UW Biostatistics Working Paper Series

6-1-2006

# Integrating the Predictiveness of a Marker with its Performance as a Classifier

Margaret S. Pepe
*University of Washington*, mspepe@u.washington.edu

Ziding Feng
*University of Washington & Fred Hutchinson Cancer Research Center*, zfeng@fhcrc.org

Ying Huang
*University of Washington*, ying@u.washington.edu

Gary M. Longton
*Fred Hutchinson Cancer Research Center*, glongton@fhcrc.org

Ross Prentice
*University of Washington*, rprentic@whi.org

***See next page for additional authors***

**Authors**
Margaret S. Pepe, Ziding Feng, Ying Huang, Gary M. Longton, Ross Prentice, Ian M. Thompson, and Yingye Zheng

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) using, for example, logistic (or Cox) regression. The value of a marker is measured by its effect on risk. The second summarizes marker performance with classification performance measures such as sensitivity and specificity, predictive values and ROC curves. There is controversy about which approach is most 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 subjects detected, i.e., sensitivity, and the fraction of non-diseased subjects falsely identified as diseased, i.e., 1-specificity.

Both statistical approaches are frequently applied, often to the same data. However, the relationship between them is unclear. Of particular concern, the two approaches frequently yield apparently contradictory results. A marker that is strongly related to risk is often a poorly performing classifier. 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 commentary, we present a new graphic, the Predictiveness Curve. It is useful for assessing the value of a risk 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 with data from the Prostate Cancer Prevention Trial (PCPT) recently reported in this journal *(3).* 5519 men on the placebo arm of the study underwent prostate biopsy and had at least 2 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 exam (DRE), age, ethnicity and prior biopsy were used to model the risk of finding prostate cancer and the risk of high grade disease (Gleason ≥ 7) at the time of prostate biopsy. Since 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 5519 men, 4.7% ultimately were found to have high grade disease. Table 1 shows the results of the logistic regression analysis. For the diagnosis of high grade disease, PSA, DRE, age and prior negative biopsy appeared to be predictive of risk.

**The Predictiveness Curve**

We can calculate an individual's risk given data on his risk factors using the fitted risk model. For the prostate cancer example,the calculation *(3)* is

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

where

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

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 risk for each of the individuals in the PCPT study. 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% on the x-axis the risk value is 0.104. This indicates that 90% of subjects in the cohort have risks below 0.104 and only 10% have risks 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, at risk = 0.20 we see that the percent is 97.8%. That is, 2.2% of subjects in the cohort have risks above 0.20. At risk =0.02 the percent is 39.0%, indicating that 39.0% of subjects in the cohort have risks below 0.02.

What does the graph offer that is not summarized in Table 1? It shows the range and distribution of 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 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% risk of high grade disease is sufficiently high to recommend a biopsy and that 2% 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% or be less than 2%. We see from Figure 1 however that the prostate cancer risk model leaves the majority of men, 58.8%, 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% risk of high grade disease may defer further evaluation while a man with a >10% risk would prefer an evaluation, the corresponding  indecisive risk region would contain only 25% 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.

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, family history and DRE. Both models calculate risks less than the 0.02 low risk threshold for 36% and 39% of the population, respectively. At the high risk end of the scale, the PSA model puts 1.2% of subjects above the 0.20 risk level while the more comprehensive model puts 2.2% of subjects in the high risk range. For comparison we also include a simulated marker with much better performance. The simulated marker (SIM) identifies 68.1% of subjects as low risk, 6.0% as high risk and leaves 25.9% with calculated risks in the equivocal (0.02,0.20) range. 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 generalized from linear to logistic regression *(4).* The values 0.053, 0.066 and 0.310 for PSA alone, for PSA and other factors and for SIM, 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. We have recently provided a more understandable interpretation as the difference in the average risks between diseased and non-diseased subjects *(5).* 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 $\rho$ = disease prevalence in the study population and Pred$(v)$ is the value of the risk at the $v^{th}$ percentile. The denominator term in $R^2$ is a standardization factor leading to values in the range 0 (useless prediction) to 1 (perfect prediction). We find the display of the Predictiveness curve more 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 the horizontal line a helpful visual aid.

## Classification Based on Risk

Clinical decision criteria are often of the form “marker ≥ threshold.” For example, the criterion “PSA ≥ 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 ≥ 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, a notion defined precisely in *(6).*

The performance of decision rules based on a risk model can be calculated from the model's Predictivness Curve. We illustrate in Figure 3. For example, the positive predictive value of the criterion "risk > 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.

## An Integrated Approach

The plot shown in Figure 4a is a comprehensive summary of the population performance of the risk model based on SIM, 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 risks above 0.20, 6% of the population proceed to biopsy, 57% of subjects with high-grade disease are detected, while 3.4% of subjects without high-grade disease are unnecessarily biopsied. The choice of threshold might be dictated by controlling one or more of the performance measures. Maintaining the false positive fraction (1-specificity) at a low level is paramount in primary screening, while a high true positive fraction (sensitivity) 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 fraction (TPF) =0.95 say, the corresponding risk threshold is 0.013. Sending individuals for biopsy when their risks are as low as 0.013 may be inappropriate. In addition we see that the corresponding false positive fraction is unacceptably high, FPF=37%.

## 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 reasonable thresholds for high and low risk, thresholds that depend on the clinical context. The Predictiveness Curve then 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.

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 representative subsample. As noted by Thompson et al *(3),* PCPT participants may not reflect the general U.S. population. Subjects in the PCPT study 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 it serves well. Nevertheless it raises 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 population.

The analyses we applied to the PCPT 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 PCPT, 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 DRE accounted for most of these (92%). 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 *(5).* Bootstrap techniques can also be applied. 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, i.e., the probability of an event in a time interval (0,t). Predictiveness curves would be plotted for different time intervals.

Using the same dataset 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 PCPT 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 bias was unnecessary.

We note that fitting an adequate risk model 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 datasets where risk modeling is not feasible one might proceed to simply evaluating the usual classification performance measures such as the ROC curve. With larger datasets, 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.

## REFERENCES

*(1)* Moons KGM, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Academ Radiol 2003;10:670–672.

*(2)* Pepe MS, Janes H, Longton G, Wendy Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidem 2004;159:882–890.

*(3)* Thompson IM, Pauler Ankerst D, Chi C, Goodman P, Tangen C, Lucia MS, Feng Z, Coltman CA. Screen-based prostate cancer risk: results from the prostate cancer prevention trial. J Natl Ca Inst 2006; 98: 529-534.

*(4)* Mittlebock M, Schemper M. Explained variation for logistic regression. Stat Med 1996;15:1987–1997.

*(5)* Ying Huang, Margaret S. Pepe, and Ziding Feng, Evaluating the Predictiveness of a Continuous Marker. UW Biostatistics Working Paper Series. Working Paper 282. http://www.bepress.com/uwbiostat/paper282.

*(6)* McIntosh M, Pepe MS. Combining several screening tests: optimality of the risk score. Biometrics 2002;58:657–664.

## **Captions**.

**Table 1.** Logistic regression analysis of risk for high grade disease  as reported in Thompson et al *(3)*.

**Figure 1.** Predictiveness curve for the risk model shown in Table 1 that includes PSA, age, prior biopsy, family history and DRE as risk factors for high grade prostate cancer.

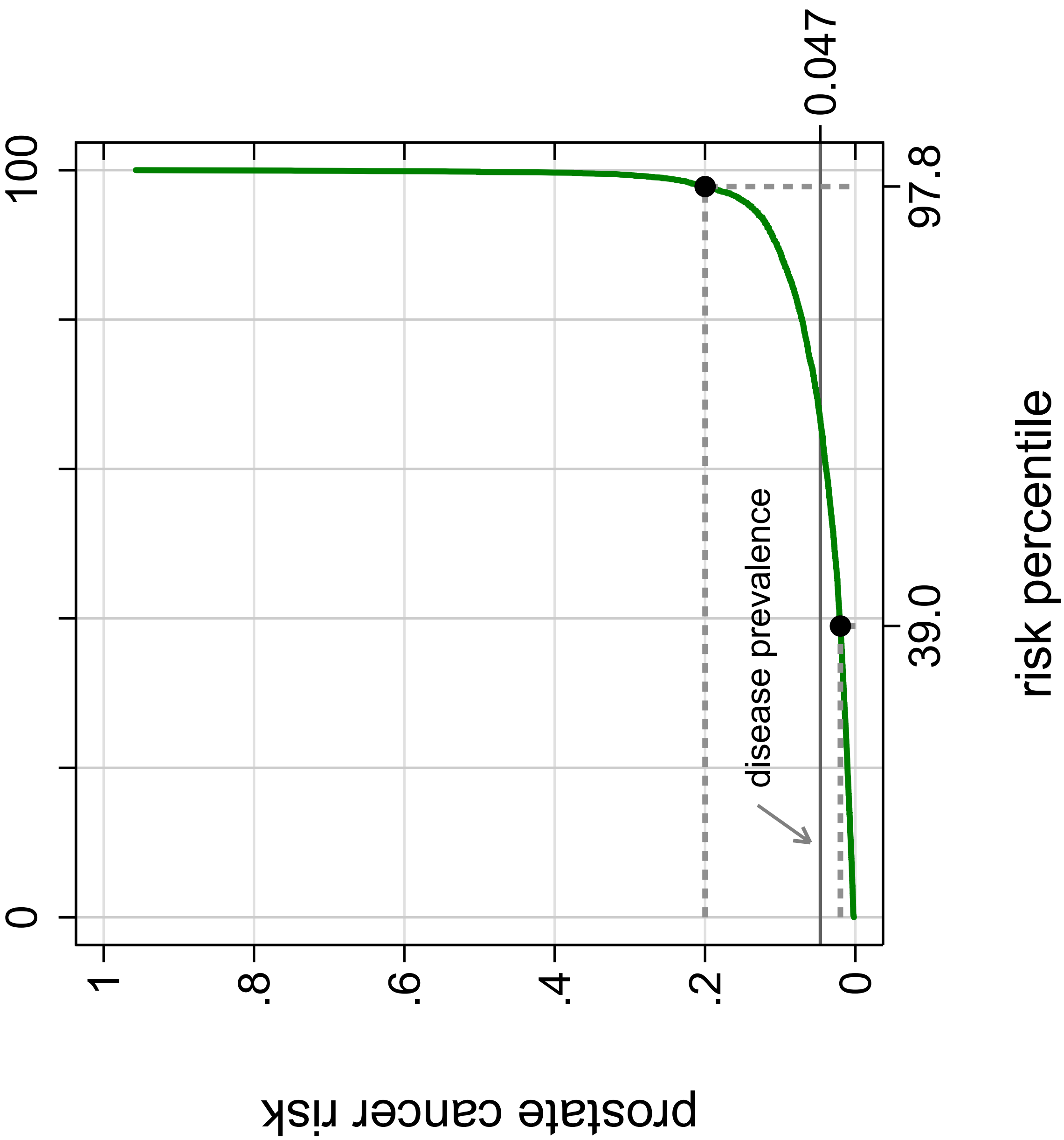
**Figure 2.** Predictiveness Curves for PSA alone, PSA and other factors and the simulated (SIM) marker.

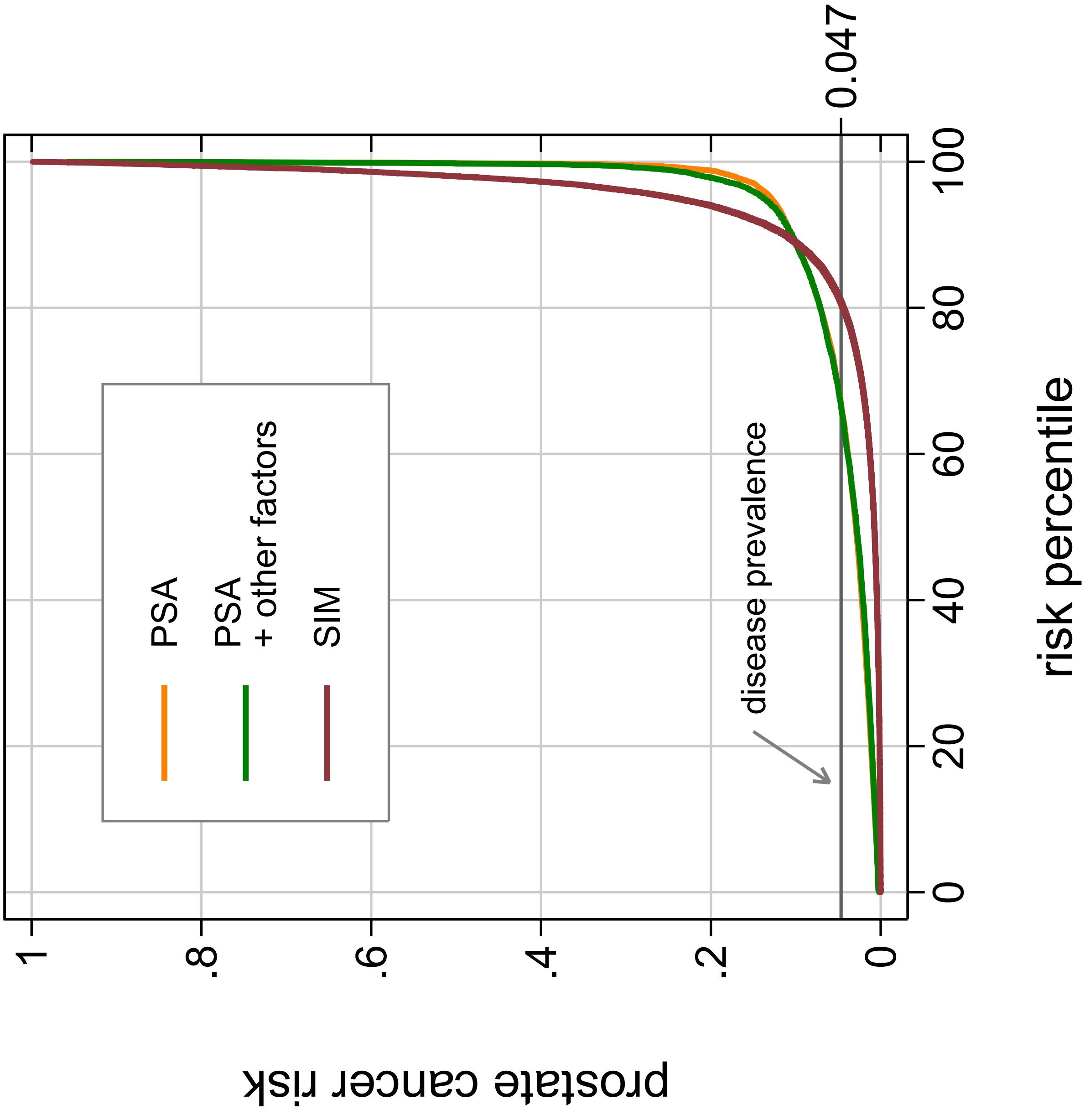
**Figure 3.** Schematic diagram showing how classifier performance parameters relate to the Predictiveness Curve. Positive Predictive Value=dark/shade dark + intermediate; Negative Predictive Value = white area/white + light shade; True Positive Fraction = dark shade/dashed box; False Positive Fraction=intermediate shade/1-dashed box.

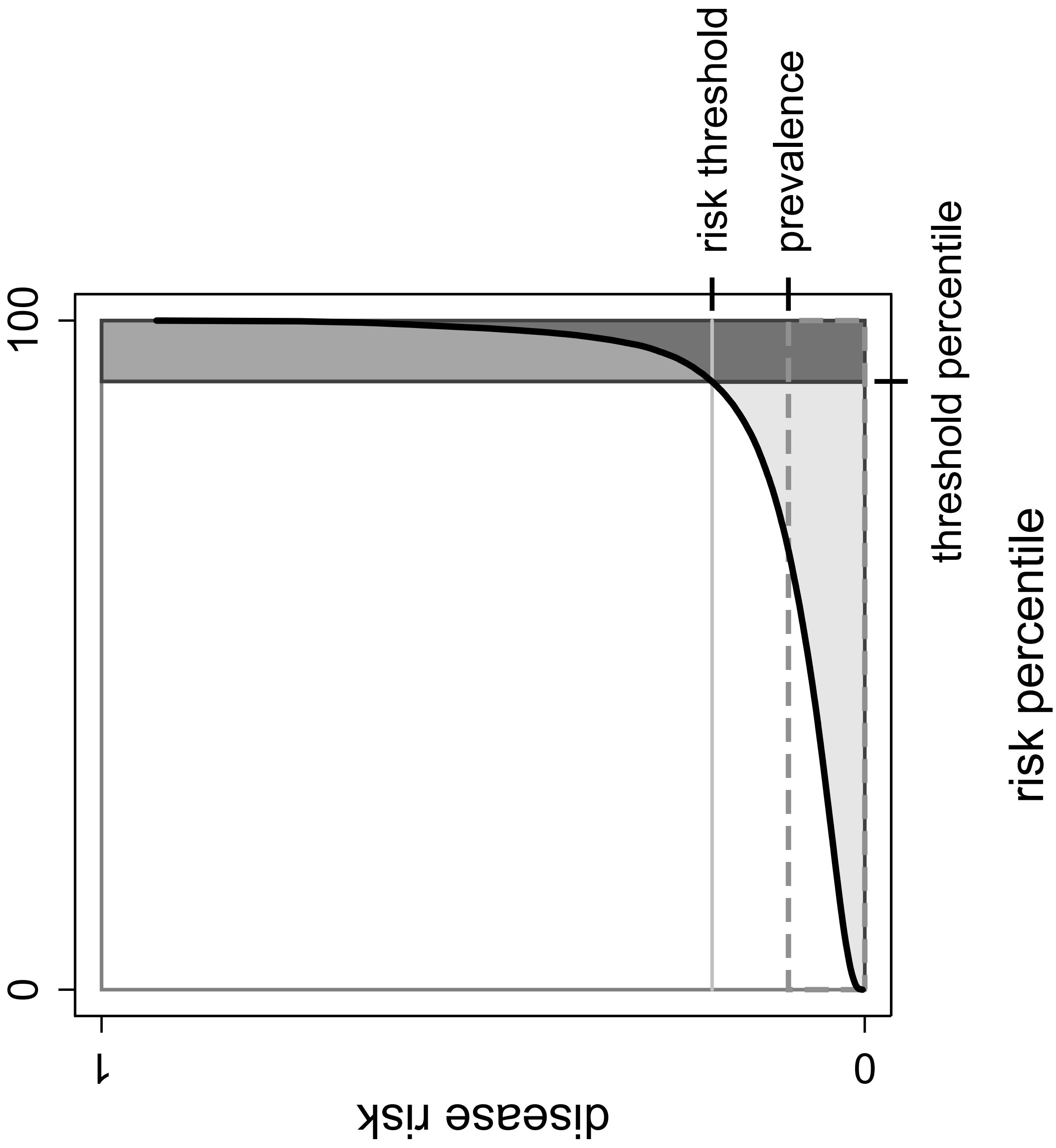
**Figure 4.** The integrated predictiveness and classification plot for the simulated marker using two criteria for defining a positive biomarker result. Criterion (a) is risk >0.20. Criterion (b) is TPF=0.95.

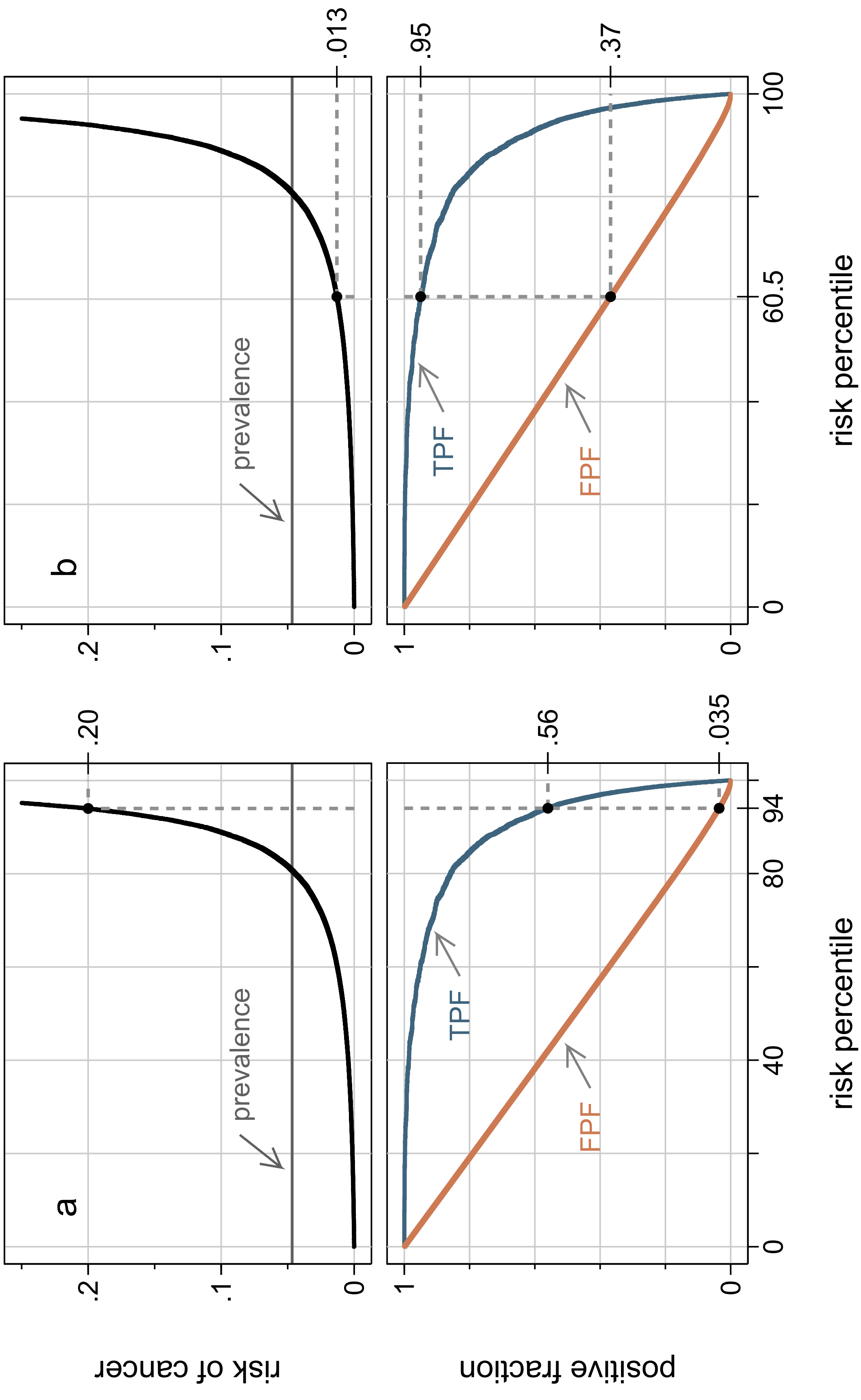
Table 1

| Factor | Log Odds Ratio | *P*-value |
|---|---|---|
| logPSA | 1.30 | <0.001 |
| Age (years) | 0.03 | 0.02 |
| DRE | 0.99 | <0.001 |
| Prior biopsy | –0.37 | 0.04 |
| Constant | –5.94 | — |


prostate cancer risk
1
.8
.6
.4
.2
0
risk percentile
0
100
39.0
97.8
0.047
disease prevalence


0.047
PSA
PSA
+ other factors
SIM
disease prevalence
prostate cancer risk
0
.2
.4
.6
.8
1
risk percentile
0
20
40
60
80
100


risk threshold
prevalence
threshold percentile
risk percentile
100
0
1
0
disease risk


a
b
risk of cancer
positive fraction
prevalence
prevalence
.20
.013
.56
.035
.95
.37
.2
.1
0
1
0
TPF
FPF
TPF
FPF
0
40
80
94
0
60.5
100
risk percentile
risk percentile