clin Oncol. 2011 Sep 20;29(27):3669-76.
13. van den Bergh RC, Roemeling S, Roobol MJ, Roobol W, Schroder FH, Bangma CH. Prospective validation of active surveillance in prostate cancer: the PRIAS study. Eur Urol. 2007 Dec;52(6):1560-3.

14. Konski A, Watkins-Bruner D, Brereton H, Feigenberg S, Hanks G. Long-term hormone therapy and radiation is cost-effective for patients with locally advanced prostate carcinoma. Cancer. 2006 Jan 1;106(1):51-7.

15. Moeremans K, Caekelbergh K, Annemans L. Cost-effectiveness analysis of bicalutamide (Casodex) for adjuvant treatment of early prostate cancer. Value Health. 2004 Jul-Aug;7(4): 472-81.

16. Penson DF, Ramsey S, Veenstra D, Clarke L, Gandhi S, Hirsch M. The cost-effectiveness of combined androgen blockade with bicalutamide and luteinizing hormone releasing hormone agonist in men with metastatic prostate cancer. J Urol. 2005 Aug;174(2):547-52; discussion 52.

17. Ramsey S, Veenstra D, Clarke L, Gandhi S, Hirsch M, Penson D. Is combined androgen blockade with bicalutamide cost-effective compared with combined androgen blockade with flutamide? Urology. 2005 Oct;66(4):835-9.

18. Damber JE, Aus G. Prostate cancer. Lancet. 2008 May 17;371(9625):1710-21.

19. Beemsterboer PM, de Koning HJ, Birnie E, van der Maas PJ, Schroder FH. Advanced prostate cancer: course, care, and cost implications. Prostate. 1999 Jul 1;40(2):97-104.

20. Heijnsdijk EA, Wever EM, Auvinen A, Hugosson J, Ciatto S, Nelen V, *et al.* Quality-of-life effects of prostate-specific antigen screening. N Engl J Med. 2012 Aug 16;367(7):595-605.

21. Heijnsdijk EA, der Kinderen A, Wever EM, Draisma G, Roobol MJ, de Koning HJ. Overdetection, overtreatment and costs in prostate-specific antigen screening for prostate cancer. Br J Cancer. 2009 Dec 1;101(11):1833-8.

22. Draisma G, Boer R, Otto SJ, van der Cruijsen IW, Damhuis RAM, Schröder FH, *et al.* Lead times and overdetection due to prostate-specific antigen screening: Estimates from the European randomized study of screening for prostate cancer. Journal of the National Cancer Institute. 2003;95(12):868-78.

PART IV

Late evaluation

Chapter 5

The added value of percentage of free to total Prostate-Specific Antigen, PCA3, and a kallikrein panel to the ERSPC risk calculator for prostate cancer in prescreened men

Vedder MM, de Bekker-Grob EW, Lilja HG, Vickers AJ, van Leenders GJLH, Steyerberg EW, Roobol MJ

Eur Urol. 2014 Dec;66(6):1109-15.

ABSTRACT

Background: PSA testing has limited accuracy for early detection of prostate cancer (PCa).

Objective: To assess the added value of %freePSA, Prostate Cancer Antigen 3 (PCA3), and a kallikrein panel (4k-panel) to the European Randomized study of Screening for Prostate Cancer (ERSPC) multivariable prediction models: risk calculators (RCs) 4, including trans rectal ultrasound, and 4+DRE, for pre-screened men.

Design, setting, and participants: Participants were invited for rescreening between October 2007 and February 2009 within the Dutch part of the ERSPC study. Biopsies were taken in men with PSA level ≥3.0ng/ml or PCA3 score ≥10. Additional analyses of 4k-panel were done on serum samples.

Outcome measurements and statistical analysis: Outcome was defined as sextant biopsy detectable PCa. ROC curve and decision curve analyses were performed to compare the predictive capabilities of %freePSA, PCA3, 4k-panel, the ERSPC RCs, and their combinations in logistic regression models.

Results and limitations: PCa was detected in 119 out of 708 men. %freePSA did not perform better univariately or added to the RCs compared to the RCs alone. In 202 men with elevated PSA, the 4k-panel discriminated better than PCA3 when modelled univariately (AUC 0.78 vs. 0.62; p=0.01). The multivariable models with PCA3 or 4k-panel were equivalent (AUC of 0.80 for RC 4+DRE). In the total population, PCA3 discriminated better than 4k-panel (univariate AUC 0.63 vs. 0.56, p=0.05). There was no statistically significant difference between the multivariable model with PCA3 (AUC=0.73) vs. the model with 4k-panel (AUC=0.71, p=0.18). The multivariable model with PCA3 performed better than the reference model (0.73 vs. 0.70, p=0.02). Decision curves confirmed these patterns, although numbers were small.

Conclusion: Both PCA3 and, to a lesser extent, a 4k-panel have added value to the DRE based ERSPC RC in detecting PCa in pre-screened men.

5.1 INTRODUCTION

PSA testing is the mainstay of early detection of prostate cancer (PCa) [1]. However, PSA has limited specificity and sensitivity in determining the presence of prostate cancer, which leads to unnecessary biopsies and diagnosis of potentially indolent PCa [2, 3]. PSA-based multivariable prediction tools have been developed to improve the prediction of having a biopsy detectable PCa. Well known externally validated models are the European Randomized Study of Prostate Cancer (ERSPC) risk calculators (http://www.prostatecancer-riskcalculator.com/) [4], the Prostate Cancer Prevention Trial (PCPT) calculator (http://deb.uthscsa.edu/URORiskCalc/Pages/calcs.jsp) [5] and the Montreal Model [6]. The addition of new biomarkers to an existing prediction tool may increase the accuracy. Novel and promising markers in the field of PCa include the Prostate Cancer Specific Antigen 3 (PCA3), a non-coding mRNA, highly over-expressed in PCa tissue [7, 8] which can be assessed using urine obtained after digital rectal exam (DRE). A promising serum-based biomarker is the kallikrein panel (4k-panel), which consists of total-PSA, free-PSA, intact-PSA, and human-kallikrein-related peptidase-2 (hK2) [9, 10]. The 4k-panel has been shown to increase predictive capability as compared to PSA and DRE alone. In this study, we aimed to assess the added value of %freePSA, PCA3, and 4k-panel to the ERSPC risk calculators (RCs) for pre-screened men.

5.2 METHODS

Participants

Participants were recruited from the Dutch part of the ERSPC study [11, 12]. We included 965 men who were invited for rescreening (3rd, 4th or 5th time) between October 2007 and February 2009. The serum based PSA level and PCA3 were measured in all men. The PCA3 score is the ratio of PCA3:PSA mRNAs multiplied by 1,000 [8]. Men with a PSA level ≥3.0ng/ml and/or a PCA3 score ≥10 were invited to undergo a DRE, trans rectal ultrasound (TRUS) and a lateral sextant biopsy. We set the cut-off for PCA3 on ≥10 to evaluate performance characteristics of the PCA3 in comparison to a biopsy indication driven by PSA values of ≥3.0 ng/ml [13]. Assessed prostate volume was categorised with cut-points of <30 cc, 30-50 cc, and ≥50 cc [14]. In case of a hypoechogenic lesion, a seventh biopsy was taken. Permission for the present study (ISBN 978-90-5549-653-2) was granted by the Medical Ethics Committee, University Medical Center Rotterdam and the Dutch Ministry of Health.

Tests to predict PCa

The PSA test (Hybritech, Beckman Coulter Inc., Fulleron, CA, USA) was carried out in a standard fashion at the clinical laboratory of the Erasmus University Medical Center, the Netherlands. The PCA3 test (Progensa™, Gen-Probe Inc., San Diego, CA, USA) was done at the laboratory of experimental urology at Radboud University Nijmegen Medical Center. Measurements of the 4k-panel, consisting of four markers (total-PSA, free-PSA, intact-PSA, and human-kallikrein-related peptidase-2 (hK2)), were performed in the Department of Laboratory Medicine (Lund University, Malmo, Sweden) on stored serum samples [15]. Separate marker values as well as an overall 4k-panel predictor were derived using a pre-specified formula, i.e. the study is an independent validation of a previously specified model [9]. The formula was a mix of linear terms and non-linear spline transformations of the four markers. A specialised pathologist (GvL) handled the histologic examinations of the biopsy specimens.

Reference model

Two models from the ERSPC Rotterdam RCs (http://www.prostatecancer-riskcalculator.com/, RC4+DRE and RC4, including TRUS) were used as reference models:

1. RC 4+DRE: A model including total PSA (ng/ml), DRE (normal/abnormal), DRE assessed volume of the prostate (<30 cc, 30-50 cc, and ≥50 cc), and whether or not there was a previous (negative) biopsy;
2. RC4: A model including total PSA (ng/ml), DRE (normal/abnormal), TRUS (normal/abnormal), TRUS assessed prostate volume (ml) and a whether or not there was a previous (negative) biopsy.

Both models are used for men who have previously had PSA screening and a previous biopsy, if indicated according to the ERSPC Rotterdam screening algorithm [16]. It predicts the chance of a positive sextant biopsy and its degree of aggressiveness; the RC4+DRE model including information on prostate volume without the necessity of a TRUS [17].

Statistical analyses

The primary outcome measure was any form of PCa vs. no cancer, detected by a sextant biopsy, in men with elevated PSA levels (≥3.0ng/ml). Secondary, we assessed the predictive value of %freePSA, PCA3, and the 4k-panel in the total population and in the population with PSA<3.0ng/ml. We assessed the predictive value of %freePSA, PCA3, and the 4k-panel, using univariate and multivariable regression models. We refitted the original RCs: RC4 and RC4+DRE to use as the reference. We subsequently refitted the models including %freePSA, PCA3 and/or the 4k-panel. We used the area under the ROC curve (AUC) to quantify the predictive accuracy of five models: (i) the first reference model (RC 4+DRE), (ii) the reference model + PCA3, (iii) the reference model + 4k-panel, (iv) the reference model + PCA3 and the 4k-panel, and (v) the reference model + %freePSA.

We used the original RC4 (i.e. including information from TRUS) as the second reference model and used the likelihood ratio test for differences between models. We applied decision curve analysis (DCA) [18, 19] to evaluate the potential clinical usefulness of making decisions based on the models including the markers. We estimated net benefit (NB) for prediction models by summing the benefits (true positive biopsies) and subtracting the harms (false positive biopsies). The harms were weighted by a factor related to the relative harm of a missed cancer versus an unnecessary biopsy. This weighting was derived from the threshold probability (p_t) of PCa at which a patient would opt for a biopsy. This threshold can vary between men; we used a p_t between 0% and 40% [20]. The interpretation of a decision curve is straightforward; a model with the highest net benefit at a particular threshold should be chosen over alternative models. The net benefit was used to calculate for the reduction in numbers of biopsies per 100 men with a PSA level of ≥3.0ng/ml [9] and/or a PCA3 score ≥10. We used the following formula:

$$\text{reduction in biopsy per 100 men} = (\Delta NB/(p_t/(1-p_t)))*100.$$

Standard statistical software was used (SPSS v 18.0, SPSS Inc., Chicago, Ill; R version 2.15.2, R Foundation for Statistical Computing, Vienna, Austria; Stata v 12.0, StataCorp. 2011. College Station, TX: StataCorp LP).

5.3 RESULTS

Of 965 invited men, 721 (75%) underwent a biopsy. 163 (17%) men did not meet the PSA or PCA3 inclusion criteria, 39 (4%) could not have a biopsy because of contraindications, and 42 (4%) men refused biopsy. Records of 708 out of 721 (98%) biopsied participants were complete, including PCA3 and 4k-panel results.

These 708 men were invited for rescreening: 339 originated from the 3rd, 357 originated from the 4th and 12 originated from the 5th screening round. Participants were aged 64-75 years at time of the visit. A previous biopsy was taken from 206 (29%) of all men. PCa was found in 119 (17%) of the 708 biopsied men, of which 40 in the 202 men with elevated PSA levels (table 5.1). Few men had an abnormal TRUS or DRE. Of 708 men, 503 had a PCA3 score ≥10 and a PSA score <3.0 ng/ml. Total PSA and PCA3 levels differed significantly between men with and without PCa (table 5.1).

In men with PSA levels ≥3.0ng/ml the 4k-panel had a higher AUC value as compared to PCA3 when studied univariately (AUC 0.78 vs. 0.62, p=0.01; table 5.2; supplementary

Table 5.1 Characteristics of men rescreened in the ERSPC trial

	PSA ≥3.0ng/ml (N=202)					Total set (N=708)				
	No Cancer N=162 (80%)		Cancer N=40 (20%)		P-value	No Cancer N=589 (83%)		Cancer N=119 (17%)		P-value
Age[1]	70.3	(68.1;72.7)	70.2	(68.6;72.4)	0.98	70.3	(68.1;72.5)	70.3	(68.4;72.3)	0.97
Previous Biopsy					<0.01					<0.01
No	41	25%	26	65%		403	68%	99	83%	
Yes	121	75%	14	35%		186	32%	20	17%	
Total PSA (ng/ml)	4.6	(3.7;6.4)	4.4	(3.6;6.9)	0.95	1.7	(0.9;3.2)	2.1	(1.4;3.7)	<0.01
DRE[3]					0.51					<0.01
Normal	133	82%	31	77.5%		504	86%	88	74%	
Abnormal	29	18%	9	22.5%		85	14%	31	26%	
Volume classes DRE					0.03					0.53
<30 cc	9	6%	6	15%		115	20%	23	19%	
30-50 cc	51	31%	17	42.5%		263	45%	60	50%	
≥50 cc	102	63%	17	42.5%		204	35%	36	30%	
TRUS[4]					0.85					0.38
Normal	155	96%	38	95%		573	97%	114	96%	
Abnormal	7	4%	2	5%		16	3%	5	4%	
4k-panel										
Free PSA	1.14	(0.86;1.62)	0.93	(0.68;1.39)	0.02	0.47	(0.28;0.84)	0.56	(0.39;0.86)	0.06
Intact PSA	0.42	(0.32;0.60)	0.40	(0.25;0.58)	0.40	0.20	(0.12;0.34)	0.23	(0.16;0.39)	0.04
hK2[5]	0.05	(0.04;0.07)	0.05	(0.04;0.07)	1.00	0.03	(0.02;0.05)	0.04	(0.03;0.05)	<0.01
4k-panel score	-2.81	(-3.37;-2.18)	-1.69	(-2.45;-1.09)	<0.01	-1.33	(-2.27;-0.98)	-1.28	(-1.76;-0.97)	0.04
Probability 4k-panel	0.06	(0.03;0.10)	0.16	(0.08;0.25)	<0.01	0.21	(0.09;0.27)	0.22	(0.15;0.28)	0.04
PCA3 score[6]	29.5	(14.0;57.5)	44.0	(20.0;118.3)	0.01	31.0	(18.0;58.5)	46.0	(28.0;97.0)	<0.01
Stage										
T1C			31	78%				87	73%	
T2A			8	20%				28	24%	
T2B			1	3%				2	2%	
T2C			0	0%				1	1%	
T3A			0	0%				1	1%	
Grade										
Gleason 6			31	78%				99	83%	
Gleason 7			5	13%				13	11%	
Gleason 8			3	8%				5	4%	
Gleason 9			1	3%				2	2%	
Serious cancer[2]			9	23%				22	18%	

[1] Continuous variables are noted as median (interquartile range)
[2] Nominal variables are noted as number and percentage
[3] DRE = digital rectal exam
[4] TRUS = Trans rectal ultrasound
[5] hK2 = kallikrein protein 2
[6] PCA3 score = the ratio of PCA3: PSA mRNAs multiplied by 1,000

Table 5.2 Incremental enhancement in discrimination for the subgroup of 202 men rescreened in the ERSPC trial with PSA ≥3.0ng/ml

	Univariate	Added to original risk calculator 4[1]	Added to original risk calculator 4+DRE[2]
	C[3] (95% CI)	C (95% CI)	C (95% CI)
Reference value[4]	0.53 (0.44-0.64)	0.78 (0.69-0.86)	0.76 (0.68-0.83)
Kallikrein panel	0.78 (0.69-0.85)	0.80 (0.71-0.87)	0.79 (0.71-0.86)
PCA3	0.62 (0.52-0.73)	0.80 (0.71-0.87)	0.78 (0.70-0.85)
Kallikrein panel AND PCA3	0.75 (0.65-0.84)	0.81 (0.72-0.88)	0.80 (0.72-0.87)
%freePSA	0.65 (0.55-0.75)	0.80 (0.71-0.88)	0.79 (0.71-0.85)

[1] A model including total PSA (ng/ml), DRE (normal/abnormal), assessed DRE volume of the prostate (<30 cc, 30-50 cc, and ≥50 cc)

[2] A model including total PSA (ng/ml), DRE (normal/abnormal), TRUS (normal/abnormal), and TRUS assessed prostate volume (ml)

[3] Area under the receiver operator curve

[4] The reference value for the univariate analysis is total PSA (ng/ml) and DRE (normal/abnormal), for the multivariate analyses it is the original risk calculator

figures.). The multivariable models with PCA3 or 4k-panel were equivalent (AUC 0.80 for RC 4+DRE, 0.78 vs. 0.79 for RC 4 with PCA3 and the 4k-panel respectively).

In the total population, PCA3 discriminated better than the 4k-panel (univariate AUC 0.63 vs. 0.56, p=0.05, table 5.3). There was no statistically significant difference between the multivariable model with PCA3 (AUC=0.73) vs. the model with 4k-panel (AUC=0.71, p=0.18). The multivariable model with PCA3 performed better than the reference model (0.73 vs. 0.70, p=0.02). A multivariable model with both markers did not perform better than the multivariable model with PCA3 alone (AUC 0.73 vs. 0.73) in the total dataset. %freePSA did not perform better univariately or added to the RCs compared to the RCs alone in the total population (table 5.3).

Analyses in men with PSA levels <3.0 ng/ml showed no value for the 4k-panel, but some added value of PCA3 (univariate AUC 0.64 (0.58-0.70), AUC 0.70 vs. 0.66 when added to the reference models, p=0.01 for RC4 and p<0.01 for RC4+DRE) (see supplementary table 5.1).

In men with elevated PSA levels, the net benefits of all models were higher than in the total dataset (figure 5.1). In this subgroup the use of a model was clinically useful from a threshold of 5%. The reduction in biopsies per 100 men differed between a threshold of 10 to 30% in the total dataset, in favour of the multivariable model with PCA3 and PCA4 + 4k-panel. In the subgroup of men with elevated PSA, different models were in

Table 5.3 Incremental enhancement in discrimination in 708 men rescreened in the ERSPC trial

	Univariate	Added to original risk calculator 4[1]	Added to original risk calculator 4+DRE[2]
	C[3] (95% CI)	C (95% CI)	C (95% CI)
Reference value[4]	0.61 (0.56-0.67)	0.70 (0.64-0.75)	0.70 (0.64-0.75)
Kallikrein panel	0.56 (0.50-0.61)	0.71 (0.65-0.76)	0.71 (0.65-0.76)
PCA3	0.63 (0.58-0.69)	0.73 (0.67-0.78)	0.73 (0.67-0.77)
Kallikrein panel AND PCA3	0.66 (0.61-0.70)	0.73 (0.68-0.78)	0.73 (0.68-0.78)
%freePSA	0.57 (0.51-0.63)	0.70 (0.65-0.76)	0.70 (0.64-0.75)

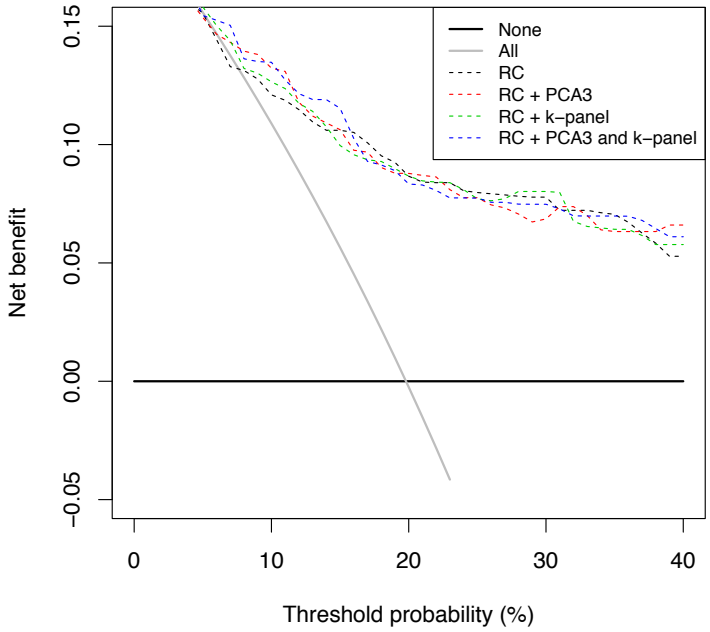
[1] A model including total PSA (ng/ml), DRE (normal/abnormal), assessed DRE volume of the prostate ((<30 cc, 30-50 cc, and ≥50 cc)

[2] A model including total PSA (ng/ml), DRE (normal/abnormal), TRUS (normal/abnormal), and TRUS assessed prostate volume (ml)

[3] Area under the receiver operator curve

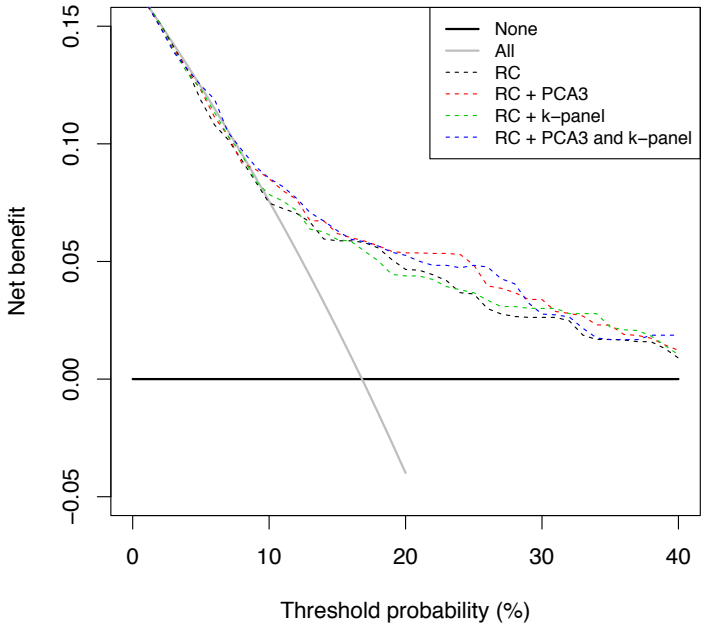
[4] The reference value for the univariate analysis is total PSA (ng/ml) and DRE (normal/abnormal), for the multivariate analyses it is the original risk calculator

Figure 5.1 Net benefit of prediction models with PCA3 and/or the 4k-panel in the subgroup of men with PSA ≥3.0ng/ml (N=202)



favour depending on the specific threshold, which also reflected the low number of PCa cases at these thresholds (figure 5.2).

Figure 5.2 Net benefit of prediction models with PCA3 and/or the 4k-panel in all men (N=708)



The prediction models had added value over biopsy in all men if the threshold for performing a biopsy exceeded 9% (figure 5.1 and figure 5.2). Between thresholds of 9 and 40% the multivariable model with PCA3 orPCA3 + 4k-panel had the highest net benefit and performed better than the reference model at all thresholds. With a cut-point of PSA ≥3.0 ng/m and PCA3>10, reduction in the number of biopsies per 1000 men at a threshold probability of 12.5% was 89 when PCA3 was added, 50 when the 4k-panel was added, and 124 when both the PCA3 and the 4k-panel marker were added to the original RC. At a threshold probability of 20%, there was a reduction of 11 biopsies per 1000 men when PCA3 was added to the original RC, and 7 per 1000 men when both PCA3 and the 4k-panel were added. In contrast, no reduction in the number of biopsies was noted in men with PSA level ≥3.0 ng/ml. Results were similar for each of the considered reference models (RC4 with DRE or RC4 with TRUS, data not shown).

5.4 DISCUSSION

In the current study, adding the 4k-panel to a previously developed PCa risk prediction model increased the predictive value in participants with PSA ≥3.0ng/ml. Adding PCA3 to the previously developed PCa risk prediction model increased the AUC in prescreened men regardless of total PSA level at time of biopsy. This was equally seen in

reference models with and without the inclusion of TRUS and TRUS assessed volume. Therefore, we advise for the model with DRE to estimate prostate volume.

In the past, %freePSA has been shown to significantly increase the accuracy of DRE and total PSA [21]. Its limited cost and wide availability in labs that run total PSA values are attractive attributes for clinical use. We found very limited predictive value of %freePSA alone or combined with the RCs.

The usefulness of PCA3 testing for the detection of PCa and possible reduction of unnecessary biopsies has been shown before [22, 23]. These studies assessed the added value of PCA3 after selecting men for biopsy solely on the basis of a PSA cut-off level. This implies that PCa in men with PSA values below the threshold will be missed. In addition, assessing the added value of PCA3 in men with a previous negative biopsy, initially selected on the basis of an elevated PSA level, is by definition biased. The benefit from PCA3 as compared to PSA is then overoptimistic. To overcome this attribution bias in the current study, men with a PCA3 score ≥10 were biopsied, even if their PSA level was <3.0 ng/ml [13, 24].

Predictions based on the 4k-panel did not differ significantly between cancer and non-cancer cases in the total study group, while some markers such as intact-PSA and Hk2 did differ. In the subgroup analyses of men with PSA level ≥3.0, the PCA3 and 4k-panel scores differed significantly between men with and without PCa, whereas intact-PSA and hK2 did not (table 5.1). Free-PSA differed significantly among those in the subgroup men with PSA level ≥3.0. Free PSA may hence be the most relevant element in the 4k-panel for rescreened men with elevated PSA levels.

The 4k-panel is developed in men with elevated PSA levels and has up to now only been tested in that particular but clinically most relevant setting. Previous studies showed that predictions based on levels of four kallikrein markers in blood distinguish between pathologically insignificant and aggressive PCa with good accuracy [15, 25]. We confirmed these results with an increase in predictive capability in addition to a risk prediction model that already had an AUC ≥0.7, albeit in a relatively low number of patients.

With respect to cost-effectiveness, data suitable for a direct comparison with our study are scarce. While data on the cost effectiveness of PCA3 are weak [26], another comparable but cheaper combination of serum-based sub forms of PSA, the Prostate Health Index (PHI) has been found to be cost-effective for screening purposes [27]. For the current study, we assessed cost-effectiveness with arbitrarily assumed costs

for the PCA3 test and for prostate biopsy (€300 and €249, [28]). The 4k-panel is not commonly available, and may be cheaper than a PCA3 test [9]. When adding PCA3 and/or the 4k-panel to previously developed PCa risk prediction model, less biopsies are needed to find the same amount of cancers (increased net benefit, figure 5.1 and figure 5.2). However, this did not result into a substantial reduction in prostate biopsies as compared to the original RCs alone for p_ts between 0 and 40%, making it very unlikely that the extended risk model will be cost-effective.

One limitation of this study was the pre-screened nature of our study cohort. Therefore we compared the performance of models with PCA3 or the 4k-panel to reference models developed for pre-screened men, allowing for a fair comparison. This, and the fact that all men were from the Netherlands, may affect external validity. However, elevated PCA3 scores have particularly been demonstrated to increase the probability of a positive repeat biopsy in men with a prior negative biopsy result, independent of PSA [29, 30].

Another limitation of this study is the small number of men included, specifically men with PSA ≥ 3.0ng/ml. The relative utility of PCA3 and the 4k-panel need to be confirmed. The number of serious cancers was low (N=22, of which 9 in men with PSA levels ≥3.0ng/ml), limiting separate analyses for this group of patients. In men with PSA ≥3.0ng/ml (N=202, of whom 40 had cancer), we used the original RC consisting of 4 variables and extended this with 1 or 2 variables – giving an events per variable (EPV) ratio of 8 or 6.7 – which could lead to overfitting of the model. Ideally the EPV would be higher, but EPV values from 5 have been shown to be valid in the context of statistical adjustment for baseline risk factors [31].

We used sextant biopsying in a repeat screening setting and found a 17% cancer detection rate (N=119), and it is likely that we missed some cases. Even using sextant biopsy for repeat screening, deaths due to PC occurred at a rate of only 0.03%, compared to 0.35% overall [32].

5.5 CONCLUSION

Both the PCA3 and, to a lesser extent, a 4k-panel have added value in detecting PCa to the DRE based ERSPC Rotterdam RC for pre-screened men. Further validation is however needed, and should focus on biomarkers capable of identifying men at elevated risk for potentially aggressive PCa. This is most relevant for men with a previous negative biopsy, where such markers may especially be useful.

REFERENCES

1. Heidenreich A, Bellmunt J, Bolla M, Joniau S, Mason M, Matveev V, *et al.* EAU guidelines on prostate cancer. Part 1: screening, diagnosis, and treatment of clinically localised disease. Eur Urol. 2011 Jan;59(1):61-71.
2. Draisma G, Boer R, Otto SJ, van der Cruijsen IW, Damhuis RA, Schroder FH, *et al.* Lead times and overdetection due to prostate-specific antigen screening: estimates from the European Randomized Study of Screening for Prostate Cancer. J Natl Cancer Inst. 2003 Jun 18;95(12): 868-78.
3. Heijnsdijk EA, der Kinderen A, Wever EM, Draisma G, Roobol MJ, de Koning HJ. Overdetection, overtreatment and costs in prostate-specific antigen screening for prostate cancer. Br J Cancer. 2009 Dec 1;101(11):1833-8.
4. 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. J Urol. 2007 Jan;177(1):107-12; discussion 12.
5. Thompson IM, Ankerst DP, Chi C, Goodman PJ, Tangen CM, Lucia MS, *et al.* Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial. J Natl Cancer Inst. 2006 Apr 19;98(8):529-34.
6. Karakiewicz PI, Benayoun S, Kattan MW, Perrotte P, Valiquette L, Scardino PT, *et al.* Development and validation of a nomogram predicting the outcome of prostate biopsy based on patient age, digital rectal examination and serum prostate specific antigen. J Urol. 2005 Jun;173(6):1930-4.
7. Bussemakers MJ, van Bokhoven A, Verhaegh GW, Smit FP, Karthaus HF, Schalken JA, *et al.* DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. Cancer Res. 1999 Dec 1;59(23):5975-9.
8. Hessels D, Schalken JA. The use of PCA3 in the diagnosis of prostate cancer. Nat Rev Urol. 2009 May;6(5):255-61.
9. Vickers AJ, Cronin AM, Aus G, Pihl CG, Becker C, Pettersson K, *et al.* A panel of kallikrein markers can reduce unnecessary biopsy for prostate cancer: data from the European Randomized Study of Prostate Cancer Screening in Goteborg, Sweden. BMC Med. 2008;6:19.
10. Gupta A, Roobol MJ, Savage CJ, Peltola M, Pettersson K, Scardino PT, *et al.* A four-kallikrein panel for the prediction of repeat prostate biopsy: data from the European Randomized Study of Prostate Cancer screening in Rotterdam, Netherlands. Br J Cancer. 2010 Aug 24; 103(5):708-14.
11. Schroder FH, Denis LJ, Roobol M, Nelen V, Auvinen A, Tammela T, *et al.* The story of the European Randomized Study of Screening for Prostate Cancer. BJU Int. 2003 Dec;92 Suppl 2:1-13.
12. Schroder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, *et al.* Screening and prostate-cancer mortality in a randomized European study. N Engl J Med. 2009 Mar 26; 360(13):1320-8.

13. Roobol MJ, Schroder FH, van Leeuwen P, Wolters T, van den Bergh RC, van Leenders GJ, *et al.* Performance of the prostate cancer antigen 3 (PCA3) gene and prostate-specific antigen in prescreened men: exploring the value of PCA3 for a first-line diagnostic test. Eur Urol. 2010 Oct;58(4):475-81.

14. Roobol MJ, van Vugt HA, Loeb S, Zhu X, Bul M, Bangma CH, *et al.* Prediction of prostate cancer risk: the role of prostate volume and digital rectal examination in the ERSPC risk calculators. Eur Urol. 2012 Mar;61(3):577-83.

15. Vickers A, Cronin A, Roobol M, Savage C, Peltola M, Pettersson K, *et al.* Reducing unnecessary biopsy during prostate cancer screening using a four-kallikrein panel: an independent replication. J Clin Oncol. 2010 May 20;28(15):2493-8.

16. Roobol MJ, Zhu X, Schroder FH, van Leenders GJ, van Schaik RH, Bangma CH, *et al.* A Calculator for Prostate Cancer Risk 4 Years After an Initially Negative Screen: Findings from ERSPC Rotterdam. Eur Urol. 2013 Apr;63(4):627-33.

17. Roobol MJ, Schroder FH, Kranse R, Erspc R. A comparison of first and repeat (four years later) prostate cancer screening in a randomized cohort of a symptomatic men aged 55-75 years using a biopsy indication of 3.0 ng/ml (results of ERSPC, Rotterdam). Prostate. 2006 May 1;66(6):604-12.

18. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006 Nov-Dec;26(6):565-74.

19. 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 Jan;21(1):128-38.

20. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak. 2008;8:53.

21. Steuber T, Vickers A, Haese A, Kattan MW, Eastham JA, Scardino PT, *et al.* Free PSA isoforms and intact and cleaved forms of urokinase plasminogen activator receptor in serum improve selection of patients for prostate cancer biopsy. Int J Cancer. 2007 Apr 1;120(7):1499-504.

22. Auprich M, Haese A, Walz J, Pummer K, de la Taille A, Graefen M, *et al.* External validation of urinary PCA3-based nomograms to individually predict prostate biopsy outcome. Eur Urol. 2010 Nov;58(5):727-32.

23. Auprich M, Chun FK, Ward JF, Pummer K, Babaian R, Augustin H, *et al.* Critical assessment of preoperative urinary prostate cancer antigen 3 on the accuracy of prostate cancer staging. Eur Urol. 2011 Jan;59(1):96-105.

24. Roobol MJ, Schroder FH, van Leenders GL, Hessels D, van den Bergh RC, Wolters T, *et al.* Performance of prostate cancer antigen 3 (PCA3) and prostate-specific antigen in Prescreened men: reproducibility and detection characteristics for prostate cancer patients with high PCA3 scores (>/= 100). Eur Urol. 2010 Dec;58(6):893-9.

25. Carlsson S, Maschino A, Schroder F, Bangma C, Steyerberg EW, van der Kwast T, *et al.* Predictive Value of Four Kallikrein Markers for Pathologically Insignificant Compared With Aggressive Prostate Cancer in Radical Prostatectomy Specimens: Results From the European Randomized Study of Screening for Prostate Cancer Section Rotterdam. Eur Urol. 2013 May 2.

26. Malavaud B, Cussenot O, Mottet N, Rozet F, Ruffion A, Smets L, *et al.* Impact of adoption of a decision algorithm including PCA3 for repeat biopsy on the costs for prostate cancer diagnosis in France. J Med Econ. 2013;16(3):358-63.

27. Nichol MB, Wu J, Huang J, Denham D, Frencher SK, Jacobsen SJ. Cost-effectiveness of Prostate Health Index for prostate cancer detection. BJU Int. 2012 Aug;110(3):353-62.

28. Fandella A. Analysis of costs of transrectal prostate biopsy. Urologia. 2011 Oct-Dec;78(4): 288-92.

29. Haese A, de la Taille A, van Poppel H, Marberger M, Stenzl A, Mulders PF, *et al.* Clinical utility of the PCA3 urine assay in European men scheduled for repeat biopsy. Eur Urol. 2008 Nov; 54(5):1081-8.

30. Gittelman M, Hertzman B, Bailen J, Williams T, Koziol I, Henderson RJ, *et al.* PROGENSA(R) PCA3 molecular urine test as a predictor of repeat prostate biopsy outcome in men with previous negative biopsies: A prospective multicenter clinical study. J Urol. 2013 Feb 14.

31. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol. 2007 Mar 15;165(6):710-8.

32. Schroder FH, van den Bergh RC, Wolters T, van Leeuwen PJ, Bangma CH, van der Kwast TH, *et al.* Eleven-year outcome of patients with prostate cancers diagnosed during screening after initial negative sextant biopsies. Eur Urol. 2010 Feb;57(2):256-66.

SUPPLEMENTARY TABLES AND FIGURES

Supplementary table 5.1 Incremental enhancement in discrimination in 506 men rescreened in the ERSPC trial with PSA <3.0ng/ml

	Univariate	Added to original risk calculator 4[1]	Added to original risk calculator 4+DRE[2]
	C[3] (95% CI)	C (95% CI)	C (95% CI)
Reference value[4]	0.63 (0.56-0.69)	0.66 (0.59-0.73)	0.66 (0.58-0.73)
Kallikrein panel	0.50 (0.43-0.56)	0.66 (0.59-0.73)	0.66 (0.59-0.73)
PCA3	0.64 (0.58-0.70)	0.70 (0.62-0.76)	0.70 (0.63-0.77)
Kallikrein panel AND PCA3	0.63 (0.57-0.69)	0.70 (0.63-0.76)	0.70 (0.64-0.77)

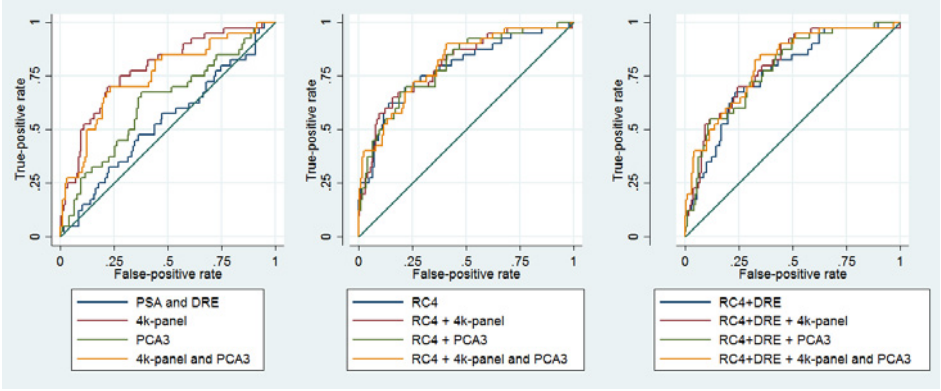
[1] A model including total PSA (ng/ml), DRE (normal/abnormal), assessed DRE volume of the prostate ((<30 cc, 30-50 cc, and ≥50 cc)

[2] A model including total PSA (ng/ml), DRE (normal/abnormal), TRUS (normal/abnormal), and TRUS assessed prostate volume (ml)

[3] Area under the receiver operator curve

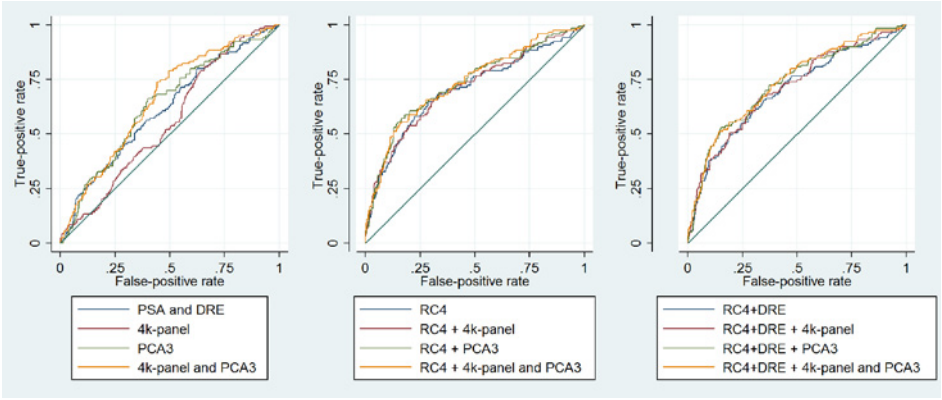
[4] The reference value for the univariate analysis is total PSA (ng/ml) and DRE (normal/abnormal), for the multivariate analyses it is the original risk calculator

Supplementary figure 5.1A-C ROC curves for the subgroup of 202 men rescreened in the ERSPC trial with PSA ≥3.0ng/ml (table 5.2).



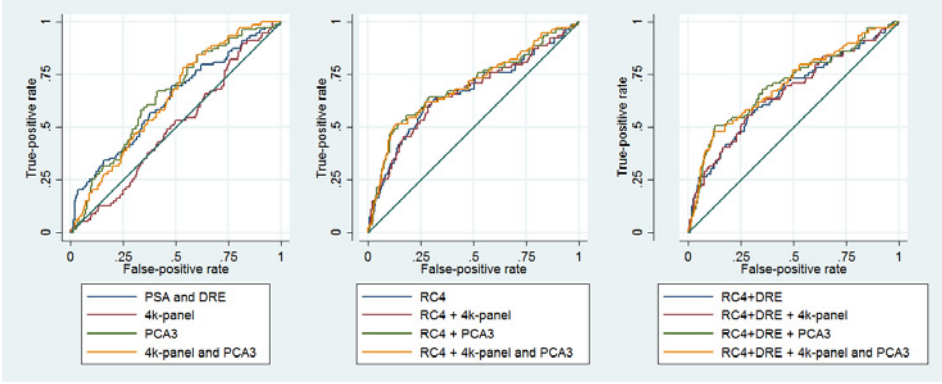
A. Univariate analysis, with PSA (ng/ml) and DRE (normal/abnormal) as a reference
B. Multivariate analysis, with risk calculator 4, a model including total PSA (ng/ml), DRE (normal/abnormal), TRUS (normal/abnormal), and TRUS assessed prostate volume (ml), as a reference
C. Multivariate analysis, with risk calculator 4+DRE, a model including total PSA (ng/ml), DRE (normal/abnormal), assessed DRE volume of the prostate ((<30 cc, 30-50 cc, and ≥50 cc), as a reference

Supplementary figure 5.2A-C ROC curves for the subgroup of 708 men rescreened in the ERSPC trial (table 5.3).



A. Univariate analysis, with PSA (ng/ml) and DRE (normal/abnormal) as a reference
B. Multivariate analysis, with risk calculator 4, a model including total PSA (ng/ml), DRE (normal/abnormal), TRUS (normal/abnormal), and TRUS assessed prostate volume (ml), as a reference
C. Multivariate analysis, with risk calculator 4+DRE, a model including total PSA (ng/ml), DRE (normal/abnormal), assessed DRE volume of the prostate ((<30 cc, 30-50 cc, and ≥50 cc), as a reference

Supplementary figure 5.3A-C ROC curves for the subgroup of 506 men rescreened in the ERSPC trial with PSA <3.0ng/ml (supplementary table 5.1).



A. Univariate analysis, with PSA (ng/ml) and DRE (normal/abnormal) as a reference
B. Multivariate analysis, with risk calculator 4, a model including total PSA (ng/ml), DRE (normal/abnormal), TRUS (normal/abnormal), and TRUS assessed prostate volume (ml), as a reference
C. Multivariate analysis, with risk calculator 4+DRE, a model including total PSA (ng/ml), DRE (normal/abnormal), assessed DRE volume of the prostate ((<30 cc, 30-50 cc, and ≥50 cc), as a reference

Chapter 6

Comparison of two prostate cancer risk calculators that include the Prostate Health Index (PHI)

Vedder MM*, Roobol MJ*, Nieboer D, Houlgatte A, Vincendeau S, Lazzeri M, Guazzoni G, Stephan C, Semjonow A, Haese A, Graefen M, Steyerberg EW

*These authors contributed equally to this work

Eur Urol Focus. 2015;1(2):185-190

ABSTRACT

Background: Risk prediction models for prostate cancer (PC) have become important tools for reducing unnecessary prostate biopsies. The Prostate Health Index (PHI) may increase their predictive accuracy .

Objectives: To compare two PC risk calculators (RCs) that include PHI.

Design, setting, and participants: We evaluated the predictive performance of a previously developed PHI based nomogram and updated versions of the European Randomized study of Screening for Prostate Cancer (ERSPC) digital rectal examination (DRE) based RCs: #3 (no prior biopsy) and #4 (having had a prior biopsy). For the ERSPC updates, the original RCs were recalibrated and PHI added as a predictor. The PHI-updated ERSPC RCs were compared with the Lughezzani nomogram in 1185 men from four European sites. Outcomes were biopsy detectable PC and potentially advanced or aggressive PC defined as a clinical stage >T2B and/or a Gleason score ≥7 ('clinically relevant PC').

Results and limitation: The PHI-updated ERSPC models had a combined AUC of 0.72 for all cancer and 0.68 for clinically relevant cancer. For the Lughezzani model, AUCs were 0.75 for all cancer and 0.69 for clinically relevant cancer. For men without a prior biopsy, the PHI-updated RC#3 resulted in AUCs of 0.73 predicting PC and 0.66 for predicting clinically relevant PC. Decision curves confirmed these patterns although numbers for clinically relevant cancer were small.

Conclusion: Differences between RCs that include PHI are small. Considering PHI in an RC leads to further reductions in rates of unnecessary biopsies as compared to a PSA based strategy.

6.1 INTRODUCTION

Prostate cancer (PC) is the most common form of cancer in men in Europe [1]. Prostate Specific Antigen (PSA) testing is the mainstay of early detection of PC [2]. However, PSA has a limited specificity in determining the presence of PC, which leads to unnecessary biopsies and the diagnosis of potentially indolent PC [3, 4]. A prostate biopsy is an invasive procedure, and apart from costs and anxiety, not without risk of complications [5].

PSA based multivariable prediction tools have been developed to improve the prediction of having a biopsy detectable PC. Well known and externally validated models include the European Randomized Study of Prostate Cancer (ERSPC) risk calculators (RCs) (http://www.prostatecancer-riskcalculator.com/) [6], the Prostate Cancer Prevention Trial (PCPT) calculator (http://deb.uthscsa.edu/URORiskCalc/Pages/calcs.jsp) [7] and the Montreal model [8]. Risk prediction models have become an important tool for reducing unnecessary prostate biopsies [9]. The addition of new biomarkers to an existing prediction tool may increase the accuracy. Novel and promising markers in the field of PC include the Prostate Health Index (PHI), based on data on total PSA (tPSA), free PSA (fPSA), and -2proPSA (p2PSA). PHI has been approved for use by the US Food and Drug Administration (http://www.accessdata.fda.gov/cdrh_docs/pdf9/p090026a.pdf).

Lughezzani *et al.* previously developed and validated a nomogram including the Prostate Health Index (PHI) [10]. We aimed to compare prostate cancer RCs that include PHI, the Lughezzani model and PHI-updated digital rectal examination (DRE) based ERSPC models.

6.2 PATIENTS AND METHODS

Participants

We studied 1185 men from four sites in Europe (Paris, Rennes, Hamburg and Muenster). Data on tPSA, fPSA, p2PSA, PHI, DRE, prostate volume and biopsy outcome (PC detected yes/no) were collected in all men. Participants in this study underwent a biopsy according to the standard clinical practice routinely used at each participating site, which was a ≥10 core biopsy. We calculated PHI using the equation: (p2PSA/fPSA)*√tPSA [11]. tPSA was between 2.0-10.0 ng/ml (Beckman-Coulter Access Hybritech assay, corresponding to 1.6-8.0 ng/ml according to WHO calibration). Outcomes were a sextant biopsy detectable PC and potentially advanced or aggressive PC (defined as a clinical stage >T2B and/or a biopsy Gleason score ≥7, 'clinically relevant PC').

Nomograms

Lughezzani *et al.* developed a PHI based nomogram including age (years), DRE (normal/abnormal), prior biopsy (yes/no), TRUS measured prostate volume (ml), and PHI [10]. Two models from the total of 8 European Randomized study of Screening for Prostate Cancer (ERSPC) Rotterdam prostate cancer risk calculators (RCs) (http://www.prostatecancer-riskcalculator.com/) were used as reference models:

1. RC 3+DRE: A model including total PSA (ng/ml), DRE (normal/abnormal), and DRE assessed prostate volume (25/40/60 ml), developed for men without a prior biopsy;
2. RC4+DRE: A model including total PSA (ng/ml), DRE (normal/abnormal), and DRE assessed prostate volume (25/40/60 ml), developed for men with a prior (negative) biopsy.

For the ERSPC models the less patient-invasive DRE assessed volume in categories of 25/40/60 ml was used. For this study, TRUS assessed volume was therefore categorized into these volume classes with cut-off values of ≤30 ml, 30-50 ml, or >50 ml. Both models predict the chance of a positive sextant biopsy and its degree of aggressiveness. We used logistic regression analyses to estimate the coefficients for p2PSA, percentage p2PSA of fPSA (%p2PSA), and PHI, in addition to the ERSPC DRE based risk calculator (RCs) #3 (no prior biopsy) and RC#4 (having had a prior biopsy) [12]. These models included the linear predictor of the ERSPC RC#3 and RC#4 as a covariate. We subsequently developed updated versions of the RC#3 and RC#4 using the original model in combination with proPSA, %proPSA and PHI.

Validation in independent, external data is the best way to compare the performance of a model with and without a new marker [13]. We developed updated versions of the RCs that include PHI by recalibrating the original models (re-estimation of the intercept and slope of the linear predictor) [14]. Subsequently, we added fPSA, p2PSA, and PHI independently of each other in separate logistic regression models [14, 15].

Comparison of the models

We evaluated the predictive performance of a previously developed PHI based nomogram from Lughezzani *et al.* and updated versions of the ERSPC based RCs #3 and #4 using the area under the curve (AUC) of the receiver operating characteristic (ROC). Inclusion of PHI instead of PSA in the ERSPC model was also evaluated by refitting the original ERSPC model without PSA while including PHI, considering that PSA is included in PHI and hence involves some collinearity. We also evaluated the added value of age as a covariate by adding age to the PHI-updated ERSPC model.

We used repeated cross-validations of large calibration samples and smaller validation samples for optimal use of the available data [16]. We split the data in three subsets for men from Hamburg, Muenster and France (Rennes/Paris). For the first cross-validation, we removed men from Muenster from the population, recalibrated the ERSPC models, and updated these with PHI. These scores were allocated to men from Muenster. The same steps were taken for men from France and Hamburg. We then combined the scores for men from Muenster, France and Hamburg to estimate model performance in the total set. Multiple imputation was performed to substitute any missing values of the predictors included in the model (five repetitions).

Furthermore, we applied decision curve analysis (DCA) [13, 17] to evaluate the potential clinical usefulness of making decisions based on the Lughezzani and PHI-updated ERSPC models. We estimated a net benefit (NB) for prediction models by summing the benefits (true positive biopsies) and subtracting the harms (false positive biopsies). The harms were weighted by a factor related to the relative harm of a missed cancer versus an unnecessary biopsy. This weighting was derived from the threshold probability (p_t) of PC at which a patient would opt for a biopsy (range considered: 0% and 40%) [18]. A model with the highest net benefit at a particular threshold should be chosen over alternative models. The potential reduction in biopsies was calculated with the following formula: reduction in biopsies per 100 men = $(\Delta NB/(p_t/(1-p_t))) *100$, where p_t is the risk threshold, defined as the probability of disease at which an attending physician is indifferent between performing a biopsy and withholding the biopsy.

Standard statistical software was used (SPSS v 18.0, SPSS Inc., Chicago, Ill; R version 2.15.2, R Foundation for Statistical Computing, Vienna, Austria).

6.3 RESULTS

Participants

Among the 1185 men studied, 797 (67%, with 453 having PC) were not previously biopsied while 388 men (170 with PC) had a previous negative biopsy (table 6.1). Median PSA was 5.0 ng/ml for men with no prior biopsy, and 5.6 ng/ml for men with prior biopsy, whereas median values for PHI were 47 and 41, respectively. Men without prior biopsy were more likely to have (clinically relevant) cancer compared to men with prior biopsy.

Updating the ERSPC model

PSA testing alone resulted in an AUC of 0.53 (95% confidence interval: 0.50-0.57) for total cancer and 0.54 (0.51-0.58) for relevant cancer. Applying PHI alone resulted in

Table 6.1 Characteristics of the validation dataset of 1,185 men

	All men (N=1185)		Prior biopsy (N=388, 33%)		No prior biopsy (N=797, 67%)	
Age (years)	64.0	(58.9-70.0)	64.2	(58.6-70.0)	64.0	(59.0-69.9)
Total PSA (ng/ml)	5.2	(4.0-6.7)	5.6	(4.2-7.2)	5.0	(3.9-6.5)
PHI	44.8	(32.3-64.4)	41.1	(30.7-58.1)	46.9	(33.3-67.3)
TRUS assessed volume (ml)	42	(30-58)	46	(33-64)	40	(30-54)
DRE assessed volume (ml)						
<30 ml	256	(22%)	73	(19%)	183	(23%)
30-50 ml	512	(43%)	152	(39%)	360	(45%)
>50 ml	417	(35%)	163	(42%)	254	(32%)
DRE normal	859	(73%)	323	(83%)	536	(67%)
Biopsy outcome						
Cancer yes	623	(53%)	170	(44%)	453	(57%)
Clinically relevant cancer	324	(27%)	80	(21%)	244	(31%)

Continuous variables are noted as median (interquartile range); nominal variables are noted as number (percentage)
PSA = Prostate Specific Antigen, PHI = Prostate Health Index, TRUS = trans rectal ultrasound, DRE = digital rectal examination
Clinically relevant prostate cancer is defined as a clinical stage >T2B and/or a biopsy Gleason score ≥7

substantially better discriminative ability compared to PSA alone, with AUCs of 0.72 (0.69-0.75) and 0.68 (0.64-0.71) respectively. Applying the ERSPC RC#3 and RC#4 also resulted in higher AUCs as compared to PSA alone (AUC: 0.65 (0.62-0.68) and 0.66 (0.62-0.69) respectively, table 6.2).

Inclusion of p2PSA, %p2PSA and PHI in the updated ERSPC model resulted in further increases in predictive capability (results not shown). The inclusion of PHI gave the largest increase in AUC (to 0.72 (0.69-0.75) in all men and 0.72 (0.67-0.77) in men with prior biopsy) (table 6.2).

For total cancer, there was a potential reduction of biopsies without missing additional cancers for risk thresholds over 25% with the original ERSPC model, and from 35% with the PHI-extended ERSPC model compared to a biopsy all strategy. For clinically relevant cancer, the original ERSPC model potentially reduced biopsies from a 20% risk threshold, whereas the PHI-updated model reduced biopsies from a 10% threshold onward (table 6.3).

Comparison of the models

The Lughezzani model had similar performance as the PHI-updated ERSPC model, with AUCs of 0.75 (0.72-0.78) for all cancer and 0.69 (0.66-0.72) for clinically relevant

Table 6.2 Discriminative ability (Area under the receiver operator curve, AUC) of the PSA test, the PHI test, the ERSPC original model not including PHI, and the PHI-updated ERSPC model for the prediction of total and clinically relevant prostate cancer in 1185 men

		PSA alone		PHI alone		ERSPC		ERSPC+PHI	
Total cancer									
All men	N=1185	0.53	(0.50-0.57)	0.72	(0.69-0.75)	0.65	(0.62-0.68)	0.72	(0.69-0.75)
Prior biopsy	N=388	0.50	(0.43-0.56)	0.71	(0.66-0.77)	0.64	(0.58-0.69)	0.72	(0.67-0.78)
No prior biopsy	N=797	0.56	(0.52-0.60)	0.71	(0.68-0.75)	0.68	(0.64-0.71)	0.73	(0.69-0.76)
Clinically relevant cancer									
All men	N=1185	0.54	(0.51-0.58)	0.68	(0.64-0.71)	0.62	(0.59-0.66)	0.68	(0.65-0.71)
Prior biopsy	N=388	0.57	(0.50-0.65)	0.74	(0.69-0.80)	0.67	(0.60-0.73)	0.74	(0.68-0.80)
No prior biopsy	N=797	0.48	(0.44-0.52)	0.64	(0.60-0.68)	0.63	(0.59-0.67)	0.66	(0.62-0.70)

PSA = Prostate Specific Antigen, PHI = Prostate Health Index, DRE = digital rectal examination
ERSPC = a risk prediction model including total PSA (ng/ml), DRE (normal/abnormal), and DRE assessed prostate volume (25/40/60 ml)
ERSPC+PHI = a risk prediction model including total PSA (ng/ml), DRE (normal/abnormal), DRE assessed prostate volume (25/40/60 ml), and PHI
Clinically relevant prostate cancer is defined as a clinical stage >T2B and/or a biopsy Gleason score ≥7

cancer. For men without a prior biopsy, the PHI-updated RC#3 resulted in AUCs of 0.73 (0.69-0.76) predicting PC and 0.66 (0.62-0.70) for predicting clinically relevant PC. For men with a prior biopsy, this was 0.72 (0.67-0.78) and 0.74 (0.68-0.80). Inclusion of TRUS assessed volume (ml) instead of DRE assessed volume (categories) did not change the AUC for the PHI-updated ERSPC model (table 6.4).

The Lughezzani model had a significantly higher AUC than the PHI-updated ERSPC model with DRE assessed volume (categories) (0.75 (0.72-0.78) vs. 0.72 (0.69-0.75), $p<0.05$) for total cancer and for total cancer in met with a prior biopsy (0.75 (0.70-0.80) vs. 0.72 (0.67-0.78), $p<0.05$). The PHI-updated ERSPC models had better abilities to find clinically relevant cancer in groups of men with and without a prior biopsy, although differences were not significant (table 6.4).

Inclusion of PHI instead of PSA had no effect on the AUC, while adding age gave the largest increase in AUC compared to the original ERSPC model, with an AUC of 0.74 (0.71-0.77) for total cancer and 0.69 (0.66-0.72) for clinically relevant cancer for total PC (supplementary table 6.1).

The decision curve analyses showed that any PHI based models performed much better than PSA alone. A potential net reduction in biopsies was seen at PC risk thresholds from approximately 30% for total cancer (figure 6.1 for total population, supplementary figures 6.1A-B for men with and without a prior biopsy) and from approximately

Table 6.3 Reductions in biopsies compared to a biopsy all strategy per 1000 men of the ERSPC risk calculator and the PHI-updated ERSPC risk calculator for all participants (N=1185)

	ERSPC	ERSPC+PHI	Reduction with addition of PHI
Total cancer			
Risk thresholds			
5%	0	0	0
10%	0	0	0
15%	0	0	0
20%	0	0	0
25%	3	0	0
30%	7	0	0
35%	11	31	20
40%	16	75	59
45%	29	116	87
50%	84	150	67
Clinically relevant cancer			
Risk thresholds			
5%	0	0	0
10%	0	6	6
15%	0	27	27
20%	18	122	104
25%	67	149	81
30%	156	194	38

ERSPC = a risk prediction model including total PSA (ng/ml), DRE (normal/abnormal), and DRE assessed prostate volume (25/40/60 ml)
ERSPC+PHI = a risk prediction model including total PSA (ng/ml), DRE (normal/abnormal), DRE assessed prostate volume (25/40/60 ml), and PHI
Clinically relevant prostate cancer is defined as a clinical stage >T2B and/or a biopsy Gleason score ≥7

20% for clinically relevant cancers (figure 6.2 for total population, supplementary figures 6.2A-B for men with and without a prior biopsy). At a risk threshold (p_t) of 20% for PC, or at a p_t of 10% for clinically relevant cancer, the updated model including PHI did not result in a net reduction in biopsies compared to a biopsy-all strategy. Adding PHI to the RCs would reduce the number of biopsies only at higher values of the risk threshold p_t, compared to using PSA alone.

6.4 DISCUSSION

The Prostate Health Index (PHI) and its PSA components add important diagnostic information to separate PC from normal prostate tissue, also when considered in addition

Table 6.4 Discriminative ability (area under the receiver operator curve, AUC) of the ERSPC and Lughezzani models including PHI for the prediction of total and clinically relevant prostate cancer in 1185 men

		Lughezzani		ERSPC+PHI volume in categories		ERSPC+PHI volume in ml	
Total cancer							
All men	N=1185	0.75*	(0.72-0.78)	0.72	(0.69-0.75)	0.72	(0.69-0.75)
Prior biopsy	N=388	0.75*	(0.70-0.80)	0.72	(0.67-0.78)	0.71	(0.66-0.76)
No prior biopsy	N=797	0.74	(0.70-0.77)	0.73	(0.69-0.76)	0.73	(0.69-0.76)
Clinically relevant cancer							
All men	N=1185	0.69	(0.66-0.72)	0.68	(0.65-0.71)	0.68	(0.65-0.71)
Prior biopsy	N=388	0.73	(0.67-0.79)	0.74	(0.68-0.80)	0.73	(0.67-0.79)
No prior biopsy	N=797	0.65	(0.61-0.69)	0.66	(0.62-0.70)	0.66	(0.62-0.70)

*statistically significant higher AUC (P<0.05) compared to the ERSPC model with DRE assessed volume (categories)

PSA = Prostate Specific Antigen, PHI = Prostate Health Index, DRE = digital rectal examination, TRUS = trans rectal ultrasound

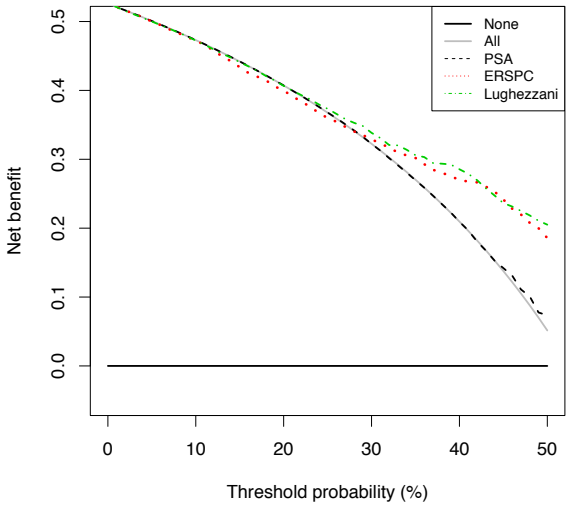
Lughezzani = a risk prediction model including age, DRE (normal/abnormal), TRUS assessed prostate volume (ml), prior biopsy (yes/no), and PHI

ERSPC+PHI (volume in categories) = a risk prediction model including total PSA (ng/ml), DRE (normal/abnormal), DRE assessed prostate volume (25/40/60 ml), and PHI

ERSPC+PHI (volume in ml) = a risk prediction model including total PSA (ng/ml), DRE (normal/abnormal), TRUS assessed prostate volume (ml), and PHI

Clinically relevant prostate cancer is defined as a clinical stage >T2B and/or a biopsy Gleason score ≥7

Figure 6.1 Net Benefit of the PHI-updated ERSPC model and the Lughezzani model for total prostate cancer

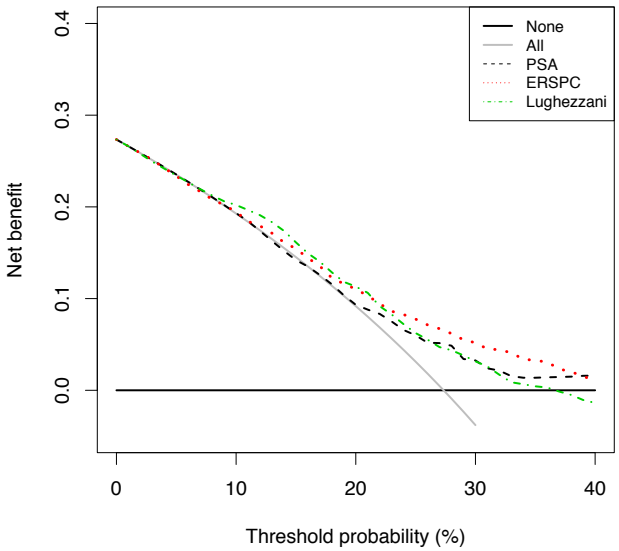


PSA = Prostate Specific Antigen, PHI = Prostate Health Index, DRE = digital rectal examination, TRUS = trans rectal ultrasound

ERSPC = a risk prediction model including total PSA (ng/ml), DRE (normal/abnormal), DRE assessed prostate volume (25/40/60 ml), and PHI

Lughezzani = a risk prediction model including age, DRE (normal/abnormal), TRUS assessed prostate volume (ml), prior biopsy (yes/no), and PHI

Figure 6.2 Net Benefit of the PHI-updated ERSPC model and the Lughezzani model for clinically relevant prostate cancer



Clinically relevant prostate cancer is defined as a clinical stage >T2B and/or a biopsy Gleason score ≥7
PSA = Prostate Specific Antigen, PHI = Prostate Health Index, DRE = digital rectal examination, TRUS = trans rectal ultrasound
ERSPC = a risk prediction model including total PSA (ng/ml), DRE (normal/abnormal), DRE assessed prostate volume (25/40/60 ml), and PHI
Lughezzani = a risk prediction model including age, DRE (normal/abnormal), TRUS assessed prostate volume (ml), prior biopsy (yes/no), and PHI

to previously developed PC risk prediction models. The net reduction in biopsies is however limited, and only seen at PC risk thresholds of approximately 20-30% using the two risk calculators we studied.

The PHI-updated models and the Lughezzani model performed similarly in discriminating between men with and without cancer. The Lughezzani nomogram includes TRUS assessed volume, while for the ERSPC models the less patient-invasive DRE assessed volume in categories of 25/40/60 ml is used [19]. No differences were observed between the PHI-updated ERSPC models with TRUS assessed volume and DRE assessed volume in categories, and both were equivalent to the Lughezzani model (table 6.4). We therefore prefer a model without the need for TRUS to assess prostate volume.

We confirmed that PHI performs better in predicting prostate cancer than conventional PSA measurement alone [20, 21]. However, the increases in performance from the

original ERSPC models to the PHI-updated ERSPC models were small. Moreover, the PHI-updated models in this study will be very well calibrated to this cohort, and external validation of the models would be required prior to clinical use. Further studies on the incremental value of PHI to multivariable models are required.

In addition to gain in discrimination, we considered the net benefit, but we did not fully consider cost-effectiveness. According to previous analyses, cost-effectiveness is dependent on the risk threshold used for measuring PHI, and on the specific range of PSA values [22, 23]. Using the updated model resulted in a relatively small reduction in prostate biopsies compared to the original RCs for risk thresholds between 0 and 40% and in no reduction in biopsies at a risk threshold of 20% for PC, or at a threshold of 10% for clinically relevant cancer. A comparative analysis of how much is gained and at what costs with the additional PHI test compared to the multivariate RC approach as indication for biopsy is required to study whether this is cost-effective.

One limitation of this study was the low number of patients with clinically relevant cancer. No strategy was hence found clearly dominant over the whole range of risk thresholds according to net benefit. Total cancer was predicted better in men with no prior biopsy compared to men with a prior biopsy, whereas clinically relevant cancer was predicted better in men with a prior biopsy compared to men without a prior biopsy. This could be due to the fact that high risk patients are no longer present in the group of men that have had a prior biopsy, and discrimination for total cancer is more difficult once these patients have already been diagnosed.

Furthermore, there is a potential risk of misclassification due to the fact that the ERSPC RCs were based on sextant biopsies, while the validation cohorts used ≥10 core biopsies. Previous validations in a clinical setting showed practically no underestimation of cancer risks [24]. We do recognize that the prevalence of cancer in the current validation cohorts was far higher than in the ERSPC setting. PHI may be more clinically useful if used in a setting with a lower risk of (clinically relevant) cancer. More men can then be spared a biopsy because of low cancer risk.

In conclusion, PHI increases the predictive ability of previously developed RCs for detection of cancer. However, only limited reductions in rates of unnecessary biopsies are possible for both the Lughezzani and the updated ERSPC models.

REFERENCES

1. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, *et al.* Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. Eur J Cancer. 2013;49(6):1374-403. Epub 2013/03/15.
2. Heidenreich A, Bastian PJ, Bellmunt J, *et al.* EAU guidelines on prostate cancer. part 1: screening, diagnosis, and local treatment with curative intent-update 2013. Eur Urol. 2014; 65(1):124-37. Epub 2013/11/12.
3. Draisma G, Boer R, Otto SJ, *et al.* Lead times and overdetection due to prostate-specific antigen screening: estimates from the European Randomized Study of Screening for Prostate Cancer. J Natl Cancer Inst. 2003;95(12):868-78. Epub 2003/06/19.
4. Heijnsdijk EA, der Kinderen A, Wever EM, Draisma G, Roobol MJ, de Koning HJ. Overdetection, overtreatment and costs in prostate-specific antigen screening for prostate cancer. Br J Cancer. 2009;101(11):1833-8. Epub 2009/11/12.
5. Loeb S, Vellekoop A, Ahmed HU, *et al.* Systematic review of complications of prostate biopsy. Eur Urol. 2013;64(6):876-92. Epub 2013/06/22.
6. 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. J Urol. 2007;177(1):107-12; discussion 12. Epub 2006/12/13.
7. Thompson IM, Ankerst DP, Chi C, *et al.* Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial. J Natl Cancer Inst. 2006;98(8):529-34. Epub 2006/04/20.
8. Karakiewicz PI, Benayoun S, Kattan MW, *et al.* Development and validation of a nomogram predicting the outcome of prostate biopsy based on patient age, digital rectal examination and serum prostate specific antigen. J Urol. 2005;173(6):1930-4. Epub 2005/05/10.
9. Heidenreich A, Bastian PJ, Bellmunt J, *et al.* EAU guidelines on prostate cancer. part 1: screening, diagnosis, and local treatment with curative intent-update 2013. Eur Urol. 2014; 65(1):124-37. Epub 2013/11/12.
10. Lughezzani G, Lazzeri M, Larcher A, *et al.* Development and internal validation of a Prostate Health Index based nomogram for predicting prostate cancer at extended biopsy. J Urol. 2012;188(4):1144-50. Epub 2012/08/21.
11. Jansen FH, van Schaik RH, Kurstjens J, *et al.* Prostate-specific antigen (PSA) isoform p2PSA in combination with total PSA and free PSA improves diagnostic accuracy in prostate cancer detection. Eur Urol. 2010;57(6):921-7. Epub 2010/03/02.
12. Roobol MJ, Kirkels WJ, Schroder FH. Features and preliminary results of the Dutch centre of the ERSPC (Rotterdam, the Netherlands). BJU Int. 2003;92 Suppl 2:48-54. Epub 2004/02/27.
13. Steyerberg EW, Vickers AJ, Cook NR, *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38. Epub 2009/12/17.

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):2567-86. Epub 2004/08/03.

15. Steyerberg EW. Clinical Prediction Models: Springer; 2009.

16. Browne MW. Cross-Validation Methods. J Math Psychol. 2000;44(1):108-32. Epub 2000/03/29.

17. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-74. Epub 2006/11/14.

18. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak. 2008;8:53. Epub 2008/11/28.

19. Roobol MJ, van Vugt HA, Loeb S, *et al*. Prediction of prostate cancer risk: the role of prostate volume and digital rectal examination in the ERSPC risk calculators. Eur Urol. 2012;61(3): 577-83. Epub 2011/11/23.

20. Loeb S, Catalona WJ. The Prostate Health Index: a new test for the detection of prostate cancer. Ther Adv Urol. 2014;6(2):74-7. Epub 2014/04/02.

21. Filella X, Gimenez N. Evaluation of [-2] proPSA and Prostate Health Index (phi) for the detection of prostate cancer: a systematic review and meta-analysis. Clin Chem Lab Med. 2013;51(4):729-39. Epub 2012/11/17.

22. Nichol MB, Wu J, Huang J, Denham D, Frencher SK, Jacobsen SJ. Cost-effectiveness of Prostate Health Index for prostate cancer detection. BJU Int. 2012;110(3):353-62. Epub 2011/11/15.

23. Heijnsdijk EAM, Huang JT, Denham D, De Koning HJ. The cost-effectiveness of prostate cancer detection using Beckman Coulter Prostate Health Index. Eur Urol Suppl. 2012;11(1): E260-E.

24. van Vugt HA, Kranse R, Steyerberg EW, *et al*. Prospective validation of a risk calculator which calculates the probability of a positive prostate biopsy in a contemporary clinical cohort. Eur J Cancer. 2012;48(12):1809-15. Epub 2012/03/13.

SUPPLEMENTARY TABLES AND FIGURES

Supplementary table 6.1 Discriminative ability (Area under the receiver operator curve, AUC) of a PHI-extended ERSPC model without PSA and a PHI-extended ERSPC model including age for the prediction of total and clinically relevant prostate cancer in 1185 men

		ERSPC+PHI -PSA		ERSPC+PHI +age	
Total cancer					
All men	N=1185	0.73	(0.70-0.76)	0.74	(0.71-0.77)
Prior biopsy	N=388	0.70	(0.65-0.76)	0.75	(0.70-0.80)
No prior biopsy	N=797	0.73	(0.69-0.76)	0.74	(0.71-0.78)
Clinically relevant cancer					
All men	N=1185	0.68	(0.65-0.71)	0.69	(0.66-0.72)
Prior biopsy	N=388	0.73	(0.66-0.79)	0.74	(0.69-0.80)
No prior biopsy	N=797	0.65	(0.61-0.69)	0.67	(0.63-0.71)

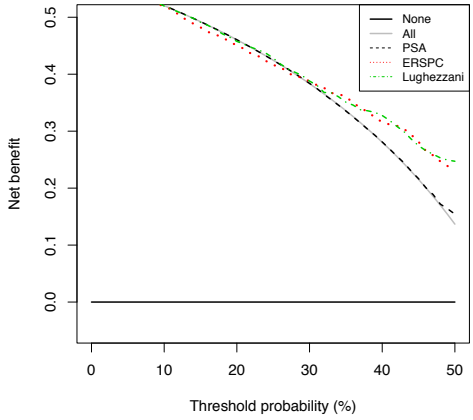
PSA = Prostate Specific Antigen, PHI = Prostate Health Index, DRE = digital rectal examination, TRUS = trans rectal ultrasound

ERSPC+PHI -PSA= a risk prediction model including DRE (normal/abnormal), DRE assessed prostate volume (25/40/60 ml), and PHI

ERSPC+PHI +age= a risk prediction model including total PSA (ng/ml), DRE (normal/abnormal), DRE assessed prostate volume (25/40/60 ml), PHI and age (years)

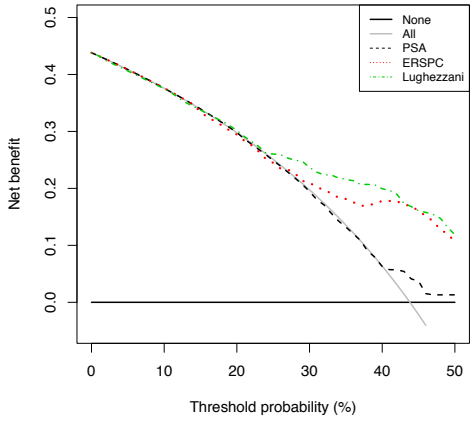
Clinically relevant prostate cancer is defined as a clinical stage >T2B and/or a biopsy Gleason score ≥7

Supplementary figure 6.1A Net Benefit of the PHI-updated ERSPC model #3 and the Lughezzani model for total prostate cancer in men with no prior biopsy



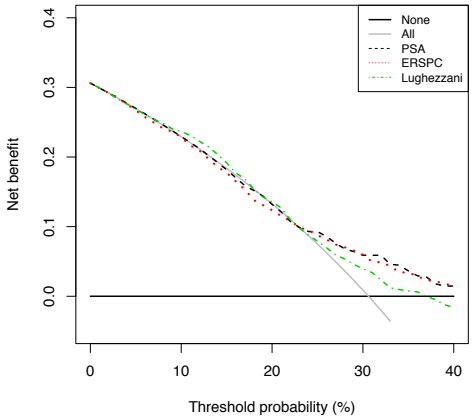
PSA = Prostate Specific Antigen, PHI = Prostate Health Index, DRE = digital rectal examination, TRUS = trans rectal ultrasound
ERSPC = a risk prediction model including total PSA (ng/ml), DRE (normal/abnormal), DRE assessed prostate volume (25/40/60 ml), and PHI
Lughezzani = a risk prediction model including age, DRE (normal/abnormal), TRUS assessed prostate volume (ml), prior biopsy (yes/no), and PHI

Supplementary figure 6.1B Net Benefit of the PHI-updated Roobol model #4 and the Lughezzani model for total prostate cancer in men with a prior biopsy



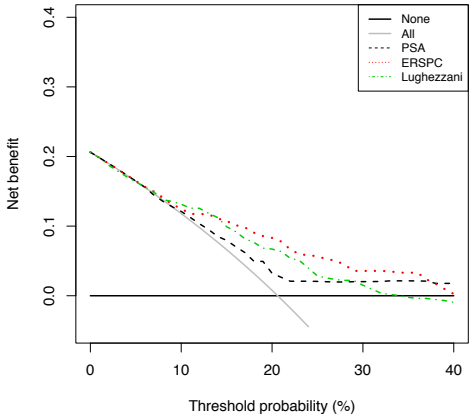
PSA = Prostate Specific Antigen, PHI = Prostate Health Index, DRE = digital rectal examination, TRUS = trans rectal ultrasound
ERSPC = a risk prediction model including total PSA (ng/ml), DRE (normal/abnormal), DRE assessed prostate volume (25/40/60 ml), and PHI
Lughezzani = a risk prediction model including age, DRE (normal/abnormal), TRUS assessed prostate volume (ml), prior biopsy (yes/no), and PHI

Supplementary figure 6.2A Net Benefit of the PHI-updated ERSPC model #3 and the Lughezzani model for clinically relevant prostate cancer in men with no prior biopsy



Clinically relevant cancer was defined as a clinical stage >T2B and/or a Gleason score ≥7
PSA = Prostate Specific Antigen, PHI = Prostate Health Index, DRE = digital rectal examination, TRUS = trans rectal ultrasound
ERSPC = a risk prediction model including total PSA (ng/ml), DRE (normal/abnormal), DRE assessed prostate volume (25/40/60 ml), and PHI
Lughezzani = a risk prediction model including age, DRE (normal/abnormal), TRUS assessed prostate volume (ml), prior biopsy (yes/no), and PHI

Supplementary figure 6.2B Net Benefit of the PHI-updated ERSPC model #4 and the Lughezzani model for clinically relevant prostate cancer in men with a prior biopsy



Clinically relevant cancer was defined as a clinical stage >T2B and/or a Gleason score ≥7
PSA = Prostate Specific Antigen, PHI = Prostate Health Index, DRE = digital rectal examination, TRUS = trans rectal ultrasound
ERSPC = a risk prediction model including total PSA (ng/ml), DRE (normal/abnormal), DRE assessed prostate volume (25/40/60 ml), and PHI
Lughezzani = a risk prediction model including age, DRE (normal/abnormal), TRUS assessed prostate volume (ml), prior biopsy (yes/no), and PHI

Chapter 7

Risk prediction scores for recurrence and progression of non-muscle invasive bladder cancer: An international validation in primary tumours

Vedder MM, Marquez M, de Bekker-Grob EW, Calle ML, Dyrskjot L, Kogevinas M, Segersten U, Malmstrom PU, Algaba F, Beukers W, Ørntoft TF, Zwarthoff E, Real FX, Malats N, Steyerberg EW

PLoS One. 2014 Jun 6;9(6):e96849.

ABSTRACT

Objective: We aimed to determine the validity of two risk scores for patients with non-muscle invasive bladder cancer in different European settings, in patients with primary tumours.

Methods: We included 1,892 patients with primary stage Ta or T1 non-muscle invasive bladder cancer who underwent a transurethral resection in Spain (n=973), the Netherlands (n=639), or Denmark (n=280). We evaluated recurrence-free survival and progression-free survival according to the European Organisation for Research and Treatment of Cancer (EORTC) and the Spanish Urological Club for Oncological Treatment (CUETO) risk scores for each patient and used the concordance index (c-index) to indicate discriminative ability.

Results: The 3 cohorts were comparable according to age and sex, but patients from Denmark had a larger proportion of patients with the high stage and grade at diagnosis ($p<0.01$). At least one recurrence occurred in 839 (44%) patients and 258 (14%) patients had a progression during a median follow-up of 74 months. Patients from Denmark had the highest 10-year recurrence and progression rates (75% and 24%, respectively), whereas patients from Spain had the lowest rates (34% and 10%, respectively). The EORTC and CUETO risk scores both predicted progression better than recurrence with c-indices ranging from 0.72 to 0.82 while for recurrence, those ranged from 0.55 to 0.61.

Conclusion: The EORTC and CUETO risk scores can reasonably predict progression, while prediction of recurrence is more difficult. New prognostic markers are needed to better predict recurrence of tumours in primary non-muscle invasive bladder cancer patients.

7.1 INTRODUCTION

Bladder cancer is the most common malignancy of the urinary tract and a major health issue [1]. Most patients with bladder cancer are diagnosed with non-muscle invasive disease (NMIBC: stage Ta or T1) [2]. After transurethral resection (TUR), recurrence of disease occurs in 30-60% of patients and, approximately, 10–15% develop progression to muscle-invasive disease in 5-year after diagnosis [3]. Therefore, regular cystoscopy is carried out for surveillances after TUR. To better target surveillance, risk scores for recurrence and progression prediction have been developed. The best known are the European Organisation for Research and Treatment of Cancer (EORTC) [4] and the Spanish Urological Club for Oncological Treatment (CUETO) [5] risk scores; the latter focusing on BCG treated patients. Despite their potential usefulness in daily practice, few studies have externally validated these models [6-11] and no study focussed on primary NMIBC. In addition, since the EORTC score was based on a cohort of patients included in 7 clinical trials, the question arises whether these scores are still valid in a broader set of NMIBC patients for predictive purposes. The EORTC and CUETO scores were based on specimens evaluated by central pathologies and specialized pathologists, whereas the specimens included in the present study had been evaluated by routine pathology. In the present study, we investigated the external validity of these risk scores in patients with primary NMIBC across European centres in an everyday routine setting.

7.2 METHODS

Study population

We included 1,892 patients with primary NMIBC from three countries; Spain, Denmark, and the Netherlands. Patients from Spain were recruited between 1998 and 2001 from 18 general and University hospitals as part of the Spanish Bladder Cancer/EPIdemiology of Cancer of the UROthelium (EPICURO) study [12]. All centres are outlined in supplementary table 7.1. Patients from Denmark were selectively included based on being at higher risk of progression from patient records of the Aarhus University Hospital between 1979 and 2007 [13]. For the Netherlands, we included consecutive patients from the Erasmus MC who underwent a TUR between 1990 and 2012. Patient and tumour characteristics and data on recurrence and progression after TUR of the primary NMIBC were extracted from hospital records up till November 2012. All patients had histologically confirmed NMIBC and were treated according to the centres' usual procedures. At the Erasmus MC in the Netherlands, follow-up of patients was according to the EAU guidelines at the time, and risk-adapted according to the EORTC risk

scores outcome. At the Aarhus University Hospital in Denmark, the common follow-up strategy for all patients was every three months. In Spain, protocols for the follow-up of bladder cancer patients were developed within each centre. For non-muscle invasive bladder cancers, follow-up for these patients consisted of bladder endoscopy every three months the first year, every six months the second year and then annually bladder endoscopy to complete five years of monitoring. White light cystoscopy was used in all centres participating in our study.

Disease progression was defined as cystoscopically detected tumour relapse with histological confirmation at tumour stage T2 or higher (progression to a muscle invasive tumour stage); it was assumed that a tumour progression always precedes death because of cancer. Patients that died because of bladder cancer without a progression were recorded as having had a progression at the time of death. Recurrence was defined as cystoscopically detected tumour relapse with histological confirmation. Data from the 3 cohorts were harmonized, anonymized, and combined in one data set for statistical analyses, stratified by cohort.

All Danish and Spanish patients gave their written informed consent, and the study was approved by the Central Denmark Region Committees on Biomedical Research Ethics (1994/2920) and by the Ethics Committees of each Spanish participating centre and the Institutional Review Board of the U.S. National Cancer Institute, NIH, USA. This observational study was exempted from formal ethical approval in the Netherlands. All data is anonymized before being used in this study.

Risk scores

The EORTC scores for recurrence and progression were based on data from 2,596 patients diagnosed with Ta/T1 tumours from seven EORTC trials [4]. A limitation of the EORTC scores was the low number of patients treated with bacillus Calmette Guérin (BCG). Therefore, the CUETO group developed a scoring model in 1,062 BCG-treated patients [5]. The EORTC score incorporated the number of tumours (single, 2-7 or ≥8), tumour size (<3 cm or ≥3 cm), prior recurrence rate (primary, ≤1 recurrence/year, >1 recurrence/year), T stage (Ta or T1), concomitant carcinoma in situ (yes/no), and grade (1, 2, or 3). The CUETO model incorporated gender, age (<60, 60-70, >70 years), recurrent tumour (yes/no), number of tumours (≤3 or >3), T stage (Ta or T1), concomitant carcinoma in situ (yes/no), and grade (1, 2, or 3).

Validation

For all patients, we calculated risks for recurrence and progression according to the EORTC and CUETO scores based on the primary tumour. Standard pathologic proce-

dures were followed in each cohort. Tumour grade was scored according to the 1973 system, and pathological stage was according to the 2002 staging system. The presence of concomitant carcinoma in situ was incomplete (CIS, n=990, 52% missing), as well as data on the number of tumours (n=346, 18% missing). We used a multiple imputation strategy [14] resulting in five sets of complete data to compute risk scores. We subsequently averaged these risk scores for each patient. Patient scores were then categorized into four risk groups, i.e. low, intermediate low, intermediate high, and high risk for recurrence or progression, as originally specified for the EORTC and CUETO scores. The two highest risk groups were combined because of low numbers. Observed recurrence-free survival (RFS) and progression-free survival (PFS) were calculated from the date of TUR of the primary tumour. An event for RFS was defined as recurrence or progression, if progression occurred as the first event during follow-up. Follow-up was censored at either the last date of follow-up, the date of death, or 120 months. We used standard Kaplan–Meier plots to visualize recurrence and progression patterns in relation to risk groups. This cause-specific analysis was not adjusted for the competing risk of death before recurrence or progression, since we focused on the discriminative ability of the 2 risk scores (quantified by a concordance measure, c-index) [15]. We conducted subgroup analyses for patients receiving only BCG treatment after TUR. Furthermore, we refitted the scores with a Cox regression analysis stratified by cohort by recalculating risk scores with EORTC and CUETO coefficients based on our data, to obtain further insight in the validity of the scores. We used likelihood ratio statistics to determine the statistical significance of predictors. For comparability with the original EORTC and CUETO scores, we scaled the refitted regression coefficients by the inverse of the Cox regression coefficient for the original scores in our data. For example, the refitted score for T1 vs Ta in the EORTC model for recurrence was calculated as: multivariable coefficient for T1 vs Ta * 1/(coefficient for EORTC score for recurrence). SPSS (version 20.0, SPSS Inc, Chicago, Illinois, USA) and R (Version R-2.15.2 for Windows, http://www.r-project.org/) were used for data analysis.

7.3 RESULTS

Study population

We included 1,892 patients; 280 patients from Denmark, 639 from the Netherlands, and 973 from Spain. During 10 years of follow-up, 209 (11%) patients died before a recurrence occurred, 839 (44%) patients had a recurrence and 258 (14%) a progression. Median follow-up for those without recurrence was 74 months. There were 98 patients (N=90 from the Netherlands, N=8 from Denmark) without follow-up because of loss to follow-up immediately after TUR. CIS (yes/no) and number of tumours was

imputed in patients with missing data, based on 902 patients with information on CIS and 1546 patients with information on the number of tumours, as well as complete information on tumour stage, grade, and size, and progression and recurrence free survival (time and yes/no). The mean age was 66 years and the majority was male (table 7.1). We do not present totals over all cohorts because of the substantial differences in settings between cohorts. Danish patients presented a larger proportion of patients with high stage and grade (P<0.01), and relatively more recurrences and progressions. The distribution of patients over the risk groups is shown in table 7.2.

Table 7.1 Patient characteristics of 1,892 patients with non-muscle invasive bladder cancer in the participating cohorts*

		Denmark (n=280)		Netherlands (n=639)		Spain (n=973)	
Age	Mean (SD)	66.4	(10.2)	65.3	(12.4)	65.7	(10.0)
Gender	Male	219	78.2%	503	78.7%	850	87.4%
Stage	Ta	177	63.2%	432	67.6%	818	84.1%
	T1	103	36.8%	207	32.4%	155	15.9%
Grade	G1	78	27.9%	189	29.6%	419	43.1%
	G2	83	29.6%	283	44.3%	327	33.6%
	G3	119	42.5%	167	26.1%	227	23.3%
Carcinoma in situ	CIS	89	31.8%	52	8.1%	0	0.0%
	No CIS	189	67.5%	572	89.5%	0	0.0%
	Missing	2	0.7%	15	2.4%	973	100.0%
Tumour size	<3 cm	175	62.5%	238	37.2%	564	58.0%
	≥3 cm	73	26.1%	114	17.9%	133	13.6%
	Missing	32	11.4%	287	44.9%	276	28.4%
Number of tumours	1	82	29.3%	349	54.6%	647	66.5%
	>1	13	4.6%	178	27.9%	277	28.5%
	Missing	185	66.1%	112	17.5%	49	5.0%
Treatment	TUR alone	227	81.0%	140	21.9%	404	41.5%
	TUR+BCG	52	18.6%	108	16.9%	289	29.7%
	TUR+Chemo	0	0.0%	80	12.5%	212	21.8%
	TUR+Chemo+BCG	1	0.4%	29	4.5%	51	5.2%
	Other	0	0.0%	5	0.8%	17	1.7%
	Missing	0	0.0%	277	43.4%	0	0.0%
Status tumour**	Recurrence	209	74.6%	303	47.4%	327	33.6%
	Progression	66	23.6%	99	15.5%	93	9.6%
Vital status**	Alive	72	25.7%	321	50.2%	700	71.9%
	Cancer death	12	4.3%	51	8.0%	62	6.4%
	Other death	7	2.5%	90	14.1%	211	21.7%
	Missing	189	67.5%	177	27.7%	0	0.0%

*Numbers are totals (%), unless stated otherwise

**At the end of follow-up

Table 7.2 Distribution of patients over the risk groups for all patients (n=1892) and BCG treated patients (n=449)

Risk category	CUETO		EORTC	
	Recurrence	Progression	Recurrence	Progression
All patients (N=1892)				
Low risk	1195 (63.2%)	1289 (68.1%)	383 (20.2%)	346 (18.3%)
Intermediate risk	421 (22.3%)	135 (7.1%)	1099 (58.1%)	929 (49.1%)
High risk*	276 (14.6%)	468 (24.7%)	410 (21.7%)	617 (32.6%)
BCG (N=449)				
Low risk	226 (50.3%)	241 (53.7%)	48 (10.7%)	30 (6.7%)
Intermediate risk	136 (30.3%)	36 (8.0%)	257 (57.2%)	197 (43.9%)
High risk*	87 (19.4%)	172 (38.3%)	144 (32.1%)	222 (49.4%)

*The high risk group is the combined group from intermediate-high and high-risk EORTC and CUETO scores, because of low patient numbers

Validation

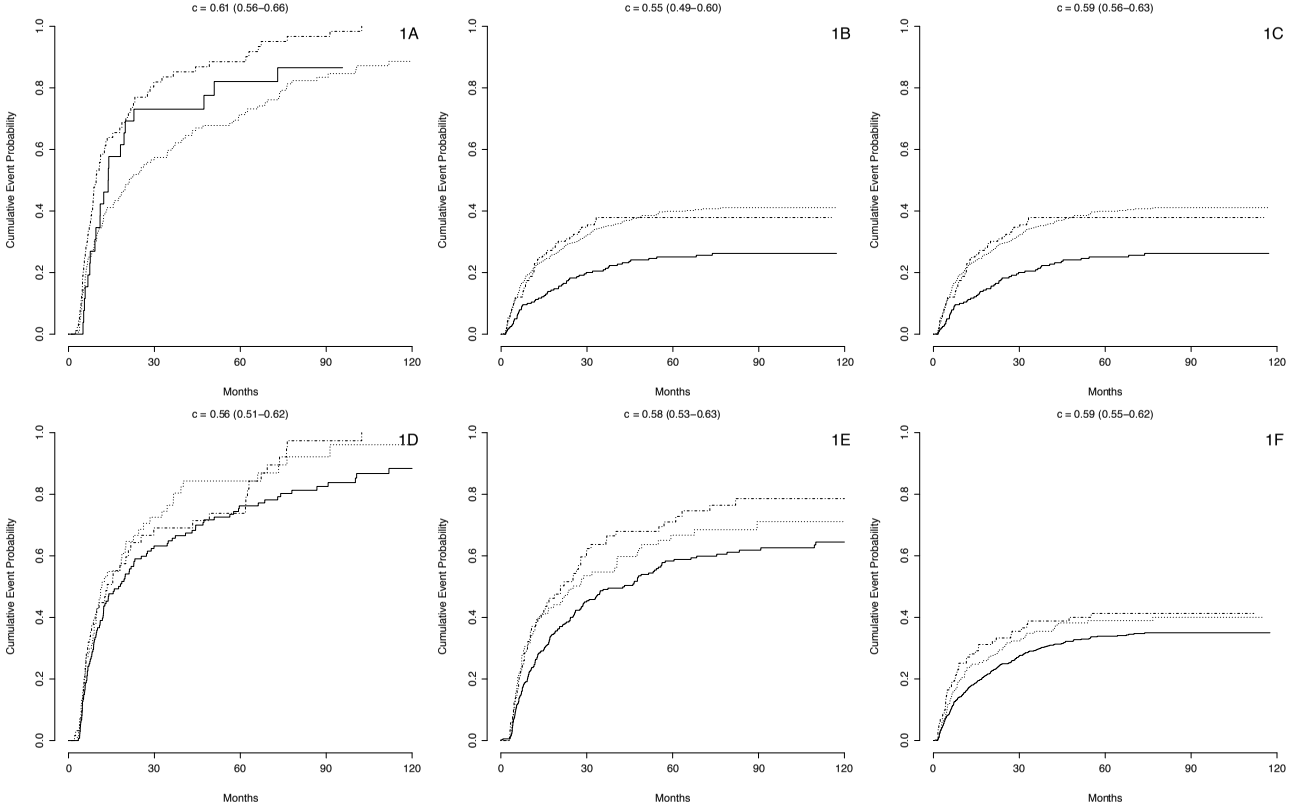
The EORTC score could not well separate low risk from high risk patients with respect to disease recurrence (figures 7.1a-c, c-indices 0.55 to 0.61). Discrimination was somewhat better for progression (figures 7.2a-c, c-indices 0.72 to 0.81). The CUETO score had a similar performance (figures 7.1d-f and 7.2d-f). Subgroup analyses in patients receiving BCG treatment (n=449) showed poorer results (supplementary figures 7.1a-f and 7.2a-f).

When we refitted the EORTC score for recurrence in Cox regression models, the prognostic effects of multiple tumours, tumour size, CIS and tumour grade were largely confirmed, but T1 tumours had no increased risk over Ta tumours (results not shown). For progression, tumour size and CIS were less predictive than in the original EORTC score, while the effect for grade was stronger. For the CUETO score, gender was confirmed to be predictive of recurrence. While older age was not predictive of recurrence, we confirmed its value for predicting progression in the refitted CUETO score ($p<0.01$).

7.4 DISCUSSION

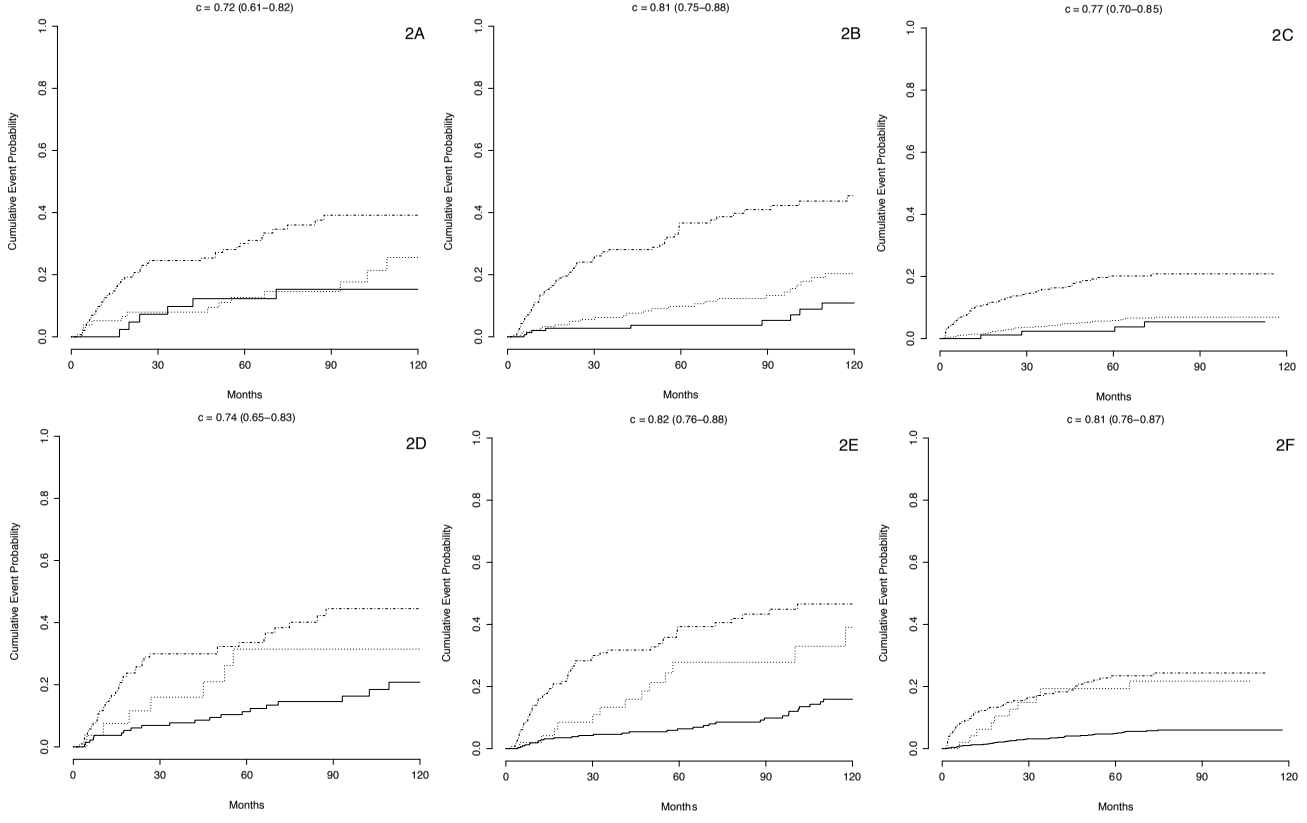
The EORTC risk tables have become a standard of care with their inclusion in European guidelines [2]. The CUETO risk model was developed more recently, with a focus on patients treated with BCG. External validation of a prognostic model on a new dataset is crucial to assess its generalizability [16]. In our study, the EORTC and CUETO risk scores showed only modest discriminative ability for the recurrence of NMIBC, with c-indices of, at most, 0.61. Prediction of progression was better with c-indices ranging

Figure 7.1 A-F Kaplan-Meier estimates of recurrence of bladder cancer in a ten-year period from transurethral resection of a non-muscle invasive bladder tumour



Full line: low risk patients, dotted line: intermediate risk patients, dashed line: high risk patients
Number of patients per country: Denmark n=280 (left figures); The Netherlands n=639 (middle figures); Spain n=973 (right figures)
*ES = EORTC risk score for the recurrence/progression of non-muscle invasive bladder cancer, figures A-C
**CS = CUETO risk score for the recurrence/progression of non-muscle invasive bladder cancer, figures D-F

Figure 7.2 A-F Kaplan-Meier estimates of progression of bladder cancer in a ten-year period from transurethral resection of a non-muscle invasive bladder tumour



Full line: low risk patients, dotted line: intermediate risk patients, dashed line: high risk patients
Number of patients per country: Denmark n=280 (left figures); The Netherlands n=639 (middle figures); Spain n=973 (right figures)
*ES = EORTC risk score for the recurrence/progression of non-muscle invasive bladder cancer, figures A-C
**CS = CUETO risk score for the recurrence/progression of non-muscle invasive bladder cancer, figures D-F

Table 7.3 Weights for predictors used in the EORTC score (ES) and CUETO score (CS) to calculate the recurrence and progression scores for patients with non-muscle invasive bladder cancer. The original scores are given first, followed by refitted scores based on a stratified Cox regression analysis in our 3 cohorts.

Recurrence	ES*	CS**	Refit ES*	Refit CS**	Progression	ES*	CS**	Refit ES*	Refit CS**
Gender					**Gender**				
M	-	0	-	0	M	-	-	-	-
F	-	3	-	4	F	-	-	-	-
Age					**Age**				
<60	-	0	-	0	<60	-	0	-	0
60–70	-	1	-	0	60–70	-	0	-	2
>70	-	2	-	0	>70	-	2	-	3
Recurrent tumour					**Recurrent tumour**				
No	0	0	-	-	No	0	0	-	-
Yes	0	4	-	-	Yes	0	2	-	-
Prior recurrence rate					**Prior recurrence rate**				
Primary	0	0	-	-	Primary	0	0	-	-
≤1 rec/yr	2	0	-	-	≤1 rec/yr	2	0	-	-
>1 rec/yr	4	0	-	-	>1 rec/yr	2	0	-	-
No. tumours					**No. tumours**				
≤3	-	0	-	-	≤3	-	0	-	-
>3	-	2	-	-	>3	-	1	-	-
Single	0	-	0	0	Single	0	-	0	0
2 to 7	3	-	-	-	Multiple	3	-	2	1
≥8	6	-	-	-					
Multiple	-	-	4	5					
Tumour size					**Tumour size**				
<3 cm	0	0	0	0	<3 cm	0	0	0	0
≥3 cm	3	0	3	3	≥3 cm	3	0	0	0
T category					**T category**				
Ta	0	0	1	1	Ta	0	0	0	0
T1	1	0	0	0	T1	4	2	6	3
CIS					**CIS**				
No	0	0	0	0	No	0	0	0	0
Yes	1	2	1	2	Yes	6	1	2	1
Grade					**Grade**				
G1	0	0	0	0	G1	0	0	0	0
G2	1	1	2	2	G2	0	2	6	3
G3	2	3	2	3	G3	5	6	10	5
Total score	0-17	0-16			**Total score**	0-23	0-14		

ES*	CS**	Recurrence	ES*	CS**	Progression
0	0-4	Low risk	0	0-4	Low risk
1-4	5-6	Intermediate low risk	2-6	5-6	Intermediate low risk
5-9	7-9	Intermediate high risk	7-13	7-9	Intermediate high risk
10-17	10-16	High risk	14-23	10-14	High risk

*ES = EORTC risk score for the recurrence/progression of non-muscle invasive bladder cancer
**CS = CUETO risk score for the recurrence/progression of non-muscle invasive bladder cancer

from 0.72 to 0.82. Our findings were consistent in the cohorts from Denmark, Spain and the Netherlands, and are in line with another external validation of the EORTC risk score [6] and with validation in primary bladder cancer cases [11].

Remarkably, the CUETO score was specifically developed for patients treated with BCG, but discriminated better in the overall population than in the selected BCG population. BCG treatment, which has become a common treatment to manage intermediate- and high-risk NMIBC [17], was used in 449 patients, of over 50% at low risk of recurrence and progression according to the CUETO risk scores. For the EORTC risk scores, we noted that BCG treatment was usually administered to higher risk patients with a relatively narrow distribution of risk scores. This homogeneity in risk may partly explain the poor discriminative ability of the scores in those treated with BCG [18]. More research in this specific group of patients needs to be done, also because of the lack of statistical power due to low numbers of BCG patients in the current study.

In the original study that presented the EORTC risk scores, prior recurrence rate was an important prognostic factor for both recurrence and progression [4]. In the clinical setting, we need to establish a surveillance plan already after TUR for the primary tumour. Therefore, it is of great importance that the EORTC risk score has predictive value also for these patients, who have not had one or multiple recurrences. We found that predicting recurrence was very difficult for primary tumours. The heterogeneity in recurrence risk becomes better known once one or more recurrences have been observed [19].

A possible explanation for the poor performance of the risk scores for the prediction of recurrence outside controlled trials is interobserver variability in bladder cancer staging and grading by pathologists. To partly overcome these issues, new methods for bladder cancer pathology have been introduced in 1998 [20] and 2004 [21]. The 1998 method has been shown to be an improvement over the 1973 method [22], which was used for our patients.

The poor predictability of recurrence may also relate to other factors, unrelated to the (observed) pathology of the disease. For example, detection of all primary tumours may be difficult at primary tumour presentation. Tumour tissue may be left behind, falsely leading to classification as a recurrent tumour. The quality of the TUR may be important but it could not be considered in our evaluation. Moreover, detection policies may vary between urologists with respect to surveillance intervals and treatment modalities (e.g. TUR vs ablation). Progression is a more robust end point, which may partly explain its better predictability with the EORTC and CUETO scores.

The retrospective analysis is a limitation of this study, and explains the presence of missing values in important variables such as CIS and tumour size. We used multiple imputation, which has been shown to be a reliable method to handle missing data [23]. We had no detailed information on treatments and surveillance policies, which may have changed over time. The treatment modalities may have led to a dilution of differences between the risk groups. On the other hand, a real life situation was considered with respect to the standard care of urologists. We furthermore note that a selected group of high risk patients was included from Denmark, which can be explained by the fact that patients originated form a specialised university medical centre. However, patients from Spain were a representative sample from standard primary NMIBC population in that country, and patients from the Netherlands, though originating from an academic centre, were similar to the general Dutch primary NMIBC patient population [24].

It is clear that the EORTC and CUETO scores need further improvement. Several markers have shown promising results, such as FGFR3 and Ki67, which improved c-indices for prediction of progression from 0.75 to 0.82 in one study [8]. Various other promising molecular and germline markers are available, which need further rigorous evaluation for their usefulness to predict recurrence and progression [25-26]. Future risk scores will again need external validation, considering discrimination and other aspects of predictive performance, such as calibration (correspondence between observed and predicted risks) and clinical usefulness (ability to make better decisions) [27-29].

We conclude that the discriminatory ability of currently available risk scores is poor for recurrence and moderate for progression in primary NMIBC. Since successful discrimination of low and high risk patients is essential to the right intensity of bladder cancer surveillance, new risk markers are urgently needed to improve risk classification in NMIBC patients.

REFERENCES

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, *et al.* (2010) GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10. Lyon, France: International Agency for Research on Cancer.
2. Babjuk M, Burger M, Zigeuner R, Shariat SF, van Rhijn BW, *et al.* (2013) EAU guidelines on non-muscle-invasive urothelial carcinoma of the bladder: update 2013. Eur Urol 64: 639-653.
3. Kirkali Z, Chan T, Manoharan M, Algaba F, Busch C, *et al.* (2005) Bladder cancer: epidemiology, staging and grading, and diagnosis. Urology 66: 4-34.
4. Sylvester RJ, van der Meijden AP, Oosterlinck W, Witjes JA, Bouffioux C, *et al.* (2006) Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. Eur Urol 49: 466-465; discussion 475-467.
5. Fernandez-Gomez J, Madero R, Solsona E, Unda M, Martinez-Pineiro L, *et al.* (2009) Predicting nonmuscle invasive bladder cancer recurrence and progression in patients treated with bacillus Calmette-Guerin: the CUETO scoring model. J Urol 182: 2195-2203.
6. Hernandez V, De La Pena E, Martin MD, Blazquez C, Diaz FJ, *et al.* (2011) External validation and applicability of the EORTC risk tables for non-muscle-invasive bladder cancer. World J Urol 29: 409-414.
7. Buethe DD, Sexton WJ (2011) Bladder cancer: validating the EORTC risk tables in BCG-treated patients. Nat Rev Urol 8: 480-481.
8. van Rhijn BW, Zuiverloon TC, Vis AN, Radvanyi F, van Leenders GJ, *et al.* (2010) Molecular grade (FGFR3/MIB-1) and EORTC risk scores are predictive in primary non-muscle-invasive bladder cancer. Eur Urol 58: 433-441.
9. Rosevear HM, Lightfoot AJ, Nepple KG, O'Donnell MA (2011) Usefulness of the Spanish Urological Club for Oncological Treatment scoring model to predict nonmuscle invasive bladder cancer recurrence in patients treated with intravesical bacillus Calmette-Guerin plus interferon-alpha. J Urol 185: 67-71.
10. Fernandez-Gomez J, Madero R, Solsona E, Unda M, Martinez-Pineiro L, *et al.* (2011) The EORTC tables overestimate the risk of recurrence and progression in patients with non-muscle-invasive bladder cancer treated with bacillus Calmette-Guerin: external validation of the EORTC risk tables. Eur Urol 60: 423-430.
11. Xylinas E, Kent M, Kluth L, Pycha A, Comploj E, *et al.* (2013) Accuracy of the EORTC risk tables and of the CUETO scoring model to predict outcomes in non-muscle-invasive urothelial carcinoma of the bladder. Br J Cancer 109: 1460-1466.
12. Porta N, Calle ML, Malats N, Gomez G (2012) A dynamic model for the risk of bladder cancer progression. Stat Med 31: 287-300.

13. Fristrup N, Ulhoi BP, Birkenkamp-Demtroder K, Mansilla F, Sanchez-Carbayo M, *et al.* (2012) Cathepsin E, maspin, Plk1, and survivin are promising prognostic protein markers for progression in non-muscle invasive bladder cancer. Am J Pathol 180: 1824-1834.

14. Rubin DB, Schenker N (1991) Multiple imputation in health-care databases: an overview and some applications. Stat Med 10: 585-598.

15. Harrel FEJ (2001) Regression Modeling Strategies: Springer-Verlag New York, Inc.

16. Justice AC, Covinsky KE, Berlin JA (1999) Assessing the generalizability of prognostic information. Ann Intern Med 130: 515-524.

17. Fahmy N, Lazo-Langner A, Iansavichene AE, Pautler SE (2013) Effect of anticoagulants and antiplatelet agents on the efficacy of intravesical BCG treatment of bladder cancer: A systematic review. Can Urol Assoc J 7: E740-749.

18. Vergouwe Y, Moons KG, Steyerberg EW (2010) External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol 172: 971-980.

19. Kompier LC, van der Aa MN, Lurkin I, Vermeij M, Kirkels WJ, *et al.* (2009) The development of multiple bladder tumour recurrences in relation to the FGFR3 mutation status of the primary tumour. J Pathol 218: 104-112.

20. Epstein JI, Amin MB, Reuter VR, Mostofi FK (1998) The World Health Organization/International Society of Urological Pathology consensus classification of urothelial (transitional cell) neoplasms of the urinary bladder. Bladder Consensus Conference Committee. Am J Surg Pathol 22: 1435-1448.

21. Montironi R, Lopez-Beltran A (2005) The 2004 WHO classification of bladder tumors: a summary and commentary. Int J Surg Pathol 13: 143-153.

22. Gonul, II, Poyraz A, Unsal C, Acar C, Alkibay T (2007) Comparison of 1998 WHO/ISUP and 1973 WHO classifications for interobserver variability in grading of papillary urothelial neoplasms of the bladder. Pathological evaluation of 258 cases. Urol Int 78: 338-344.

23. Ambler G, Omar RZ, Royston P (2007) A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. Stat Methods Med Res 16: 277-298.

24. Dutch Cancer registration (2010) www.iknl.nl. Integraal Kankercentrum Nederland.

25. van Rhijn BW (2012) Combining molecular and pathologic data to prognosticate non-muscle-invasive bladder cancer. Urol Oncol 30: 518-523.

26. Shariat SF, Lotan Y, Vickers A, Karakiewicz PI, Schmitz-Drager BJ, *et al.* (2010) Statistical consideration for clinical biomarker research in bladder cancer. Urol Oncol 28: 389-400.

27. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, *et al.* (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21: 128-138.

28. Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 26: 565-574.

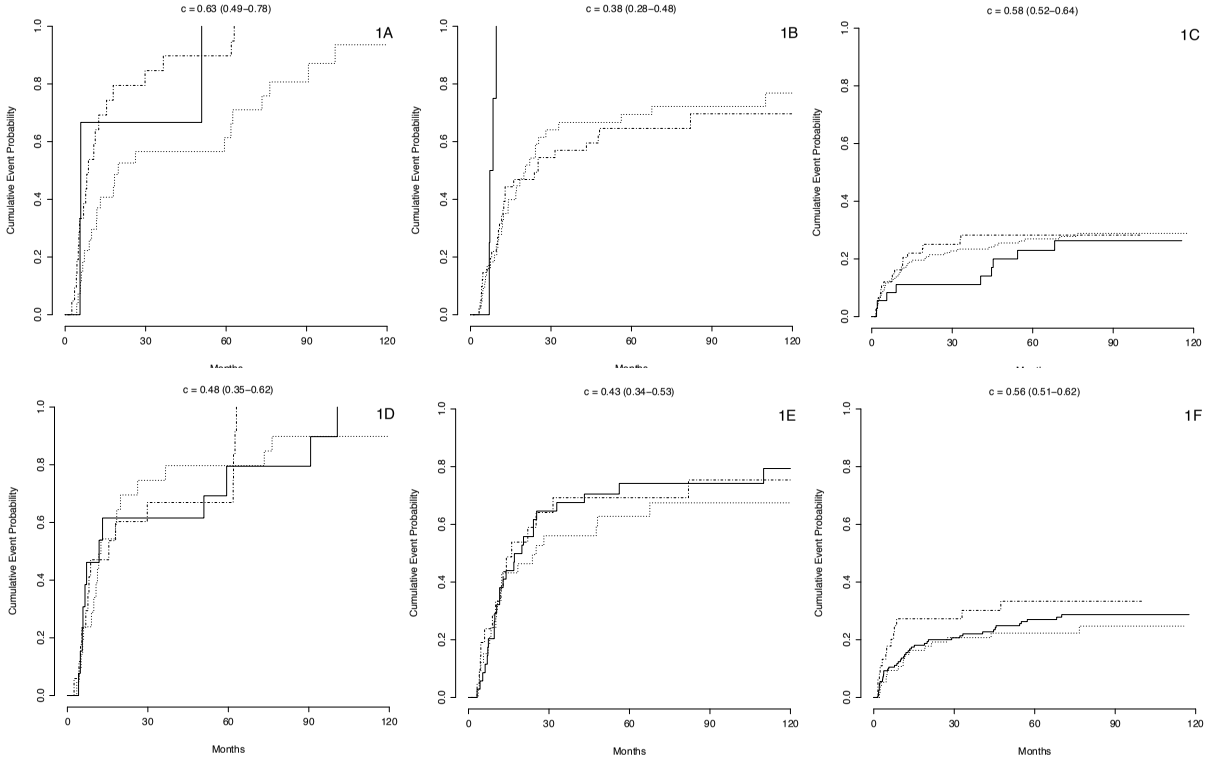
29. Vickers A (2010) Prediction models in urology: are they any good, and how would we know anyway? Eur Urol 57: 571-573; discussion 574.

SUPPLEMENTARY TABLES AND FIGURES

Supplementary table 7.1 Centres and members of the Spanish study group

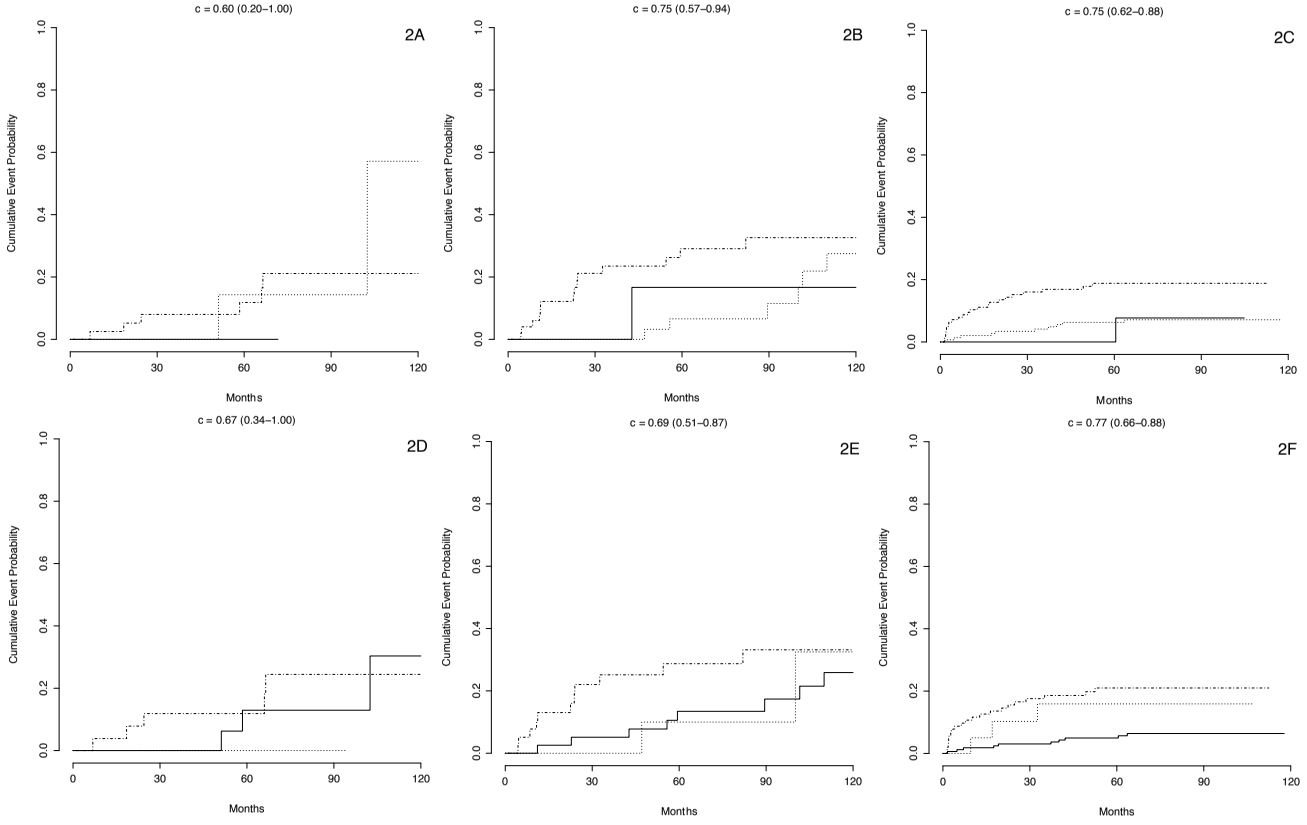
Area	Center	N
Barcelona	Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra (coordinating centre)	
Barcelona	Hospital del Mar (Barcelona)	88
Barcelona	Hospital Germans Tries i Pujol (Badalona, Barcelona)	101
Barcelona	Hospital de Sant Boi (Sant Boi, Barcelona)	27
Barcelona	Centre Hospitalari Parc Taulí (Sabadell, Barcelona)	95
Barcelona	Centre Hospitalari i Cardiològic (Manresa, Barcelona)	64
Tenerife	Hospital Universitario (La Laguna, Tenerife)	42
Tenerife	Hospital La Candelaria (Santa Cruz, Tenerife)	107
Alicante	Hospital General de Elche (Elche, Alicante)	84
Asturias	Hospital Monte Naranco (Oviedo, Asturias)	2
Asturias	Hospital San Agustin (Aviles, Asturias)	86
Asturias	Hospital Central Covadonga (Oviedo, Asturias)	71
Asturias	Hospital Central General (Oviedo, Asturias)	26
Asturias	Hospital de Cabueñes (Gijón, Asturias)	63
Asturias	Hospital de Jove (Gijón, Asturias)	35
Asturias	Hospital de Cruz Roja (Gijón, Asturias)	27
Asturias	Hospital Alvarez-Buylla (Mieres, Asturias)	21
Asturias	Hospital Jarrio (Coaña, Asturias)	24
Asturias	Hospital Carmen y Severo Ochoa (Cangas, Asturias)	10

Supplementary figure 7.1A-F Kaplan-Meier estimates of recurrence of bladder cancer in a ten-year period from transurethral resection of a bladder tumour for patients with non-muscle invasive bladder cancer treated with BCG.



Full line: low risk patients, dotted line: intermediate risk patients, dashed line: high risk patients
Number of patients per country: Denmark n=52 (left figures); The Netherlands n=108 (middle figures); Spain n=289 (right figures)
*ES = EORTC risk score for the recurrence/progression of non-muscle invasive bladder cancer, figures A-C
**CS = CUETO risk score for the recurrence/progression of non-muscle invasive bladder cancer, figures D-F

Supplementary figure 7.2A-F Kaplan-Meier estimates of progression of bladder cancer in a ten-year period from transurethral resection of a bladder tumour for patients with non-muscle invasive bladder cancer treated with BCG.



Full line: low risk patients, dotted line: intermediate risk patients, dashed line: high risk patients
Number of patients per country: Denmark n=52 (left figures); The Netherlands n=108 (middle figures); Spain n=289 (right figures)
*ES = EORTC risk score for the recurrence/progression of non-muscle invasive bladder cancer, figures A-C
**CS = CUETO risk score for the recurrence/progression of non-muscle invasive bladder cancer, figures D-F

PART V

Discussion

Chapter 8

General discussion

The aim of this thesis was to study methods to evaluate biomarkers, and to apply these methods in two urological cancers. The main focus was on methods of biomarker evaluation, early HTA of a new biomarker, and late evaluation. In this chapter, the most important findings of the studies are described first. Next, possible shortcomings and methodological considerations are discussed. This chapter ends with an overall conclusion, the practical implications of these studies, and recommendations for future research.

8.1 METHODS OF BIOMARKER EVALUATION

Main findings

In the first part of this thesis, methods of biomarker evaluation were studied. The first research question of this thesis was:

What are controversies in performing and reporting the Net Reclassification Improvement (NRI) and how can the graphical assessment of incremental value of new biomarkers be improved?

The research described in chapter 2 and 3 of this thesis shows that the NRI seems simple and intuitive to interpret, but is an easily misleading summary measure for the predictive evaluation of a biomarker. This is primarily because the NRI is the sum of two conditional probabilities. A systematic approach towards presenting NRI analysis was proposed in chapter 2, where we made the following recommendations: motivate the specific methods used for computation of the NRI, use clinically meaningful risk cut-offs for the category-based NRI, report both NRI components, address issues of calibration, and do not interpret the overall NRI as a percentage of the study population reclassified. A net reclassification risk graph provides further insights beyond the NRI and is therefore possibly more useful for clinical practice.

Methodological considerations

New biomarkers may improve prediction of binary health outcomes. The NRI is an increasingly popular measure for evaluating improvements in risk predictions when new predictors, such as biomarkers, are added. The NRI was first proposed by Pencina [1] and has quickly been adopted as a standard measure to quantify incremental value in major medical journals. The NRI intends to quantify the extent to which the new biomarker improves the classification of patients into clinically distinct categories. The NRI is determined for those with events and those without events, and a total NRI is the sum of these two numbers. This sum if often interpreted as 'the fraction of

patients correctly reclassified'. This interpretation is correct for the NRI for events and the NRI for those without events, but not for the sum of these two numbers. There is no simple and correct summary in words for the sum of these conditional probabilities. The NRI hence seems to offer a simple intuitive way of quantifying improvement by a new biomarker, but this simple interpretation is incorrect. Hence, the NRI is an easy but misleading summary measure. Its simplicity has likely contributed to its popularity among researchers. Many other criticisms have been debated for the NRI, including the fact that it does not properly take into account the weight of the harms and benefits associated with false positive or false negative test outcomes.

Besides the NRI, various graphs and summary measures can be used to assess the incremental predictive value of a biomarker, as was described in chapter 3. A density plot or a box plot is a potential graphical form to illustrate separation in predicted risks between those with and without a biomarker. The discrimination of a risk prediction model measures that model's ability to distinguish between subjects with and without events. When a biomarker well separates the predicted risks for those with and without an event, it is useful for prediction purposes. A popular measure of discrimination is the area under the receiver operating characteristic curve (AUC). It shows the relation between the sensitivity (true positive rate) and 1 minus the specificity (false positive rate). The AUC is the probability that given two subjects, the model will assign a higher probability to the subject with than the subject without an event. Therefore, an AUC of 0.5 is comparable to a coin toss while an AUC of 1 means perfect discrimination [2].

The increase in the AUC for a model where a new biomarker is added to an existing model depends on the strength of the baseline model. The increment, ΔAUC, does not have a direct interpretation of its own, other than the increase in probability of correct classification as defined above. This increment should always be reported together with the AUC of the baseline model to put it in proper context. For example, to increase the AUC from 0.50 to 0.55, a new predictor with a weak effect size will suffice; to increase the AUC from 0.85 to 0.90, we need a new predictor with a much stronger effect size [3].

Both the NRI and the AUC have the disadvantage of not taking harms of false positives and false negatives explicitly into account. A novel method for evaluating risk prediction models is Decision Curve Analysis (DCA), a method that accounts for clinical consequences of a decision by weighting the harms and benefits of under- and over-treatment [4]. This method uses the threshold probability of a disease or event, defined as the threshold at which a patient or doctor would opt for treatment considering the relative harms of a false-positive and a false-negative prediction. This theoretical rela-

tionship is used to derive the net benefit of the model across different threshold probabilities. Plotting net benefit against threshold probability yields the decision curve. This method does take the weight of false positives and false negatives into account, but it has the downside of being hard to interpret. Furthermore, we should realize that miscalibration of an existing model can result in a DCA where a treat-all strategy is better than using a risk prediction model. This behavior is consistent with the use of a prediction model for decision making, but different from a statistical perspective where we value the predictive information per se. If a model is well calibrated, the decision analytic and statistical perspectives agree [5].

8.2 EARLY HTA OF A NEW BIOMARKER

Main findings

In the second part of this thesis, the cost-effectiveness of adding a new biomarker to PSA-based screening for prostate cancer is studied. The second research question of this thesis is:

Under what conditions is adding a new biomarker to PSA testing cost-effective for prostate cancer screening?

As is described in chapter 4, Prostate Specific Antigen (PSA) combined with a novel biomarker could lead to reductions in the number of biopsies needed to find the same amount of cancers as with PSA-only screening. However, PSA combined with a novel biomarker will only be a cost-effective alternative to screening with PSA alone if costs are very low, or selectively used in those with high PSA.

Methodological considerations

It was shown in chapter 4 that a reference strategy of annual screening with PSA alone led to 290 biopsies to diagnose 74 cancers per 1,000 men. In the best-case scenario, the combination of PSA and the novel biomarker was assumed to have positive predictive value (PPV) of 100%. The number of biopsies would then decline from 290 to 113, a reduction with a factor 2.6. With the best-case novel biomarker, there were still 39 unnecessary biopsies carried out that would not have been initiated without screening, i.e. for men dying of other causes before the prostate cancer would become clinically significant. However, this is a substantial reduction of 216 unnecessary biopsies to 39, i.e. a factor of 5.5 less biopsies that result in overdiagnosis compared to the reference strategy. With a PPV of 100% (perfect specificity), screening with PSA combined with a novel biomarker was expected to be cost-effective if the costs of the biomarker test

did not exceed €20, assuming a threshold of €50.000/QALY, respectively. For screening every 4 years, this maximum biomarker cost would be higher (€50).

The difference outcomes between the annual and the four-yearly strategy can be explained by the fact that there are fewer tests that need to be carried out. Screening every four years results in more overdiagnoses than screening annually. Prostate cancer is a slowly developing disease and screening for prostate cancer is associated with lead time bias, defined as the time by which screening advances the diagnosis compared with absence of screening [6]. Screening annually will find more prostate cancer cases, but not necessarily more clinically relevant or aggressive cancers.

In a strategy were the novel biomarker test was performed only after a PSA test >3.0ng/ml, the number of avoided biopsies and QALYs saved were equal to these numbers in a test all scenario. However, because less novel biomarker tests were carried out, total costs for screening with this strategy were reduced. The maximum price of the test may be substantially higher, up to €280 for a threshold of €50,000/QALY, at perfect PPV of 100%.

MISCAN prostate cancer model might be a well validated robust method; however, many assumptions are included in the model. Therefore, the price thresholds are based on the assumption that no additional burden was caused by the biomarker measurement. For the novel biomarker test, we only assume additional costs and no additional disadvantages. Indirect costs are also not considered in the model. Additionally, there is a potential risk of misclassification due to the fact that the ERSPC study used sextant biopsy as the outcome. Sextant biopsies are expected to have lower predictive value due to missed cancers. Using sextant biopsy for repeat screening has been studied. The rate of deaths due to prostate cancer in men with an initial negative biopsy of 0.03% compared favorably to the 0.35% rate of overall prostate cancer mortality [7]. Furthermore, in this field many developments are happening. Specifically, the diagnostic work-up for prostate cancer is nowadays increasingly based on magnetic resonance imaging (MRI) [8]. However, MRI targeted biopsy implemented in a prostate cancer screening setting is still in its infancy. Besides, adding MRI to PSA-based screening for prostate cancer will be associated with more assumptions in terms of personnel and material costs and QALY loss for patients in the model.

8.3 LATE EVALUATION

Main findings

In the third part of this thesis (chapters 5 and 6), the following research question was answered:

What is the added value in predictive ability of new biomarkers to existing risk prediction models for prostate cancer?

Inclusion of the biomarkers PCA3, the 4k-panel and PHI resulted in an increase in predictive ability to previously developed risk prediction models for prostate cancer, but the reduction in the number of unnecessary biopsies was limited. We furthermore confirmed that after development of a risk score, external validation is crucial.

The second research question for this part of the thesis was answered in chapter 7:

How well can recurrence and progression of bladder cancer be predicted with current models?

It was found that the EORTC and CUETO risk scores reasonably predict progression in non-muscle invasive bladder cancer (NMIBC) patients, but that they fail in the prediction of recurrence. These risk scores predicted progression with AUC values ranging from 0.72 to 0.82 while for recurrence, these ranged from 0.55 to 0.61.

Methodological considerations

In chapters 5 and 6 it was shown that inclusion of several biomarkers resulted in an increase in predictive ability of existing risk prediction scores for prostate cancer. The decision curve analyses confirmed these patterns. DCA curves were used to study clinical value of the updated risk prediction model for prostate cancer, compared to the original ERSPC model. It was shown that considering PCA3 or PHI leads to further reductions in rates of unnecessary biopsies as compared to a biopsy all strategy and compared to the original risk prediction score for prostate cancer. These reductions were, however, limited and may not outweigh the costs of these biomarker tests. DCA is well accepted in the field of prostate cancer research, and proved valuable in the overall judgement of clinical usefulness of the markers considered in the specific context that we considered.

The ERSPC risk calculator discriminated well between those with and without prostate cancer among initially screened men, and outperformed a PSA and DRE-based ap-

proach in the decision to perform a biopsy. However, the model overestimated the risk of a positive biopsy. The early detection of prostate cancer has led to the increased incidence of tumours that are unlikely to become symptomatic during life ("indolent cancers"). The ability to predict indolent prostate cancer is needed to avoid overtreatment by unnecessary invasive therapies and to select men for active surveillance. Statistical power to study the predictive ability for risk prediction models especially for men with a previous negative biopsy or for clinically relevant cancer was compromised by the low number of patients in subgroup settings.

The addition of PCA3 to risk assessment tools leads to an increase in predictive capability for rebiopsy in other studies [9]. Based on research performed for this thesis, we can conclude that the PCA3 test is not capable of replacing the PSA test in clinical practice and an appropriate cut-off level with acceptable performance characteristics is hard to define. Its value as a first-line diagnostic test is limited.

In chapter 7, risk prediction scores for bladder cancer were externally validated in an international study population with primary NMIBC. This study evaluated the predictive ability for existing risk scores for bladder cancer in primary patients. The fact that all subjects had primary cancer influences the risk prediction. The fact that there was a previous tumour is a risk factor in the CUETO risk prediction score and the number of recurrences is a risk factor in the EORTC risk prediction score. Bladder cancer recurrence and progression is expected to be better predicted in patients with previous bladder cancer. Therefore it is important to validate the risk prediction scores in primary bladder cancer patients only. AUC values were lower than in comparable studies that also included recurrent cases [10-13].

A possible explanation for the poor performance of the risk scores for the prediction of recurrence outside controlled trials is interobserver variability in bladder cancer staging and grading by pathologists. To partly overcome these issues, new methods for bladder cancer pathology have been introduced in 1998 and 2004, but the data used in this study is collected during a previous period. Progression of bladder cancer is a more definitive end point than recurrence. It is therefore easier to discriminate between cases and non-cases for progression than for recurrence of bladder cancer.

For the prediction of recurrence and progression of NMIBC we only studied the AUC. We did not study relative weights of a false positive and a false negative finding in a DCA, because the harms of missing cancers and overdiagnosis are not well quantified. Therefore, a threshold probability at which the treating physician would opt for the

appropriate time between consecutive follow-ups is not well defined. More general, NMBIC management would benefit from personalized surveillance protocols [14].

8.4 LIMITATIONS

Methods for biomarker evaluation

Methods for biomarker evaluation are abundant, but there is no one perfect measure that combines proper information provision with easy interpretation. A first step is often an ROC curve, depicted in figure 8.1 with data described in chapter 6 of this thesis. The main downside of this type of analysis is the lack of information on the harms associated with missed cancer diagnoses on the one hand and overdiagnoses on the other.

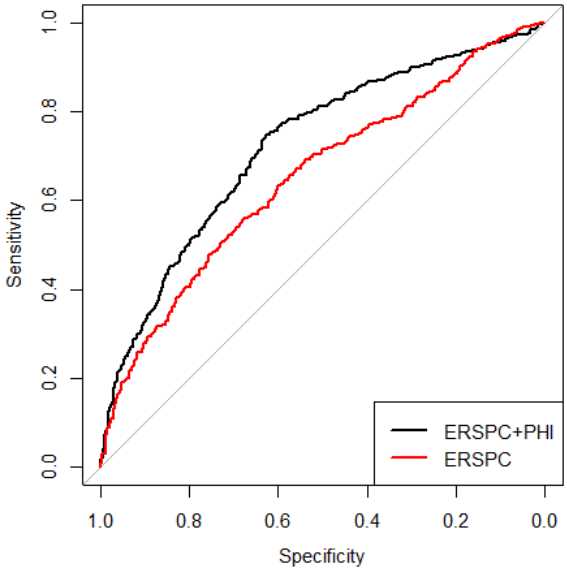


Figure 8.1 The ROC curve for the prediction of prostate cancer with the original (red line) and PHI-updated (black line) ERSPC models for total cancer in all 1185 men, with an AUC of 0.65 for the original model and 0.72 for the PHI-updated model

Decision Curve Analysis is able to overcome this problem and was advocated as a 'simple method for evaluating prediction methods, diagnostic tests, and molecular markers' [4]. In a DCA curve, the decision curve is plotted over a range of threshold probabilities, as is shown figure 8.2. This curve is based on the net benefit of the

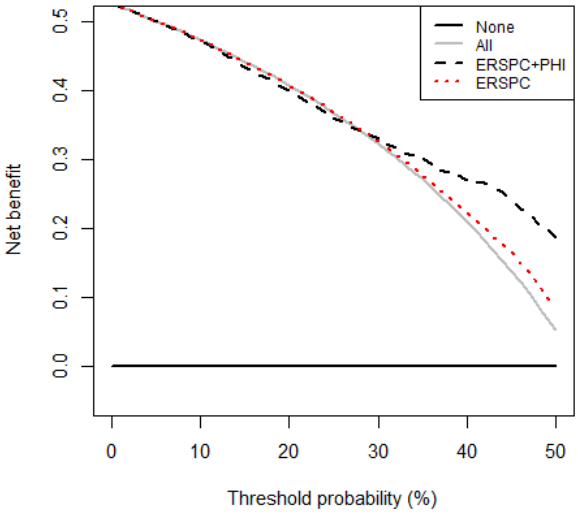


Figure 8.2 The DCA curve for the prediction of prostate cancer with the original (red, dotted line) and the PHI-updated (black, dashed line) ERSPC models for total cancer in all 1185 men

decision for each threshold probability. The key assumption is that the threshold probability of a disease or event at which a patient would opt for treatment is informative of how the patient weighs the relative harms of a false-positive and a false-negative prediction [15]. This theoretical relationship between harms and benefits of a medical decision overcomes the disadvantages of ROC analysis, but at a price of being harder to understand and interpret by clinicians.

A method that might overcome both these pitfalls is a combination of both methods where the sensitivity and specificity of a risk prediction model are plotted on the y-axes while the threshold probability increases over the x-axis, as is shown in figure 8.3. This plot clarifies the exchange between sensitivity and specificity and the threshold were the two lines for sensitivity and specificity meet where there is a maximum sum of the two. The total of harms and benefits requires the absolute numbers of patients that are correctly and incorrectly classified at different thresholds. These numbers might be shown in separate tables, or below the graph.

Early HTA of a new biomarker

The most important methodological limitation concerning our early HTA study is the theoretical nature of the data. Specifically, the diagnostic qualities of new biomarkers are speculative at early phases, where initial results are usually promising. While the underlying model was based on a large, European trial with over 160,000 men participating, it remains a modeling study in which many assumptions had to be made.

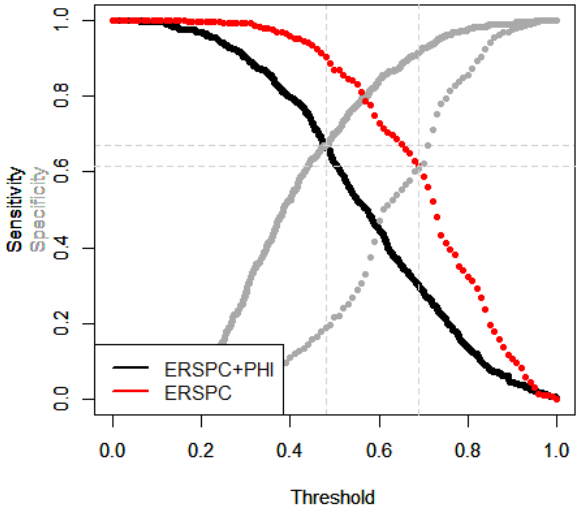


Figure 8.3 The sensitivity/specificity versus probability threshold curve for the prediction of prostate cancer with the original (red, dotted line for sensitivity and grey, dotted line for specificity) and PHI-updated (black, full line for sensitivity and grey, full line for specificity) ERSPC models for total cancer in all 1185 men

In the MISCAN prostate model, there was no distinction made between total cancer, clinically relevant cancer, and indolent cancer. Even with a perfect biomarker, with 100% sensitivity and specificity, there will be some overdiagnoses of men that would never have had complains during their life, but who will receive medical care after being diagnosed during screening. This modelling study hence provides the starting point of what a new biomarker may cost.

Late evaluation

Another limitation was that biomarkers studied in this thesis were not primarily meant to replace PSA. They were intended to be considered in subgroups of men where there is low predictive ability for PSA, such as men with a previous negative biopsy. Indeed, PCA3 is associated with an increase in predictive capability for rebiopsy [9] and this was also seen in chapter five of this thesis. However, in this chapter no comparison with men with an initial biopsy was made.

8.5 DIRECTIONS FOR FUTURE RESEARCH

Methods for biomarker evaluation

- Promising NRI findings need to be followed with decision analytic or formal cost-effectiveness evaluations.
- The best way to determine usefulness of a prediction model is with a DCA or another method that includes the relative weights of a false positive and a false negative finding.
- The sensitivity/specificity versus probability threshold curve should be further developed, since the DCA is difficult to interpret for physicians.

Biomarkers in urological cancers

- Prospective studies with multivariable analyses, including larger sample sizes and avoiding attribution bias caused by preselection on the basis of serum PSA, are required to study new biomarkers for prostate cancer for men where there is low predictive ability for PSA.
- Patient stratification between men with indolent prostate cancer and those with clinically relevant cancer is necessary to identify those men who may actually benefit from early detection.
- Since successful discrimination of low and high risk patients is essential to the right intensity of bladder cancer surveillance, new prognostic biomarkers are needed to better predict recurrence of tumours in primary NMIBC patients.

8.6 CONCLUSIONS AND RECOMMENDATIONS

Based on this thesis, the following conclusions can be drawn:

- The NRI seems simple and intuitive to interpret, but is an easily misleading summary measure for the predictive evaluation of a biomarker.
- A net reclassification risk graph provides further insights beyond the NRI and is therefore possibly more useful for clinical practice.
- PSA-based screening combined with a novel biomarker will lead to reductions in the number of biopsies needed to find the same number of cancers as with PSA-only screening. However, PSA-based screening combined with a novel biomarker will only be a cost-effective alternative to screening with PSA alone if costs are very low, or selectively used in those with high PSA.
- PSA has been shown to be the single most significant predictive factor for identifying men at increased risk of developing prostate cancer. A risk prediction model based on the ERSPC study outperforms risk stratification based on PSA alone, and should therefore always be considered when testing for prostate cancer.
- Both the PCA3 and PHI and, to a lesser extent, a 4k-panel, have added value to the DRE-based ERSPC risk calculator in detecting prostate cancer. Inclusion of these new biomarkers results in an increase in predictive ability to previously developed risk prediction models for prostate cancer, but the reduction in the number of unnecessary biopsies is limited.
- Previously developed risk scores for bladder cancer reasonably predict progression of primary NMIBC, but they fail in the prediction of recurrence.
- For the evaluation of biomarkers, a method that takes into account the relative weights of a false positive and a false negative finding should be used.
- The addition of recently discovered biomarkers to existing risk scores for the diagnosis of prostate cancer is not recommended at this moment, since these biomarkers are insufficiently cost-effective.

REFERENCES

1. 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 Jan 30;27(2):157-72; discussion 207-12.
2. Steyerberg EW. Clinical Prediction Models. Gail M, Krickeberg K, Samet J, Tsiatis A, Wong W, editors: Springer; 2009.
3. Pencina MJ, D'Agostino RB, Pencina KM, Janssens AC, Greenland P. Interpreting incremental value of markers added to risk prediction models. Am J Epidemiol. 2012 Sep 15;176(6): 473-81.
4. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006 Nov-Dec;26(6):565-74.
5. Leening MJ, Steyerberg EW, Van Calster B, D'Agostino RB, Sr., Pencina MJ. Net reclassification improvement and integrated discrimination improvement require calibrated models: relevance from a marker and model perspective. Stat Med. 2014 Aug 30;33(19):3415-8.
6. Finne P, Fallah M, Hakama M, Ciatto S, Hugosson J, de Koning H, *et al.* Lead-time in the European Randomised Study of Screening for Prostate Cancer. Eur J Cancer. 2010 Nov; 46(17):3102-8.
7. Schroder FH, van den Bergh RC, Wolters T, van Leeuwen PJ, Bangma CH, van der Kwast TH, *et al.* Eleven-year outcome of patients with prostate cancers diagnosed during screening after initial negative sextant biopsies. Eur Urol. 2010 Feb;57(2):256-66.
8. Schoots IG, Roobol MJ, Nieboer D, Bangma CH, Steyerberg EW, Hunink MG. Magnetic Resonance Imaging-targeted Biopsy May Enhance the Diagnostic Accuracy of Significant Prostate Cancer Detection Compared to Standard Transrectal Ultrasound-guided Biopsy: A Systematic Review and Meta-analysis. Eur Urol. 2014 Dec 2.
9. Roobol MJ, Schroder FH, van Leeuwen P, Wolters T, van den Bergh RC, van Leenders GJ, *et al.* Performance of the prostate cancer antigen 3 (PCA3) gene and prostate-specific antigen in prescreened men: exploring the value of PCA3 for a first-line diagnostic test. Eur Urol. 2010 Oct;58(4):475-81.
10. Hernandez V, De La Pena E, Martin MD, Blazquez C, Diaz FJ, Llorente C. External validation and applicability of the EORTC risk tables for non-muscle-invasive bladder cancer. World J Urol. 2011 Aug;29(4):409-14.
11. Buethe DD, Sexton WJ. Bladder cancer: validating the EORTC risk tables in BCG-treated patients. Nat Rev Urol. 2011 Sep;8(9):480-1.
12. Rosevear HM, Lightfoot AJ, Nepple KG, O'Donnell MA. Usefulness of the Spanish Urological Club for Oncological Treatment scoring model to predict nonmuscle invasive bladder cancer recurrence in patients treated with intravesical bacillus Calmette-Guerin plus interferon-alpha. J Urol. 2011 Jan;185(1):67-71.

13. Xylinas E, Kent M, Kluth L, Pycha A, Comploj E, Svatek RS, *et al.* Accuracy of the EORTC risk tables and of the CUETO scoring model to predict outcomes in non-muscle-invasive urothelial carcinoma of the bladder. Br J Cancer. 2013 Sep 17;109(6):1460-6.

14. Geavlete B, Stanescu F, Moldoveanu C, Jecu M, Adou L, Bulai C, *et al.* NBI cystoscopy and bipolar electrosurgery in NMIBC management - An overview of daily practice. J Med Life. 2013 Jun 15;6(2):140-5.

15. Localio AR, Goodman S. Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. Ann Intern Med. 2012 Aug 21;157(4):294-5.

16. Nichol MB, Wu J, Huang J, Denham D, Frencher SK, Jacobsen SJ. Cost-effectiveness of Prostate Health Index for prostate cancer detection. BJU Int. 2012 Aug;110(3):353-62.

PART VI

Miscellaneous

Summary

I INTRODUCTION

Biomarkers are increasingly studied in medical research. In recent history, the use of biomarkers in diagnosis and staging of cancer has increased. When a biomarker is considered for use the biomarker first has to be evaluated for its usefulness. There are several ways to evaluate new biomarkers; a systematic way to evaluate new technologies in medicine is called Health Technology Assessment (HTA). Within this framework, both the effectiveness and the cost-effectiveness of a new biomarker are assessed. A new biomarker should not be tested on its own, but within the body of knowledge that exists. Some techniques to do this include (improvement in) the area under the curve (AUC) of a receiver operating characteristic curves with all possible cut-offs for predictions, the Net Reclassification Improvement (NRI), and decision curve analysis (DCA) over a plausible range of cut-off for predictions. In this thesis, the focus is on the evaluation of biomarkers with an application in urological cancers. Both the effectiveness of biomarkers as well as the cost-effectiveness is considered, with an application in two urological cancers: prostate cancer and bladder cancer.

Aim of this thesis

The aim of this thesis is to assess the different methods of evaluating new biomarkers. The focus is on patients with bladder or prostate cancer. The main study questions are:

- What are controversies in performing and reporting the Net Reclassification Improvement (NRI) and how can the graphical assessment of incremental value of new biomarkers be improved?
- Under what conditions is adding a new biomarker to PSA testing cost-effective for prostate cancer screening?
- What is the added value in predictive ability of new biomarkers to existing risk prediction models for prostate cancer?
- How well can recurrence and progression of bladder cancer be predicted with current models?

II METHODS OF BIOMARKER EVALUATION

Chapter 2 assesses the computation, interpretation and controversies to the topic of the Net Reclassification Improvement (NRI). The net reclassification improvement (NRI) is an increasingly popular measure for evaluating improvements in risk predictions. This chapter details a review of 67 publications in high-impact general clinical journals that considered the NRI. To aid improved applications of the NRI, tis study elaborates on

several aspects of the computation and interpretation in various settings. Limitations and controversies are discussed, including the effect of miscalibration of prediction models, the use of the continuous NRI and "clinical NRI," and the relation with decision analytic measures. A systematic approach toward presenting NRI analysis is proposed: Detail and motivate the methods used for computation of the NRI, use clinically meaningful risk cutoffs for the category-based NRI, report both NRI components, address issues of calibration, and do not interpret the overall NRI as a percentage of the study population reclassified.

Chapter 3 describes the graphical assessment of incremental value of new biomarkers in prediction models. In this review, various performance measures for the incremental value of adding HDL cholesterol to a prediction model were studied. An initial assessment may consider statistical significance and distributions of predicted risks (densities or box plots) with various summary measures. A range of decision thresholds is considered in predictiveness and receiver operating characteristic curves, with the difference in AUC. We can furthermore focus on reclassification of participants with and without an event in a reclassification graph, with the continuous NRI as a summary measure. This chapter concludes with the proposition of a net reclassification risk graph, which allows focus on the number of reclassified persons and their event rates. Important insights for impact on decision making are provided by a simple graph for the net reclassification risk.

III EARLY HTA OF A NEW BIOMARKER

Chapter 4 involves the early HTA of new biomarkers. In this part, the cost-effectiveness of prostate cancer screening using a PSA test combined with a new biomarker is determined. To study this, an adapted version of the MISCAN-Prostate microsimulation-model, which includes data from the European Randomised study of Screening for Prostate Cancer (ERSPC) trial, was used. Prostate Specific Antigen (PSA) combined with a novel biomarker could lead to reductions in the number of biopsies needed to find the same amount of cancers as with PSA-only screening. However, PSA combined with a novel biomarker will only be a cost-effective alternative to screening with PSA alone if costs are very low, or selectively used in those with high PSA.

IV LATE EVALUATION

In this part of the thesis, the focus is on late evaluation of risk prediction scores and new biomarkers. In chapter 5, the added value of percentage of free to PSA, PCA3, and a kallikrein panel to the ERSPC risk calculator for prostate cancer in pre-screened men is determined. The added value of these biomarkers was assessed in participants of the ERSPC study that were invited for rescreening. The percentage of free PSA did not perform better univariately or added to the RCs compared with the RCs alone. In men with an elevated PSA, the 4k-panel discriminated better than PCA3 when modelled univariately . The multivariable models with PCA3 or the 4k-panel were equivalent. In the total population, PCA3 discriminated better than the 4k-panel. The multivariable model with PCA3 performed better than the reference model for all men. Decision curves confirmed these patterns, although numbers were small. Based on these results, the conclusion is that both PCA3 and, to a lesser extent, a 4k-panel have added value to the DRE-based ERSPC RC in detecting PCa in prescreened men.

Chapter 6 describes a comparison of the predictive performance of two prostate cancer risk calculators that include the Prostate Health Index. The Lughezzani model, a previously developed PHI based nomogram, was compared with updated versions of the ERSPC model. For the ERSPC updates, the original model was recalibrated and PHI was added as a predictor. It was found that differences between risk calculators that include PHI are small. Considering PHI in a risk calculator leads to further reductions in rates of unnecessary biopsies as compared to a PSA based strategy.

In the last chapter of this part, chapter 7, risk prediction scores for recurrence and progression of non-muscle invasive bladder cancer are validated in men and women with primary bladder cancer . Recurrence-free survival and progression-free survival according to the European Organisation for Research and Treatment of Cancer (EORTC) and the Spanish Urological Club for Oncological Treatment (CUETO) risk scores were evaluated for 1,892 patients. The concordance index (c-index) was used to indicate discriminative ability. The EORTC and CUETO risk scores both predicted progression better than recurrence with c-indices ranging from 0.72 to 0.82 while for recurrence, those ranged from 0.55 to 0.61. Based on these results, the conclusion is that the EORTC and CUETO risk scores can reasonably predict progression, while prediction of recurrence is more difficult.

V DISCUSSION

In the first part of this thesis, methods of biomarker evaluation were studied. The first research question of this thesis is: **What are controversies in performing and reporting the Net Reclassification Improvement (NRI) and how can the graphical assessment of incremental value of new biomarkers be improved?**

The research described in chapter 2 and 3 of this thesis shows that the NRI seems simple and intuitive to interpret, but is an easily misleading summary measure for the predictive evaluation of a biomarker. This is primarily because the NRI is the sum of two conditional probabilities. When the NRI is used, one must motivate the specific methods used for computation of the NRI, use clinically meaningful risk cut-offs for the category-based NRI, report both NRI components, address issues of calibration, and not interpret the overall NRI as a percentage of the study population reclassified. A net reclassification risk graph provides further insights beyond the NRI and is therefore possibly more useful for clinical practice.

In the second part of this thesis, the following research question is answered: **Under what conditions is adding a new biomarker to PSA testing cost-effective for prostate cancer screening?**

PSA combined with a novel biomarker could lead to reductions in the number of biopsies needed to find the same amount of cancers as with PSA-only screening. However, PSA combined with a novel biomarker will only be a cost-effective alternative to screening with PSA alone if costs are very low, or selectively used in those with high PSA. A limitation of this study is that many assumptions are included in the MISCAN prostate cancer model.

In the third part of the thesis, the theory of biomarker evaluation is applied to two urological cancers. In chapter 5 and 6, the following research question was answered: **What is the added value in predictive ability of new biomarkers to existing risk prediction models for prostate cancer?**

Inclusion of the biomarkers PCA3, the 4k-panel and PHI resulted in an increase in predictive ability to previously developed risk prediction models for prostate cancer, but the reduction in the number of unnecessary biopsies was limited. It should be noted that biomarkers studied in these chapters were not primarily meant to replace PSA, but that they were intended to be considered in subgroups of men where there is low predictive ability for PSA, such as men with a previous negative biopsy

The second research question for this part of the thesis was: **How well can recurrence and progression of bladder cancer be predicted with current models?**

It was found that the EORTC and CUETO risk scores reasonably predict progression in primary non-muscle invasive bladder cancer (NMIBC) patients, but that they fail in the prediction of recurrence. It is furthermore confirmed that after development of a risk score, external validation is crucial.

Based on this thesis, the following conclusions can be drawn:

- The NRI seems simple and intuitive to interpret, but is an easily misleading summary measure for the predictive evaluation of a biomarker.
- A net reclassification risk graph provides further insights beyond the NRI and is therefore possibly more useful for clinical practice.
- PSA-based screening combined with a novel biomarker will lead to reductions in the number of biopsies needed to find the same number of cancers as with PSA-only screening. However, PSA-based screening combined with a novel biomarker will only be a cost-effective alternative to screening with PSA alone if costs are very low, or selectively used in those with high PSA.
- PSA has been shown to be the single most significant predictive factor for identifying men at increased risk of developing prostate cancer. A risk prediction model based on the ERSPC study outperforms risk stratification based on PSA alone, and should therefore always be considered when testing for prostate cancer.
- Both the PCA3 and PHI and, to a lesser extent, a 4k-panel, have added value to the DRE-based ERSPC risk calculator in detecting prostate cancer. Inclusion of these new biomarkers results in an increase in predictive ability to previously developed risk prediction models for prostate cancer, but the reduction in the number of unnecessary biopsies is limited.
- Previously developed risk scores for bladder cancer reasonably predict progression of primary NMIBC, but they fail in the prediction of recurrence.
- For the evaluation of biomarkers, a method that takes into account the relative weights of a false positive and a false negative finding should be used.
- The addition of recently discovered biomarkers to existing risk scores for the diagnosis of prostate cancer is not recommended at this moment, since these biomarkers are insufficiently cost-effective.

Samenvatting

I. INLEIDING

Biomarkers worden steeds meer bestudeerd binnen het medisch onderzoek. De afgelopen jaren is het gebruik van biomarkers bij het stellen van de diagnose en het bepalen van het stadium van kanker toegenomen. Wanneer een biomarker wordt overwogen voor gebruik moet deze eerst worden geëvalueerd op zijn bruikbaarheid. Er zijn verschillende manieren om nieuwe biomarkers te evalueren; een systematische manier om nieuwe technologieën te evalueren in de geneeskunde wordt Health Technology Assessment (HTA) genoemd. Binnen HTA worden zowel de effectiviteit als de kosteneffectiviteit van een nieuwe biomarker beoordeeld. Een nieuwe biomarker dient niet op zichzelf te worden getest, maar binnen de bestaande kennis over voorspellers van de betreffende ziekte. Verschillende technieken om dit te doen zijn onder andere (het verbeteren van) het oppervlakte onder de 'receiver operating characteristic' curve (AUC genoemd); de netto herclassificatie verbetering (NRI); en beslissingscurve analyse (DCA), over een plausibel bereik van afkappunten voor voorspellingen. In dit proefschrift ligt de nadruk op de evaluatie van biomarkers waarbij deze zijn toegepast op het voorspellen van de diagnose en het verloop van urologische kankers. Zowel de effectiviteit van biomarkers als de kosteneffectiviteit worden behandeld, toegepast op twee urologische kankers: prostaat- en blaaskanker.

Doel van dit proefschrift

Het doel van dit proefschrift is het beschrijven en evalueren van de verschillende methoden voor het evalueren van nieuwe biomarkers. De focus ligt op patiënten met blaas- of prostaatkanker. De belangrijkste onderzoeksvragen zijn:

- Wat zijn de controverses in de uitvoering en de rapportage van de netto herclassificatie verbetering (Net Reclassification Index - NRI) en hoe kan de grafische evaluatie van de incrementele waarde van nieuwe biomarkers worden verbeterd?
- Onder welke omstandigheden is het toevoegen van een nieuwe biomarker aan het gebruik van de Prostaat Specifiek Antigeen (PSA) test voor bevolkingsonderzoek naar prostaatkanker kosteneffectief?
- Wat is de toegevoegde waarde in voorspellend vermogen van nieuwe biomarkers voor bestaande risicovoorspelmodellen voor prostaatkanker?
- Hoe goed kunnen terugkeer en progressie van blaaskanker worden voorspeld met de huidige risicovoorspelmodellen?

II METHODEN VOOR BIOMARKER EVALUATIE

In hoofdstuk 2 worden de berekening, interpretatie en controverses op het gebied van de NRI behandeld. De NRI is een steeds populairder wordende methode voor het evalueren van eventuele verbetering in risicovoorspelmodellen. Hoofdstuk 2 beschrijft een overzicht van 67 publicaties in klinische tijdschriften waarin de NRI wordt toegepast. Er wordt verder ingegaan op diverse aspecten van de berekening en de interpretatie om betere toepassing van de NRI te bevorderen. Beperkingen en controverses worden besproken, inclusief het effect van miscalibratie van risicovoorspelmodellen, het gebruik van de continue NRI en klinische NRI en de relatie met andere evaluatiemethoden. Verder wordt er in dit hoofdstuk een systematische benadering van de presentatie van NRI analyse voorgesteld, namelijk: motiveer hoe en waarom de NRI is gebruikt, gebruik voor de berekening van de NRI klinisch relevante afkappunten, vermeld alle componenten van de berekening en interpreteer de algehele NRI niet als het percentage van de studiepopulatie dat correct geclassificeerd is.

Hoofdstuk 3 beschrijft de grafische evaluatie van de incrementele waarde van nieuwe biomarkers in risicovoorspelmodellen. In dit overzicht worden diverse prestatie-indicatoren voor de incrementele waarde van het toevoegen van HDL-cholesterol aan een bestaand risicovoorspelmodel bestudeerd. Een eerste beoordeling kan bestaan uit een beoordeling van statistische significantie van een voorspellende biomarker en de verdeling van de voorspelde risico's. Een plausibel bereik van afkappunten van een risicovoorspelmodel wordt overwogen met een AUC. Daarnaast kan men zich focussen op herclassificatie van de studiepopulatie met en zonder een bepaalde uitkomst, bijvoorbeeld wel of geen diagnose, in een tabel met de NRI als samenvattende maat. Dit hoofdstuk sluit af met de introductie van een 'netto herclassificatie grafiek', die het mogelijk maakt om te focussen op het aantal geherclassificeerde personen met een verbeterde voorspelde uitkomst. Hiermee kan de impact op de besluitvormig over het gebruik van de nieuwe biomarker inzichtelijk worden gemaakt middels een eenvoudige grafiek.

III VROEGE HTA VAN EEN NIEUWE BIOMARKER

Hoofdstuk 4 beschrijft een vroege HTA van nieuwe biomarkers. In dit deel van het proefschrift wordt de kosteneffectiviteit van bevolkingsonderzoek voor prostaatkanker screening met de PSA-test in combinatie met een nieuwe biomarker bepaald. Hiertoe is een aangepaste versie van de MISCAN-Prostaat microsimulatie-model gebruikt. PSA in combinatie met een nieuwe biomarker kan leiden tot een vermindering van het aantal

biopten dat nodig is voor het vinden van hetzelfde aantal prostaatkankers. Echter, PSA gecombineerd met een nieuw biomarker is slechts een kosteneffectief alternatief voor bevolkingsonderzoek met alleen PSA als de kosten zeer laag zijn, of als er selectief gebruikt gebruik wordt gemaakt van de nieuwe biomarker bij mannen met een verhoogd PSA.

IV LATE EVALUATIE

In dit deel van het proefschrift ligt de nadruk op de late evaluatie van risicovoorspelmodellen en nieuwe biomarkers. In hoofdstuk 5 wordt de toegevoegde waarde van het percentage vrij PSA en genetische markers PCA3 en het 4k-panel bepaald voor een risicovoorspelmodel voor prostaatkanker in eerder onderzochte mannen. Het referentiemodel is het ERSPC model, gebaseerd op de Europese Gerandomiseerde Studie voor Bevolkingsonderzoek naar Prostaatkanker (European Randomized Study of Screening for Prostate Cancer - ERSPC). Het percentage vrij PSA presteerde niet beter in een univariate analyse en toegevoegd aan de originele risicovoorspelmodellen dan de originele modellen zelf. Bij mannen met een verhoogd PSA voorspelde het 4k-panel beter dan PCA3 in univariate analyses. De multivariabele modellen met PCA3 of het 4k-panel waren in deze populatie equivalent. In de totale studiepopulatie discrimineerde PCA3 beter dan het 4k-paneel tussen mannen met en zonder kanker. Het multivariabele model met PCA3 presteerde ook beter dan het referentiemodel. DCA curves bevestigden deze uitkomsten. Op basis van deze resultaten wordt geconcludeerd dat zowel PCA3, en in mindere mate het 4k-panel, toegevoegde waarde hebben voor het originele risicovoorspelmodel voor het detecteren van prostaatkanker bij eerder onderzochte mannen.

Hoofdstuk 6 beschrijft een vergelijking van risicovoorspelmodellen voor prostaatkanker die de Prostaat Health Index (PHI) bevatten. Het Lughezzani model, een eerder ontwikkeld model, wordt in dit hoofdstuk vergeleken met bijgewerkte versies van de ERSPC-model. Voor de bijgewerkte versies van het ERSPC model werd het oorspronkelijke model opnieuw gekalibreerd en werd PHI vervolgens toegevoegd als voorspeller. Het bleek dat verschillen tussen risicovoorspelmodellen voor prostaatkanker die PHI bevatten klein zijn. Het toevoegen van PHI zorgde wel voor een verlaging van het aantal biopten dat nodig was voor het vinden van eenzelfde hoeveelheid prostaatkanker diagnoses vergeleken met een op PSA gebaseerde strategie.

In het laatste hoofdstuk van dit deel, hoofdstuk 7, worden twee risicovoorspelmodellen voor terugkeer en progressie van blaaskanker gevalideerd. Het onderzoek focust

zich op mannen en vrouwen met primaire niet-spier-invasieve blaaskanker. Ziektevrije en progressievrije overleving werden beoordeeld voor 1892 patiënten en vergeleken met de uitkomst van twee veel gebruikte risicovoorspelmodellen: EORTC en CUETO. De concordantie index (c-index) werd gebruikt om onderscheidend vermogen te evalueren. De EORTC en CUETO risico scores voorspelden beiden de progressie beter dan de terugkeer van blaaskanker, met c-indices variërend van 0,72-0,82 voor progressie en 0,55-0,61 voor terugkeer. Op basis van deze resultaten wordt geconcludeerd dat zowel de EORTC als het CUETO risicovoorspelmodel redelijkerwijs progressie kunnen vorspellen. Het voorspellen van terugkeer van kanker is echter moeilijker.

V DISCUSSIE

In het eerste deel van dit proefschrift werden verschillende methoden om biomarkers te evalueren bestudeerd. De eerste onderzoeksvraag van dit proefschrift is: **Wat zijn de controverses in de uitvoering en rapportage van de netto herclassificatie verbetering (Net Reclassification Index - NRI) en hoe kan de grafische evaluatie van de incrementele waarde van nieuwe biomarkers worden verbeterd?**

De in hoofdstuk 2 en 3 van dit proefschrift beschreven onderzoeken tonen aan dat de NRI eenvoudig en intuïtief te interpreteren lijkt, maar dat het een gemakkelijk misleidend samenvattende maat is voor de evaluatie van een biomarker. Dit komt met name doordat de NRI de som is van twee conditionele kansen. Wanneer de NRI wordt gebruikt moet men de specifieke methoden die worden gebruikt voor de berekening van de NRI motiveren, de klinisch relevante afkappunten voor de NRI categorieën beschrijven, kalibratie kwesties melden en de interpretatie van de NRI niet als een percentage van de geherclassificeerde studiepopulatie interpreteren. Een 'netto herclassificatie grafiek' geeft meer inzicht naast de NRI en is daarom wellicht nuttig voor de klinische praktijk.

In het tweede deel van dit proefschrift wordt de volgende onderzoeksvraag beantwoord: **Onder welke omstandigheden is het toevoegen van een nieuwe biomarker aan het gebruik van de Prostaat Specifiek Antigeen (PSA) test voor bevolkingsonderzoek naar prostaatkanker kosteneffectief ?**

PSA gecombineerd met een nieuwe biomarker kan leiden tot een reductie in het aantal biopten dat nodig is om eenzelfde aantal prostaatkankers te vinden bij bevolkingsonderzoek. Deze strategie is echter slechts een kosteneffectief alternatief voor screening met PSA alleen als de kosten van de nieuwe biomarker zeer laag zijn of als deze selec-

tief gebruikt wordt bij mannen met een verhoogd PSA. Een beperking van deze studie is dat er veel aannames zijn gedaan in de MISCAN prostaatkanker model.

In het derde deel van het proefschrift wordt de theorie van de biomarker evaluatie toegepast op twee urologische kankers. In hoofdstuk 5 en 6 werd de volgende onderzoeksvraag beantwoord: **Wat is de toegevoegde waarde in voorspellend vermogen van nieuwe biomarkers voor bestaande risicovoorspelmodellen voor prostaatkanker?**

Inclusie van de biomarkers PCA3, het 4k-panel en PHI resulteerde in een toename van voorspellend vermogen van eerder ontwikkelde risicovoorspelmodellen voor prostaatkanker, maar de vermindering van het aantal onnodige biopsieën bleef beperkt. Opgemerkt dient te worden dat de biomarkers die in deze hoofdstukken onderzocht werden niet primair bedoeld zijn om PSA te vervangen, maar dat ze bedoeld zijn om te worden toegepast in subgroepen van mannen waar minder voorspellende waarde voor PSA is verwacht, bijvoorbeeld na eerdere negatieve biopten.

De tweede onderzoeksvraag voor dit deel van het proefschrift is: **Hoe goed kunnen terugkeer en progressie van blaaskanker worden voorspeld met de huidige risicovoorspelmodellen?**

Het bleek dat het EORTC en CUETO risicoscores redelijk progressie bij primaire niet-spier-invasieve blaaskanker patiënten kunnen voorspellen, maar dat ze falen in het voorspellen van terugkerende blaaskanker. Verder wordt bevestigd dat externe validatie van cruciaal belang is na de ontwikkeling van een risicovoorspelmodel.

Op basis van dit proefschrift kunnen de volgende conclusies worden getrokken:

- De 'net reclassification improvement' (NRI) is makkelijk maar misleidend als samenvattende maat voor de evaluatie van de voorspellende waarde van een biomarker.
- Een 'netto herclassificatie grafiek' geeft meer inzicht naast de NRI en is daarom wellicht nuttig voor de klinische praktijk.
- Het toevoegen van een nieuwe biomarker aan prostaat-specifiek antigeen (PSA) gebaseerde screening zal alleen kosteneffectief zijn als deze biomarker weinig kost of selectief wordt ingezet bij mannen met verhoogde PSA waardes.
- PSA is de belangrijkste voorspellende factor voor de identificatie van mannen met een verhoogd risico op prostaatkanker. Een risicovoorspelmodel gebaseerd op de ERSPC studie overstijgt risico stratificatie gebaseerd op PSA alleen, bij het testen voor prostaatkanker moet dit model daarom altijd in overweging worden genomen.

- Zowel PCA3 en PHI en, in mindere mate, het 4k-panel, zijn van toegevoegde waarde in het ERSPC risicovoorspelmodel bij het opsporen van prostaatkanker. Inclusie van deze nieuwe biomarkers resulteert in een toename van het voorspellend vermogen van eerder ontwikkelde risicovoorspelmodellen voor prostaatkanker. De vermindering van het aantal onnodige biopten is echter beperkt.
- Eerder ontwikkelde risicovoorspelmodellen voor blaaskanker voorspellen redelijk de progressie van primaire niet-spier-invasieve blaaskanker, maar falen in het voorspellen van terugkeer van kanker.
- Voor de evaluatie van biomarkers moet een methode worden gebruikt die rekening houdt met de relatieve gewichten van vals positieve en vals negatieve bevindingen.
- De toevoeging van recent gevonden biomarkers aan bestaande risicovoorspelmodellen voor de diagnose van prostaatkanker wordt op dit moment niet aanbevolen, aangezien deze biomarkers onvoldoende kosteneffectief zijn.

List of publications

LIST OF PUBLICATIONS

2012

Altorf-van der Kuil W, Engberink MF, **Vedder MM**, Boer JM, Verschuren WM, Geleijnse JM. Sources of dietary protein in relation to blood pressure in a general Dutch population. PLoS One. 2012;7(2):e30582

2014

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 Jan 21;160(2):122-31.

Vedder MM, Marquez M, de Bekker-Grob EW, Calle ML, Dyrskjot L, Kogevinas M, *et al.* Risk prediction scores for recurrence and progression of non-muscle invasive bladder cancer: an international validation in primary tumours. PLoS One. 2014;9(6):e96849.

Vedder MM, de Bekker-Grob EW, Lilja HG, Vickers AJ, van Leenders GJ, Steyerberg EW, *et al.* The added value of percentage of free to total prostate-specific antigen, PCA3, and a kallikrein panel to the ERSPC risk calculator for prostate cancer in prescreened men. Eur Urol. 2014 Dec;66(6):1109-15.

2015

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. 2015 Jul;57(4):556-70.

Roobol MJ*, **Vedder MM***, Nieboer D, Houlgatte A, Vincendeau S, Lazzeri M, Guazzoni G, Stephan C, Semjonow A, Haese A, Graefen M, Steyerberg EW. *These authors contributed equally to this work. Comparison of two prostate cancer risk calculators that include the Prostate Health Index (PHI). Eur Urol Focus. 2015;1(2):185-190.

Submitted

Vedder MM, de Bekker-Grob EW, de Koning HJ, Heijnsdijk EAM, Steyerberg EW. Cost-effectiveness of prostate cancer screening using a PSA test combined with a novel biomarker.

Dankwoord

DANKWOORD

De vaak gewichtige dankwoorden in proefschriften doen het lijken alsof 'het boekje' het magnum opus is, terwijl het vaak pas de start is van je carrière. Dit promotieonderzoek is echter een belangrijke fase in mijn leven geweest, waarvan ik op zowel wetenschappelijk als persoonlijk vlak veel geleerd heb. Eén ding is zeker: als ik alles alleen had moeten doen, had hier nu niet dit proefschrift gelegen. Ik wil daarom enkele mensen graag bedanken.

Allereerst wil ik de personen bedanken die hebben deelgenomen aan de wetenschappelijke studies waarvan de data is gebruikt in dit proefschrift. Jullie bijdrage aan medische kennis kan niet genoeg worden benadrukt.

Graag wil ik mijn promotor Prof. Dr. Ewout Steyerberg en copromotor Dr. Esther de Bekker-Grob enorm bedanken. Ewout, je denkt zo snel dat ik je soms niet kan bijbenen, maar je bent altijd bereid dingen aan me uit te leggen. Ik heb veel van je geleerd, zowel wetenschappelijk als over de manieren van de wetenschap. Bedankt dat je zo toegankelijk bent en dat je altijd betrokken en creatief meedenkt, en dat je de begeleiding actief oppakte in de perioden dat Esther afwezig was. Esther, wat moet een chaoot als ik zonder een copromotor als jij? Je hield mij strak wanneer dat nodig was, maar gaf me ook de vrijheid zelf te ontdekken hoe je dingen het best aanpakt. Ook als het even tegen zat kon ik bij jou terecht. Zonder jou had dit proefschrift er écht niet gelegen. Bedankt voor een fijne samenwerking!

Ook de overige leden van de promotiecommissie wil ik graag bedanken voor alle vragen, feedback en de uiteindelijke beoordeling van het manuscript. Jullie hebben dit onderzoek naar een hoger niveau getild. I would like to thank all my colleagues from the UROMOL and PCMM project groups for the opportunity to work with and learn from them and all their valuable input as co-authors and critiquers of the articles published in this thesis.

MGZ was een fijne plek om te werken. Ik zal de gezelligheid, het sparren, maar ook het samen klagen als het tegenzat zeker missen. Ik heb het altijd getroffen met toffe kamergenoten. Maggie, Marie-Louise en Rianne: in het oude gebouw waren jullie mijn kamergenoten met al wat meer werkervaring, jullie namen mij onder jullie hoede en ik heb veel van jullie geleerd. Frederik, Inge, Marcel, Raquel, Nana, Cherry: you're the best! Thanks for all the girls talk and fun times. Ook buiten mijn eigen kamer trof ik gezelligheid tijdens de borrels en in de lunchpauzes en (vele) koffiepauzes, waar ik ook dankbaar voor ben. Daan, jou wil ik specifiek bedanken voor je hulp met R. Sanne,

bedankt voor al je praktische hulp, je hebt me de afgelopen jaren meerdere malen werk uit handen genomen. Ik wens verder iedereen veel succes met zijn/haar onderzoek en het afronden van het promotietraject.

Specifiek wil ik graag mijn twee fantastische paranimfen bedanken. Hilde, ik vond en vind het altijd gezellig met jou en ben blij dat je mijn paranimf wil zijn! Zonder jou was het een stuk minder leuk op MGZ, getuige ook de keren dat ik hoopvol langs je plek liep en ik bedroefd was omdat je een mama-dag bleek te hebben. Ook de uitjes buiten werktijd waren top. Denise, wat fijn dat jij mijn andere paranimf bent. Het stond voor mij eigenlijk al meteen vast dat ik jou naast me wilde hebben staan op deze belangrijke dag. Je bent de meest nuchtere persoon die ik ken, dus samen met jou zijn die zenuwen wel onder controle te houden. Van bloedserieus tot buikpijn van het lachen, op naar de volgende roadtrip!

Ook buiten het werk heb ik het getroffen met alleen maar lieve mensen die vaak geïnteresseerd vroegen hoe het ging met mijn onderzoek, of juist een keertje níet. De dames van Neptunus zorgen voor de nodige sportieve ontspanning naast het werk. Ik wil de meiden van Subliem, Kasjé en de Miepen bedanken voor alle gezellige tijden die we samen hebben gehad en hopelijk nog vaak zullen hebben. Laura en Suzanne, bedankt voor de jarenlange vriendschap. Ik hoop dat ik jullie allemaal nog heel vaak zal zien om bij te kletsen met thee en chocola. En bier. "Friends are the family you choose" en dat geldt zeker voor jullie, Kim-Quy, Iris, Jena, Jordy en Denise. Dat we maar altijd bevriend mogen blijven.

Ik wil ook mijn familie en schoonfamilie bedanken voor alle getoonde interesse – sommigen van jullie hebben zelfs een poging gedaan om een artikel van mij te lezen! Hoewel het niet altijd duidelijk is wat ik nu precies doe, staan jullie in praktische zin altijd voor me klaar. Ik zeg het te weinig, maar ik hou van jullie! Moeders, ik hoop je nog heel vaak trots te mogen maken.

En tenslotte natuurlijk mijn bijna-familie, mijn verloofde, mijn grootste motivatie en tevens grootste afleiding. Jorden, dankjewel dat je bestaat.

Moniek Vedder
Rotterdam, november 2015

PhD portfolio

PHD PORTFOLIO

Name PhD student: Moniek M. Vedder
Erasmus MC Department: Public Health

PhD period: 2012-2015
Promotor: Prof. Dr. E.W. Steyerberg
Supervisor(s): Dr. E.W. de Bekker-Grob

1. PhD training	Year	Workload (ECTS)
General courses		
Scientific integrity	2014	0.3
Biomedical English Writing	2014	3.0
Scientific Writing Course MGZ	2013	1.0
Specific courses		
NIHES Erasmus Summer Programme, Rotterdam	2012	
Causal Inference		0.7
Clinical Decision Analysis		0.7
Methods of Public Health Research		0.7
Markers and Prognostic Research		0.7
Primary and Secondary Prevention Research		0.7
Health Economics		0.7
NIHES Erasmus Winter Programme, Rotterdam	2013	
Diagnostic research		0.7
Advanced topics in decision making in medicine		1.5
NIHES Erasmus Summer Programme, Rotterdam	2013	
Survival Analysis		1.5
Basiscursus oncologie, Nederlandse Vereniging voor Oncologie	2013	1.5
Seminars and workshops		
Absolute risk prediction (Netherlands Cancer Institute, Amsterdam)	2012	0.3
Seminars department of Public Health, Erasmus MC	2012-2015	3.0
Meeting Clinical Decision Making, department of Public Health, Erasmus MC	2012-2015	1.0
Conferences: visits, oral presentations and poster presentations		
Visit LolaHESG, Vereniging voor Gezondheidseconomie, Olmen	2012	0.6
Visit WEON, Nederlandse Vereniging voor Epidemiologie, Rotterdam	2012	0.6
Two posters SMDM, Society for Medical Decision Making, Baltimore, US	2013	2.0
Oral presentation WEON, Nederlandse Vereniging voor Epidemiologie, Utrecht	2013	1.0
Oral presentation SMDM Europe, Antwerpen, Belgium	2013	1.0

Oral presentation WEON, Nederlandse Vereniging voor Epidemiologie, Leiden	2014	1.0
Poster presentation WEON, Nederlandse Vereniging voor Epidemiologie, Leiden	2014	1.0
2. Teaching activities		
Supervisor medical students theme 3.C.4 (Community projects)	2013-2014	4.0
3. Other		
Reviewer for several journals	2013-2015	0.8
Total workload (ECTS)		**30.0**

Curriculum Vitae

CURRICULUM VITAE

Moniek Martine Vedder was born on september 7, 1987 in Rotterdam, the Netherlands. She grew up in Spijkenisse and graduated from secondary school 'Penta College CSG de Blaise Pascal' in 2005. She started her BSc Nutrition and Health at the Wageningen University and graduated in 2009. At the end of 2010, she finished her MSc in Epidemiology and Public Health at that same university. Her MSc thesis focused on the role of dietary protein of different sources in relation to blood pressure. After her graduation, Moniek worked as a data manager at the Leiden University Medical Hospital at the department of Clinical Epidemiology.

When she realized she wanted to study the data herself, she started as Junior Researcher at the Department of Public Health at the Erasmus MC Rotterdam in 2012. This thesis is the result of her work at this department, studying the evaluation of biomarkers and the application of biomarkers in risk prediction scores for urological cancers. At this moment, Moniek works as a Senior Data Analist at Stedin Meetbedrijf.