Biomarker Evaluation and Clinical Development

Melissa Assel, Andrew J. Vickers✉

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, United States

Abstract

Most candidate biomarkers are never adopted into clinical practice. The likelihood that a biomarker with good predictive properties will be incorporated into urologic decision-making and will improve patient care can be enhanced by following established principles of biomarker development. Studies should follow the REMARK guidelines, should have clinically relevant outcomes, and should evaluate the biomarker on the same patients to whom the biomarker would be applied in practice. It is also important to recognize that biomarker research is comparative: the question is not whether a biomarker provides information, but whether it provides better information than is already available. Continuous biomarkers should not be categorized above or below a fixed cutpoint: risk prediction allows for individualization of care. The risk predictions must be calibrated, that is, close to a patient's true risk, and decision analysis is required to determine whether using the biomarker in clinical practice would change decisions and improve outcomes. Finally, impact studies are needed to evaluate how use of the biomarker in the real world affects outcomes.

Introduction

Biomarkers are used either to assess the risk of a current diagnostic state, such as having biopsy-detectable cancer, or to predict the risk of a future event, such as prostate cancer death. In the former case, the biomarker gives the clinician information at less cost, risk, and inconvenience than the diagnostic test; in the latter case, it provides an estimate of probability for occurrence of a future outcome in an individual patient. In this paper, we review methodologic considerations for biomarker development using serum biomarkers in prostate cancer as an example. We do not discuss how biomarkers are discovered or how they can best be measured accurately and reproducibly. We start from early phase studies in humans evaluating the association between the biomarker and the outcome, and move to the later phase trials (ie, impact studies) examining the effects of the biomarker when used in the clinic.

To be used most effectively, biomarkers need to be integrated into other information available to the clinician, such as a patient's age or the stage of the tumor. This can be done informally by "clinical judgment," using cutpoints and clinical rules, or by using a prediction model. In the case of PSA for prostate cancer early detection, an early approach was to use the clinical rule of "PSA > 4 or positive digital rectal examination (DRE)." This subsequently evolved to the more informal clinical judgment approach, in which the urologist considers the age of the patient, the recent clinical history (such as symptoms of benign prostate disease) and the DRE, as well as the absolute level of PSA. In the last 10 to 15 years, there has been a move to statistical methods of risk prediction. Using statistical models such as the "PCPT risk calculator [1]" or the "PBCG model [2]," the urologist enters clinical data about the patient age, race, DRE, family history, and history of prior negative biopsy, as well as the level of PSA, and obtains a percentage risk of high-grade cancer. The advantage of prediction models is that they give more accurate predictions than either informal clinical judgment—numerous studies have demonstrated that computer models outperform clinicians [3–5]—or the risk groupings used for clinical prediction rules [6–9]. Moreover, use of prediction models allows greater individualization of care. A man who is older, has comorbidities, or is averse to medical procedures, but has a PSA

Key Words

Biomarkers, prediction modeling, prostate cancer, clinical utility, decision analysis, discrimination, calibration, net benefit

Competing Interests

Dr Vickers reports grants from National Institutes of Health during the conduct of the study; personal fees from Opko outside the submitted work. In addition, Dr Vickers has a patent Arctic Partners issued. Dr Assel reports grants from National Institutes of Health during the conduct of the study.

Article Information

Received on June 30, 2020
Accepted on August 7, 2020

Soc Int Urol J. 2020;1(1):16–22

Abbreviations

AUC	area under the curve
DRE	digital rectal examination
EPCA	early prostate cancer antigen
PCPT	Prostate Cancer Prevention Trial
PCPTRC	Prostate Cancer Risk Calculator
ROC	receiver operating characteristic
EPCA	early prostate cancer antigen

level just above 4 might reasonably ask whether his PSA warrants a biopsy; comparably, a man anxious about prostate cancer who has a PSA just below 4 might want reassurance that his risk is indeed low. It is only by using predicted probabilities that urologists can have a rational conversation about risk that takes into account patient preferences and characteristics.

Statistical methods for building models are described at length in various publications and are not further discussed here [10]. Instead, we focus on approaches to assess the predictiveness of a biomarker in 2 different scenarios: when the biomarker is used independently and when it is incorporated into a prediction model.

Can the Biomarker Predict the Outcome of Interest?

Choose the right outcome

The appropriate clinical endpoint for a biomarker is sometimes more complex that it appears. Well-known studies such as the PRACTICAL collaboration have developed polygenic risk scores for the endpoint of incident prostate cancer [11]. But incident prostate cancer is not synonymous with cancer-related mortality or morbidity. The central problem of prostate cancer early detection is overdiagnosis, reflecting that cancers are diagnosed that would never cause symptoms during the course of the patient's natural life. It is thus not as useful to know a man's risk of a prostate cancer diagnosis as it is to know the risk of prostate cancer metastasis or death: a man at high risk of prostate cancer death might want to consider screening to find a cancer early before it spreads; it is not at all clear what a man should do if he is at higher risk of prostate cancer. Biomarkers or models that predict the risk of any grade cancer on prostate biopsy can be subject to a similar criticism: we want to find cancers that we would consider treating (eg, grade 2 or higher disease); we do not need to know about the risk of all cancers, including grade group 1 disease, the most appropriate management strategy for which is to order a second biopsy.

Naturally, the ideal endpoint for any biomarker to predict is cancer-specific morbidity (ie, metastasis) or mortality. Given that such endpoints may occur 10 or 20 years after diagnosis, this is challenging and has been attempted only for a handful of prostate cancer biomarkers, including the 4Kscore [12,13], the DECIPHER score [14], and, of course, PSA [15,16].

Does the biomarker distinguish between samples of clearly distinguishable patients?

Investigators can test whether biomarker levels differ in clearly distinguishable groups of people. These studies can be performed relatively quickly as samples can be obtained from patients on the basis of an outcome status already achieved as opposed to following a cohort of patients prospectively until the outcome of interest occurs, or waiting to accrue patients undergoing a procedure such as biopsy. For example, in the now retracted EPCA study, levels of EPCA in men with prostate cancer were compared with those in healthy men, healthy women, and patients with other diseases, such as liver cancer or benign lung disease.

Diagnostic accuracy should instead be assessed using a sample representative of the population, as shown in Table 1. In scenarios A and B, we have a biomarker with high sensitivity and specificity (both 90%) for advanced disease, but unable to distinguish localized disease (sensitivity and specificity of 50%). Scenario A represents a sample with an equal number of patients in each disease group, in which the sensitivity and specificity in the entire population for detecting cancer is 70%. However, if the distribution of patients were more reflective of the population, as in scenario B, the sensitivity and specificity drop to 58% and 63%, respectively.

Is the biomarker associated with the outcome in patients the biomarker would be applied to in practice?

Just as drugs are studied in the patients who would receive the drug were it shown to be effective, biomarkers should be studied in the patients to whom they would be applied in practice. The development of free-to-total PSA ratio is a good example of a marker that moved from research on convenience samples to the target population of men considering biopsy. First, Christensson et al. demonstrated an association between the ratio of a PSA isoform, free PSA, and the total amount of PSA in serum (free-to-total PSA ratio) is significantly lower among men with prostate cancer than in men with benign prostate hyperplasia [18]. Catalona et al. then determined that free-to-total PSA can enhance the specificity of prostate cancer screening by confirming that the association of free-to-total PSA

TABLE 1.

Two hypothetical studies that sample from individuals without disease, with benign disease, with localized disease, or with advanced cancer*

	Cancer (Advanced or Localized)	No Cancer (No Disease or Benign Condition)
Scenario A		
Biomarker Positive	(90% x 250) + (50% x 250) = 350 (true positives)	(10% x 250) + (50% x 250) = 150 (false positives)
Biomarker Negative	(10% x 250) + (50% x 250) = 150 (false negatives)	(90% x 250) + (50% x 250) = 350 (true negatives)
Total	500	500
Scenario B		
Biomarker Positive	(90% x 50) + (50% x 200) = 145 (true positives)	(10% x 250) + (50% x 500) = 275 (false positives)
Biomarker Negative	(10% x 50) + (50% x 200) = 105 (false negatives)	(90% x 250) + (50% x 500) = 475 (true negatives)
Total	250	750

and prostate cancer on biopsy remains significant among men with total PSA values of 4.1 to 10 ng/mL indicated for prostate biopsy in clinical practice [19].

How well does the biomarker predict the outcome of interest compared to information available to the clinician?

A useful biomarker should make additional information available to the clinician. In the Catalona et al. example above, measurement of free-to-total PSA ratio added information about prostate cancer risk over and above total PSA and DRE [19]. Assessments of discrimination or clinical utility (explained in detail below) can be used to compare the performance of a new biomarker with the performance of an existing model or existing biomarker.

Alternatively, a biomarker can be combined with other clinical factors by building a prediction model. For example, Klein et al. assess the added value of the Genomic Prostate Score by demonstrating that it is significantly associated with the risk of adverse pathology on multivariable logistic regression analysis when added to model containing standard clinical predictors (age, PSA, clinical stage, and biopsy) or established prediction models, including Cancer of the Prostate Risk Assessment score [20] and the National Comprehensive Cancer Network risk groupings [21].

Assessing Predictiveness

Discrimination

The area under the receiver operating characteristic curve (AUC), also referred to as the concordance statistic (or C index), is commonly used to assess discrimination: the probability that a randomly selected patient with the disease will have higher predicted probability of having the disease according to the test compared to a randomly selected subject without the disease [22-24].

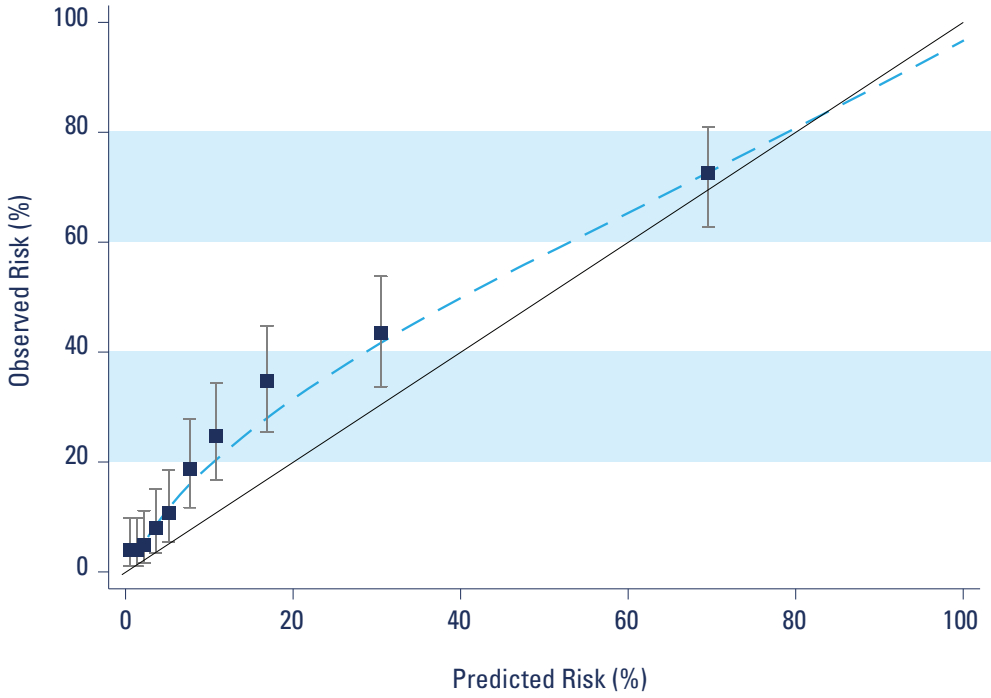
When comparing the discrimination of different models or biomarkers, investigators are encouraged to report the difference in discrimination along with 95% confidence intervals. Approaches to assess whether there is a significant difference in discrimination depend on whether the models being compared are "nested." A nested model is created when, for example, a new biomarker is added to an existing model, for instance, when the 2 models are PSA, DRE, and age versus PSA, DRE, age, and free-to-total-ratio. A comparison of 2 existing models, such as the PCPT and the PBCG model, would be non-nested. In these cases, the Delong test can be used to test for a difference in discrimination [25]. When models are nested, the *P*-value from the Wald test corresponding to the biomarker should be reported, the Delong test being invalid [26,27].

Calibration

To be clinically useful, a prediction model must not only be able to discriminate between patients with and without the disease but also provide an accurate risk prediction. The degree to which predictions are in agreement with the observed outcomes is known as calibration [28]. A calibration plot visualizes the agreement between model predictions on the x-axis and the actual outcome on the y-axis. This is typically done by splitting the data into equal sized groups of increasing predicted probabilities (deciles) and plot the mean of the observed outcome by the decile of prediction [23]. See Figure 1 for an example of a calibration plot. A model with poor calibration in ranges of

FIGURE 1.

A calibration plot for a model predicting the risk of high-grade prostate cancer on prostate biopsy. The dots show the average risk (and 95% CI) of patients divided into 10 groups of increasing risk. The dots and dashed regression line fall above the 45-degree line for good calibration, demonstrating that patients had higher risk than that predicted by the model. This is particularly a problem for risks around 10%, the sort of risk at which a patient might opt for prostate biopsy. Such a calibration plot would raise questions about whether the model should be used to inform prostate biopsy decision-making.



probabilities in which treatment decisions can reasonably differ is likely to be of limited clinical value, even if discrimination is excellent: it is difficult to make a good decision if information about patient risk is wrong.

Some biomarkers, such as the prostate health index [29] and ExoDx Prostate IntelliScore, provide a score, and decisions are made by comparing the score with a proposed cutpoint but these scores do not represent risk of disease. Therefore, it is not possible to assess calibration in the traditional sense, although investigators can report the probability of the outcome above and below the previously proposed cutpoints to assess clinical value.

Clinical utility

A new biomarker is of value only if its use leads to improvement in patient outcome via a change in treatment decision patterns. A full assessment of the prognostic value of a biomarker or model must incorporate clinical consequences of the resulting decisions made. Table 2 shows a hypothetical study of 1000 men with elevated PSA levels. Risk of cancer on biopsy was calculated on the basis of a prediction model including a new biomarker. This shows that 300 men had high-grade cancer and that among the 510 men with a predicted risk of 10% or greater (our threshold to indicate biopsy) cancer was detected in 210 of these men (Table 2). The clinical consequences shown in Table 2 indicate that to determine whether it is better to biopsy all men or to use the statistical model and biopsy those with a 10% risk of high-grade cancer, we need to consider whether it is worth missing 90 cancers to avoid 490 biopsies.

In some cases, the results will be fairly obvious: if, for instance, there were only 10 high-grade cancers missed for a reduction of 490 biopsies, the value of the biomarker would be apparent. When results are not immediately clear, decision analysis can be of value. One of the simplest approaches, and the most widely used according to the urologic literature, is "net benefit," which incorporates the consequences of clinical decisions of a prediction model or biomarker in the analysis [30]. Net benefit incorporates both discrimination (AUC) and calibration, making it an ideal statistic for comparing prognostic value [31]. A key aspect of net benefit is that the level of risk at which a patient opts to undergo a biopsy is informative of how a

TABLE 2.

Hypothetical results of a biomarker study for prostate biopsy illustrating clinical consequences and decision analysis.

Strategy	Biopsied	Biopsy avoided	High-grade cancers caught	High-grade cancers missed	Unnecessary biopsies	Net benefit
Biopsy all men with elevated PSA	1000	0	300	0	700	$300 - \left(700 \times \frac{0.10}{0.90}\right) = 222$
Biopsy all men with elevated PSA	510	490	210	90	300	$210 - \left(300 \times \frac{0.10}{0.90}\right) = 177$

patient weighs the relative harms of a false-positive (an unnecessary biopsy with risks of side-effects including infectious complications and hospitalization) versus a false-negative (missing or delaying the detection of a high-grade cancer) result. This level of risk is termed threshold probability [30]. The threshold probability chosen in Table 2 was 10%, corresponding to odds of 10:90, implying that missing a cancer is 9 times worse than performing an unnecessary biopsy [32]. A threshold of 10% corresponds to a "number-needed-to-test" of 1/10% = 10, meaning that 10 men need to be biopsied to find 1 cancer [32,33]. Applying this 9:1 ratio to the study results gives the findings in Table 2, where it can be seen that the biomarker is actually harmful. Even though the marker has reasonable sensitivity and specificity (~70% and ~60%), too many high-grade cancers are missed for the decrease in unnecessary biopsy achieved [34]. One obvious issue is that the threshold can vary between patients or doctors: a patient worried about cancer might opt of a threshold of 6%, whereas one nervous about medical procedures might demand a 15% risk before considering biopsy. In decision curve analysis, the threshold probability is varied over a reasonable range and net benefit plotted against threshold probability [30]. By visualizing the decision curve, one can readily ascertain whether one strategy or model is optimal for across the full range of threshold probabilities of interest. For more on decision curves, which are very widely used in urology research, a selection of further reading is available at www.decisioncurveanalysis.org.

Impact Studies

Decision analytic techniques provide hypothetical assessments of clinical consequences. Impact studies assess the real-world consequences of a new biomarker- or model-based strategy. For example, an impact study might assess whether the results of the biomarker translated to changes in decisions. For instance, a typical study of the 4Kscore would conclude that, hypothetically, were doctors to use the 4Kscore to make biopsy decisions based on a cut-off of 10%, then the biopsy rates would fall by about 50%. In a study designed to determine what happens in actual practice, Konety et al. reported a 65% reduction in prostate biopsies in men receiving the 4Kscore [35]. However, not all impact studies are consistent with clinical biomarker studies: White et al. found that use of the PHI in practice led to a very large decrease in the capture of high-grade cancers, with an approximate 30% risk of high-grade cancer amongst men who avoided biopsy [36]. Impact studies are also undertaken because some endpoints are not entirely predictable from clinical research. Early research on PSA did find that it detected prostate cancer at an early stage, but it was unclear if prostate cancer screening regimens based on PSA led to reductions in mortality. The European Randomized Study of Screening for Prostate Cancer followed men for 16 years and demonstrated a reduction in mortality with PSA screening, and can therefore be considered an impact study [37,38].

Study Design Issues

The REMARK guidelines discuss study design considerations at length [39]. For instance, one key point is that assessors of the outcome should be blinded from the biomarker status. Another key concept is that of internal versus external validation. Interval validation occurs when a multivariable regression model is developed or a new cutpoint for a biomarker is selected and evaluated for performance on the same dataset. When a prediction model or biomarker cutpoint is developed and assessed on the same dataset estimates of performance are over-optimistic, a phenomenon known as overfitting [40,41]. Harrell et al. describe methods for obtaining optimism-corrected internal assessments of performance including data splitting, cross validation, and bootstrapping [42].

External validation not only solves the problem of overoptimism but evaluates genuine differences between

cohorts. A model predicting recurrence after radical prostatectomy, for instance, may be affected by surgeon skill—less skilled surgeons having higher recurrence rates—or by differences in pathologic grading. An excellent practical example of external validation was a study showing that the risk of prostate cancer among Chinese men with a given PSA had been shown to be lower than for European men, the most likely explanation being that Chinese men have higher rates of benign disease. This true difference between cohorts will mean that prediction models using PSA will likely have poor properties when applied in China [43].

Recommendations

In this paper, we have outlined the evaluation of prostate cancer biomarkers. Our key "take-aways" can be summarized as follows:

1. Biomarkers should predict risk rather than be categorized as being above or below a fixed cutpoint: risk prediction allows individualization of care.
2. Choose a clinically relevant outcome: many endpoints commonly used in biomarker studies, such as incident prostate cancer or advanced surgical pathology, are problematic.
3. Evaluate the biomarker on the patients to whom the biomarker would be applied in practice.
4. Follow the REMARK guidelines for the conduct and reporting of biomarker studies.
5. Biomarker research is comparative: the question is not whether a biomarker provides us with information, but whether it provides us better information than we already have, from clinical features or a currently used biomarker.
6. Report discrimination, calibration, and net benefit: a biomarker must be able to discriminate better than existing predictors, but risk predictions must be close to a patient's true risk; decision analysis is required to determine whether using the biomarker in clinical practice would change decisions and whether doing so would improve outcomes.
7. Conduct impact studies: evaluate how use of the biomarker in the real world affects outcomes.

Conclusions

It has often been noted that biomarker research has a poor track record of getting biomarkers into clinical practice. Following established principles of biomarker development increases the chances that a biomarker with good predictive properties will be incorporated into urologic decision-making and ultimately improve patient care.

References

1. Thompson IM, Ankerst DP, Chi C, et al. Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial. *J Ntl Cancer Inst*. 2006;98(8):529-34.
2. Ankerst DP, Straubinger J, Selig K, et al. A Contemporary Prostate Biopsy Risk Calculator Based on Multiple Heterogeneous Cohorts. *Eur Urol*. 2018;74(2):197-203.
3. Kattan MW, Yu C, Stephenson AJ, Sartor O, Tombal B. Clinicians versus nomogram: predicting future technetium-99m bone scan positivity in patients with rising prostate-specific antigen after radical prostatectomy for prostate cancer. *Urology*. 2013;81(5):956-61.
4. Jelovsek JE, Chagin K, Brubaker L, et al. A model for predicting the risk of de novo stress urinary incontinence in women undergoing pelvic organ prolapse surgery. *Obstet Gynecol*. 2014;123 (2 Pt 1):279-87.
5. Ross PL, Gerigk C, Gonen M, et al. Comparisons of nomograms and urologists' predictions in prostate cancer. *Semin Urol Oncol*. 2002;20(2):82-8.
6. Peeters KC, Kattan MW, Hartgrink HH, et al. Validation of a nomogram for predicting disease-specific survival after an R0 resection for gastric carcinoma. Cancer. 2005;103(4):702-7.
7. Novotny AR, Schuhmacher C, Busch R, Kattan MW, Brennan MF, Siewert JR. Predicting individual survival after gastric cancer resection: validation of a U.S.-derived nomogram at a single high-volume center in Europe. *Ann Surg*. 2006;243(1):74-81.
8. Weiser MR, Landmann RG, Kattan MW, et al. Individualized prediction of colon cancer recurrence using a nomogram. *J Clin Oncol*. 2008;26(3):380-5.
9. Weiser MR, Gönen M, Chou JF, Kattan MW, Schrag D. Predicting survival after curative colectomy for cancer: individualizing colon cancer staging. *J Clin Oncol*. 2011;29(36):4796-802.
10. Steyerberg EW. Clinical prediction models: a practical approach to development, validation and updating. *New York: Springer;* 2019.
11. Schumacher FR, Al Olama AA, Berndt SI, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet*. 2018;50(7):928-36.
12. Vertosick EA, Häggström C, Sjoberg DD, et al. Prespecified 4 Kallikrein Marker Model (4Kscore) at Age 50 or 60 for Early Detection of Lethal Prostate Cancer in a Large Population Based Cohort of Asymptomatic Men Followed for 20 Years. *J Urol*. 2020:101097ju0000000000001007.

13. Sjoberg DD, Vickers AJ, Assel M, et al. Twenty-year risk of prostate cancer death by midlife prostate-specific antigen and a panel of four Kallikrein markers in a large population-based cohort of healthy men. *Eur Urol.* 2018;73(6):941-8.

14. Marascio J, Spratt DE, Zhang J, et al. Prospective study to define the clinical utility and benefit of Decipher testing in men following prostatectomy. *Prostate Cancer Prostatic Dis.* 2020;23(2):295-302.

15. Vickers AJ, Ulmert D, Sjoberg DD, et al. Strategy for detection of prostate cancer based on relation between prostate specific antigen at age 40-55 and long term risk of metastasis: case-control study. *BMJ.* 2013;346:f2023.

16. Vickers AJ, Cronin AM, Björk T, et al. Prostate specific antigen concentration at age 60 and death or metastasis from prostate cancer: case-control study. *BMJ.* 2010;341:c4521.

17. Leman ES, Cannon GW, Trock BJ, et al. EPCA-2: a highly specific serum marker for prostate cancer. *Urology.* 2007;69(4):714-20.

18. Christensson A, Björk T, Nilsson O, et al. Serum prostate specific antigen complexed to alpha 1-antichymotrypsin as an indicator of prostate cancer. *J Urol.* 1993;150(1):100-5.

19. Catalona WJ, Smith DS, Wolfert RL, et al. Evaluation of Percentage of Free Serum Prostate-Specific Antigen to Improve Specificity of Prostate Cancer Screening. *JAMA.* 1995;274(15):1214-20.

20. Cooperberg MR, Broering JM, Carroll PR. Risk Assessment for Prostate Cancer Metastasis and Mortality at the Time of Diagnosis. *J Ntl Cancer Inst.* 2009;101(12):878-87.

21. Network NCC.

22. Bamber P. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psychol.* 1975;12(4):387-415.

23. 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.

24. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143(1):29-36.

25. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44(3):837-45.

26. Demler OV, Pencina MJ, D'Agostino RB, Sr. Misuse of DeLong test to compare AUCs for nested models. *Stat Med.* 2012;31(23):2577-87.

27. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. *BMC Med Res Methodol.* 2011;11(1):13.

28. Hilden J, Habbema JD, Bjerregaard B. The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. *Methods Inf Med.* 1978;17(4):227-37.

29. 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.

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

31. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making.* 2015;35(2):162-9.

32. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res.* 2019;3(1):18.

33. 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.

34. Vickers AJ, Cronin AM, Gönen M. A simple decision analytic solution to the comparison of two binary diagnostic tests. *Stat Med.* 2013;32(11):1865-76.

35. Konety B, Zappala SM, Parekh DJ, et al. The 4Kscore® Test Reduces Prostate Biopsy Rates in Community and Academic Urology Practices. *Rev Urol.* 2015;17(4):231-40.

36. White J, Tutrone RF, Reynolds MA. Second Reply to Letter to the Editor re: "Clinical utility of the Prostate Health Index (phi) for biopsy decision management in a large group urology practice setting". *Prostate Cancer Prostatic Dis.* 2019;22(4):639-40.

37. Schröder FH, Hugosson J, Carlsson S, et al. Screening for prostate cancer decreases the risk of developing metastatic disease: findings from the European Randomized Study of Screening for Prostate Cancer (ERSPC). *Eur Urol.* 2012;62(5):745-52.

38. Hugosson J, Roobol MJ, Månsson M, et al. A 16-yr Follow-up of the European Randomized study of Screening for Prostate Cancer. *Eur Urol.* 2019;76(1):43-51.

39. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumour MARKer prognostic studies (REMARK). *Br J Cancer.* 2005;93(4):387-91.

40. Smith GC, Seaman SR, Wood AM, Royston P, White IR. Correcting for optimistic prediction in small data sets. *Am J Epidemiol.* 2014;180(3):318-24.

41. Steyerberg E. Overfitting and optimism in prediction models. New York: *Springer Verlag;* 2009:83-100.

42. Harrell FE, Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stats Med.* 1996;15(4):361-87.

43. Chen R, Sjoberg DD, Huang Y, et al. Prostate Specific Antigen and Prostate Cancer in Chinese Men Undergoing Initial Prostate Biopsies Compared with Western Cohorts. *J Urol.* 2017;197(1):90-6