HHS Public Access

Author manuscript

Eur Urol. Author manuscript; available in PMC 2019 December 01.

Published in final edited form as:

Eur Urol. 2018 December ; 74(6): 796–804. doi:10.1016/j.eururo.2018.08.038.

Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators

Ben Van Calster[a,b,†,*], **Laure Wynants**[a,†], **Jan F.M. Verbeek**[c], **Jan Y. Verbakel**[d,e], **Evangelia Christodoulou**[a], **Andrew J. Vickers**[f], **Monique J. Roobol**[c], and **Ewout W. Steyerberg**[b,c]

[a] Department of Development and Regeneration, KU Leuven, Leuven, Belgium [b] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands [c] Department of Public Health, Erasmus MC, Rotterdam, The Netherlands [d] Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium [e] Nuffield Department of Primary Care Health Sciences, University of Oxford, UK [f] Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Abstract

Context: Urologists regularly develop clinical risk prediction models to support clinical decisions. In contrast to traditional performance measures, decision curve analysis (DCA) can assess the utility of models for decision making. DCA plots net benefit (NB) at a range of clinically reasonable risk thresholds.

Objective: To provide recommendations on interpreting and reporting DCA when evaluating prediction models.

Evidence acquisition: We informally reviewed the urological literature to determine investigators' understanding of DCA. To illustrate, we use data from 3616 patients to develop risk models for high-grade prostate cancer (n = 313, 9%) to decide who should undergo a biopsy. The

[*]Corresponding author. Department of Development and Regeneration, KU Leuven, Herestraat 49 box 805, 3000 Leuven, Belgium. Tel. +3216377788, Ben.vancalster@kuleuven.be (Ben Van Calster).

[†]These authors are joint first authors.

Author contributions: Ben Van Calster had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Van Calster, Wynants, Vickers, Steyerberg.

Acquisition of data: Verbeek, Roobol.

Analysis and interpretation of data: All authors.

Drafting of the manuscript: Van Calster, Wynants.

Critical revision of the manuscript for important intellectual content: All authors.

Statistical analysis: Van Calster, Wynants, Verbakel, Christodoulou, Vickers.

Obtaining funding: None.

Administrative, technical, or material support: None.

Supervision: None.

Other: None.

Financial disclosures: Ben Van Calster certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (eg, employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: None.

baseline model includes prostate-specific antigen and digital rectal examination; the extended model adds two predictors based on transrectal ultrasound (TRUS).

Evidence synthesis: We explain risk thresholds, NB, default strategies (treat all, treat no one), and test tradeoff. To use DCA, first determine whether a model is superior to all other strategies across the range of reasonable risk thresholds. If so, that model appears to improve decisions irrespective of threshold. Second, consider if there are important extra costs to using the model. If so, obtain the test tradeoff to check whether the increase in NB versus the best other strategy is worth the additional cost. In our case study, addition of TRUS improved NB by 0.0114, equivalent to 1.1 more detected high-grade prostate cancers per 100 patients. Hence, adding TRUS would be worthwhile if we accept subjecting 88 patients to TRUS to find one additional high-grade prostate cancer or, alternatively, subjecting 10 patients to TRUS to avoid one unnecessary biopsy.

Conclusions: The proposed guidelines can help researchers understand DCA and improve application and reporting.

Patient summary: Decision curve analysis can identify risk models that can help us make better clinical decisions. We illustrate appropriate reporting and interpretation of decision curve analysis.

Keywords

Clinical utility; Decision curve analysis; Net benefit; Risk prediction models; Risk threshold; Test tradeoff

1. Introduction

Clinical risk prediction models are commonly developed in urology and other medical fields to predict the probability or risk of a current disease (eg, biopsy-detectable aggressive prostate cancer), or a future state (eg, cancer recurrence) [1–3]. Such models are usually evaluated with statistical measures for discrimination and calibration. Discrimination evaluates how well the predicted risks distinguish between patients with and without disease. The c-statistic is the most commonly used measure for discrimination. Calibration evaluates the reliability of the estimated risks: if we predict 10%, on average 10 out of 100 patients should have the disease [1,4]. Assessments of calibration may include graphs and statistics such as observed versus expected ratios or calibration slopes. Although a model with better discrimination and calibration should theoretically be a better guide to clinical management [4–6], statistical measures fall short when we want to evaluate whether the risk model improves clinical decision making. Such measures cannot inform us whether it is beneficial to use a model to make clinical decisions or which of two models leads to better decisions, especially if one model has better discrimination and the other better calibration [7].

To overcome this limitation, decision-analytic measures have been developed to summarize the performance of the model in supporting decision making. We focus on net benefit (NB) as the key part of decision curve analysis (DCA), which was introduced in 2006 [8]. Editorials supporting DCA have been published in leading medical journals including *JAMA*, *Lancet Oncology*, *Journal of Clinical Oncology*, *BMJ*, *PLoS Medicine*, and *Annals of Internal Medicine* [9–17]. Importantly, evaluating NB is recommended by the TRIPOD guidelines for prediction models [18]. DCA is widely used within urology and many other

clinical fields. A Web of Science search (July 22, 2018) revealed that the 2006 paper was cited 682 times in total. DCA was most often cited in journals from urology and nephrology (173 citations), oncology (140), and general and internal medicine (72). *European Urology* is the journal with most citations (44).

However, based on various personal discussions, we notice that researchers struggle with the interpretation and reporting of NB. We therefore aim to provide an investigators' guide to NB and DCA. A case study on prediction of high-grade prostate cancer is used as an illustrative example.

2. Evidence acquisition

We informally reviewed the urological literature to determine investigators' understanding of DCA. To illustrate, we use data from 3616 patients to develop risk models for high-grade prostate cancer (n = 313, 9%) to decide who should undergo a biopsy. The baseline model includes prostate-specific antigen (PSA) and digital rectal examination; the extended model adds two predictors based on transrectal ultrasound (TRUS).

3. Evidence synthesis

3.1. Case study: prediction of high-grade prostate cancer to decide who to biopsy

Screening with PSA results in overdiagnosis of indolent prostate cancer [19]. Risk calculators have been developed for high-grade prostate cancer [20]. Using these models to decide who to biopsy can reduce unnecessary biopsies, which are aversive procedures with a risk of sepsis and lead to detection of indolent disease. Detecting high-grade prostate cancer is important, because early detection of these potentially lethal cancers can lead to curative treatment [21]. The Rotterdam Prostate Cancer Risk Calculator (RPCRC) predicts the risk of high-grade cancer in an individual patient based on PSA, abnormal digital rectal examination (DRE), abnormal TRUS findings, and TRUS-based prostate volume [22]. The RPCRC was developed from the European Randomized Study of Screening for Prostate Cancer, Rotterdam section. Men between ages 54 and 74 yr and with PSA ≥3.0 ng/ml received lateral sextant biopsy between November 1993 and March 2000 (n = 3616). The outcome was high-grade prostate cancer (n = 313, 9%), defined as Gleason score 3 + 4 or higher on biopsy and/or tumor stage >T2b.

We focused on a baseline model containing two predictors: PSA value and abnormal DRE. Then we fitted an extended model to investigate the additional value of abnormal TRUS and TRUS-based prostate volume (Table 1). We had 313 events for two or four model coefficients (ie, at least 78 events per variable), substantially limiting the risk of overfitting. To check this, we calculated the calibration slope using bootstrapping [1,4]. Where a slope of 1 indicates no overfitting, we found slopes of 0.998 for the baseline model and 0.995 for the extended model, suggesting a marginal 0.2–0.5% overfitting. The c-statistic was 0.814 (95% confidence interval 0.785–0.840) for the baseline model and 0.866 (0.841–0.888) for the extended model. Thus, based on the traditional metrics of discrimination and calibration, most researchers would agree that the extended model is clearly better.

3.2. Risk thresholds

To use a risk model for treatment decisions, we specify a risk threshold T above which we would treat. In our example, treatment refers to biopsy; however, depending on the application, "treatment" can refer to a wide range of interventions, such as additional diagnostic workup, referral to specialized care, a procedure (eg, lymph node resection), delaying surgery (eg, in patients at high risk of complications), medical treatment, or lifestyle changes. In our prostate biopsy example, we could recommend biopsy if the predicted risk of high-grade cancer was 10% or more (T = 10%) and otherwise advise monitoring without biopsy. Correct classifications are labeled true positives (for patients with the event) or true negatives (for patients without the event). Incorrect decisions are labeled false negatives and false positives.

Many investigators select a threshold that maximizes the sum of the true positive and true negative rates [23]. However, this assumes that sensitivity and specificity are equally important. Relevant thresholds incorporate clinical considerations for decision making. In our case, it is more important to find an aggressive cancer than to avoid unnecessary biopsy. According to decision theory, the risk threshold reflects the risk at which we are indifferent about treatment [24]. Assume that we are willing to biopsy no more than 10 men in order to find one high-grade prostate cancer. Then we consider the benefit of detecting one high-grade prostate cancer to be nine times larger than the harm of an unnecessary biopsy: the "harm-to-benefit" ratio is 1:9. This ratio is hard to specify directly. Fortunately, it has a direct relationship with the risk threshold T: the odds of T equal the harm-to-benefit ratio [24]. For example, a risk threshold of 10% implies a harm-to-benefit ratio of 1:9 (odds [10%] = 10/90). A reasonable risk threshold for decision making involves a holistic assessment of all possible outcomes. A biopsy can be painful and inconvenient, and entails a risk of infection; therefore, it is preferable to avoid a biopsy when deemed unnecessary. In case the patient has high-grade prostate cancer, the biopsy can lead to cancer treatment, which may improve prognosis but may cause side effects. Hence, different strategies have their benefits and harms, which may also be of financial or organizational nature. Balancing of all benefits and harms determines which risk thresholds are reasonable.

3.3. NB and DCA

The utility of risk models may be evaluated with cost-effectiveness studies [25], supported by empirical evaluations of the impact of using a model in clinical practice. Such studies are difficult to conduct. Instead, there are simpler measures to evaluate the potential clinical utility of risk models [26]. We focus on NB, which combines the number of true positives and false positives into a single "net" number [8,9]. NB is a concept similar to that of net profit in business: income minus expenditure. In the prostate cancer example, the "income" represents true positives—cases of aggressive prostate cancer found early; the "expenditure" represents false positives—unnecessary biopsies. In most medical scenarios, income and expenditure are on different scales. Therefore, we need an "exchange rate" to reflect the balance between the benefit of a true positive and the harm of a false positive (the harm-to-benefit ratio). Going back to our example, with a risk threshold of 10%, we would weigh each false positive by the odds of 10 (10 ÷ 90 = 0.1111). The baseline model at T = 10% yields 211 detected high-grade prostate cancers and 621 unnecessary biopsies. Then, 211

true positives minus (10 ÷ 90) × 621 false positives gives 142 “net” true positives. Correction for the harm of the unnecessary biopsies adjusts the observed 211 detected high-grade prostate cancers to a net number of 142. The net result is positive because there were only 2.9 false positives per true positive (= 621 ÷ 211) at the 10% risk threshold, whereas this threshold implies that we are willing to accept much more unnecessary biopsies (ie, nine) per detected high-grade prostate cancer. NB is obtained by dividing the net true positives by the sample size, which gives 0.0393 for the baseline model (Table 2). This means that there are 3.9 net detected high-grade prostate cancers per 100 patients. The division by sample size avoids that the magnitude of NB depends on the size of a dataset. Several measures have been proposed that are closely related to NB and that lead to identical conclusions (see the Supplementary material) [16,26–28]. Usually, there is no single risk threshold that is universally acceptable and so it is important to evaluate NB over a range of reasonable thresholds [9,29]. In the case of prostate biopsy, for example, a patient averse to the risk of untreated cancer may prefer a lower risk threshold, whereas a patient less tolerant of invasive procedures such as biopsy may choose a higher threshold. The clinical decision for which the model is used is pivotal to set the relevant threshold range. For example, using a risk model to select patients with suspicious bladder tumors for general urological surgery will require a different threshold compared with using the model to select patients for specialized oncological surgery. A decision curve plots NB for a range of relevant risk thresholds (Fig. 1). In our example, we focused on thresholds between 5% and 20%.

3.4. Are model-based decisions useful? Comparison with default strategies

To interpret NB properly, we introduce two default strategies where patients are managed without the use of a model [8]. We can biopsy either all patients (“treat all”) or no one (“treat none”). NB of treat none is always 0 because this strategy has no true or false positives. Treat all does not imply that T has been set to 0. Rather, the decision to treat everyone is evaluated at all reasonable values of T (see the Supplementary material for formula). For risk thresholds below prevalence, treat all has a higher NB than treat none. For thresholds above prevalence, the opposite holds true, which implies a negative NB for treat all. At the 10% risk threshold, treat all has an NB of –0.0149.

A model is only clinically useful at threshold T if it has a higher NB than treat all and treat none. If a model has a lower NB than any default strategy, we consider the model clinically harmful: one of the default strategies leads to better decisions. Importantly, when models are calibrated, they cannot be harmful [4,5]. Only miscalibrated models can be harmful. For example, if we underestimate the risk of high-grade prostate cancer, we would too often advise against biopsy, missing more cancers than anticipated, leading to poorer NB. When applying DCA, we first evaluate whether the model(s) under study has (have) a higher NB than the default strategies. When comparing two models, we check which model has the highest NB. When one of the models is harmful, further model comparison is redundant. The baseline and extended models of our case study outperform the default strategies across the relevant threshold range, and the extended model outperforms the baseline model.

To interpret DCA results, we illustrate various hypothetical scenarios in Figure 2. We show decision curves for an application where the threshold probability is typically about 20%,

but a reasonable range of thresholds is determined to be 10–30%. We show threshold probabilities outside this range for didactic purposes. In Figure 2A, the model (black dashed line) has a higher NB than both treat all (thin gray line) and treat none (thin black line) only for threshold probabilities above 20%. As the range of reasonable thresholds is 10–30%, that is, some patients would choose treatment if their risk was only 10% or 15%, the model is not of value. Indeed, for patients with these types of thresholds, NB of the model is worse than the strategy of "treating all", that is, opting for treatment irrespective of the risk from the model. The lower NB at these thresholds is because the model is miscalibrated, slightly underestimating the risk. In Figure 2B, we show a well-calibrated model with a relatively high area under the curve. However, the prevalence of disease in the study is very high (~60%). With baseline risk being very high, it is very difficult for a model to push the risk low enough for a patient to refuse treatment. The model has a higher NB for only a small part of the range of reasonable thresholds, and therefore the model is not of value. In Figure 2C, the model is of benefit for almost, but not quite, the whole of the reasonable range 10–30%: the curves diverge only at the threshold probability of about 13%. However, NB of the model is about the same as the NB of treat all below 13%. Therefore, if the investigators believed that it was not common to have such low threshold probabilities, they could probably justify clinical use. In Figure 2D, either the model or the competing binary test (grey dashed line) has a higher NB than treating all or no patients across the entire range of reasonable threshold probabilities. However, the curves cross in the middle of the reasonable range. In general, the conclusion would be that no strategy is optimal across the whole range of reasonable threshold probabilities, and hence further research is required. However, depending on the clinical situation, there might be calls to favor the model or the test. For instance, NB for each is similar in the key range of thresholds, so if one approach is superior in terms of costs and risks of convenience, then that might be the approach chosen. In Figure 2E, the model is well calibrated with a c-statistic of 0.75. The competing model has a c-statistic of 0.80 but is miscalibrated (risks are underestimated). As a result, the model with the lowest c-statistic is superior to the entire reasonable range of threshold probabilities. The miscalibrated model is even harmful for thresholds up to 15%.

3.5. Interpretation of NB

NB gives the proportion of "net" true positives in the dataset: the observed number of true positives is corrected for the observed proportion of false positives weighted by the odds of the risk threshold, and the result is divided by the sample size. This "net" proportion is equivalent to the proportion of true positives in the absence of false positives (ie, perfect specificity). The baseline model has an NB of 0.0393 at the 10% risk threshold, which is equivalent to detecting 3.93 (≈4) high-grade prostate cancers and suggesting zero unnecessary biopsies per 100 patients (ie, four true positives and zero false positives). In fact, this is a direct comparison with treat none, which has zero true positives and zero false positives by default. Even though a model may compare well with treat none (ie, NB is positive), it may still be worse than treat all. This is possible when the risk threshold is below prevalence, because then the NB of treat all is higher than the NB of treat none.

To interpret the NB difference between models, consider that the extended model yielded 236 true positives and 475 false positives at the 10% risk threshold (NB = 0.0507). The

difference in NB for the extended versus baseline model is 0.0507 – 0.0393 = 0.0114. The extended model has 1.14 more net detected high-grade prostate cancers per 100 patients. This is equivalent to having1.14 more detected high-grade prostate cancers per 100 patients for the same number of unnecessary biopsies.

3.6. Test tradeoff

NB does not directly account for the cost and harms associated with measuring the predictors in the model. This is a reasonable assumption where the model includes only routinely available data (such as in our base model of PSA and DRE), but if a predictor requires an invasive or expensive test (such as TRUS), we should account for the harm or cost of measurement. We may specify the harms of a model upfront: we ask clinicians "how many of these tests would you do to find one case (eg, high-grade prostate cancer) if the test were 100% perfect"; the reciprocal of that number is the "test harm," which is subtracted from NB [8]. Test harm may be difficult to specify directly. Alternatively, we can focus on the difference in NB (ΔNB) to derive the "test tradeoff" [30–32].

3.6.1. Evaluation of a single model—When validating a single model, ΔNB refers to the difference between the NB of the model and the NB of the best default strategy. The test tradeoff, 1/ΔNB, is the minimum number of tests per true positive that we have to accept to make the model worthwhile given its cost. For the baseline model at 10%, ΔNB is 0.0393 and the test tradeoff is 25.4 (=1/0.0393). If we are willing to use the baseline model on 25 patients to detect one high-grade prostate cancer, this model is worthwhile.

3.6.2. Model comparison—The test tradeoff for the comparison of two models refers to the minimum number of tests for one additional true positive with the best model to make this model worthwhile given its additional cost. At the 10% risk threshold, ΔNB of the extended versus the baseline model is 0.0114, and the test tradeoff is 87.7 (=1/0.114). If we consider it acceptable to subject 88 patients to TRUS to detect one additional high-grade prostate cancer compared with the model without TRUS, the utility of the extended model is worth the cost of TRUS.

3.6.3. Test tradeoff in terms of true negatives—NB is based on the numbers of true and false positives. From these numbers, it is easy to derive the numbers of true and false negatives. It is therefore possible to obtain an alternative expression of NB, which corrects the number of true negatives for the weighted number of false negatives (see the formula in the Supplementary material) [33]. As a result, we can express the test tradeoff in terms of true negatives as well. This test tradeoff, obtained as odds (T)/ΔNB, gives the number of patients we should be willing to classify with the best model per additional true negative.

For evaluation of a single model (the baseline model for the case study), the test tradeoff in terms of true negatives equals 2.8 (=odds [10%]/0.0393). If we are willing to use the baseline model on three patients to avoid one unnecessary biopsy, this model is worthwhile. When comparing the extended model with the baseline model at 10%, we find a test tradeoff of 9.7 patients per additional true negative (=odds [10%]/0.0114). The extended model is

preferable over the baseline model if we accept doing TRUS on 10 patients to avoid one additional unnecessary biopsy.

3.6.4. Interpretation of the test tradeoffs for the case study—When evaluating the baseline model with the default strategies, the test tradeoff indicates that the baseline model is clearly of value given that the model only requires data that the urologist already has at hand.

TRUS could be invasive and unpleasant; hence, the test tradeoffs can be considered high. Some urologists would not agree to subject 88 patients to TRUS to find one high-grade cancer or perform 10 TRUS to avoid one biopsy (Table 2), despite an increase in the c-statistic of 0.052. Other urologists may accept the test tradeoffs given that TRUS has almost no complications. Nevertheless, we might consider alternative sources, such as magnetic resonance imaging or DRE, to measure volume [34–36].

3.7. Recommendations for practice

3.7.1. Interpreting the results of NB—The first step in DCA is to determine whether any model is superior to all other models, and the default strategies of treating all or no patient, across the full range of reasonable threshold probabilities. If so, we can declare that the use of that model would improve patient outcome irrespective of patient or doctor preference. The second step is to consider whether there are important risks, harms, or costs to using the model. If so, we need to interpret the magnitude of the increase in NB versus best default or the competing model, and evaluate the test tradeoff, or use test harm, to check whether the increase in NB is worth the additional cost and harm of using the model.

3.7.2. Defining the treatment decision clearly—DCA evaluates the utility of a model or test to decide who should receive treatment, which can be any diagnostic or therapeutic intervention depending on the application. It is therefore important to unambiguously define the decision. If a model serves mainly to counsel patients (eg, survival probabilities), the meaning of decision curves becomes debatable since the range of personal decisions is wide. For example, a model predicting probability of death at 1 yr in patients with advanced cancer might be used to inform decisions ranging from sorting out legal affairs to travel plans or retirement.

3.7.3. Defining a reasonable range of risk threshold—For a particular treatment decision, utility should be evaluated for a reasonable range of thresholds only. “Reasonable” means that no one would reasonably use a threshold outside that range to decide upon treatment. We therefore recommend showing and interpreting decision curves only for the adopted reasonable range. The ideal situation is when one model shows the highest utility over the entire range. Elsewhere, we have given further details of how researchers can develop ideas about the suitable range of threshold probabilities [9]. When researchers decide to use DCA for a model used for patient counseling, decision curves might be plotted for wider ranges of risk thresholds, even for the full range between 0 and 1.

3.7.4. Not using DCA to choose a risk threshold—We cannot use DCA to choose an optimal risk threshold. NB depends on the adopted risk threshold, not the other way

around. More generally, the choice of a clinically appropriate threshold should not depend on the results of a study of a prediction model [16].

3.7.5. Reporting the test tradeoffs where appropriate—An increase in NB may not be worth the additional cost of using the best model. Investigators should consider reporting the test tradeoff, in particular when there are significant harms or costs associated with obtaining data for the model. When comparing two models, we can express the test tradeoff in terms of true positives and true negatives. We recommend reporting both.

4. Conclusions

DCA is a statistical method to evaluate whether a model has utility in supporting clinical decisions, and which of two models leads to the best decisions. It is therefore an essential validation tool on top of measures such as discrimination and calibration.

$$NB = \frac{TP - odds(T) \times FP}{N} = P - odds \times , \ . RU = \ .$$

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

***Funding/Support and role of the sponsor*:** B.V.C, L.W., J.Y.V., and E.C. are supported by the Research Foundation–Flanders (FWO) project G0B4716N, and Internal Funds KU Leuven (project C24/15/037). A.J.V. is supported by funds from the Sidney Kimmel Center for Prostate and Urologic Cancers, P50-CA92629 SPORE grant from the National Cancer Institute to Dr. H. Scher, the P30-CA008748 NIH/NCI Cancer Center Support Grant to MSKCC, and R01 CA179115 to Dr. A. Vickers. E.W.S. is supported by Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1606–35555) and by the National Institutes of Health (U01NS086294). All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors, or Methodology Committee. None of the funders had a role in the study design, data collection, data analysis, data interpretation, or the writing of the manuscript.

References

[1]. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York, NY: Springer; 2009.

[2]. Shariat SF, Kattan MW, Vickers AJ, Karakiewicz PI, Scardino PT. Critical review of prostate cancer predictive tools. Future Oncol 2009;5:1555–84. [PubMed: 20001796]

[3]. Vickers AJ, Cronin AM. Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). Urology 2010;76:1298–301. [PubMed: 21030068]

[4]. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167–76. [PubMed: 26772608]

[5]. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Making 2015;35:162–9. [PubMed: 25155798]

[6]. Olchanski N, Cohen JT, Neumann PJ, Wong JB, Kent DM. Understanding the value of individualized information: the impact of poor calibration or discrimination in outcome prediction models. Med Decis Making 2017;37:790–801. [PubMed: 28399375]

[7]. Vickers AJ. Incorporating clinical considerations into statistical analyses of markers: a quiet revolution in how we think about data. Clin Chem 2016;62:671–2. [PubMed: 26988582]

[8]. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565–74. [PubMed: 17099194]

[9]. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6. [PubMed: 26810254]

[10]. Vickers AJ. Prediction models: revolutionary in principle, but do they do more good than harm? J Clin Oncol 2011;29:2951–2. [PubMed: 21690474]

[11]. Localio AR, Goodman S. Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. Ann Intern Med 2012;157:294–5. [PubMed: 22910942]

[12]. Holmberg L, Vickers AJ. Evaluation of prediction models for decision-making: beyond calibration and discrimination. PLoS Med 2013;10:e1001491. [PubMed: 23935462]

[13]. Pepe MS, Janes H, Li CI. Net risk reclassification p values: valid or misleading? J Natl Cancer Inst 2014;106:dju041. [PubMed: 24681599]

[14]. Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in oncology: more than meets the eye. Lancet Oncol 2015;16:e173–80. [PubMed: 25846097]

[15]. Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. JAMA 2015;313:409–10. [PubMed: 25626037]

[16]. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. J Clin Oncol 2016;34:2534–40. [PubMed: 27247223]

[17]. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128–38. [PubMed: 20010215]

[18]. Collins GS, Reitsma JB, Altman DG, Moons KGM; Members of the TRIPOD group. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Eur Urol 2015;67:1142–51. [PubMed: 25572824]

[19]. Schröder FH, Hugosson J, Roobol MJ, 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;384:2027–35. [PubMed: 25108889]

[20]. Van Neste L, Hendriks RJ, Dijkstra S, et al. Detection of high-grade prostate cancer using a urinary molecular biomarker-based risk score. Eur Urol 2016;70:740–8. [PubMed: 27108162]

[21]. Mottet N, Bellmunt J, Bolla M, et al. EAU-ESTRO-SIOG guidelines on prostate cancer. Part 1: screening, diagnosis, and local treatment with curative intent. Eur Urol 2017;71:618–29. [PubMed: 27568654]

[22]. Roobol MJ, Steyerberg EW, Kranse R, et al. A risk-based strategy improves prostate-specific antigen-driven detection of prostate cancer. Eur Urol 2010;57:79–85. [PubMed: 19733959]

[23]. Perkins NJ, Schisterman EF. The inconsistency of “optimal” cut points obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol 2006;163:670–5. [PubMed: 16410346]

[24]. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. New Engl J Med 1975;293:229–34. [PubMed: 1143303]

[25]. Hlatky MA, Greenland P, Arnett DK, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation 2009;119:2408–16. [PubMed: 19364974]

[26]. 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;33:490–501. [PubMed: 23313931]

[27]. Baker SG, Cook NR, Vickers AJ, Kramer BS. Using relative utility curves to evaluate risk prediction. J R Stat Soc A 2009;172:729–48.

[28]. Pencina MJ, D’Agostino RB, Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med 2011;30:11–21. [PubMed: 21204120]

[29]. Hunink MGM, Weinstein MC, Wittenberg E, et al. Decision making in health and medicine: integrating evidence and values. ed. 2 Cambridge, UK: Cambridge University Press; 2014.

[30]. Baker SG, Van Calster B, Steyerberg EW. Evaluating a new marker for risk prediction using the test tradeoff: an update. Int J Biostat 2012;8:1.

[31]. Baker SG, Schuit E, Steyerberg EW, et al. How to interpret a small increase in AUC with an additional risk prediction marker: decision analysis comes through. Stat Med 2014;33:3946–59. [PubMed: 24825728]

[32]. Baker SG, Kramer BS. Evaluating prognostic markers using relative utility curves and test tradeoffs. J Clin Oncol 2015;33:2578–80. [PubMed: 26124476]

[33]. Rousson V, Zumbrunn T. Decision curve analysis revisited: overall net benefit, relationships to ROC curve analysis, and application to case-control studies. BMC Med Inform Decis Making 2011;11:45.

[34]. 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;61:577–83. [PubMed: 22104592]

[35]. Pereira-Azevedo N, Braga I, Verbeek JF, et al. Prospective evaluation on the effect of interobserver variability of digital rectal examination on the performance of the Rotterdam Prostate Cancer Risk Calculator. Int J Urol 2017;24:826–32. [PubMed: 28901582]

[36]. Radtke JP, Wiesenfarth M, Kesch C, et al. Combined clinical parameters and multiparametric magnetic resonance imaging for advanced risk modeling of prostate cancer-patient-tailored risk stratification can reduce unnecessary biopsies. Eur Urol 2017;72:888–96. [PubMed: 28400169]

[37]. Hilden J Prevalence-free utility-respecting summary indices of diagnostic power do not exist. Stat Med 2000;19:431–40. [PubMed: 10694728]

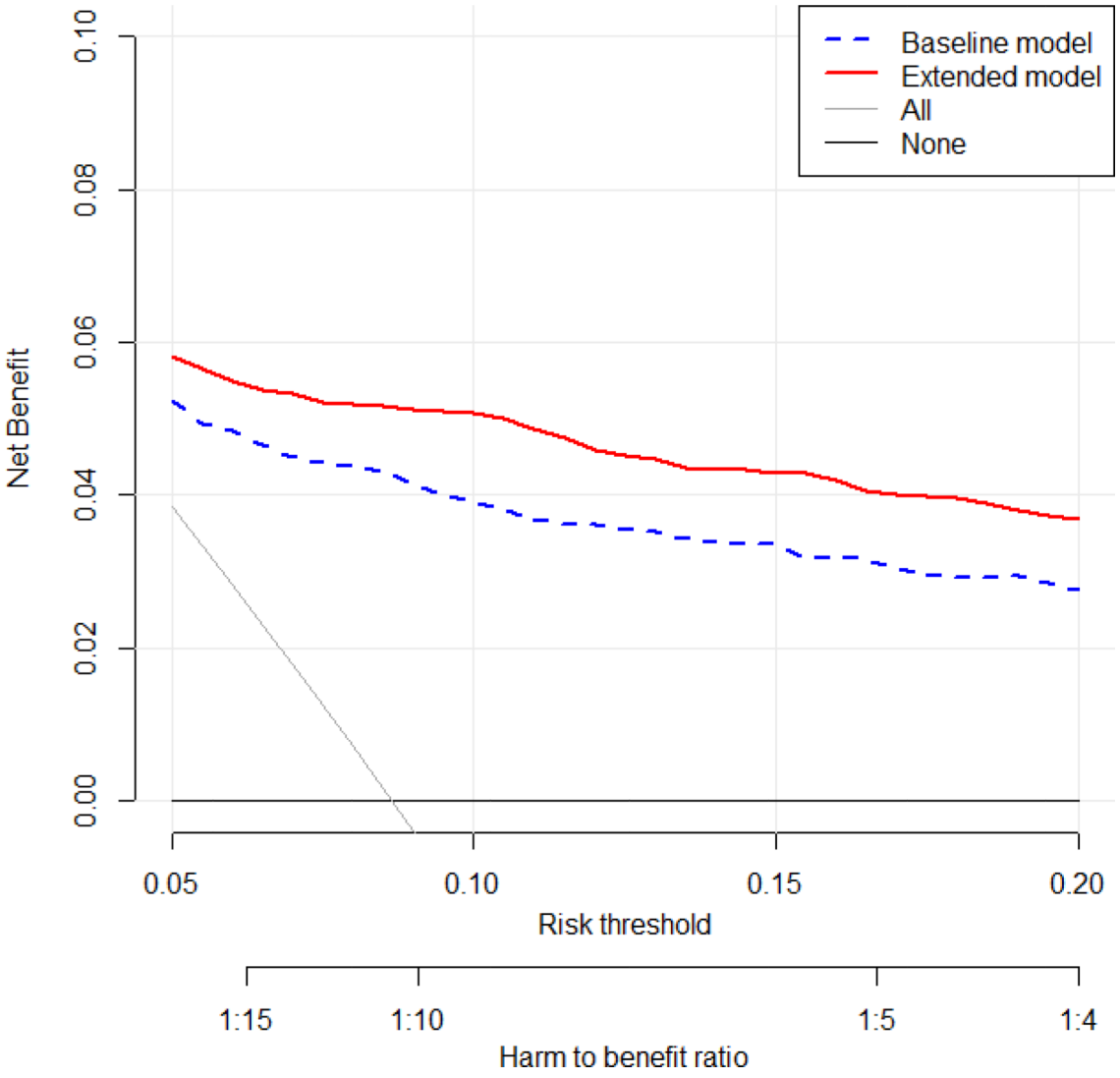

Fig. 1 –.

Decision curves for the default strategies and for the baseline and extended models.

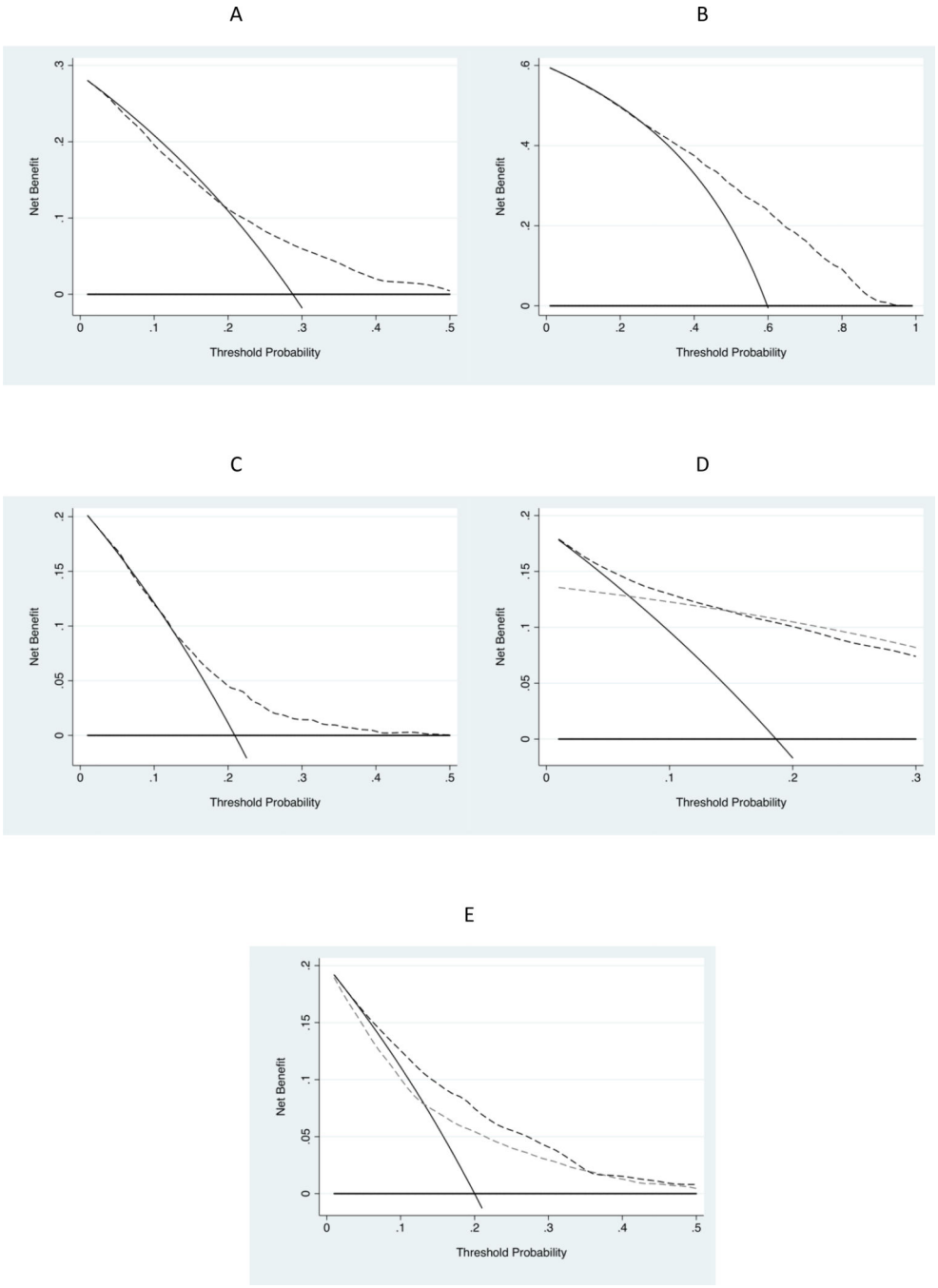

Fig. 2 –.
Hypothetical decision curves illustrating several possible scenarios.

Table 1 –

Baseline and extended models to predict high-grade prostate cancer

Predictor	Median (IQR) or *n* (%)	Baseline model		Extended model	
		***B* (SE)**	**OR (95% CI)**	***B* (SE)**	**OR (95% CI)**
Intercept		−5.68 (0.21)		−0.20 (0.67)	
PSA [a]	4.3 ng/ml (3.1−6.4)	1.03 (0.063)	2.79 per doubling (2.47−3.16)	1.21 (0.072)	3.36 per doubling (2.92−3.87)
Abnormal DRE	1279 (35%)	1.60 (0.14)	4.95 (3.79−6.46)	1.03 (0.15)	2.81 (2.10−3.76)
Abnormal TRUS	1229 (34%)			1.21 (0.15)	3.35 (2.50−4.48)
Tumor volume [a]	41 ml (32−55)			−1.16 (0.13)	0.31 per doubling (0.24−0.41)

B = regression coefficient; CI = confidence interval; DRE = digital rectal examination; IQR = interquartile range; OR = odds ratio; PSA = prostate-specific antigen; SE = standard error; TRUS = transrectal ultrasound.

[a]PSA and tumor volume are modeled with log2 transformation, such that the odds ratios for these variables represent the change in odds per doubling of the PSA/volume.

Table 2 –

Net benefit and test tradeoff results for the baseline and extended models to predict high-grade prostate cancer at a risk threshold of 10%

Statistic	Result
Default strategies	
NB if all men subject to biopsy (NB_{TrA})	−0.0149
NB if no one subject to biopsy (NB_{TrN})	0
Baseline model	
NB if baseline model is used to select patients for biopsy	0.0393
Detected HG-PCa without unnecessary biopsies	3.9 per 100 patients
Test tradeoff, patients biopsied per detected HG-PCa	25.4
Extended versus baseline model	
NB if extended model is used to select patients for biopsy	0.0507
NB difference between extended and baseline models	0.0114
Additional HG-PCa detected (without change in unnecessary biopsies) when using the extended model rather than the baseline model	1.14 per 100 patients
Test tradeoff, patients undergoing TRUS per additionally detected HG-PCa	87.7
Test tradeoff, patients undergoing TRUS per avoided unnecessary biopsy	9.7

HG-PCa = high-grade prostate cancer; NB = net benefit; TrA = treat all; TrN = treat none; TRUS = transrectal ultrasound.