Updating risk prediction tools: A case study in prostate cancer

Donna P. Ankerst*,[1,2,3], **Tim Koniarski**[4], **Yuanyuan Liang**[2,3], **Robin J. Leach**[2], **Ziding Feng**[5], **Martin G. Sanda**[6], **Alan W. Partin**[7], **Daniel W. Chan**[7], **Jacob Kagan**[8], **Lori Sokoll**[7], **John T. Wei**[9], and **Ian M. Thompson**[2]

[1] Department of Mathematics, Technische Universitaet Muenchen, Unit M4, Boltzmannstr 3, 85748 Garching b. Munich, Germany
[2] Department of Urology, University of Texas Health Science Center at San Antonio (UTHSCSA), 7703 Floyd Curl Dr., San Antonio, TX, USA 78229, USA
[3] Department of Epidemiology/Biostatistics, University of Texas Health Science Center at San Antonio (UTHSCSA), 7703 Floyd Curl Dr., San Antonio, TX, USA 78229, USA
[4] International Real Estate Business School, University of Regensburg, 93040 Regensburg, Germany
[5] Program in Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, PO Box 19024, Seattle, WA 98109, USA
[6] Department of Urology, Harvard Medical School, and Prostate Center, Urology, Beth Israel Deaconess Medical Center, 330 Brookline MA 02215, USA
[7] Departments of Pathology and Urology, Johns Hopkins Medical Institution, Baltimore, MD, USA
[8] Cancer Biomarkers Research Group, Division of Cancer Prevention, National Cancer Institute, Rockville, MD, USA
[9] Department of Urology, University of Michigan, Ann Arbor, MI, USA

Received 10 March 2011, revised 9 June 2011, accepted 23 August 2011

Online risk prediction tools for common cancers are now easily accessible and widely used by patients and doctors for informed decision-making concerning screening and diagnosis. A practical problem is as cancer research moves forward and new biomarkers and risk factors are discovered, there is a need to update the risk algorithms to include them. Typically, the new markers and risk factors cannot be retrospectively measured on the same study participants used to develop the original prediction tool, necessitating the merging of a separate study of different participants, which may be much smaller in sample size and of a different design. Validation of the updated tool on a third independent data set is warranted before the updated tool can go online. This article reports on the application of Bayes rule for updating risk prediction tools to include a set of biomarkers measured in an external study to the original study used to develop the risk prediction tool. The procedure is illustrated in the context of updating the online Prostate Cancer Prevention Trial Risk Calculator to incorporate the new markers %freePSA and [-2]proPSA measured on an external case–control study performed in Texas, U.S.. Recent state-of-the art methods in validation of risk prediction tools and evaluation of the improvement of updated to original tools are implemented using an external validation set provided by the U.S. Early Detection Research Network.

Keywords: Calibration; Discrimination; Net benefit; Prostate cancer prevention trial; Risk prediction; Validation.

Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1002/bimj.201100062.

*Corresponding author: e-mail: ankerst@ma.tum.de, Phone: +49-89-289-17443, Fax: +49-89-289-17435

1 Introduction

Risk prediction tools for diagnosis, prognosis and treatment of disease are now widely available on the Internet. Two of the more prominently used online risk tools are the Framingham study 10-year risk calculator for cardiovascular disease (Grundy et al., 2004) and the Gail model Breast Cancer Risk Assessment Tool (BRCAT, Gail et al., 1989, 2007). Emergence of such tools on the Internet has expedited translational medicine, more quickly bringing scientific discoveries from the laboratories to the clinic, as well as increased the practice of informed joint decision-making between doctors and their patients concerning individual health management.

In the realm of early cancer detection, high-throughput technologies are now the routine and have brought about mass discoveries of potential cancer markers. Networks of laboratory research centers, such as the Early Detection Research Network (EDRN), have mobilized to expedite laboratory discoveries through validation phases (Pepe et al., 2001). As new cancer biomarkers are discovered and validated, the question arises as to how they can be incorporated into the existing online risk prediction tools, tools which have been created on strong foundations based on extensive analyses of prior large cohorts. Typically, the newly discovered markers cannot be retrospectively measured on sera or other samples stored from participants of the original study or no such biological samples were stored in the first place. Hence, fitting an expanded prediction model that includes the new markers on the same set of data used to develop the original prediction tool is not an option. An additional challenge is that due to cost considerations relatively new biomarkers are typically only measured on smaller retrospective case–control studies. Although the same markers and risk factors from the original risk prediction tool may be measured alongside the new markers in the smaller study so that an expanded model could in principle be constructed on the smaller study, it would seem imprudent to discard the large foundation on which the original risk prediction tool was built.

Fortunately, the Bayesian paradigm is exactly suited for updating prior knowledge with newly available data through the transformation of prior odds to posterior odds via the likelihood ratio. Likelihood ratios have been intricately investigated as a means in and of themselves for evaluating the diagnostic performance of a single marker in the case of a dichotomous marker by Janssens et al. (2005) and a continuous marker by Gu and Pepe (2009, 2011). Steyerberg (2010) reviewed these cases as part of a general paradigm. A fully Bayesian approach for updating prior risks through likelihood ratios to obtain posterior risks was implemented by Ankerst et al. (2008) to incorporate the urine marker PCA3 into the Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) and by Skates et al. (2001) to estimate risk of ovarian cancer based on longitudinal CA125 measurements. The case study to be evaluated in this report adds to the prior literature by considering the incorporation of two or more possibly dependent biomarkers into a risk tool. It assumes that the markers can be transformed to follow multivariate Normal distributions and specifies distinct variance-covariance matrices for the joint marker distributions in the cancer and non-cancer participants. Viewed as a decision rule for classifying subjects as diseased versus non-diseased, likelihood ratios modeled in this fashion correspond to quadratic discriminant analysis as opposed to linear discriminant analysis, which specifies that the variance–covariance matrices of the two populations are the same (Izenman, 2008, pp. 257–258).

The motivation for the case study was a need to update an existing online tool, the PCPTRC, for two markers that have recently emerged in early prostate cancer detection research. The PCPTRC had been published online in 2006 following completion of a large prevention trial, and provides a simple-to-use accessible device for urologists and patients to calculate their risk of prostate cancer based on the established risk factors prostate-specific antigen (PSA), digital rectal exam (DRE), first-degree family history of prostate cancer and history of a prior negative prostate biopsy (Thompson et al., 2006). Since its publication, the PCPTRC has been widely implemented and validated, and in 2010 the American Cancer Society recommended it in its guidelines for prostate cancer screening (Wolf et al., 2010). Using data from a recent case–control study, the San Antonio

Biomarkers Of Risk of prostate cancer (SABOR) study, reporting on two promising new prostate cancer biomarkers, %freePSA and [-2]proPSA, (Liang et al., 2011), this case study first shows how the PCPTRC is updated to incorporate new markers via likelihood ratios modeled via multivariate normal distributions. The updated PCPTRC is then validated on a separate EDRN validation set arising from patients visiting university hospitals in Michigan, Massachusetts and Maryland, U.S. using state-of-the-art evaluation criteria for risk prediction tools including measures of discrimination, calibration, clinical net benefit and the integrated discrimination index (IDI).

2 Bayesian updating of risk models to incorporate new risk factors

The Bayesian paradigm begins with an existing risk model for the presence or absence of cancer, called the prior model, which is based on some established risk factors. The prior risk of cancer from the prior model is converted into the prior odds of cancer. A likelihood ratio for new markers conditional on the established risk factors is calculated based on some external study to that used in constructing the prior model, which is then multiplied by the prior odds to obtain the posterior odds of cancer. The general concept applies to the prediction of any binary endpoint with corresponding appropriate model, and to any types of new markers with the corresponding appropriate joint models. However, to keep to the issues at hand, the specific context and models of the case study will be used for the definition of the method.

A risk model for cancer constructed by logistic regression yields the estimate of the prior odds of cancer

$$\frac{\mathrm{P}(\mathrm{Cancer}|X)}{\mathrm{P}(\mathrm{No\ Cancer}|X)} = \exp(\beta' X)$$

dependent upon a vector of log odds ratios β for a group of established risk factors X. In this case study as in much of the applications surrounding online prediction tools, the prior risk model has been developed on a large prospective population so that β accurately estimates the population prevalence and relationship of risk factors to cancer with small variance.

Next, an external case–control or prospective study provides information on the association between cancer and the same risk factors X, or subset thereof, and additionally, new markers Y. It is assumed that Y is a vector of continuous markers that can be transformed to approximately follow a multivariate normal distribution among the cancer cases and controls so that multivariate regression of Y on X can be performed to estimate the numerator and denominator of the likelihood ratio (LR):

$$\mathrm{LR} = \frac{\mathrm{P}(Y|\mathrm{Cancer}, X)}{\mathrm{P}(Y|\mathrm{No\ Cancer}, X)} = \frac{|\Sigma_{\mathrm{cancer}}|^{-1/2}\exp\{-\frac{1}{2}(Y-\mu_{\mathrm{cancer}})'\Sigma_{\mathrm{cancer}}^{-1}(Y-\mu_{\mathrm{cancer}})\}}{|\Sigma_{\mathrm{no\ cancer}}|^{-1/2}\exp\{-\frac{1}{2}(Y-\mu_{\mathrm{no\ cancer}})'\Sigma_{\mathrm{no\ cancer}}^{-1}(Y-\mu_{\mathrm{no\ cancer}})\}},$$

where μ_{cancer} and $\mu_{\mathrm{no\ cancer}}$ are the least-square estimates of the linear regression means, $\mathrm{E}(Y|\mathrm{Cancer}, X) = X\gamma_{\mathrm{cancer}}$ and $\mathrm{E}(Y|\mathrm{No\ Cancer}, X) = X\gamma_{\mathrm{no\ cancer}}$, and Σ_{cancer} and $\Sigma_{\mathrm{no\ cancer}}$ are unbiased estimators of the variance–covariance matrices from the multivariate regression applied to cases and controls, respectively. Multivariate regression can be performed in `R` using the `lm` command and model selection via the Wilks' lambda test using the `anova.mlm` command with `test = ''Wilks''`.

Bayes rule then applies for updating the prior odds to the posterior odds through the likelihood ratio:

$$\text{Posterior Odds} = \text{Likelihood Ratio} \times \text{Prior Odds},$$

$$\frac{\mathrm{P}(\mathrm{Cancer}|X, Y)}{\mathrm{P}(\mathrm{No\ Cancer}|X, Y)} = \frac{\mathrm{P}(Y|\mathrm{Cancer}, X)}{\mathrm{P}(Y|\mathrm{No\ Cancer}, X)} \times \frac{\mathrm{P}(\mathrm{Cancer}|X)}{\mathrm{P}(\mathrm{No\ Cancer}|X)}.$$

The posterior odds are converted into posterior risks with their confidence intervals calculated from the variance covariance matrices of the component models: Var(all parameters) = diag(Var(β), Var(γ_{cancer}), Var(Σ_{cancer}), Var($\gamma_{\text{no cancer}}$), Var($\Sigma_{\text{no cancer}}$)), using the delta method (implementable via the `deltamethod` function in the `R` package `msm`) or the parametric bootstrap.

3 Validation

As noted in extensive philosophical discussions by Anscombe (1967) and Chatfield (1995) and realized extensively in medical practice, the only real validation of a prediction model is confirmation by a completely independent set of observations collected on a different set of patients from different centers by different investigators. Recently, Steyerberg et al. (2010) provided a clarifying review of evaluation methods for risk prediction models, separating evaluation metrics according to whether they measure discrimination, calibration or both. The purpose of this section is to review the latest state-of-the-art in validation principles for risk prediction tools and for comparing updating risk tools to existing tools.

Typically the way risk prediction tools, or diagnostic markers in general, are applied in practice is to choose an arbitrary cutpoint c and take further action if the prediction exceeds c, referred to as a positive test, and no action if the prediction falls below c, referred to as a negative test. The discrimination accuracy of the test is reported separately for cancer cases and controls, as the sensitivity (proportion of cancer cases testing positive) and specificity (number of controls testing negative), for various choices of the cutpoint c. The receiver operating characteristic curve plots the sensitivity versus the false-positive rate (1-specificity) for all cutpoints c. The area underneath the receiver operating characteristic curve (AUC) ranges from 0.50 for a test with no discriminative power to 1.0 for a test with 100% sensitivity at all possible cutpoints c. The AUC holds an alternative intuitively appealing definition as the probability that for a randomly drawn cancer case and randomly drawn control, the case has a higher risk than the control. Defined as such, it is a rank-based metric equivalent to the non-parametric Wilcoxon test statistic for comparing distributions in two populations (`wilcox.test` in `R`), and can be implemented to test the null hypothesis that AUC = 0.50 versus the alternative that AUC > 0.50. The U-statistic approach of DeLong et al. (1988) can be used for a formal statistical test of the null hypothesis that two risk tools or markers have the same AUC on a validation set versus the two-sided alternative that they differ (`roc.test` function with `method ''Delong''` and `paired = T` option, in `R` package `pROC`). The AUC is practically the ubiquitous endpoint in diagnostic medicine and has been nearly the sole performance criterion for evaluating the PCPTRC (Parekh et al., 2006; Eyre et al., 2009; Hernandez et al., 2009). However, it only measures one dimension of performance, discrimination, and even there, recent statistical reports have criticized its use for placing too much weight on the clinically irrelevant portion of the receiver operating characteristic curve (ROC) (Pencina et al., 2008; Greenland, 2008).

Calibration assesses how closely predicted risks match actual risks in the population and can also be assessed among subgroups to better identify where the prediction model is failing. A formal test of calibration can be implemented by splitting a validation set into k groups, typically $k = 10$ groups defined by deciles of the distribution of evaluated risks on the validation set, and using an approximation to Pearson's chi-square goodness-of-fit test recommended by Lemeshow and Hosmer (1982):

$$\mathrm{X}^2 = \sum_{i=1}^{k} \frac{(O_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)},$$

where O_i is the observed number of cancer cases, n_i the number of individuals, and π_i the average risk for the ith group, for $i = 1,\ldots k$.

Discrimination and calibration metrics objectively summarize accuracy but do not provide information as to which thresholds of a prediction model might be useful for basing clinical decisions. Towards this end, Vickers and Elkin (2006) proposed a measure of net benefit justified through a layman's decision analysis framework that does not rely on user-specified costs associated with various outcomes as full-blown decision analyses typically do. The approach relies on assigning weights to the relative harms of false-positive and negative decisions and then evaluating the net benefit as the average benefit of the decisions. Some decision theoretic arguments show that if a threshold c of a risk prediction is chosen for deciding to take action, such as to get a prostate biopsy, and the value of a true positive decision is set to 1 for identifiability, then the value of a false-positive decision becomes $-c/(1-c)$ (Vickers and Elkin, 2006). As with the other accuracy measures, net benefit is evaluated on an external cohort to the one on which the risk model was developed. The net benefit is defined as the average benefit value over the true and false-positive counts:

$$\text{NetBenefit}(c) = \frac{\text{TruePositiveCount}(c)}{\text{SampleSize}} - \frac{\text{FalsePositiveCount}(c)}{\text{SampleSize}}\left(\frac{c}{1-c}\right),$$

where for emphasis dependency on the user-selected threshold c is included in the definition. The expression for the net benefit can be rewritten to show that it is also a function of the discrimination measures sensitivity, TPR(c), and 1-specificity, FPR(c), evaluated on the external validation set and weighted by the proportions of cancer cases (%Cancer) and non-cancer cases (%Non-Cancer) and their benefit values in the validation set:

$$\text{NetBenefit}(c) = \text{TPR}(c) \times \%\text{Cancer} - \text{FPR}(c) \times \%\text{Non-Cancer} \times \left(\frac{c}{1-c}\right).$$

The discrimination metrics TPR and FPR already tend to vary by validation set. As the above expression shows, net benefit further relies on the cancer prevalence in the validation set. In other words, for two validation sets with the same operating characteristics of a prediction model, the one with higher cancer prevalence will demonstrate higher net benefit.

Vickers and Elkin (2006) suggested evaluating the net benefit over all possible thresholds c of the prediction model ranging from 0 to 1, but in the specific application as the case study here, of determining whether or not to proceed to prostate biopsy for determination of prostate cancer, Steyerberg and Vickers (2008) discussed that most men would reasonably be uncertain as to the course of action with risks in the 10–40% range. Specific values of the net benefit can be difficult to interpret in isolation so Vickers and Elkin (2006) also recommended overlaid decision curves for the strategies of referring no patients to biopsy or all patients to biopsy regardless of the threshold c selected. For these curves the last expression, $c/(1-c)$, remains the same but the TPR and FPR are calculated based on the test rule that assigns no patients to test positive (in other words, $c>1$) and all patients to test positive (in other words, $c<0$). For referring no patients to biopsy, the TPR and FPR are identically 0 so the net benefit curve for this rule is the horizontal line at 0 across all thresholds c. For the decision rule referring all patients to prostate biopsy the TPR and FPR are 1 and the net benefit curve becomes %Cancer$-$%Non-Cancer $\times$ $c/(1-c)$.

For comparing risk predictions from a new model to risk predictions from an old model, Pencina et al. (2008) proposed the IDI that is simply the difference in discrimination slopes between the new and old predictions as proposed by Yates (1982):

$$\text{IDI} = \left(\frac{1}{n_{\text{events}}}\sum_{i=1}^{n_{\text{events}}} p_{\text{new},i} - \frac{1}{n_{\text{non-events}}}\sum_{i=1}^{n_{\text{non-events}}} p_{\text{new},i}\right) - \left(\frac{1}{n_{\text{events}}}\sum_{i=1}^{n_{\text{events}}} p_{\text{old},i} - \frac{1}{n_{\text{non-events}}}\sum_{i=1}^{n_{\text{non-events}}} p_{\text{old},i}\right),$$

where n_{events} are the number of events, in this case prostate cancer cases, and $n_{\text{non-events}}$ are the number of non-events, in this case non-cancer cases, and the summations sum over the predicted probabilities from the new and old models as subscripted among the cancer cases and non-cases as subscripted on n's. The logic of the IDI is clear, a good prediction model should provide higher estimated risks among the cancer cases in the validation set compared with the controls, how good is determined by the discrimination slopes of the models. A positive IDI would indicate a new model has better discrimination slope than the old.

As a final note, all the measures defined above require no missing values for all covariates appearing in the risk prediction tool, but this is often not the reality for externally collected validation sets. Janssen et al. (2010) recently showed by simulation that imputation for missing covariates results in less biased estimates of validation metrics than other practices of either excluding the entire patient from analysis, or the covariate out of a model. The current state of the art in imputation is based on specification of full conditional distributions for missing covariates and termed Multivariate Imputation by Chained Equations (MICE), implementable with the `mice` package in `R` (van Buuren, 2007). MICE can be used without additional model specifications to impute missing data under a missing-at-random (MAR) mechanism and with additional specifications to impute missing data assumed to be not-MAR (NMAR). Briefly the method works by fitting an appropriate conditional model, such as a logistic model for dichotomous variables, for all missing variables conditional on all other variables, outcomes and covariates, in the model. The `R mice` package recommends using all measured variables in a data set to build the imputation model even if they are not part of the analysis.

4 Case study: updating the PCPTRC

The online PCPTRC was established based on logistic regression on data from 5519 placebo arm participants in the PCPT; full details on the patient inclusion and model selection procedures can be found in Thompson et al. (2006). Let $X = (1, \log(\mathrm{PSA}), \mathrm{DRE}, \mathrm{FamHist}, \mathrm{PriorBiop})'$ be a five-dimensional vector of covariates plus intercept term for an individual man, with PSA recorded in ng/mL, log denoting the natural logarithm, DRE = 1 for an abnormal digital rectal exam that is suspicious for cancer versus 0 otherwise, FamHist = 1 for a recorded history of prostate cancer in a father, brother or son and 0 otherwise, and Priorbiop = 1 if ever a prior biopsy was performed that was negative for prostate cancer and 0 otherwise (note the PCPTRC is not valid for persons with either a prior positive prostate biopsy or any prior diagnosis of prostate cancer). Let the vector β represent the estimated log odds ratios from the logistic regression used to fit the PCPTRC: $\beta = (-1.797, 0.849, 0.905, 0.269, -0.448)'$ (Thompson et al., 2006). From the signs of these log odds ratios, higher PSA, an abnormal DRE and a positive family history increased the odds of prostate cancer, while a prior negative prostate biopsy decreased the odds; for interpretation of the magnitudes see Thompson et al. (2006).

Liang et al. (2011) reported on the operating characteristics of 10 potential prostate cancer biomarkers recorded on 227 prostate cancer cases and 247 age- and race-matched controls identified in the SABOR screening cohort. Of the 10 markers studied only two were statistically significantly and independently differentiated between cases and controls:[-2]proPSA was higher among cancer cases and %freePSA was lower (both p-values<0.001, Table 1). This evidence suggested these two markers might be worthwhile candidate markers for augmenting the PCPTRC.

Figure 1 shows a joint scatterplot of the two markers for the 227 prostate cancer cases and 247 controls on the logarithmic scale. Although there is overlap, prostate cancer cases tended to have higher levels of [-2]proPSA and lower levels of %freePSA. Chi-square plots, the generalization of normal probability plots for multivariate outcomes, suggested some potential departures from multivariate normality caused by a handful of outliers for both the cases and controls (data not shown). These are seen on the left side of Fig. 1, with the four prostate cancer cases with [-2]proPSA values exceeding 100 pg/mL, a handful of controls with %freePSA very low and one with an outlying low [-2]proPSA value in addition. As the number of potential outliers was small and there

Table 1 Characteristics of participants in the SABOR and EDRN case–control studies.

Characteristic		SABOR		EDRN	
		Cases $N = 227$	Controls $N = 247$	Cases $N = 251$	Controls $N = 324$
Age (years)$^{\mathrm{V}}$	Mean (SD)	64.1 (8.4)	64.2 (8.6)	63.4 (9.3)	60.5 (7.8)
	Range	44–89	45–84	41–93	42–80
Race, N (%)$^{\mathrm{D}}$	White	129 (56.8)	167 (67.6)	216 (86.1)	276 (85.2)
	Black	32 (14.1)	35 (14.2)	21 (8.4)	24 (7.4)
	Other	66 (29.1)	45 (18.2)	8 (3.2)	18 (5.6)
	Unknown	0 (0.0)	0 (0.0)	6 (2.4)	6 (1.9)
Prior negative biopsy, N (%)	Never	177 (78.0)	203 (82.2)	251 (100.0)	324 (100.0)
	At least one	50 (22.0)	44 (17.8)	0 (0.0)	0 (0.0)
Digital rectal exam, N (%)$^{\mathrm{D}}$	Normal	160 (70.5)	240 (97.2)	190 (75.7)	261 (80.6)
	Abnormal	67 (29.5)	7 (2.8)	61 (24.3)	59 (18.2)
	Unknown	0 (0.0)	0 (0.0)	0 (0.0)	4 (1.2)
Family history, N (%)$^{\mathrm{D}}$	No	171 (75.3)	219 (88.7)	173 (68.9)	233 (71.9)
	Yes	56 (24.7)	28 (11.3)	65 (25.9)	75 (23.1)
	Unknown	0 (0.0)	0 (0.0)	13 (5.2)	16 (4.9)
PSA, ng/mL$^{\mathrm{D,V}}$	Mean (SD)	5.1 (8.8)	1.5 (1.3)	10.6 (26.3)	4.5 (3.1)
	Range	0.3–93.8	0.1–8.4	0.7–310.6	0.3–18.2
[-2]proPSA, pg/mL$^{\mathrm{D,V}}$	Mean (SD)	18.1 (34.5)	8.9 (5.7)	32.7 (128.0)	12.1 (10.4)
	Range	2.3–447.9	0.8–39.0	3.9–1831.7	2.2–133.9
%freePSA$^{\mathrm{D,V}}$	Mean (SD)	21.6 (11.6)	31.9 (11.6)	17.4 (10.4)	23.4 (10.6)
	Range	4.4–72.0	6.6–73.0	3.7–78.9	6.4–64.8

$^{\mathrm{D}}$if p-value<0.05 between cases and controls in SABOR. $^{\mathrm{V}}$if p-value < 0.05 between cases and controls in EDRN.

were no recorded reasons they might be in error it was decided to pursue the multivariate normal regression under a potential slight violation of the normal assumption rather than a more robust modeling approach to handle the outliers.

Separate model selection procedures based on Wilk's lambda test were performed for the bivariate regression of marker pairs (log%freePSA, log[-2]proPSA) on the risk factors logPSA, DRE, FamHist (family history), PriorBiop (prior biopsy), Age and Race (African american versus other) in the SABOR cases and controls. In other words, the additional predictors Age and Race were considered for the models in the likelihood ratio compared with the four predictors in the prior risk model (X) and it was not required that all components in the prior risk model be included in models for the likelihood ratio. Component models of the likelihood ratio and prior odds have different endpoints and it is not expected that the same factors would affect both. This procedure resulted in the covariates logPSA and Age being included in the multivariate regressions for the controls and cases. The multivariate generalization of the squared sum of residuals divided by the degrees of freedom was used to estimate the variance-covariance matrices for the distribution of (log%freePSA, log[-2]proPSA) in cases and controls, resulting in the following expressions for

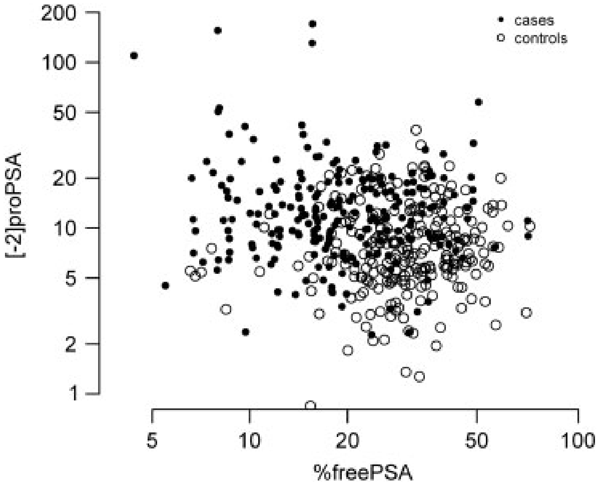

Figure 1 Scatterplots of %freePSA (in percent) and [-2] proPSA (in pg/mL) for the 227 SABOR prostate cancer cases and 247 SABOR controls.

terms comprising the likelihood ratio LR:

$$\begin{aligned}
Y &= (\log \%\text{freePSA}, \log[-2]\text{proPSA})' \\
\mu_{\text{cancer}} &= \begin{bmatrix} 2.667 - 0.365\log \text{PSA} + 0.011\text{Age} \\ 1.385 + 0.627\log \text{PSA} + 0.006\text{Age} \end{bmatrix} \\
\Sigma_{\text{cancer}} &= \begin{bmatrix} 0.179 & 0.121 \\ 0.121 & 0.231 \end{bmatrix} \\
\mu_{\text{no cancer}} &= \begin{bmatrix} 3.276 - 0.235\log \text{PSA} + 0.002\text{Age} \\ 2.438 + 0.571\log \text{PSA} - 0.008\text{Age} \end{bmatrix} \\
\Sigma_{\text{no cancer}} &= \begin{bmatrix} 0.128 & 0.097 \\ 0.097 & 0.188 \end{bmatrix}.
\end{aligned}$$

With these parameter values substituted in, LRs become a function of a patient's age, PSA, %freePSA and [-2]proPSA values. LRs greater than 1 represent configurations of these characteristics that inflate posterior risks of prostate cancer greater than PCPTRC risks, LRs less than 1 represent configurations that diminish posterior risks compared with prior risks, and LRs equal to 1, situations where the addition of the new markers %freePSA and [-2]proPSA do not alter prior risks. LR surfaces as a function of %freePSA and [-2]proPSA for a 65-year-old man with PSA 2.0 ng/mL versus a 65-year-old man with PSA 5.0 ng/mL are shown in Fig. 2, where the axes have been chosen to reflect contours of the marker pairs near values of LRs of 1, meaning no preference for cancer case versus control assignment. For both scenarios, LRs are highest for low %freePSA and high [-2]proPSA values, but specific pairs of %freePSA and [-2]proPSA yield different LRs for lower versus higher PSA values as seen by comparison of Fig. 2A and B.

5 Case study: Validation of the updated PCPTRC

The EDRN validation cohort is also listed in Table 1 and comprised 575 men who presented at multiple urologic facilities in the northeastern part of the U.S. with clinical symptoms, and was previously reported on by Sokoll et al. (2010). As seen in Table 1, none of the men in the cohort ever

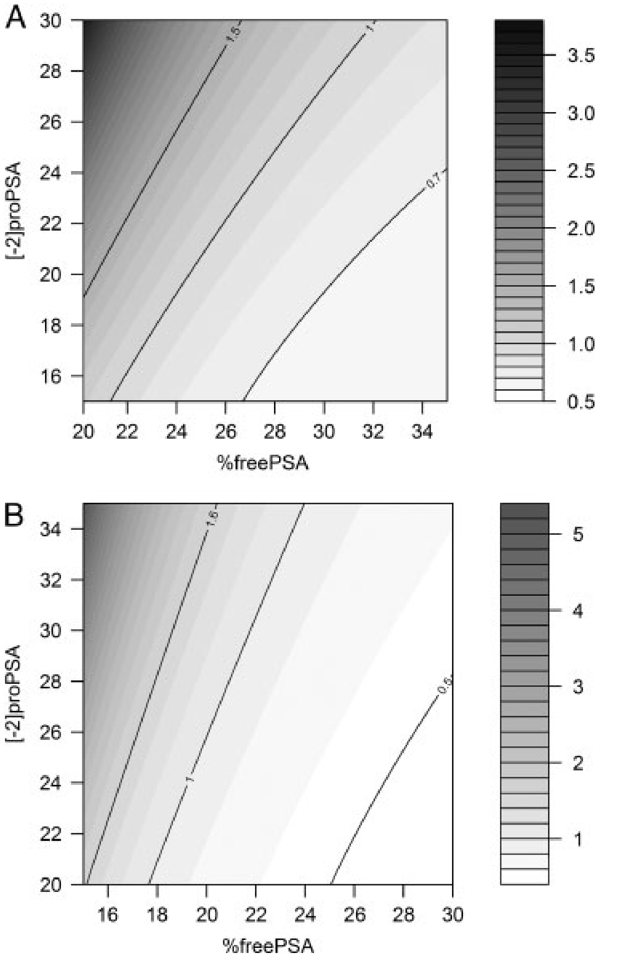

Figure 2 Contour plots of estimated likelihood ratios for a 65-year-old man with PSA value 2.0 ng/mL (A) and 5.0 ng/mL (B).

had a prior biopsy. Only the discrete risk factors, race, DRE, and family history, had any missing values in the EDRN validation set, and these were very modest (approximately 2% for race and DRE and 5% for family history). Multiple imputation via chained equations was peformed under a MAR hypothesis for missing data, by specifying logistic regressions for race, DRE, and family history conditional on all other variables in Table 1 except for prior biopsy, along with prostate cancer status as covariates. The number of iterations was set to 20 and five imputed data sets were retained for analysis. Any validation summaries requiring imputation were averaged over these five data sets, thus resulting in ranges for some sample sizes.

Figure 3 shows ROCs for updated posterior PCPTRC risks (AUC = 0.698, 95% CI = 0.655–0.741), %freePSA (AUC = 0.693, 95% CI = 0.649–0.736), PCPTRC risks (AUC = 0.677, 95% CI = 0.634–0.721), PSA (AUC = 0.663, 95% CI = 0.619–0.707), and [-2]proPSA (AUC = 0.648, 95% CI = 0.604–0.693) evaluated on the EDRN validation set. There were no statistical differences between any of the AUCs as assessed by paired AUC tests (all p-values > 0.05). Restricting to the clinically relevant part of the AUC curve, the portion with FPRs less than 20%, revealed that posterior PCPTRC risks had the highest sensitivities for FPRs $< 10\%$, followed by prior PCPTRC risks, but in the region of FPRs ranging from 10 to 20%, %freePSA obtained the

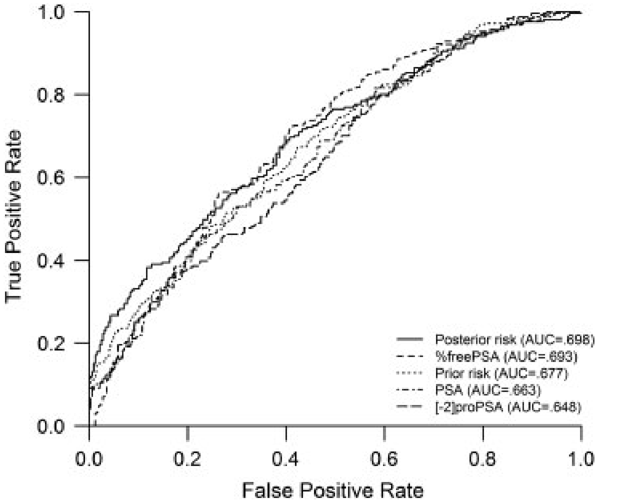

Figure 3 ROCs for the EDRN validation set.

Table 2 Sensitivities in percent [thresholds of marker or risk tool in respective units] obtained in the EDRN validation set for false-positive rates (FPR) 5, 10, 15 and 20%.

Marker or tool (units)	FPR 5%	FPR 10%	FPR 15%	FPR 20%
Posterior PCPTRC risk (%)	26.7 [75.1]	33.1 [66.9]	39.4 [58.7]	45.0 [54.7]
PCPTRC risk (%)	21.9 [62.2]	28.7 [55.8]	33.5 [52.1]	40.6 [48.4]
PSA (ng/mL)	14.7 [10.8]	21.9 [8.4]	32.7 [6.8]	39.8 [6.1]
[-2]proPSA (pg/mL)	13.9 [28.3]	25.1 [21.6]	32.7 [17.7]	37.8 [15.7]
%freePSA (%)	19.5 [9.8]	33.9 [11.9]	43.4 [13.4]	49.4 [14.6]

For all markers except %freePSA, values above the threshold indicate a positive test; for %freePSA values at or below the threshold indicate a positive test.

same sensitivities as both PCPTRC and posterior PCPTRC risks (Table 2). McNemar's test performed at the 0.05 level of statistical significance revealed no differences between sensitivities between the posterior PCPTRC risks and PCPTRC risks, nor between posterior PCPTRC risks and %freePSA, at all FPRs listed in Table 2.

The average PCPTRC risk and updated PCPTRC posterior risk among the 575 participants in the EDRN validation set were 43.1% (95% CI 41.8–44.4%) and 45.9% (95% CI 44.0–47.8%), respectively, indicating that posterior risks were higher than prior risks on average in the cohort. The actual proportion of the EDRN cohort that had prostate cancer, 43.7% (95% CI 39.6–47.8%), indicated the possibility however that posterior PCPTRC risks could be overestimating and PCPTRC risks underestimating actual risks, a problem with calibration-in-the-large as termed in Steyerberg (2010). Table 3 reveals that this tendency of the posterior PCPTRC risk to overestimate actual risks occurred across many subpopulations. For most of the risk categories, average posterior PCPTRC risks tended to overestimate actual risks, although the 95% CIs did overlap. For the subgroup with [-2]proPSA 15 pg/mL or less, average posterior PCPTRC risks were statistically significantly higher than the actual rate of prostate cancer in the subgroup: 44.7% (95% CI 42.6–46.8%) compared with 37.6% (95% CI 32.9–42.3%), respectively.

The poor calibration of posterior PCPTRC risks to actual risks is more formally borne out in Table 4, where the 575 members of the EDRN cohort have been grouped according to deciles of

Table 3 Comparisons of prior and posterior PCPTRC risks to observed prostate cancer rates in the EDRN for various subgroups

Subgroup (sample size range across imputed data sets)	Prostate cancer (%) [95% CI]	Average PCPTRC risk (%) [95% CI]	Average posterior PCPTRC risk (%) [95% CI]
All (575)	43.7 [39.6, 47.8]	43.1 [41.8, 44.4]	45.9 [44.0,47.8]
PSA 4 ng/mL or less (242)	32.2 [26.3, 38.1]	31.7 [30.4, 32.9]	37.4 [35.1, 39.8]
PSA greater than 4 ng/mL (333)	52 [46.6, 57.4]	51.4 [49.9, 52.9]	52.1 [49.5, 54.7]
DRE normal (454–455)	41.9 [37.3, 46.4]	40.1 [38.9, 41.3]	42.7 [40.7, 44.7]
DRE abnormal (120–121)	50.6 [41.6, 59.5]	54.2 [50.7, 57.7]	58.2 [54.1, 62.3]
Race other (527–530)	43.3 [39.1, 47.6]	42.7 [41.4, 44.0]	45.7 [43.8, 47.7]
Race black (45–48)	47.4 [33.0, 61.8]	48 [42.3, 53.8]	48.5 [41.3, 55.8]
Age 65 years or older (214)	52.3 [45.6, 59.0]	47.5 [45.5, 49.6]	46 [43.1, 48.9]
Age younger than 65 (361)	38.5 [33.5, 43.5]	40.5 [38.9, 42.0]	45.9 [43.4, 48.4]
Family history (147–155)	46.9 [38.9, 54.9]	44.7 [42.3, 47.1]	48 [44.5, 51.6]
No family history (420–428)	42.5 [37.8, 47.2]	42.5 [41.0, 44.1]	45.2 [43.0, 47.4]
[-2]proPSA 15 pg/mL or less (402)	37.6 [32.9, 42.3]	38.1 [36.9, 39.4]	44.7 [42.6, 46.8]
[-2]proPSA greater than 15 pg/mL (173)	57.8 [50.4, 65.2]	54.6 [52.1, 57.0]	48.8 [44.7, 52.9]
%freePSA 20 or less (309)	56 [50.5, 61.5]	46.7 [45.0, 48.4]	58.9 [56.5, 61.3]
%freePSA greater than 20 (266)	29.3 [23.8, 34.8]	38.9 [37.0, 40.7]	30.9 [29.2, 32.6]

their PCPTRC posterior risk values. In only 4 out of the 10 decile groups did actual proportions of prostate cancer fall within the bounds of the predicted PCPTRC risks, a statistically significant rejection of goodness-of-fit according to the Hosmer–Lemeshow test (p-value = 0.003).

The discrimination slope or average difference between predicted risks between cancer and non-cancer cases for the upgraded PCPTRC model on the EDRN validation set was 16.8%, a modest improvement over the 10.5% for the original PCPTRC. The difference, or IDI, was statistically significantly different from 0 (6.3%, 95% CI 3.0–9.6%). The net benefit curves in Fig. 4 indicated benefit of using both the PCPTRC prior and posterior risks over the blanket rule of referring all men in the EDRN cohort to prostate biopsy for thresholds exceeding 20%, but no clear benefit of the more complicated posterior PCPTRC risks over the prior PCPTRC risks in the region of risks between 10 and 40% of most clinical relevance to men deciding on whether to pursue prostate

Table 4 Comparisons of posterior PCPTRC risks to observed prostate cancer rates in the EDRN by deciles of posterior PCPTRC risks used in calculation of the Lemeshow–Homer test of goodness-of-fit.

Posterior PCPTRC risk range (%)	Proportion of EDRN cohort with prostate cancer
2.1–21.7	15.5
21.7–25.7	26.3
25.7–30.4	41.4
30.4–34.7	26.3
34.7–40.3	41.4
40.3–46.4	43.9
46.4–55.6	49.1
55.6–66.8	48.3
66.8–82.9	57.2
82.9–99.9	86.2

Sample sizes in each category were either 57 or 58.

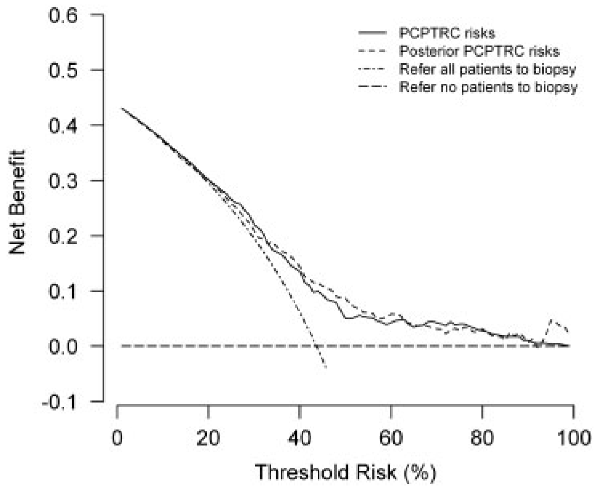

Figure 4 Net benefit curves for the EDRN validation set.

biopsy as referenced by Steyerberg and Vickers (2008). Across all thresholds, both the PCPTRC and posterior PCPTRC rules provided benefit over the rule referring no patients to biopsy.

6 Concluding remarks

This case study illustrated a simple analytical update to the original PCPTRC incorporating %freePSA and [-2]proPSA, and accompanying R code for the formulas are available as Supporting Information and posted online at the PCPTRC website. The methodology was based on estimating likelihood ratios based on multivariate normal distributions applied separately to data from cancer cases and controls obtained from an external case–control study to that used to build the prior PCPTRC model. Retrospective case-control designs are much more efficient than prospective cohort studies for studying relationships between markers and cancer, so the methodology optimizes incorporation of new information for upgrading the PCPTRC. Towards a different objective of evaluating the diagnostic accuracy of new markers, Janssens et al. (2005) and Gu and Pepe (2009, 2011) have proposed an alternative method for estimating log likelihood ratios as the difference between the log posterior and prior odds, with logistic regressions used for both. This approach has some simplicity in modeling assumptions since only the disease outcome is modeled but is tailored to the situation where the new and old markers are measured on the same data set, not where a considerably sized cohort has been used to build the prior odds and a smaller more efficient case–control study is all that is available for the new markers as is the case here.

Formulated as a more general problem for updating an existing risk model founded on logistic regression to apply to new settings, for example, to adapt risk tools to changes of grading systems for tumors, or new geographic regions, and/or to include new markers, Chapters 20 and 21 of Steyerberg (2010) contain an excellent review of the many frequentist and empirical Bayes methodologies that have recently come under study. For example, if one risk model is to be extended to a population known to have a higher prevalence of disease, an updating methodology for the intercept in the logistic regression is proposed, incorporating prior information from other sources, and with shrinkage in order to not overfit to the new data. Due to the small SABOR data set used to form the likelihood ratio, some shrinkage might also improve the poor calibration performance, and special fixes among subpopulations might be warranted for some of the subgroups of patients in Table 3. These very interesting directions warrant further detailed study through simulation and collection of more data. Additionally, continual dynamic updating of risk models as new data flow in are another topic of increased attention due to their practical need (Steyerberg, 2010). For example, we expect new case–control studies for %freePSA and [-2]proPSA to soon become available, as these markers are becoming more widely used. It might be prudent then to shift the EDRN validation set to become a joint development data set with SABOR and make space for the more recently available data to be used as the new validation set. This incurs new methodologic challenges for combining multiple data sets, potentially requiring the introduction of random effects (Steyerberg, 2010).

A limitation to likelihood ratios modeled as ratios of probability density functions is that they are unstable in areas where the control density in the denominator has a smaller tail than the case density in the numerator, causing the denominator to approach 0 and the likelihood ratio and hence posterior risks to explode. This occurs for some configurations of the covariates and markers. Similar problems occur with other risk prediction models so it is routine practice to bound risks reported on the PCPTRC website at 75% (the returned output is a message that the risk exceeds 75% with no confidence interval reported). The problem can be alleviated somewhat by constraining the variances for cancer cases and controls to be the same, which also will help monotonicity of the likelihood ratio (that for a marker elevated in cancer cases, the likelihood ratio will monotonically increase). Unfortunately this constraint does not match science, however, as even after transformation, markers are typically much more dispersed in cancer cases compared to controls. More research is needed into taming the likelihood ratios to find a balance between accurate modeling of marker distributions among cancer cases and controls and stability of outcome.

Our method assumed marker distributions could be transformed to multivariate normality in cancer cases and controls, which retains a nice analytical formula for posterior risks. However, typically, as was the case in the data set of this study, the observed markers contain potential outliers and/or skewness in their distribution. Furthermore there could conceivably be clusters among the diseased and non-diseased populations, identifiable with sufficiently large sample sizes. Multivariate skew t mixture distributions pose a nice framework for handling these issues and estimation-maximization algorithms are available for fitting (Frühwirth-Schnatter and Pyne, 2010). Additionally they contain multivariate normal, skew normal, t and skew t as special cases. Due to the small sample sizes in this study it was not anticipated the extensions would make much of a difference, but it is a worthy route of investigation for future studies to upgrade the PCPTRC. Additional extensions are needed for scenarios where the markers to be incorporated are not all continuous, but rather mixtures of categorical and continuous variables, or where multiple external case–control studies are combined, potentially warranting a random-effects or meta-analysis type approach.

Validation on the EDRN validation cohort, a more clinical cohort with higher PSA values than the SABOR cohort, indicated however, no clear advantages of posterior PCTPRC risks that incorporate the two less easily obtainable markers to the standard PCTPRC in terms of net benefit, calibration nor discrimination, and also no clear improvement in discrimination over just using one of the markers, %freePSA. Interestingly, a referee pointed out that the strong correlation between logarithms of the two markers in the estimated multivariate normal models for cancer cases (correlation 0.595) and controls (0.789) might actually be diminishing performance of the posterior PCPTRC as compared with just using the stronger of the two correlated markers, %freePSA, to upgrade the PCPTRC. We investigated this possibility by just using the marginal univariate normal distribution for log %freePSA implied by the fitted bivariate normal model and evaluated marginal posterior risks based on just this marker. We were surprised by the results: the estimated AUC of marginally upgraded PCPTRC was 0.722, just a minor 0.02 increase above the AUC of the posterior PCPTRC (AUC = 0.698), but enough to be statistically significantly better than the posterior PCPTRC (p-value = 0.02) and %freePSA (p-value = 0.03). Clinically, this is a pleasing finding since %freePSA is a much more widely available marker than [-2]proPSA, and the resulting calculator is simpler to implement. Statistically, it points out that the same parsimonious model-building principles that would automatically be applied when introducing new markers as covariates into a logistic model would also apply for their incorporation into risk prediction tools through likelihood ratios. In other words, one should consider model selection on the number of new markers modeled as outcomes in the likelihood ratios by investigating subsets of markers.

Reports of a single validation cohort should not be taken as the final word on a risk prediction tool as even for the most studied marker of prostate cancer, PSA, reports of its AUC have ranged from no predictive power 0.525 (Nguyen et al., 2010), to decent predictive power, 0.678 (Thompson et al., 2005), to remarkable power, 0.840 (Liang et al., 2011). For this reason, it is critical that risk tools be published online to facilitate a spectrum of validations. An advantage to transparent analytical formulas underpinning risk models, such as the likelihood ratio approach adopted in this article, is that they foster multiple external validations, the collective experience of which can ultimately judge whether or not and for which populations a tool should be used. An additional advantage of posting risk tools online is that they transport state-of-the-art medicine to even remote corners of the world, to small community clinics and to their patients.

Acknowledgements This research was supported by National Institutes of Health grants CA086402, CA054174, CA86323, UO1CA113913, EDRN grant CA115102, and a National Cancer Institute Cancer Center support grant to the UTHSCSA CTRC: 5P30 CA0541474-18. The authors thank the SABOR and EDRN participants.

Conflict of interest

The authors have declared no conflict of interest.

References

Ankerst, D. P., Groskopf, J., Day, J. R., Blase, A., Rittenhouse, H., Pollock, B. H., Tangen, C., Parekh, D., Leach, R. J. and Thompson, I. (2008). Predicting prostate cancer risk through incorporation of prostate cancer gene 3. *Journal of Urology* **180**, 1303–1308.

Anscombe, F. J. (1967) Topics in the investigation of linear relations fitted by the method of least squares (with discussion). *Journal of the Royal Statistical Society Series B* **29**, 1–52.

Chatfield, C. (1995). Model uncertainty, data mining, and statistical inference (with discussion). *Journal of the Royal Statistical Society Series A* **158**, 419–466.

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845.

Eyre, S. J., Ankerst, D. P., Wei, J. T., Nair, P. V., Regan, M. M., Bueti, G., Tang, J., Rubin, M. A., Kearney, M., Thompson, I. M. and Sanda, M. G. (2009). Validation in a multiple urology practice setting of the Prostate Cancer Prevention Trial calculator for predicting prostate cancer detection. *Journal of Urology* **182**, 2653–2658.

Frühwirth-Schnatter, S. and Pyne, S. (2010). Bayesian interface for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics* **11**, 317–336.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C. and Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* **81**, 1879–1886.

Gail, M. H., Costantino, J. P., Pee, D., Bondy, M., Newman, L., Selvan, M., Anderson, G. L., Malone, K. E., Marchbanks, P. A., McCaskill-Stevens, W., Norman, S. A., Simon, M. S., Spirtas, R., Usin, G. and Bernstein, L. (2007). Projecting individualized absolute invasive breast cancer risk in African American women. *Journal of the National Cancer Institute* **99**, 1782–1792.

Greenland, S. (2008). 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. et al., Statistics in Medicine. *Statistics in Medicine* **27**, 199–206.

Grundy, S. M., Cleeman, J. I., Merz, C. N. B., Brewer, B. Jr., Clark, L. T., Hunninghake, D. B., Pasternak, R. C., Smith, S. C. and Stone, N. J. for the Coordinating Committee of the National Cholesterol Education Program (2004). Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III Guidelines. *Circulation* **110**, 227–239.

Gu, W. and Pepe, M. S. (2009). Estimating the capacity for improvement in risk prediction with a marker. *Biostatistics* **10**, 172–186.

Gu, W. and Pepe, M. S. (2011). Estimating the diagnostic likelihood ratio of a continuous marker. *Biostatistics* **12**, 87–101.

Hernandez, D. J., Han, M., Humphreys, E. B., Mangold, L. A., Taneja, S. S., Childs, S. J., Bartsch, G. and Partin, A. W. (2009). Predicting the outcome of prostate biopsy: comparison of a novel logistic regression-based model, the prostate cancer risk calculator, and prostate-specific antigen level alone. *British Journal Urology International* **103**, 609–614.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. Springer, New York.

Janssen, K. J. M., Donders, A. R. T., Harrell, F. E. Jr., Vergouwe, Y., Chen, Q., Grobbee, D. E. and Moons, K. G. M. (2010). Missing covariate data in medical research: To impute is better than to ignore. *Journal of Clinical Epidemiology* **63**, 721–727.

Janssens, A. C. J. W., Deng, Y., Borsboom, G. J. J. M., Eijkemans, M. J. C., Habbema, J. D. F. and Steyerberg, E. W. (2005). A new logistic regression approach for the evaluation of diagnostic test results. *Medical Decision Making* **25**, 168–177.

Lemeshow, S. and Hosmer, D. W. Jr., (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology* **115**, 92–106.

Liang, Y., Ankerst, D. P., Ketchum, N. S., Ercole, B., Shah, G., Shaughnessy, J. D. Jr., Leach, R. J. and Thompson, I. M. (2011). Prospective evaluation of operating characteristics of prostate cancer detection biomarkers. *Journal of Urology* **185**, 104–110.

Nguyen, C. T., Yu, C., Moussa A., Kattan, M. W. and Jones, J. S. (2010). Performance of prostate cancer prevention trial risk calculator in a contemporary cohort screened for prostate cancer and diagnosed by extended prostate biopsy. *Journal of Urology* **183**, 529–533.

Parekh, D. J., Ankerst, D. P., Higgins, B. A., Hernandez, J., Canby-Hagino, E., Brand, T., Troyer, D., Leach, R. and Thompson, I. M. (2006). External validation of the Prostate Cancer Prevention Trial risk calculator. *Urology* **68**, 1152–1155.

Pencina, M. J., D'Agostino Sr., R. B., D'Agostino Jr., R. B., Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27**, 157–172.

Pepe, M. S., Etzioni, R., Feng, Z., Potter, J. D., Thompson, M. L., Thornquist, M., Winget, M. and Yasui, Y. (2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* **93**, 1054–1061.

Skates, S. J., Pauler, D. K. and Jacobs, I. J. (2001). Screening based on the risk of cancer calculation from Bayesian hierarchical change-point models of longitudinal markers. *Journal of the American Statistical Association* **96**, 429–439.

Sokoll, L. J., Sanda, M. G., Feng, Z., Kagan, J., Mizrahi, I. A., Broyles, D. L., Partin, A. W., Srivasta, S., Thompson, I. M., Wei, J. T., Zhang, Z. and Chan, D. W. (2010). A prospective, multicenter National Cancer Institute Early Detection Research Network study of [-2]proPSA: improving prostate cancer detection and correlating with cancer aggressiveness. *Cancer Epidemiology, Biomarkers and Prevention* **19**, 1193–1200.

Steyerberg, E. W. and Vickers, A. J. (2008). Decision curve analysis: a discussion. *Medical Decision Making* **28**, 146–149.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J. and Kattan, M. W. (2010). Assessing the performance of prediction models, a framework for traditional and novel measures. *Epidemiology* **21**, 128–138.

Steyerberg, E. W. (2010). *Clinical Prediction Models*. Springer, New York.

Thompson, I. M., Ankerst, D. P., Chi, C., Lucia, M. S., Goodman, P., Crowley, J. J., Parnes, H. L. and Coltman Jr., C. A. (2005). The operating characteristics of prostate-specific antigen in a population with initial PSA of 3. 0 ng/mL or lower. *Journal of the American Medical Association* **294**, 66–70.

Thompson, I. M., Ankerst, D. P., Chi, C., Goodman, P. J., Tangen, C. M., Lucia, M. S., Feng, Z., Parnes, H. L. and Coltman Jr., C. A. (2006). Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial. *Journal of the National Cancer Institute* **98**, 529–534.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* **16**, 219–242.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* **26**, 565–574.

Wolf, A. M. D., Wender, R. C., Etzioni, R., Thompson, I. M., D'Amico, A. V., Volk, R. J., Brooks, D. D., Dash, C., Guessous, I., Andrews, K., DeSantis, C. and Smith, R. A. (2010). American Cancer Society guideline for the early detection of prostate cancer: update 2010. *CA Cancer Journal for Clinicians* **60**, 70–98.

Yates, J. F. (1982). External correspondence: decomposition of the mean probability score. *Organizational Behavior and Human Performance* **30**, 132–156