Research Article

Cancer Epidemiology, Biomarkers & Prevention

The Potential of Genes and Other Markers to Inform about Risk

Margaret S. Pepe[1,2], Jessie W. Gu[1,2], and Daryl E. Morris[1,2]

Abstract

Background: Advances in biotechnology have raised expectations that biomarkers, including genetic profiles, will yield information to accurately predict outcomes for individuals. However, results to date have been disappointing. In addition, statistical methods to quantify the predictive information in markers have not been standardized.

Methods: We discuss statistical techniques to summarize predictive information, including risk distribution curves and measures derived from them, that relate to decision making. Attributes of these measures are contrasted with alternatives such as receiver operating characteristic curves, R^2, percent reclassification, and net reclassification index. Data are generated from simple models of risk conferred by genetic profiles for individuals in a population. Statistical techniques are illustrated, and the risk prediction capacities of different risk models are quantified.

Results: Risk distribution curves are most informative and relevant to clinical practice. They show proportions of subjects classified into clinically relevant risk categories. In a population in which 10% have the outcome event and subjects are categorized as high risk if their risk exceeds 20%, we identified some settings where more than half of those destined to have an event were classified as high risk by the risk model. Either 150 genes each with odds ratio of 1.5 or 250 genes each with odds ratio of 1.25 were required when the minor allele frequencies are 10%. We show that conclusions based on receiver operating characteristic curves may not be the same as conclusions based on risk distribution curves.

Conclusions: Many highly predictive genes will be required to identify substantial numbers of subjects at high risk. *Cancer Epidemiol Biomarkers Prev; 19(3); 655–65. ©2010 AACR.*

Background

Predicting risk is a natural part of human life. In the context of cancer research, we seek markers that can predict the risk of developing cancer, predict the chance of responding to treatment for cancer, predict the risk of recurrence after treatment for cancer, and so forth. We have greeted advances in genomic, proteomic, and imaging technologies with enthusiasm in part because of their potential to help predict outcomes for individuals. The first goal of this article is to explore the extent to which we can we expect genes or other factors to be predictive of individual risk.

Statistical measures used to quantify the predictive information in a marker are often difficult to understand and not directly relevant to clinical practice. For example, the AUC [area under the receiver operating characteristic (ROC) curve] has been used to quantify the potential of newly discovered single nucleotide polymorphisms and genes to improve risk prediction (1, 2). However, there is no direct relationship between increments in the AUC and clinically meaningful improvements in risk prediction. A second goal of this article is to give guidance on clinically relevant ways to measure the predictive information provided by a genetic profile or other risk predictor.

Authors' Affiliations: [1]Biostatistics and Biomathematics Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center and [2]Department of Biostatistics, School of Public Health, University of Washington, Seattle, Washington

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Corresponding Author: Margaret S. Pepe, Biostatistics and Biomathematics Program, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-B500, Seattle, WA 98105. Phone: 206-667-7398; Fax: 206-667-7004. E-mail: mspepe@u.washington.edu

doi: 10.1158/1055-9965.EPI-09-0510

Measures that Quantify Predictive Capacity of a Marker

Context

Suppose two risk prediction calculators have been developed for predicting an outcome event, such as contracting cancer or dying from cancer within a specified period, using different sets of genes and possibly other risk factors. For each subject in the data set, his risk of a bad outcome can be calculated using both models, model A and model B. For example, according to the values of genes in model A, the risk of an event for a subject may be calculated as 30%. On the other hand, when

the information on genes in model B is used to calculate risk, his risk is calculated to be 50%. The topic of this section concerns how to quantify and compare the predictive capacities of models A and B. In other words, in this population, which set of genes does the better job of predicting risk.

We use a very large simulated data set to illustrate statistical approaches to quantifying predictive information. On purpose, we do not provide details of the data here to focus on the statistical method. Details are provided in Appendix A (see online supplementary data). Overall, 10% of subjects have an event and both risk calculators are "correct" in the following sense: the calculated risk value for a subject with genetic profile y_A, in which y_A are the genes in model A, reflects the proportion of events among all individuals who have genetic profile y_A and similarly for model B. In standard statistical terminology, this means that both models are well calibrated. The issue is that one set of genes may be more informative than the other. Note that a risk model with completely uninformative genes would "correctly" assign risk equal to 10% to everyone because uninformative markers tell us nothing about individual risk. We now discuss various ways to describe the predictive information provided by models A and B.

Risk Distributions

Figure 1 (top) shows the population distributions of risk calculated according to the models. Displaying risk distributions is a fundamental step in evaluating the performance of a risk prediction model (3, 4), a step that is often overlooked in practice. We can see from the risk distributions the proportions of subjects identified as high risk or as low risk according to the risk models. For example, suppose we want to select for preventive intervention or treatment subjects at high risk for the event in which risk levels at or above 20% are considered high. Only 1% of the population is identified as high risk according to model A, whereas 10% are identified as high risk according to model B. In this sense, model B is better at identifying high-risk subjects. Model B would be more useful as a screening tool for selecting patients to a clinical trial. The curves would also be of interest to individuals who are deciding whether to have their genetic information measured. Suppose an individual will opt for an intervention only if his risk is >20%. In the absence of genetic information, his risk is calculated as 10% and he declines intervention. There is only a 1% chance that his decision about intervention will change if he ascertains the genetic information required for model A, but a 10% chance for model B. That is, the information in model A is

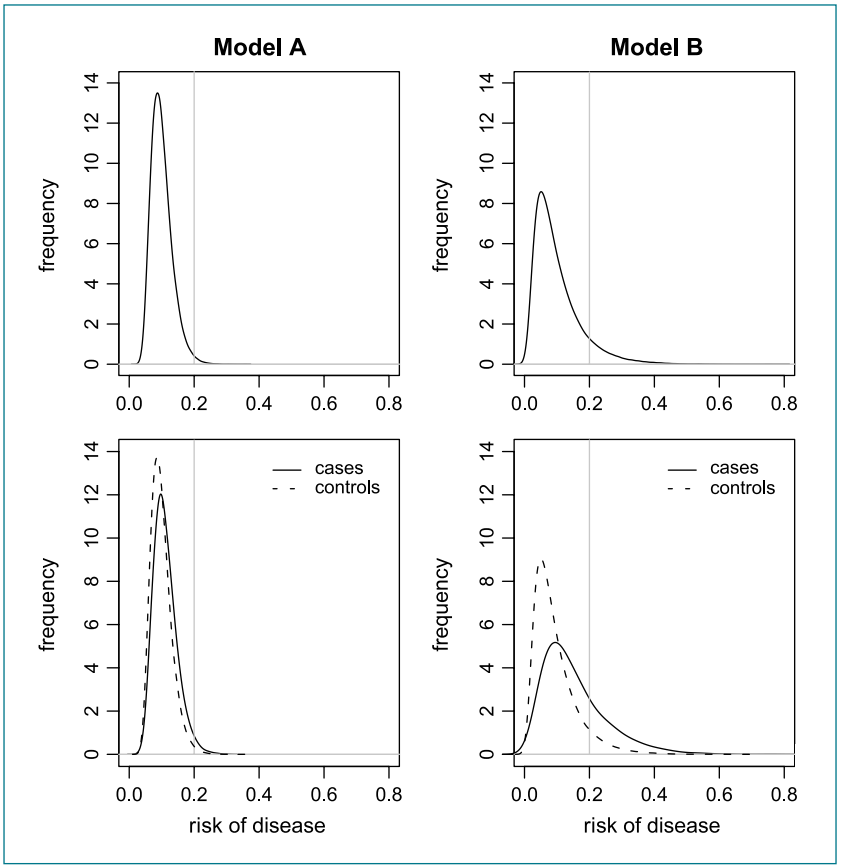

Figure 1. Distributions of risks for two risk models involving different sets of genes, models A and B. Distributions are displayed using probability density functions. Probability densities are shown for the population as a whole in the top panel and separately in the bottom panel for cases, those destined to have events and controls, and those destined not to have events. Vertical lines are displayed at the high-risk threshold, $risk_H$, equal to 0.2. Subjects whose risks exceed 20% are regarded as being at high risk.

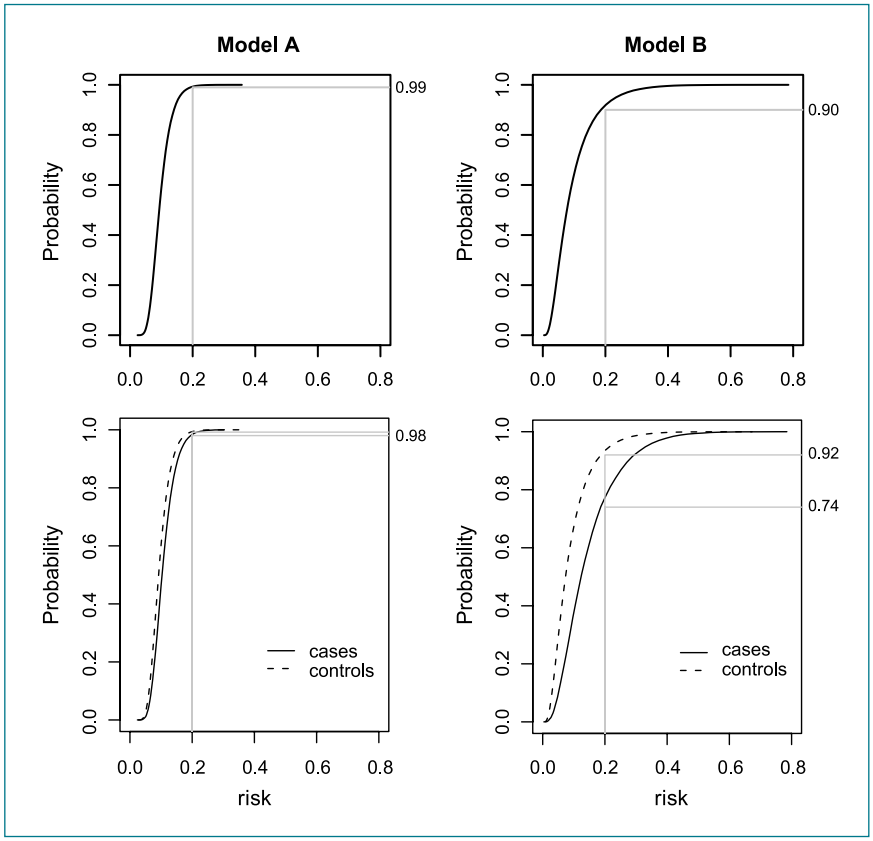

Figure 2. Distributions of risks for models A and B are displayed with cumulative distribution function curves. The high-risk category is defined as risk exceeding 20%. Approximately 10% of the population is classified as high risk with model B but only 1% with model A. Approximately 26% of cases (i.e., those that had an event in the absence of intervention) are classified as high risk with model B but only 2% with model A. Of control subjects (i.e., those who did not have an event), 8% and 1% are classified as high risk according to models B and A, respectively.

unlikely to alter medical decisions, whereas the information in model B is more likely to have an effect.

The information displayed in Fig. 1 (top) is also displayed in Fig. 2 (top) but using cumulative distribution curves rather than probability density curves. The cumulative distribution curves are more useful because one can directly read from them the proportions of subjects whose risk values lie below or above a threshold of interest.

In considering the effect of risk models on individual decision making, one should consider the costs and benefits consequent to those decisions. Subjects who would have an event in the absence of intervention, whom we call cases, will benefit on average from high-risk designation because they will receive the potentially beneficial intervention. On the other hand, subjects who would not have an event in the absence of intervention, whom we call control subjects, will not benefit from intervention but will only suffer its negative effects, including monetary costs. Figures 1 and 2 (bottom) display risk distributions separately for case subjects and for control subjects. These displays are also important and useful in evaluating risk prediction models. One observes, for example, that the proportions of case subjects placed in the high-risk category are 26% with use of model B but only 2% with use of model A. From this, we conclude that there is more benefit to be gained with use of model B. However, we also see that, unfortunately, 8% of controls are designated as high risk with model B, which is substantially more than the corresponding proportion, 1%, for model A. In this sense, there is more cost associated with model B as well as more benefit. An alternative but equivalent way to display the information in Figs. 1 and 2 was previously described (4) and is displayed in Fig. 3.

The ROC curve is derived from the case and control risk distributions. It plots the proportion of case risk values exceeding a threshold, $TPR(risk_H)$, versus the proportion of control risk values exceeding that same threshold, $FPR(risk_H)$, in which $risk_H$ is the risk threshold (see Fig. 4). Unfortunately, the risk thresholds themselves are not visible from the ROC curve so one cannot see the correspondence between risk threshold and case and control proportions that one can see from the risk distributions in Figs. 2 and 3. Moreover, from the ROC curves alone, one cannot compare two risk models with regard to total, case, or control population proportions that exceed a risk threshold of interest. Therefore, we suggest displaying the risk distributions because they are more informative than the ROC curves.

Standard Statistical Summary Indices

A single number is often used to summarize the predictive performance of a risk prediction model. The AUC is the most popular statistical index. However, it has been criticized (6) and recent criticisms in the

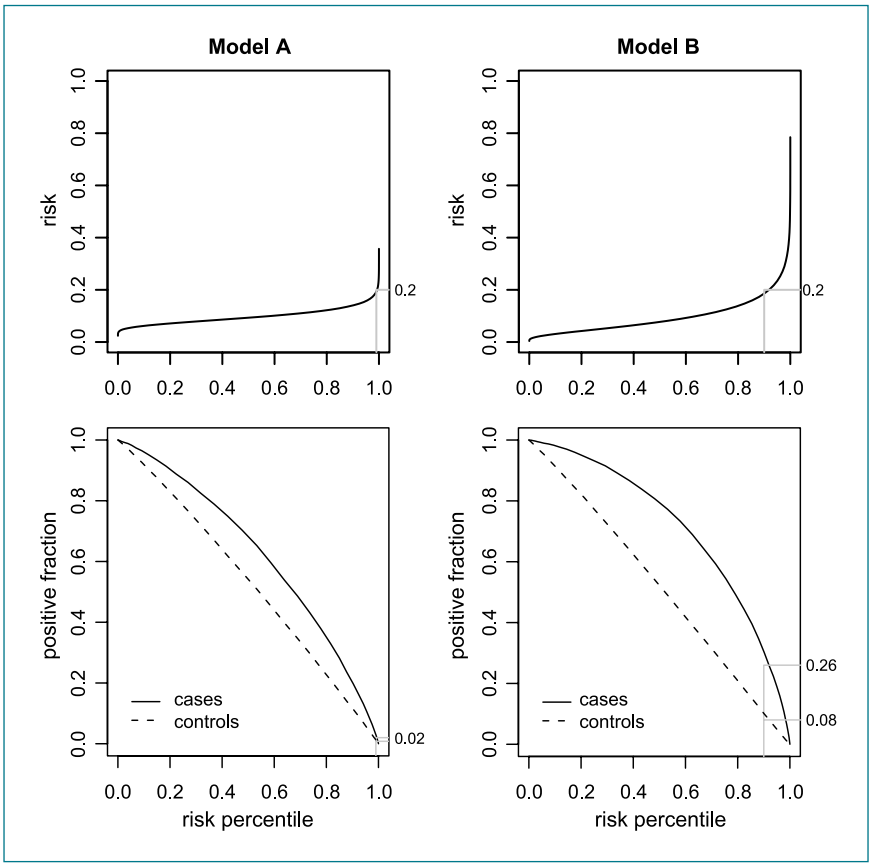

Figure 3. An alternative way to display the risk distributions shown in Figs. 1 and 2 is with the integrated plot (4). The integrated plot displays the risk quantiles in the top panels together with true- and false-positive rates associated with corresponding risk thresholds in the bottom panels. Values corresponding to the risk threshold $risk_H = 0.20$ are shown.

cardiovascular literature (7) have led to much debate about alternative approaches to summarizing predictive performance (3, 8–10). The AUC is

AUC = probability (the risk value of a random case > the risk value of a random control).

This entity is not of interest in practice because the practical problem is not to determine the case and control identities in a random case-control pair. Rather, the problem is to correctly flag subjects at high risk of an event. AUC values of 0.6 for model A and 0.7 for model B (Table 1) suggest that model B is superior but do not quantify their predictive capacities in a clinically useful way.

Most standard statistical indices of predictive performance summarize the difference between case and control risk distributions. The AUC is the Mann-Whitney-Wilcoxon statistic for testing for differences between these two risk distributions. The R^2 statistic, which is familiar from linear regression of continuous outcomes, generalizes to dichotomous event outcomes as

$$R^2 = \text{mean risk in cases} - \text{mean risk in controls}.$$

Again, this entity, the difference in mean risks, does not relate directly to the task of identifying subjects at high risk. The average risk for cases – average risk for controls is 0.11 – 0.10 = 0.01 for model A and 0.15 – 0.09 = 0.06 for model B. Interestingly, this version of R^2 is the same (11) as the integrated discrimination improvement statistic recently proposed by Pencina et al. (8) as an alternative to the AUC. However, it does not solve the main problem with AUC, namely the lack of clinical usefulness. Other versions of R^2 that average functions of the risk values (12) lack easy interpretation as well as practical relevance.

Because the AUC and R^2 do not represent quantities of clinical relevance, what should their roles be in evaluating the population performance of risk prediction models? We recommend that they be de-emphasized in reporting study results. Rather than focusing on these single numerical summaries, the risk distributions themselves should be given greater prominence. Using a few risk thresholds of interest, one should report proportions of the case, control, and overall populations that have risk values exceeding those thresholds. When no specific risk thresholds or risk ranges are of interest, one could complement the risk distribution displays with AUC or R^2 summary statistics and compare risk models by basing hypothesis tests on them (13).

New Summary Indices Using Risk Categories

Two new indices, the reclassification percent and the net reclassification improvement (NRI), have been

proposed recently and are both based on the idea of categorizing risk. When two risk categories are defined, high versus low, based on a single risk threshold value, these statistics are directly related to the population proportions at high risk discussed earlier. We use our example with risk threshold equal to 20% to illustrate.

Consider first comparing a model, for example, model B, with no model. The reclassification percent (7) is the proportion of the whole population that is classified as high risk by the model (10%) because in the absence of genetic data, all subjects are classified as low risk by assigning them risk values equal to the population prevalence $\rho = 10\%$. The NRI index proposed by Pencina et al. (8) is the difference between the proportions of cases and controls classified as high risk (TPR − FPR = 18%) in the notation used above. We recommend reporting the two components, TPR = 26% and FPR = 8%, separately because it is more informative than just reporting the difference and offers the flexibility to weight differently the benefits and costs of high-risk designations for cases and controls, respectively.

Now consider NRI and reclassification percent when comparing two models, model B versus model A. In this case, NRI = $(TPR_B - TPR_A) - (FPR_B - FPR_A) = 17\%$. Again, reporting the components, namely the change in TPR [$(TPR_B - TPR_A) = 26\% - 2\% = 24\%$] and the change in FPR [$(FPR_B - FPR_A) = 8\% - 1\% = 7\%$], seems much more informative than reporting the single composite 17% number. The reclassification percent for comparing models A and B is the proportion of subjects who are classified in different risk categories according to the two models. This measure can be large or small even if the two models have exactly the same performance because it is a function of the correlation between the genes in the two models. Therefore, it has been argued (10) that reclassification percent is not well suited to the task of comparing models.

The reclassification percent and the NRI are both defined for settings involving more than two risk categories but suffer additional weaknesses in these settings. They do not distinguish between small and large changes in risk. Moreover, they are highly dependent on the number and nature of risk categories chosen. In Table 1, using

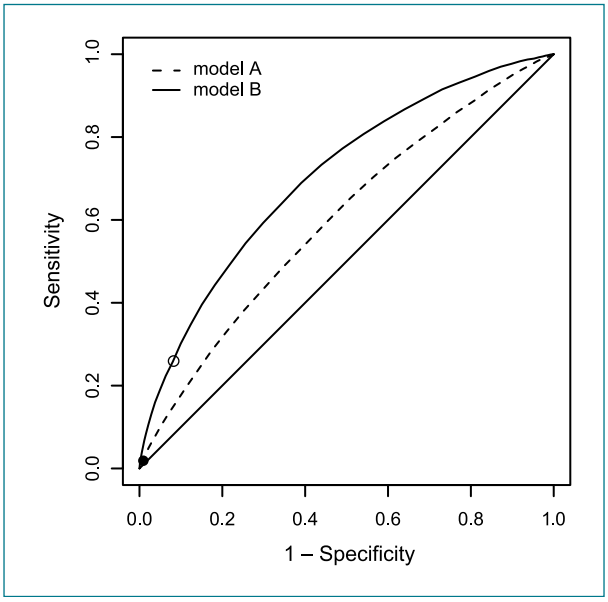

Figure 4. ROC curves for risk values calculated according to models A and B. The solid and filled circles are the true- and false-positive rates corresponding to the high-risk threshold of $risk_H = 20\%$. Sensitivity = TPR and FPR = 1 − specificity. The AUCs are 0.60 for model A and 0.70 for model B.

Table 1. Statistical measures summarizing and comparing models A and B

	Model A	Model B
No risk categories		
AUC	0.601	0.705
R^2	0.012	0.058
Two risk categories (high and low)		
Proportion of subjects classified as high risk		
All subjects	0.009	0.098
Case subjects	0.020	0.258
Control subjects	0.008	0.081
Reclassification percent	0.9%	9.8%
Net reclassification index	0.012	0.177
Four risk categories		
Proportion of subjects classified as high risk		
All subjects	0.009	0.098
Case subjects	0.020	0.258
Control subjects	0.008	0.081
Proportion of subjects classified as medium-high risk		
All subjects	0.403	0.275
Case subjects	0.519	0.382
Control subjects	0.390	0.263
Proportion of subjects classified as medium-low risk		
All subjects	0.569	0.376
Case subjects	0.453	0.275
Control subjects	0.582	0.387
Proportion of subjects classified as low risk		
All subjects	0.019	0.251
Case subjects	0.009	0.085
Control subjects	0.020	0.269
Reclassification percent	59.7%	72.5%
Net reclassification index	0.15	0.47

NOTE: The two risk categories, high and low risk, are defined as risk above or below 20%. The four risk categories are defined by risk thresholds at 5%, 10%, and 20%. Overall, 10% of subjects in the population have an event. Reclassification percent and net reclassification index compare each of models A and B with no model.

four risk categories defined by the risk thresholds 0.05, 0.10, and 0.20, we report values for reclassification percent of 60% when comparing model A with no model and 73% when comparing model B with no model. Corresponding NRI values are 0.15 and 0.47, respectively. Observe the dramatic changes in values by use of four versus two risk categories. In our opinion, reclassification percent and NRI are not compelling measures of predictive performance. More informative than the single number summaries that accumulate data over multiple risk categories are the proportions of the total, case, and control populations whose risks are in each of the four risk categories (Table 1). These values can be read from the risk distribution curves in Fig. 2 as well. The curves have the advantage that the reader can specify his own risk categories of interest.

Cost-Benefit Analysis for Decision Making

A subject will need to consider several factors simultaneously in evaluating the potential benefit associated with using a risk model. These factors are his overall chance of having an event in the absence of intervention, denoted by ρ, and equal to 10% in our example; the chance that the model will assign him a high-risk status if he is destined to be a case in the absence of treatment, denoted by TPR; and the chance that the model will assign him a high-risk status if he is destined to be a control in the absence of treatment, denoted by FPR. These values are TPR = 26% and FPR = 8%, respectively, for model B in our example (Fig. 2). Finally, he will need to consider the relative values of the potential benefit and of the potential cost associated with high-risk designation. Interestingly, Vickers and Elkin (5) and several other articles drawing on results from decision theory note that use of a high-risk threshold, risk_H, is equivalent to considering the cost-benefit ratio to be cost/benefit = $\text{risk}_H/(1 - \text{risk}_H)$. In our example, the risk threshold is risk_H = 20%, which is equivalent to a cost-benefit ratio of 20%/(100% − 20%) = 0.25. That is, the use of the 20% risk threshold implies that the net benefit associated with intervention for a subject who would have an event in the absence of intervention is considered four times the net cost of intervention to subjects who would not have an event.

Formally, the expected benefit associated with the use of the risk model and assigning high-risk status to those with risks exceeding risk_H is calculated as:

$$\begin{aligned}\text{expected benefit } &= \rho \times \text{Benefit} \times \text{TPR} - (1-\rho) \times \text{Cost} \times \text{FPR} \\ &= \text{Benefit} \times \{\rho \times \text{TPR} - (1-\rho) \times (\text{risk}_H/(1-\text{risk}_H)) \times \text{FPR}\},\end{aligned}$$

in which "Benefit" denotes the benefit associated with a case being designated as high risk; this is the unit in which benefit is measured. In our example, with risk threshold risk_H = 20%, the expected benefit associated with use of model B is (0.1 × 0.258 − 0.9 × 0.25 × 0.081) = 0.0076. That is, the expected benefit is positive and equal to 0.0076 times the benefit associated with a case being designated as high risk. For model A, the expected benefit is nearly zero (0.0002). Figure 5 displays the expected benefit for the two models using risk thresholds ranging from 0 to 0.4. We see that no matter what risk threshold is used, or equivalently no matter what cost-benefit ratio is entertained, model B yields more expected benefit. The calculations presented here ignore the costs associated with obtaining the information needed to calculate the modeled risks. Therefore, the expected benefit associated with any model is at least as good as not using any model (expected benefit $\geq$0). The expected benefit could be negative if the cost of genetic testing were taken into consideration.

The decision curves (5) shown in Fig. 5 are useful in deciding whether to obtain genetic information for an individual who has in mind a specific risk threshold that would lead to an action. Risk thresholds may vary from individual to individual so the expected benefit for one individual may not be the same as that expected for another. Higher expected benefits correspond to lower risk thresholds in Fig. 5 because subjects with low-risk thresholds perceive low cost compared with benefit. To summarize the expected benefit of applying a risk model across the population, one needs to integrate with the decision curve the probability distribution of risk thresholds likely to be used in practice. For example, suppose that 50% of individuals in the target population use a risk threshold equal to 0.20, but that 25% use the lower threshold of 0.10 and 25% use the larger threshold of 0.30. The average benefit in the population is then calculated as the weighted average of expected benefits associated with each of the three thresholds: for model A, the average benefit is 0.003, whereas for model B, the average benefit is 0.011.

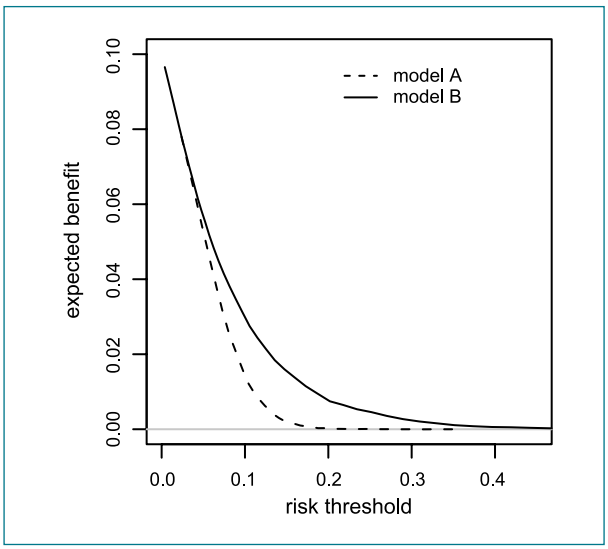

Figure 5. Decision curves for models A and B. The expected benefit of using the risk model is calculated for each risk threshold value, risk_H, where the ratio of the cost of high-risk designation for a control to the benefit of high-risk designation for a case is $\text{risk}_H/(1 - \text{risk}_H)$. The unit for expected benefit is the benefit of high-risk designation for a case.

In conclusion, we promote displays of case and control risk distribution curves (Figs. 1, 2, or 3) in conjunction with decision curves (Fig. 5) for evaluating and comparing risk prediction models. The case- and control-specific risk distribution curves that display the TPR and FPR values associated with the subject's risk threshold are easy to understand and are key to his decision about ascertaining his genetic profile and other risk factors in the risk prediction model. Decision curves provide additional insights by formalizing the cost-benefit analysis. Because subjects vary in their tolerance for risk, having the distributions displayed is convenient because it allows the use of various risk threshold values. The overall benefit associated with the use of a risk model in the population can only be summarized into a single meaningful number if one specifies a population distribution for risk thresholds used by individuals.

Predictive Potential of Genes

Scenarios Evaluated

Janssens et al. (2) evaluated the predictive potential of genetic profiling by simulating a wide variety of scenarios. We investigate the same scenarios as Janssens did. The illustrative data set shown in Figs. 1 to 3 was simulated using two such scenarios. Our simulation program is publicly available so that an investigator can simulate specific scenarios of interest for themselves. The use of the program is described in Appendix A.

A scenario is specified by the number of subjects, the overall event rate in the population, ρ, the number of genes that confer risk, the allele frequencies for the genes, and the association of each allele with risk. We consider simple settings in which each gene has two alleles, with genotypes and allele frequencies in the Hardy-Weinberg equilibrium, and no linkage disequilibrium between genes. The true risk of an event for a subject is derived from a standard additive model on the logistic scale. That is, the logarithm of the odds of having an event is a sum of terms associated with each high-risk allele—high-risk homozygotes contributing two equal terms, one for each allele—and there are no statistical interactions between genes. The lower frequency allele for each gene is associated with higher risk. The magnitude of the association between a gene and risk is quantified by the odds ratio (OR) for the high-risk allele: OR = odds of an event for heterozygotes/odds of an event for homozygotes with the dominant lower-risk allele variant. Details of how data are simulated are given in Appendix A.

Very large sample sizes were used in our simulation studies. Consequently, the results in Tables 1, 2, and 3 show the true values (precise to two decimal places) of the prediction performance measures for each risk model, not estimates. We evaluate predictive performance by focusing on the proportions of high-risk subjects identified from the information in their genetic profiles and expected benefit. We use the high-risk threshold equal to 20% for illustration. Tables for other risk thresholds and other scenarios are provided in Appendix A. In contrast to our approach, Janssens (2) reported AUCs and R^2 summary statistics. These are provided here as well for completeness. In addition to generating data, investigators can use our programs to calculate all of the summary indices shown in Tables 2 and 3 after specifying a risk threshold that defines the high-risk (or low-risk) category.

Results for Equally Predictive Genes

In the first set of simulations (Table 2), all genes in a gene profile have the same minor allele frequency and are equally predictive. We investigated settings in which the number of genes associated with risk ranged from 50 to 350, the frequency of the minor allele varied from 5% to 30%, and the OR associated with the heterozygous genotype ranged from 1.05 to 1.5. Subjects whose risks are 20% or more are considered at high risk. This contrasts with the overall event rate of 10%.

The proportion of high-risk subjects identified is generally low in the scenarios we studied. The maximum value for the proportion of high-risk subjects identified was ~17%. For example, when the gene profile consists of 350 predictive genes each with a minor allele frequency of 5% and OR equal to 1.5, 17% of the population have calculated risk values exceeding 20%.

The high-risk population proportion typically increases with larger numbers of predictive genes, with stronger associations of genes with risk and with higher minor allele frequencies. However, counter examples abound. For example, with a common OR value of 1.5, the proportion of the population at high risk is 17.5% when 150 genes are predictive but smaller and 13.1% when 250 genes are predictive. The overall reduction in the proportion at high risk in this example is due to the facts that fewer controls are deemed at high risk by the more predictive 250-gene model and that the bulk of the population is composed of controls.

The sensitivity of risk models is low especially when genetic associations are weak. We see that less than half the cases are classified as high risk when ORs are <1.1, regardless of the number of genes in the profile. Even when the common OR is 1.25, to classify >50% of cases, at least 250 genes with allele frequencies of 10% or 150 genes with allele frequencies of 30% are required in the model. When only 50 genes are in the model, the proportion of cases classified as high risk only exceeded 50% in one scenario, namely for common genes with allele frequencies of 30% and large ORs equal to 1.5.

In Table 2, there are tendencies for improvements in proportions of cases and controls classified as high risk by the models with inclusion of larger numbers of predictive genes, with stronger associations of genes with risk and with higher minor allele frequencies. However, there are no absolute rules evident in this regard. On the other hand, the expected benefit due to the use of the risk model always improved with these three factors: with inclusion of larger numbers of predictive genes, with stronger

Table 2. The predictive capacity of genetic profiling under different scenarios defined by the number of genes involved, the OR associated with the high-risk allele, and the population frequency of the risk allele

	No. of genes	Total at high risk	Cases at high risk (TPR)	Controls at high risk (FPR)	Expected net benefit	Net reclassification index	R^2	AUC
A. Risk allele frequency 5%								
OR 1.05	50	0	0	0	0	0	0.001	0.533
	150	0	0	0	0	0	0.003	0.550
	250	0	0.001	0	0	0.001	0.005	0.568
	350	0.002	0.004	0.002	0	0.002	0.007	0.574
OR 1.10	50	0	0.001	0	0	0	0.004	0.560
	150	0.009	0.024	0.008	0.001	0.016	0.013	0.602
	250	0.030	0.073	0.026	0.002	0.047	0.021	0.629
	350	0.054	0.127	0.045	0.003	0.082	0.029	0.648
OR 1.25	50	0.028	0.075	0.023	0.002	0.052	0.026	0.639
	150	0.119	0.327	0.097	0.011	0.230	0.075	0.723
	250	0.131	0.407	0.100	0.018	0.307	0.116	0.768
	350	0.132	0.459	0.096	0.024	0.363	0.157	0.801
OR 1.50	50	0.128	0.371	0.101	0.014	0.271	0.089	0.738
	150	0.175	0.612	0.127	0.033	0.485	0.217	0.842
	250	0.131	0.603	0.077	0.043	0.526	0.302	0.885
	350	0.170	0.736	0.107	0.050	0.629	0.364	0.908
B. Risk allele frequency 10%								
OR 1.05	50	0	0	0	0	0	0.002	0.541
	150	0.001	0.002	0.001	0	0.001	0.006	0.568
	250	0.006	0.014	0.005	0	0.009	0.010	0.588
	350	0.016	0.032	0.014	0	0.018	0.013	0.602
OR 1.10	50	0.005	0.010	0.004	0	0.005	0.008	0.582
	150	0.038	0.090	0.032	0.002	0.058	0.024	0.634
	250	0.061	0.160	0.050	0.005	0.110	0.039	0.667
	350	0.096	0.250	0.078	0.007	0.172	0.054	0.694
OR 1.25	50	0.074	0.195	0.060	0.006	0.136	0.047	0.684
	150	0.144	0.454	0.110	0.021	0.344	0.130	0.781
	250	0.133	0.503	0.091	0.030	0.411	0.197	0.829
	350	0.144	0.590	0.095	0.038	0.496	0.253	0.862
OR 1.50	50	0.125	0.437	0.090	0.024	0.348	0.150	0.796
	150	0.144	0.662	0.087	0.047	0.575	0.328	0.897
	250	0.166	0.777	0.098	0.056	0.679	0.420	0.929
	350	0.144	0.792	0.073	0.063	0.719	0.490	0.948
C. Risk allele frequency 30%								
OR 1.05	50	0	0	0	0	0	0.005	0.563
	150	0.015	0.033	0.013	0	0.021	0.014	0.607
	250	0.037	0.087	0.031	0.002	0.056	0.024	0.639
	350	0.054	0.130	0.046	0.003	0.084	0.032	0.655
OR 1.10	50	0.021	0.048	0.018	0.001	0.030	0.018	0.620
	150	0.094	0.247	0.077	0.007	0.170	0.055	0.700
	250	0.111	0.329	0.086	0.014	0.244	0.089	0.744
	350	0.132	0.417	0.101	0.019	0.316	0.121	0.777
OR 1.25	50	0.116	0.358	0.088	0.016	0.269	0.100	0.756
	150	0.142	0.578	0.094	0.037	0.484	0.246	0.862
	250	0.152	0.685	0.092	0.048	0.593	0.340	0.904
	350	0.151	0.738	0.085	0.055	0.652	0.407	0.926

(*Continued on the following page*)

Table 2. The predictive capacity of genetic profiling under different scenarios defined by the number of genes involved, the OR associated with the high-risk allele, and the population frequency of the risk allele (Cont'd)

	No. of genes	Total at high risk	Cases at high risk (TPR)	Controls at high risk (FPR)	Expected net benefit	Net reclassification index	R^2	AUC
OR 1.50	50	0.163	0.646	0.108	0.040	0.538	0.268	0.874
	150	0.142	0.779	0.071	0.062	0.708	0.480	0.947
	250	0.152	0.867	0.073	0.070	0.794	0.576	0.967
	350	0.132	0.863	0.051	0.075	0.812	0.631	0.976

NOTE: The proportion of subjects in the population who have an event is 10%. Subjects with risks exceeding 20% are classified as high risk.
Abbreviations: TPR, true-positive rate; FPR, false-positive rate.

associations of genes with risk, and with higher minor allele frequencies.

Note that the expected benefit values displayed in Table 2 are weighted averages of the proportions of cases and controls classified as high risk. The weighting acknowledges that use of 0.20 as the high-risk threshold implies that the cost for a control classified as high risk is equivalent to one fourth of the benefit for a case classified as high risk. Let us consider how to interpret the expected benefit values shown in Table 2 with a concrete example. Suppose a policy maker is deciding if ascertaining information such as genotype is economically advantageous. Assume some hypothetical monetary costs for treatment, for example, $20,000 for treating a subject diagnosed with disease and $1,000 for interventions to prevent disease occurring in the first place. If prevention interventions reduce the risk of disease by 25%, then the expected benefit for a subject that would be a case in the absence of intervention is 0.25 × ($20,000) – $ 1,000 = $4,000, whereas the expected cost for a subject that would be a control in the absence of intervention is $1,000. The cost-benefit ratio is therefore $1,000/$4,000 = 1/4 in this setting, leading to the use of the risk threshold 0.2. The expected benefit values in Table 2 are in units corresponding to the benefit of high-risk designation for a case. That is, to convert the values in Table 2 to monetary values in this hypothetical setting, we multiply by $4000. Thus, for example, the expected monetary benefit associated with the model in the last row in Table 2A is 0.05 × $4000 = $200 per person. If testing costs more than $200, there is no gain in financial terms by using this risk model. However, nonmonetary aspects must be factored into policy making as well.

Results for Heterogeneously Predictive Genes

In the second set of simulations summarized in Table 3, the genetic profiles are such that the ORs and minor allele frequencies both vary. The ORs for the strongest 20 genes vary uniformly from a maximum value displayed in Table 3 to 1.15, whereas the OR decreases uniformly from 1.15 to 1.05 over the remaining genes. The minor allele frequency starts at 0.05 and increases by 0.005 for each gene over the first 50 genes, then by 0.0005 for each of the remaining genes. A key feature in these scenarios is that the strong genes are uncommon whereas the genes weakly associated with risk are relatively more common. Again, our scenarios mimic those reported by Janssens et al. (2).

We see that the population proportions at high risk, overall for cases and for controls, and the expected net benefit are determined to a large extent by the relatively few genes in the strong set, especially when their ORs are high.

Use of Risk Distributions versus AUC

Tables 2 and 3 display values of the AUC for each risk model. Janssens et al. (2) use the criterion AUC ≥0.80 to indicate high discriminative accuracy. Others use similar criteria. However, a model may have AUC as large as 0.80, yet it may not be useful in practice. For example, the model in row 12 of Table 2 has an expected benefit of 0.024. Assuming the hypothetical values mentioned earlier for monetary costs and benefits as well as risk reductions afforded by prevention interventions, the expected monetary benefit of using this test is $96 per person. If the cost of testing is $96, there is no net benefit despite the fact that the AUC for the risk model is 0.801. On the other hand, Gail and Pfeiffer (12) have shown that the modified Gail model for breast cancer risk (model 2 in Costantino et al., 1999) is useful for selecting women for prevention treatment with tamoxifen despite the fact that its AUC is 0.66. As another example, consider that the expected monetary benefit for the model in Table 2A with 50 genes each and with ORs of 1.25 is 0.002 × $4000 = $8 per person, which is derived from its capacity to classify 7.5% cases and 2.3% controls as high risk using the risk

Table 3. The predictive capacity of genetic profiling when there is a mixture of strongly predictive genes and weakly predictive genes

Max OR	No. of genes	Total at high risk	Cases at high risk (TPR)	Controls at high risk (FPR)	Expected net benefit	Net reclassification index	R^2	AUC
1.5	20	0.050	0.126	0.042	0.003	0.084	0.029	0.642
	50	0.079	0.210	0.064	0.007	0.146	0.048	0.684
	150	0.129	0.394	0.099	0.017	0.295	0.108	0.767
	250	0.140	0.463	0.103	0.023	0.360	0.143	0.794
	350	0.145	0.492	0.105	0.026	0.388	0.163	0.812
2.0	20	0.106	0.312	0.082	0.013	0.229	0.080	0.723
	50	0.117	0.358	0.090	0.016	0.268	0.098	0.746
	150	0.141	0.472	0.104	0.024	0.368	0.151	0.802
	250	0.145	0.516	0.103	0.028	0.413	0.180	0.820
	350	0.148	0.539	0.104	0.031	0.436	0.198	0.834
3.0	20	0.140	0.515	0.098	0.029	0.417	0.191	0.819
	50	0.142	0.540	0.098	0.032	0.442	0.202	0.830
	150	0.150	0.590	0.101	0.036	0.489	0.241	0.856
	250	0.150	0.610	0.098	0.039	0.511	0.258	0.865
	350	0.15	0.622	0.098	0.04	0.525	0.274	0.873

NOTE: ORs for the weakly predictive genes vary from 1.05 to 1.15, whereas ORs for the 20 strongly predictive genes vary from 1.15 to max OR. Genes with higher ORs are infrequent, and genes with lower ORs are common. The minor allele frequency of the 50 strongest genes varies from 0.05 to 0.15 and increases by 0.0005 after the 50th gene. The proportion of subjects in the population who have an event is 10%. Subjects with risks exceeding 20% are classified as high risk.

threshold of 0.20, which is deemed clinically relevant in our hypothetical example. If the corresponding genetic test costs are less than $8 per person, then it will be cost effective to offer it to people. Yet, the AUC for this model is only 0.64.

The crucial issue is that one cannot assess the value of a risk model according to AUC, which ignores the population and clinical context in which the model is to be applied. For example, the AUC does not incorporate the case prevalence in the population. Another problem with the AUC is that it does not take into consideration risk thresholds that motivate intervention in the clinical context. Consider the setting in row 12 of Table 2A again. If the benefit of treating a case is constant but the cost of treating a control is high, so that only subjects at very high risk, say >30%, should receive intervention, the benefit of using the model will be different than if the cost of treating a control is less where subjects with risks, say >10%, should be intervened upon. With risk threshold equal to 30%, only 32% of cases and 5% of controls satisfy the criterion for high risk, and the expected benefit is 0.014. The corresponding numbers using the lower risk threshold equal to 10% are 73% of cases and 28% of controls, and have an expected benefit of 0.045. Clearly, the implications of the risk model are different in these two scenarios. Yet, AUC makes no distinction. Indeed, it accumulates over all possible risk thresholds, considering all values between 0 and 1 as plausible.

The R^2 summary statistic and the NRI, also shown in Tables 2 and 3, share many of the same drawbacks as AUC. They do not incorporate the clinical context into their calculations. Interestingly, R^2 varies with population prevalence and NRI varies with the high-risk threshold. However, neither is incorporated in ways that make the resulting measure clinically relevant for evaluating the risk prediction model.

Discussion

Our simulations indicate that to identify a sizable number of subjects at substantially increased risk for an event, large numbers of independent genes that confer at least moderately elevated relative risks or, alternatively, a few genes that are strongly associated with risk are required. To date, whole genome analyses have yielded genes and single nucleotide polymorphisms in particular that are only weakly associated with outcome. These are unlikely to be helpful in identifying large groups of individuals at substantially elevated risk.

Our conclusions are limited to the set of scenarios studied here. Tables for additional settings are provided in Appendix A, and alternative scenarios can be investigated using the general programs we have developed. A key feature of the scenarios simulated is that genes are in linkage equilibrium and that they have statistically independent effects on disease risk. Correlations between

genes are likely to give rise to prediction models with poorer performance. On the other hand, it is possible that certain types of interactions between genes and interactions between environmental factors and genes may yield better capacities to predict risk.

In addition to exploring the potential predictive capacities of specific genetic profiles, we have argued for using clinically relevant, easy to understand ways of quantifying the capacity of genes, markers, and other factors to predict risk. We promote the use of risk distribution plots because they are both easy to understand and because they give clinically useful information. Moreover, all statistical summaries of predictive capacity are derived from them. In addition, decision curves that are relatively simple and useful for formal cost-benefit analyses are derived from them.

We showed that risk distribution curves are preferable to ROC curves. In particular, criteria based on AUC can be misleading. A risk model that is beneficial in a particular population may not have an AUC that indicates good discrimination. A risk model that is not beneficial in a particular population may have an excellent AUC. Although Gail (1) previously evaluated an addition of seven single nucleotide polymorphisms to a breast cancer risk model using AUC, he has more recently used criteria based on risks and benefits for this evaluation (14).

In this article, we do not provide technical discussion about using data to fit risk prediction models or to assess their performance. We investigated performance in the ideal setting in which the true risk values can be calculated from an individual's genetic profile and a very large data set is available to assess the true performance of the risk model in the population. Methods for estimating performance from study data along with confidence interval construction and hypothesis testing are under development (13, 15–17).

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Grant Support

NIH grants R01 GM054438 and R01 CA129934.
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received 06/01/2009; revised 12/09/2009; accepted 12/23/2009; published OnlineFirst 02/16/2010.

References

1. Gail MH. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. J Natl Cancer Inst 2008; 100:1037–41.
2. Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, van Duijn CM. Predictive testing for complex diseases using multiple genes: fact or fiction? Genet Med 2006;8:395–400.
3. Stern RH. Evaluating new cardiovascular risk factors for risk stratification. J Clin Hypertens (Greenwich) 2008;10:485–8.
4. Pepe MS, Feng Z, Huang Y, et al. Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol 2008; 167:362–8.
5. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565–74.
6. Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York: Oxford University Press; 2003. p. 78.
7. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928–35.
8. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27: 157–72.
9. McGeechan K, Macaskill P, Irwig L, Liew G, Wong TY. Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. Arch Intern Med 2008;168:2304–10.
10. Janes H, Pepe MS, Gu J. Assessing the value of risk predictions using risk stratification tables. Ann Int Med 2008;149:751–60.
11. Pepe MS, Feng Z, Gu JW. Invited commentary on "Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond". Stat Med 2008;27: 173–181.
12. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. Biostatistics 2005;6:227–39.
13. Gu JW, Pepe MS. Measures to summarize and compare the predictive capacity of markers. Int J Biostat 2009;5; article 27.
14. Gail MH. Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. J Natl Cancer Inst 2009;101: 959–963.
15. Huang Y, Pepe MS, Feng Z. Evaluating the predictiveness of a continuous marker. Biometrics 2007;63:1181–88.
16. Huang Y, Pepe MS. A parametric ROC model based approach for evaluating the predictiveness of continuous markers in case-control studies. Biometrics 2009;65:1133–44.
17. Huang Y, Pepe MS. Semiparametric methods for evaluating risk prediction markers in case-control studies. Biometrika 2009;96: 991–7.
18. Costantino JP, Gail MH, Pee D, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst 1999;91:1541–8.