American Journal of Epidemiology
Vol. 167, No. 3
DOI: 10.1093/aje/kwm305
Advance Access publication November 2, 2007

Practice of Epidemiology

Integrating the Predictiveness of a Marker with Its Performance as a Classifier

Margaret S. Pepe[1,2], Ziding Feng[1], Ying Huang[2], Gary Longton[1], Ross Prentice[1], Ian M. Thompson[3], and Yingye Zheng[1]

[1] Fred Hutchinson Cancer Research Center, Seattle, WA.
[2] University of Washington, Seattle, WA.
[3] University of Texas Health Sciences Center, San Antonio, TX.

Received for publication March 15, 2007; accepted for publication September 18, 2007.

There are two popular statistical approaches to biomarker evaluation. One models the risk of disease (or disease outcome) with, for example, logistic regression. A marker is considered useful if it has a strong effect on risk. The second evaluates classification performance by use of measures such as sensitivity, specificity, predictive values, and receiver operating characteristic curves. There is controversy about which approach is more appropriate. Moreover, the two approaches can give contradictory results on the same data. The authors present a new graphic, the predictiveness curve, which complements the risk modeling approach. It assesses the usefulness of a risk model when applied to the population. Although the predictiveness curve relates to classification performance measures, it also displays essential information about risk that is not displayed by the receiver operating characteristic curve. The authors propose that the predictiveness and classification performance of a marker, displayed together in an integrated plot, provide a comprehensive and cohesive assessment of a risk marker or model. The methods are demonstrated with data on prostate-specific antigen and risk factors from the Prostate Cancer Prevention Trial, 1993–2003.

biological markers; classification analysis; diagnostic tests, routine; epidemiologic methods; predictive value of tests; prostate-specific antigen; risk assessment; risk model

Abbreviations: CI, confidence interval; FPF, false positive fraction; PSA, prostate-specific antigen; ROC, receiver operating characteristic; TPF, true positive fraction.

Biomarker development is a major focus of research in cancer as well as in other diseases. We seek biomarkers for many purposes, including risk assessment, screening, diagnosis, and prognosis. New molecular technologies, in particular, promise to provide biomarkers that can inform about risk and help guide clinical decisions.

There are two basic statistical approaches for evaluating such biomarkers. The first models the risk of disease (or disease outcome) as a function of the biomarker(s) with, for example, logistic (or Cox) regression. The value of a marker is measured by its effect on risk conditional on other predictors. This is adequate in etiologic research but does not address the capacity of the marker to correctly classify or predict risk in the population. The second summarizes marker performance with classification performance measures, such as sensitivity and specificity, predictive values, and receiver operating characteristic (ROC) curves. There is controversy about which approach is more appropriate. Moons and Harrell (1) argue in favor of risk models, since ultimately the patient wants to know his risk given his biomarker measurement. On the other hand, Pepe et al. (2) emphasize that the public health value of a marker lies in the fraction of diseased (or destined to be diseased) subjects detected, that is, sensitivity, and the fraction of

Correspondence to Dr. Margaret Sullivan Pepe, Biostatistics and Biomathematics Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-B500, Seattle, WA 98109 (e-mail: mspepe@u.washington.edu).

nondiseased subjects falsely identified as diseased, that is, 1 – specificity.

Both statistical approaches are frequently applied, often to the same data. However, the relation between them is unclear. Of particular concern, the two approaches frequently yield apparently contradictory results. A marker that is strongly related to risk may be a poorly performing classifier (2). A marker that is a strong predictor of risk after controlling for other risk factors often adds little to them in terms of improving classification performance.

In this paper, we present a new graphic, the ''predictiveness curve,'' which combines concepts from both risk modeling and population performance approaches to analysis. In particular, it is useful for assessing the fit of a risk model and the clinical utility of the model when applied to the population. We also extend the plot to simultaneously evaluate the risks associated with a marker and the marker's performance as a classifier. This integrated approach provides a more complete and comprehensive analysis than current practice.

DATA FOR ILLUSTRATION

We illustrate this new approach with data from the recently reported Prostate Cancer Prevention Trial (3). A total of 5,519 men on the placebo arm of the study underwent prostate biopsy and had at least two prostate-specific antigen (PSA) measurements in the 3 years prior to biopsy. Along with PSA and PSA change over time, data on family history of prostate cancer, results of digital rectal examination, age, ethnicity, and prior biopsy were used to model the risk of finding prostate cancer and the risk of high-grade disease (Gleason score of ≥ 7) at the time of prostate biopsy. Because the data are used only for illustrating a statistical method, in the interests of being relatively brief, we restrict the analysis to high-grade disease, although a similar approach could be used for all prostate cancer. Of the 5,519 men, 4.7 percent ultimately were found to have high-grade disease. Table 1 shows the results of the logistic regression analysis. Procedures for selecting variables and fitting models are as described by Thompson et al. (3), who found that, for the diagnosis of high-grade disease, PSA, digital rectal examination, age, and prior negative biopsy appeared to be predictive of risk.

THE PREDICTIVENESS CURVE

We can calculate an individual's estimated risk given data on his risk factors by use of the fitted risk model. For the prostate cancer example, the calculation (3) is as follows:

$$\text{Risk of high-grade disease} = \exp(\text{Y})/\{1+\exp(\text{Y})\}$$

where DRE is digital rectal examination, and

$$Y = -5.94 + 1.30[\log(\text{PSA})] + 0.03(\text{age}) + 0.99 \times (\text{DRE positive}) - 0.37(\text{prior biopsy}).$$

This risk calculator of Thompson et al. (3) is available online at http://www.compass.fhcrc.org/edrnnci/bin/calculator/main.asp. We calculated the estimated risk for each of the individuals in the Prostate Cancer Prevention Trial. The predictiveness curve in figure 1 shows the distribution of risks. To create the curve, we ordered the risks from lowest to highest and plotted their values. We see that, at 90 percent on the x-axis, the risk value is 0.104. This indicates that, on the basis of the predictors in the model, 90 percent of subjects in the cohort have calculated risks below 0.104 and only 10 percent have risks at or above 0.104.

Another way of using the graph is to start at a risk value on the y-axis and to read the corresponding percent on the x-axis. For example, with ''risk = 0.20,'' we see that the percent is 97.8 percent. That is, we estimate that 2.2 percent of the subjects in the cohort have estimated risks at or above 0.20. With ''risk = 0.02,'' the percent is 39.0 percent, indicating that 39.0 percent of the subjects in the cohort have risks below 0.02.

TABLE 1. Logistic regression analysis* of risk for high-grade disease, Prostate Cancer Prevention Trial, 1993–2003

Factor	Log odds ratio	*p* value†
Log(PSA‡)	1.30	<0.001
Age (years)	0.03	0.02
DRE‡	0.99	<0.001
Prior biopsy	−0.37	0.04
Constant	−5.94	

* Analysis as reported by Thompson et al. (3).
† *p* values are based on two-sided Wald tests.
‡ PSA, prostate-specific antigen; DRE, digital rectal examination.

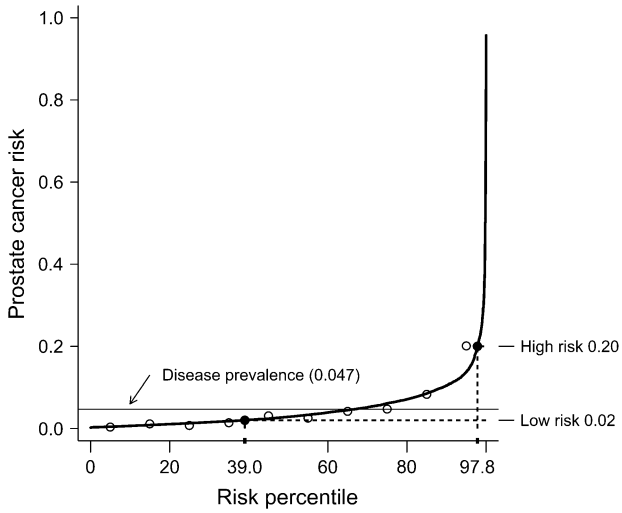

FIGURE 1. Predictiveness curve for the risk model shown in table 1 that includes prostate-specific antigen, age, digital rectal examination, and prior biopsy as risk factors for high-grade prostate cancer, Prostate Cancer Prevention Trial, 1993–2003. Shown on right are the risk thresholds of ≥ 0.20 for high risk and < 0.02 for low risk. Open circles display observed proportions of high-grade cancers within risk deciles.

What does the graph offer that is not summarized in table 1? It shows the range and distribution of estimated risk levels associated with the model when it is applied to the population from which the cohort was drawn. Consider that an individual wants to use his calculated risk in deciding whether or not to have a biopsy. The decision is more straightforward if his estimated risk of disease is close to 0 or 1. If his calculated risk is in an equivocal range, it is not helpful. Suppose, for illustration, that 20 percent risk of high-grade disease is sufficiently high to recommend a biopsy and that 2 percent risk is sufficiently low to decide against biopsy. Individuals whose risks are calculated in the range 0.02–0.20 are unsure about whether or not they should have a biopsy obtained. (A formal cost-benefit analysis that incorporates their risk of disease might be helpful, although specifying costs and benefits is always difficult.) A risk model will be most useful for individual decision making if calculated risks of having high-grade disease tend to exceed 20 percent or be less than 2 percent. We see from figure 1, however, that the prostate cancer risk model leaves the majority of men, 58.8 percent, in the indecisive risk region. Alternative thresholds might be chosen for defining high and low risk. If it is reasonable to assume that a man with a <5 percent risk of high-grade disease may defer further evaluation while a man with a >10 percent risk would prefer an evaluation, the corresponding indecisive risk region would contain only 25 percent of the population. It is important to keep in mind, however, that individuals typically do not distinguish between minor variations in risk, so we prefer to use the more extreme definitions of low and high risk in our illustrations.

A risk calculator should be derived from a risk model that fits the data well. The standard approach to evaluating model fit, that is, calibration, is to categorize subjects according to deciles (or other quantiles) of risk according to the model and to compare average predicted risk with the observed proportion of events in each category. The Hosmer-Lemeshow statistic (4) uses this approach to formally test for goodness of fit. Interestingly, the predictiveness curve offers a graphical approach to assessing goodness of fit in this sense. At the midpoint of each decile of risk in figure 1, we superimpose the corresponding observed proportions of high-grade cancer. Visually, one can compare these observed proportions with the predictiveness curve, noting that the curve averaged over the decile category is the average model predicted risk. An equivalent display often seen in practice is to plot the observed proportion versus the average risk (5, section 14.6). For the model in table 1, the Hosmer-Lemeshow statistic is 9.11 ($p = 0.33$), indicating that it fits the data rather well. However, the graphical display offers a more complete description of how observed and modeled risks compare. It shows the components of the test statistic. In addition, when there is particular interest in model fit in low- and high-risk groups, figure 1 allows one to focus accordingly. We obtained similar results when the data were split into halves, with the model fit on one half and assessed with the Hosmer-Lemeshow statistic and corresponding graphic on the second half. This avoids issues with fitting a model and assessing its fit with the same data.

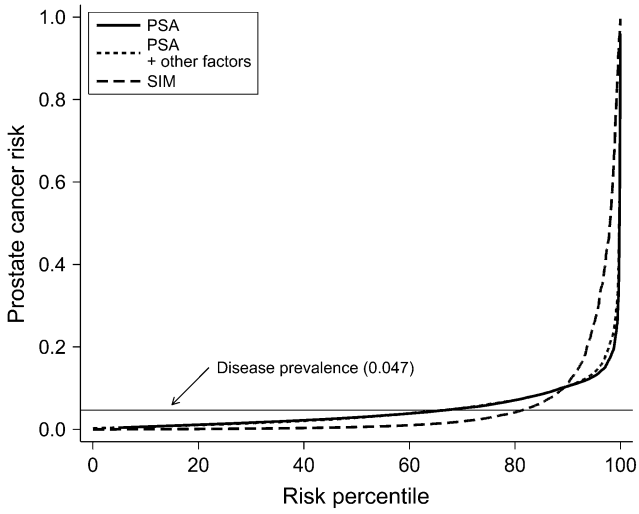

FIGURE 2. Predictiveness curves for prostate-specific antigen (PSA) alone, PSA and other factors, and the simulated marker (SIM), Prostate Cancer Prevention Trial, 1993–2003.

In viewing the predictiveness curve, one must also be cognizant of sampling variability. Neither the risks nor the distributions of predictors in the population are known with certainty, and both components enter into the predictiveness curve. This can be addressed by using bootstrapping techniques (6) to calculate confidence intervals and p values. We used the simple bootstrap, resampling 5,519 subjects with replacement from the original data set, fitting the risk model with the four selected covariates and calculating fitted risks for resampled subjects. When confidence intervals were calculated, a confidence level of 95 percent was used throughout. As an example, we noted that only 10 percent of the subjects have risks in excess of 0.104 (95 percent confidence interval (CI): 0.090, 0.120), indicating that the risk quantile is estimated rather precisely, at least assuming correct form for the risk model. Similarly, the estimates and confidence intervals for the proportions of subjects with risks at or above 0.20 and below 0.02 are 0.022 (95 percent CI: 0.014, 0.034) and 0.390 (95 percent CI: 0.318, 0.467), respectively.

Different risk models can be compared through their predictiveness curves. In figure 2, we see that the predictiveness curve for PSA alone is almost identical to that of the more comprehensive model that includes the additional risk factors of age, prior biopsy, and digital rectal examination. Both models calculate risks at or less than the 0.02 low-risk threshold for 36 percent and 39 percent of the population, respectively. Although the p value for this comparison, $p = 0.05$, is marginally statistically significant, the magnitude of the difference, 3 percent, is clinically insignificant. At the high-risk end of the scale, the PSA model puts 1.2 percent (95 percent CI: 0.7, 2.2) of subjects at or above the 0.20 risk level, while the more comprehensive model puts 2.2 percent (95 percent CI: 1.4, 3.4) of subjects in the high-risk range ($p = 0.007$). For comparison, we also include a simulated

marker with much better performance. The simulated marker identifies 70.4 percent (95 percent CI: 66.7, 73.9) of the subjects as low risk and 6.3 percent (95 percent CI: 5.5, 7.2) as high risk, but it leaves 23.3 percent with calculated risks in the equivocal range between 0.02 and 0.20. This marker was simulated as a standard normal random variable for controls and a normal (mean = 2, standard deviation = 1) random variable for cases.

Another approach to comparing risk models is with the *R*-squared statistic, the proportion of explained variation generalized from linear to logistic regression (7). The values 0.053, 0.066, and 0.310 for PSA alone, for PSA and other factors, and for the simulated marker, respectively, corroborate the results depicted in the predictiveness curves. However, the interpretation of the *R*-squared value as the proportion of the variance in disease explained by the model is not very intuitive. Interestingly, R^2 can be calculated as a summary index from the predictiveness curve:

$$R^2 = \int_0^1 (\text{pred}(v) - \rho)^2 dv / \rho(1-\rho),$$

where ρ = disease prevalence in the study population, and pred(v) is the value of the risk at the vth percentile. The denominator term in R^2 is a standardization factor leading to values in the range from 0 (useless prediction) to 1 (perfect prediction). We find the display of the predictiveness curve more clinically useful than simply reporting its R^2 summary index.

In our plots, we include a horizontal line located at the risk level equal to the prevalence. This corresponds to the predictiveness curve for a completely uninformative risk model, one that assigns all subjects equal risk. It serves as a reference curve. Moreover, mathematically, the positive area above the horizontal line but below the predictiveness curve must equal the negative area below the horizontal line but above the predictiveness curve. Better markers will show larger positive and negative areas, and we find that the horizontal line is a helpful visual aid.

CLASSIFICATION BASED ON RISK

Clinical decision criteria are often of the form ''marker $\geq$ threshold.'' For example, the criterion ''PSA $\geq$ 4.0 ng/ml'' has been used to recommend biopsy. However, decision criteria might be better formulated in terms of risk. For example, the criterion ''risk $\geq$ 0.20'' could be used to recommend biopsy. Criteria formulated in terms of risk are natural and intuitive. In addition, they are statistically optimal in the sense that they minimize false positive and false negative error rates (8). In particular, with PSA, age, digital rectal examination, and prior biopsy as predictors, existing decision theory based on the Neyman-Pearson lemma (9) states that the best trade-off between true positive fraction (sensitivity) and false positive fraction (1 – specificity) in the population as a whole is achieved with criteria of the form ''risk $\geq$ threshold.''

The performance of decision rules based on a risk model can be calculated from the model's predictiveness curve. We illustrate this in figure 3. For example, the positive predictive value of the criterion ''risk $\geq$ threshold'' is the proportion of the dark area in the shaded rectangle that lies under the curve. The true positive fraction corresponding to this criterion is the same dark area under the curve divided by the prevalence of disease. Although exact calculations will be made directly from the data, approximate calculations can be made by simply viewing the predictiveness curve.

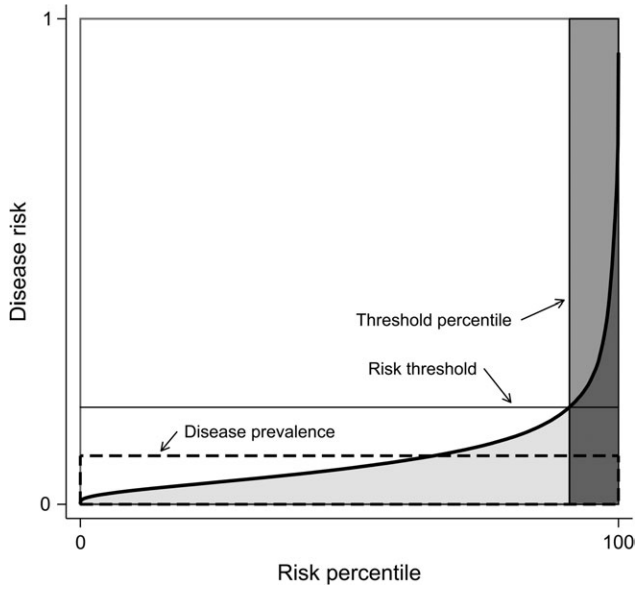

FIGURE 3. Schematic diagram showing how classifier performance parameters relate to the predictiveness curve. Positive predictive value = dark/shade dark + intermediate; negative predictive value = white area/white + light shade; true positive fraction = dark shade/dashed box; false positive fraction = intermediate shade/1 − dashed box.

The plot shown in figure 4A is a comprehensive summary of the population performance of the risk model based on the simulated marker. It allows one to assess decision criteria from multiple points of view. For example, we see that, by recommending biopsy for subjects with estimated risks at or above 0.20, 6.3 percent (95 percent CI: 5.5, 7.2) of the population proceed to biopsy and 60 percent (95 percent CI: 53.2, 65.6) of subjects with high-grade disease are detected, while 3.7 percent (95 percent CI: 3.2, 4.3) of subjects without high-grade disease are unnecessarily biopsied. These calculations do not depend on correctness of the risk model and, indeed, it is possible that, in some applications, one might achieve adequate classification based on estimated risk even if the model suffers from some degree of lack of fit. The choice of risk threshold for classification might be dictated by controlling one or more of the performance measures. Maintaining the false positive fraction at a low level is paramount in primary screening, while a high true positive fraction is often crucial in diagnostic settings. Yet, the corresponding risk threshold will also be an important aspect to consider in order to ensure that decisions are satisfying to individuals. To illustrate, in figure 4B, if we choose the positivity criterion on the basis of a true positive

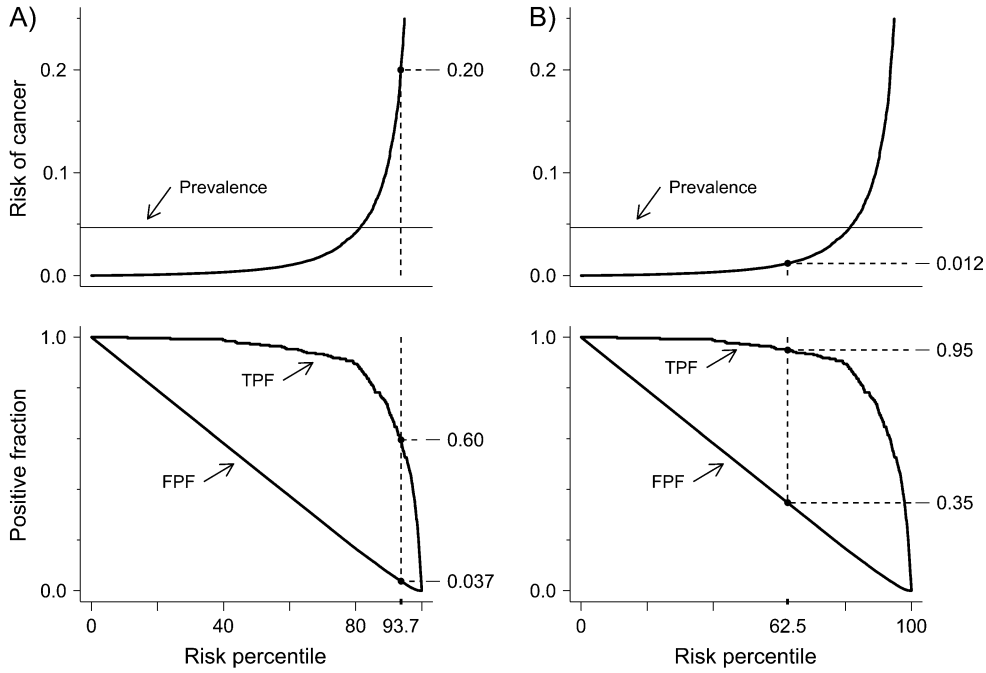

FIGURE 4. The integrated predictiveness and classification plot for the simulated marker using two criteria for defining a positive biomarker result, Prostate Cancer Prevention Trial, 1993–2003. In part A, the criterion is risk $\geq$0.20; in part B, the criterion is true positive fraction (TPF) = 0.95. FPF, false positive fraction.

fraction (TPF) = 0.95, for example, the corresponding risk threshold is 0.012 (95 percent CI: 0.005, 0.026). Sending individuals for biopsy when their risks are less than 2.6 percent may be inappropriate. In addition, we see that the corresponding false positive fraction (FPF) is unacceptably high: FPF = 0.35 (95 percent CI: 0.23, 0.49).

The ROC curve for a risk model plots the TPF associated with a risk threshold criterion versus the corresponding FPF for all possible threshold criteria. Curves are shown in figure 5 for the three risk models considered in figure 2. The problem with the ROC curve is that typically the risk thresholds are not displayed. If one wanted to compare the population performances of the different models using the risk criterion "risk $\geq$ 0.20," for example, this cannot be done using standard ROC curves. One cannot locate on an ROC curve the point that corresponds to this criterion because it displays only TPF and FPF but not risk. We show the points in figure 5 that correspond to the risk threshold at or above 0.20 and note that they are in different horizontal and vertical locations on the three ROC curves. For evaluating risk prediction markers or models, where decision criteria may be based on individual-level risk, we therefore feel that the integrated plot that aligns the models according to risk thresholds offers a more pertinent display than the ROC curve that aligns models according to TPF or FPF. Similar criticisms apply to the Lorenz curve (10, 11) that plots the TPF associated with a risk threshold against the population proportion exceeding that threshold. With note taken that the latter is a weighted average of TPF and FPF,

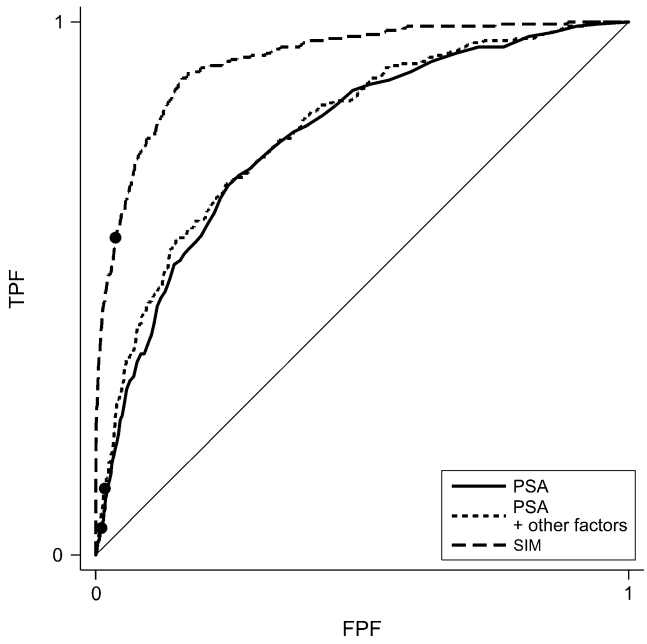

FIGURE 5. Receiver operating characteristic (ROC) curves for risk models based on prostate-specific antigen (PSA) alone, PSA and other risk factors, and the simulated marker (SIM), Prostate Cancer Prevention Trial, 1993–2003. These ROC curves correspond to the predictiveness curves in figure 2. False positive fraction (FPF) and true positive fraction (TPF) points corresponding to the high-risk designation (risk: $\geq$0.20) are displayed for each model.

namely, $\rho \times \text{TPF} + (1 - \rho) \times \text{FPF}$, the Lorenz curve shows information that is equivalent to the TPF versus FPF plot of the ROC curve, so it suffers the same defects. There is no information about absolute risk on the Lorenz curve.

DISCUSSION

The fitting of risk models to biomarker and risk factor data is a valuable exercise. However, the usefulness of the model as applied to the population is rarely evaluated. We suggest for this purpose the predictiveness curve, a display of the risk distribution revealed by the biomarkers and risk factors in the population. A desirable model performs a triage process, placing most individuals at high- or low-risk values, where decisions are more easily made. In developing a biomarker, we need to define a reasonable threshold or range of thresholds for high (and low) risk. The threshold depends on the clinical context and involves weighing the expected costs against benefits associated with a high-risk designation. This may be done with formal decision analysis, or perhaps more often it is done informally. Given such, the predictiveness curve shows the capacity of the marker to identify meaningful variations in risk. By simultaneously displaying predictiveness and classification performance with the integrated plot, we believe that biomarker researchers are better equipped to understand the potential utility of a risk model applied in the population. This practical goal motivated our research.

We noted that the curve can also be helpful in evaluating the fit of a risk model, in the same sense as the standard Hosmer-Lemeshow goodness-of-fit statistic. However, grouping individuals according to values of estimated risks is not equivalent to grouping individuals according to their predictors. Therefore, it has been noted that the Hosmer-Lemeshow approach can miss detecting lack of fit when individuals with different true risks based on their covariate patterns are grouped together by a model that erroneously assigns them similar risk values. Alternative omnibus approaches have been proposed (12–14). In addition, improvements in goodness of fit could be examined in relation to improved classification with the integrated plots. For the model in table 1, investigators compared its fit with models that included interactions and variable transformations and found no evidence of improved fit with these additional terms (3).

The methods presented in this paper are related to methods used informally and occasionally in the literature. For example, Goldman et al. (15) and, more recently, Cook et al. (16), compare risk models by evaluating the numbers of subjects whose risks exceed a therapeutic risk threshold. Predictiveness curves provide for such comparisons across *all possible* risk thresholds. In addition, new statistical techniques will now provide methods for making formal statistical inference (17, 18). We believe that, in addition to calculating the proportion of subjects exceeding a risk threshold, it is important to simultaneously evaluate the numbers of subjects that correctly and incorrectly exceed the threshold (19, 20). The TPF and FPF values in the integrated plot show exactly these entities.

We note that a histogram of population risk values is essentially equivalent to the predictiveness curve, and these have also appeared as informal displays in the literature. However, the histogram has the drawback that it requires defining intervals (or bins) for risk values, and the curve has the advantage that it numerically displays the quantities of interest, that is, the proportions of subjects exceeding risk thresholds.

One must always be cautious to interpret a risk model and its predictiveness curve in the context of the population that gave rise to the data. Strictly speaking, the ''population'' refers to a population for which the available cohort is a representative subsample. As noted by Thompson et al. (3), Prostate Cancer Prevention Trial participants may not reflect the general US population. Subjects in the Prostate Cancer Prevention Trial were participants in a clinical trial. They may differ from the general population because of eligibility criteria, characteristics related to their self-selection for the study, and their care during the course of the study. Therefore, their risk model may not apply with complete fidelity to the general population. We use the data here simply to illustrate statistical methodology and, for that purpose, the data serve well. Nevertheless, they raise questions about risk assessment using research cohorts in general and clinical trial cohorts in particular. Although they may provide a useful starting point for marker evaluation and marker comparison, ultimately risk models should be calculated on cohorts representative of the target population.

The analyses we applied to the Prostate Cancer Prevention Trial data showed that additional risk factors do not add substantially to the predictiveness of PSA alone, in that the fraction of subjects in the equivocal risk range is not appreciably decreased. A risk factor can have a large effect on risk, but if it is rare in the population, it cannot substantially influence population risk prediction. In the Prostate Cancer Prevention Trial, few subjects have risk factor levels that substantially change their risk calculated on the basis of PSA alone. For only 72 subjects did their risk change from <0.2 to ≥ 0.2 and, not surprisingly, a positive digital rectal examination accounted for most of these (92 percent). Nevertheless, the fact that the risk model has limitations on a population level does not mean that it won't contribute in a meaningful way to the biopsy decision-making process of some individuals, which was the specified purpose of developing the risk model (3).

The predictiveness curve is easy to calculate once a risk model has been fitted. We have developed procedures for constructing confidence intervals and for comparing points on two curves under cross-sectional cohort designs (17). Bootstrap techniques were applied in the current paper. For case-control study designs, it is possible to estimate risk from a fitted logistic regression if the disease prevalence is known. Corresponding procedures to estimate the predictiveness curve from case-control data are currently under development. For settings with an outcome variable that is a time to an event, such as disease or death, one can define risk as a function of time, that is, the probability of an event in a time interval $(0, t)$. Predictiveness curves would be plotted for different time intervals.

Use of the same data set to fit a risk model and to assess its performance can lead to optimistic estimates of model performance. This is an issue particularly when many

predictors are involved. Cross-validation or bootstrapping can be applied in these settings to correct for this bias. In our analysis of the Prostate Cancer Prevention Trial data, the model that included other risk factors in addition to PSA showed minimal improvement over PSA alone with uncorrected predictiveness curves, so correcting for optimistic bias was unnecessary. The conclusion about minimal improvement would remain the same.

A bona fide risk calculator must correctly and precisely calculate an individual's risk. Fitting an adequate risk model for individual risk assessment is an ambitious statistical task, more ambitious than estimating an ROC curve, for example. The former is akin to estimating a probability density, while the latter is akin to estimating a cumulative distribution function. Therefore, with small data sets where risk modeling is not feasible, one might proceed to simply evaluating the usual classification performance measures such as the ROC curve. With larger data sets, classification performance measures should also be assessed but perhaps in conjunction with the predictiveness curve in order to integrate the risk modeling and classification approaches to data analysis.

ACKNOWLEDGMENTS

This work is supported in part by grants from the National Institutes of Health (RO1 GM054438, PO1 CA053996, and UO1 CA086368).

Conflict of interest: none declared.

REFERENCES

1. Moons KGM, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Acad Radiol 2003;10:670–2.
2. Pepe MS, Janes H, Longton G, et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 2004;159:882–90.
3. Thompson IM, Ankerst DP, Chi C, et al. Screen-based prostate cancer risk: results from the Prostate Cancer Prevention Trial. J Natl Cancer Inst 2006;98:529–34.
4. Lemeshow S, Hosmer DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 1982;115:92–106.
5. Neter J, Kutner MH, Wasserman W, et al. Applied linear statistical models. Hightstown, NJ: McGraw Hill, 1996.
6. Efron B, Tibshirani RJ. An introduction to the bootstrap. Boca Raton, FL: Chapman & Hall/CRC, 1993.
7. Mittlebock M, Schemper M. Explained variation for logistic regression. Stat Med 1996;15:1987–97.
8. McIntosh M, Pepe MS. Combining several screening tests: optimality of the risk score. Biometrics 2002;58:657–64.
9. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypothesis. Philos Trans R Soc Lond A 1933;231:289–337.
10. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. Biostatistics 2005;6:227–39.
11. Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. J Natl Cancer Inst 2003;95:470–8.
12. Pulkstenis E, Robinson TJ. Two goodness of fit tests for logistic regression models with continuous covariates. Stat Med 2002;21:79–93.
13. Lin DY, Wei LJ, Ying Z. Model-checking techniques based on cumulative residuals. Biometrics 2002;58:1–12.
14. le Cessie S, van Houwelingen C. Testing the fit of a regression model via score tests in random effects models. Biometrics 1995;51:600–14.
15. Goldman L, Cook EF, Mitchell N, et al. Incremental value of the exercise test for diagnosing the presence or absence of coronary artery disease. Circulation 1982;66:945–53.
16. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. Ann Intern Med 2006;145:21–9.
17. Huang Y, Pepe MS, Feng Z. Evaluating the predictiveness of a continuous marker. Biometrics (doi:10.1111/j.1541-0420.2007.00814.x).
18. Bura E, Gastwirth JL. The binary regression quantile plot: assessing the importance of predictors in binary regression visually. Biom J 2001;43:5–21.
19. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med (http://www3.interscience.wiley.com/cgi-bin/fulltext/114278764/PDFSTART).
20. Pepe MS, Janes H, Gu JW. Re: ''Use and misuse of the receiver operating characteristic curve in risk prediction.'' (Letter). Circulation 2007;116:e132.