Putting Risk Prediction in Perspective: Relative Utility Curves

Stuart G. Baker

Risk prediction models based on medical history or results of tests are increasingly common in the cancer literature. An important use of these models is to make treatment decisions on the basis of estimated risk. The relative utility curve is a simple method for evaluating risk prediction in a medical decision-making framework. Relative utility curves have three attractive features for the evaluation of risk prediction models. First, they put risk prediction into perspective because relative utility is the fraction of the expected utility of perfect prediction obtained by the risk prediction model at the optimal cut point. Second, they do not require precise specification of harms and benefits because relative utility is plotted against a summary measure of harms and benefits (ie, the risk threshold). Third, they are easy to compute from standard tables of data found in many articles on risk prediction. An important use of relative utility curves is to evaluate the addition of a risk factor to the risk prediction model. To illustrate an application of relative utility curves, an analysis was performed on previously published data involving the addition of breast density to a risk prediction model for invasive breast cancer.


A risk prediction model is a mathematical model to predict the risk of developing disease or other health outcome on the basis of medical history and the results of tests. A primary use of risk prediction is to help make treatment decisions depending on a person’s risk for developing disease in the absence of treatment. (Throughout this commentary, the “absence of treatment” refers to the absence of the particular treatment contemplated for high-risk persons.) The risk prediction model is constructed from data on persons who did not receive the treatment and hence estimates the risk of disease in the absence of treatment. An important issue in evaluating risk prediction models is to determine whether or not it is worthwhile to include an additional risk factor in the risk prediction model, such as the inclusion of breast density in a model to predict the risk of invasive breast cancer (1). This issue is addressed by comparing risk prediction models with and without the additional risk factor.

Purely statistical measures for comparing two risk prediction models have limited use for medical decision making because they do not incorporate harms and benefits related to treatment decisions arising from the risk prediction model. Such statistical measures include the odds ratio for the additional risk factor, the difference in areas under the receiver operating characteristic (ROC) curves, the difference in fraction of persons who are in the lowest and highest risk categories (2), and the fractions of persons who are correctly and incorrectly reclassified in terms of the risk prediction model with the additional risk factor (1–3). In contrast, harms and benefits are explicitly incorporated into utility formulations for evaluating risk prediction models (4–8). However, these utility formulations are sometimes difficult to implement because of the need to specify multiple harms and benefits and the complexity of investigating the impact of varying the levels of harms and benefits.

The difficulties with the utility formulation can be reduced by using the recently developed methods of relative utility curves (9) and related decision curves (10,11). Relative utility is the fraction of the utility of perfect prediction that is achieved at the optimal cut point for a risk prediction model. The relative utility curve is a plot of relative utility vs risk threshold (ie, a summary measure of harms and benefits). Relative utility and decision curves make it easy to determine the impact of various risk thresholds. The focus in this commentary is on relative utility curves because, unlike decision curves, they provide perspective relative to perfect prediction. Evaluation of an additional risk factor in a risk prediction model is based on a comparison of relative utility curves for risk prediction models with and without the additional risk factor. These topics will be explained more fully below.

The computation and application of relative utility curves is illustrated with the evaluation of risk prediction for invasive breast cancer when there is a possible recommendation for chemoprevention, such as treatment with tamoxifen, for women at high risk for invasive breast cancer. The additional risk factor to be evaluated is breast density. Investigators have fit two risk prediction models, with and without breast density, to prospectively collected data from women who were older than 35 years, who had received at least one mammogram with a measurement of breast density, and who did not have cancer in the first 6 months of follow-up (1). Follow-up lasted approximately 5 years. Presumably, few persons in this study received chemoprevention for breast cancer, so that the risk of invasive breast cancer in the risk prediction model is in the absence of chemoprevention for breast cancer. For the risk prediction model labeled in this presentation as “Baseline Factors,”

Affiliation of author: Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD.

Correspondence to: Stuart G. Baker, ScD, Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, EPN 3131, 6130 Executive Blvd MSC 7354, Bethesda, MD 20892-7354 (e-mail: sb16i@nih.gov).

See “Funding” and “Notes” following “References.”

DOI: 10.1093/jnci/djp353

Published by Oxford University Press 2009.

Advance Access publication on October 20, 2009.
the investigators estimated the risk of invasive breast cancer as a function of the baseline variables of age, race, ethnicity, family history, and biopsy history. For the risk prediction model labeled here as “Baseline Factors and Breast Density,” the investigators estimated the risk of invasive breast cancer as a function of breast density as well as the other baseline variables. Details of the model fits, which can be found in the original publication (1), are not relevant to the evaluation of the risk prediction model. Key formulas for computing relative utility are presented in the text and in Table 1. Interested readers can find derivations of the formulas in the Supplementary Material (available online), much of which is based on Baker et al. (9).

Data Tables

The starting point for these computations is a table of data on the number of persons who develop disease in each risk group and the total number in each risk group; such tables are found in many publications on risk prediction models. Consider the example involving invasive breast cancer (Table 1). For each person in the aforementioned breast cancer study (1), a 5-year predicted risk of invasive breast cancer in the absence of chemoprevention was separately computed under each risk prediction model, “Baseline Factors” and “Baseline Factors and Breast Density.” For each risk prediction model, each person in the study was assigned to one of four risk groups on the basis of her predicted risk of invasive breast cancer: risk group 1 = predicted risks of less than 1%; risk group 2 = predicted risks of 1%–1.66%; risk group 3 = predicted risks of 1.67%–2.49%; and risk group 4 = predicted risks greater than or equal to 2.5%. For each risk prediction model, the data for computing relative utility curves are the numbers of persons in each risk group and the numbers of persons who developed invasive breast cancer in each risk group.

Estimated Risks in the Study

From the data table, it is straightforward to compute the estimated risks for each risk group and the overall probability of developing disease in the study. Let \( j = 1, 2, 3, \) or 4 (ie, according to the index of the four risk groups). Let \( x_j \) denote the number of persons in risk group \( j \) who developed disease in the absence of treatment and let \( n_j \) denote the number of persons in risk group \( j \). Then the estimated risk of developing disease in the absence of treatment among persons in risk group \( j \) of the study is
\[
r_j = x_j / n_j,
\]
the fraction of persons in the study whose predicted risk under the model falls into risk group \( j \) in the study is
\[
w_j = n / \sum n_j,
\]
and the estimated overall probability of developing disease in the absence of treatment in the study is
\[
p = \sum r_j w_j.
\]

Table 1. Data and computations for the invasive breast cancer example†

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Symbol and formula</th>
<th>( j = 1 )</th>
<th>( j = 2 )</th>
<th>( j = 3 )</th>
<th>( j = 4 )</th>
<th>( j = 1 )</th>
<th>( j = 2 )</th>
<th>( j = 3 )</th>
<th>( j = 4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. in risk group ( j ) of study who developed disease in absence of treatment‡</td>
<td>( x_j )</td>
<td>1576</td>
<td>2640</td>
<td>2723</td>
<td>1845</td>
<td>1761</td>
<td>2390</td>
<td>2513</td>
<td>2120</td>
</tr>
<tr>
<td>No. in risk group ( j ) of study†</td>
<td>( n_j )</td>
<td>215402</td>
<td>201927</td>
<td>148795</td>
<td>63105</td>
<td>249959</td>
<td>186106</td>
<td>124420</td>
<td>68744</td>
</tr>
<tr>
<td>Fraction in study who are in risk group ( j )</td>
<td>( w_j = n_j / \sum n_j )</td>
<td>0.342</td>
<td>0.321</td>
<td>0.236</td>
<td>0.100</td>
<td>0.397</td>
<td>0.296</td>
<td>0.198</td>
<td>0.109</td>
</tr>
<tr>
<td>Risk of disease in risk group ( j ) of study</td>
<td>( r_j = x_j / n_j )</td>
<td>0.0073</td>
<td>0.0131</td>
<td>0.0183</td>
<td>0.0292</td>
<td>0.0070</td>
<td>0.0128</td>
<td>0.0202</td>
<td>0.0308</td>
</tr>
<tr>
<td>Risk of disease in risk group ( j ) of target population†</td>
<td>( r'_j = \frac{a_j}{(a_j + b_j)} ), where ( a_j = r'_j \times p^* (p) ) and ( b_j = (1 - r'_j) \times (1 - p^* (1 - p)) )</td>
<td>0.0052</td>
<td>0.0094</td>
<td>0.0131</td>
<td>0.0210</td>
<td>0.0050</td>
<td>0.0092</td>
<td>0.0145</td>
<td>0.0222</td>
</tr>
<tr>
<td>False-positive rate</td>
<td>( FPR_j = \frac{\sum r_j w_j}{(1 - r_j) w_j} \times (1 - p) )</td>
<td>1.000</td>
<td>0.655</td>
<td>0.334</td>
<td>0.099</td>
<td>1.000</td>
<td>0.600</td>
<td>0.304</td>
<td>0.107</td>
</tr>
<tr>
<td>True-positive rate</td>
<td>( TPR_j = \frac{\sum r_j w_j}{p} \times (1 - p) )</td>
<td>1.000</td>
<td>0.821</td>
<td>0.520</td>
<td>0.210</td>
<td>1.000</td>
<td>0.800</td>
<td>0.527</td>
<td>0.241</td>
</tr>
<tr>
<td>Slope of ROC curve (on left side of point)</td>
<td>( ROCSlope_j = \frac{[1 - p]}{p^* (1 - p)} \times [(r_j - (1 - r_j)) \times (1 - r'_j)] )</td>
<td>0.0052</td>
<td>0.936</td>
<td>1.317</td>
<td>2.127</td>
<td>0.501</td>
<td>0.919</td>
<td>1.456</td>
<td>2.248</td>
</tr>
<tr>
<td>Relative utility if no testing harm ( (r'_j \geq p^* ) )</td>
<td>( RU_{(no\ test\ harm)} = TPR_j \times ROCSlope_j \times FPR_j )</td>
<td>Not applicable</td>
<td>0.080</td>
<td>$</td>
<td>Not applicable</td>
<td>0.085</td>
<td>$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative utility if no testing harm ( (r'_j &lt; p^* ) )</td>
<td>( RU_{(no\ test\ harm)} = \frac{1 - FPR_j}{(1 - TPR_j)} \times ROCSlope_j )</td>
<td>$</td>
<td>0.153</td>
<td>Not applicable</td>
<td>$</td>
<td>0.182</td>
<td>Not applicable</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

† \( p = 0.014 \) = estimated overall probability of developing disease in the absence of treatment in the study = \( \sum r_j w_j \) and \( p^* = 0.010 \) = overall probability of developing disease in absence of treatment in the target population. \( FPR_j = \) false-positive rate; ROC= receiver operating characteristic; \( ROCSlope_j = \) slope of the ROC curve to the left of the point \( (FPR_j, TPR_j) \); \( RU_{(no\ test\ harm)} \) = the relative utility for a person with risk threshold \( R \) if the harms of testing for or measuring risk factors are ignored; \( TPR_j = \) true-positive rate.

‡ Data are from table 6 of Tice et al. (1).

§ Relative utility is mathematically equal to zero.

jnci.oxfordjournals.org
With the four risk groups, \( p = r_1w_1 + r_2w_2 + r_3w_3 + r_4w_4 \), which is a weighted average of the estimated risks in each risk group (see Table 1 for an example of these calculations).

**ROC Curves**

An ROC curve can be computed by using the estimated risks in the study. The ROC curve plots the true-positive rate, which is the estimated probability of a positive classification among persons who developed disease, vs the false-positive rate, which is the estimated probability of positive classification among persons who did not develop disease. When calculating false-positive rate and true-positive rate, a cut point \( j \) is specified to indicate that persons in risk groups \( j \) or higher are classified as positive, and summations are over risk groups indexed by \( r \) that are greater than or equal to cut point \( j \). In terms of estimated risks, the false-positive rate (FPR) for cut point \( j \) is computed as

\[
FPR_j = \sum_{s \geq j}(1 - r)w_s / (1 - p)
\]

and the true-positive rate (TPR) for cut point \( j \) is computed as

\[
TPR_j = \sum_{s \geq j}r_s w_s / p.
\]

For example, if there are four risk groups, the false-positive rate for risk group 2 is \( FPR_2 = [(1 - r_2)w_1 + (1 - r)w_2 + (1 - r_3)w_3] / (1 - p) \). Another key quantity is the slope of the ROC curve to the left of the point \( (FPR_j, TPR_j) \), which can be written as

\[
\text{ROCSlope}_j = (TPR_j - TPR_{(j-1)}) / (FPR_j - FPR_{(j-1)}) = [(1 - p) / p] \times (r_j / (1 - r_j)).
\]

Examples of these computations are shown in Table 1.

In the application to risk prediction for invasive breast cancer, the ROC curve for the risk prediction model “Baseline Factors and Breast Density” was only slightly higher than the ROC curve for “Baseline Factors” (Figure 1), indicating that inclusion of breast density in the risk prediction model provided little improvement in classification accuracy. However, this comparison of ROC curves is limited because it does not evaluate the impact on medical decision making of including breast density in the risk prediction model. More details of the limitations of ROC curves for evaluating risk prediction will be discussed below.

**Estimated Risk in the Target Population**

Sometimes investigators want to evaluate risk prediction in a target population with an overall probability of developing disease that differs from the overall probability of developing disease in the study sample. In this situation, the estimated risk of disease in risk group \( j \) of the target population will differ from the estimated risk of disease in risk group \( j \) of the study and so needs to be computed. Let \( p^* \) = the overall probability of developing disease in the absence of treatment in the target population, a quantity that is considered to be known and \( r_j^* \) = estimated risk of developing disease in the absence of treatment among persons in risk group \( j \) of the target population.

If the ROC curve in the target population is the same as in the study, then the quantity \( r_j^* \) can be computed from \( p^*, p \), and \( r_j \) by using the following formula:

\[
r_j^* = (r_j \times p^*/p) / [(r_j \times p^*/p) + (1 - r_j) \times (1 - p^*)/(1 - p)].
\]

The slope of the ROC curve in the target population is \( \text{ROCSlope}_j = [(1 - p)/p] \times (r_j^* / (1 - r_j^*)) \), which equals the slope of the ROC curve in the study.

**Risk Threshold**

For summarizing harms and benefits, the key quantity is the risk threshold, \( R \), which is the risk of disease in the absence of treatment at which a person would be indifferent between receiving treatment for the disease and not receiving treatment for the disease. As will be discussed, the relative utility that ignores any harms of testing is a function of the harms and benefits of the possible outcomes of treatment and no treatment only through the risk threshold. This characteristic of relative utility is important because it obviates detailed specification of harms and benefits. For the application used in this commentary, the risk threshold is the 5-year risk of invasive breast cancer at which a woman would be indifferent between undertaking or not undertaking chemoprevention (ie, taking tamoxifen to prevent breast cancer).

**Relevant Region**

Whether or not treatment is given in the absence of risk prediction provides important information about the range of risk thresholds and the part of the ROC curve that is relevant to a person (ie, the relevant region). If a person would not receive treatment in the absence of risk prediction, as is typically the case with the chemoprevention of breast cancer, the risk threshold for that person must be greater than the overall probability of developing disease in the absence of treatment, namely, \( R \) is greater than \( p^* \), and the slope of the ROC curve must be greater than 1 (eg, \( j = 3 \) or 4 in this example). Conversely, if a person would receive treatment in the
absence of risk prediction, the relevant region is that in which \( R \) is less than \( p^* \) and the relevant slope of the ROC curve is less than 1 (eg, \( j = 1 \) or 2 in this example). The area under the ROC curve is generally an inappropriate measure of risk prediction because it includes the area under the part of ROC curve outside the relevant region (9).

**Relative Utility Curves**

The term “utility” refers to either harms or benefits that are measured in the same units. The term “expected utility” refers to the average utility taking into account the probability of developing disease. The relative utility is the fraction of the expected utility of perfect prediction achieved at the optimal cut point for a risk prediction model. The optimal cut point for a risk prediction model is the cut point that maximizes the expected utility of prediction. A classic result in the theory of medical decision making is that for a person with risk threshold \( R \), the optimal cut point \( j \) is associated with \( r_j = R \) or equivalently \( \text{ROCSlope}_j = [(1 - p^*)/p^*] \times [R/(1 - R)] \) (4,5,6,7,9,12). If the harms of testing for or measuring risk factors are ignored, the relative utility (RU) ignoring testing harms for a person with risk threshold \( R \) can be written simply as

\[
RU_{(\text{no test harm})j} = \text{TPR}_j - \text{ROCSlope}_j \times \text{FPR}_j, \text{if } r_j \geq p^*,
\]

and

\[
RU_{(\text{no test harm})j} = [(1 - \text{FPR}_j) - (1 - \text{TPR}_j)]/\text{ROCSlope}_j, \text{if } r_j < p^*,
\]

when cut point \( j \) is selected so that \( r_j \) equals \( R \) (ie, the optimal cut point to maximize the expected utility of prediction). The conditions that \( r_j \geq p^* \) and \( r_j < p^* \) correspond to relevant regions of the relative utility curve when, as required, each \( r_j \) corresponds to a risk threshold. As will be discussed below, the above formula facilitates computation of the maximum acceptable harm of testing for an additional risk factor.

A relative utility curve is a plot of the relative utility when harms of testing are ignored \( RU_{(\text{no test harm})j} \) against risk threshold \( R = r_j \). In other words, the relative utility curve plots the fraction of the expected utility of perfect prediction obtained by the risk prediction model at the optimal cut point associated with the risk threshold \( R \) in the scenario when testing harms are ignored (Figure 2). Linear interpolation is used to compute relative utilities for values of \( R \) between plotted points. For a smooth ROC curve with the usual concave (curving inward) shape, the relative utility curve is highest when the risk threshold \( R = p^* \) and decreases toward zero to the left and right of the point where \( R = p^* \). The theoretically highest possible value of the relative utility is 1, which corresponds to perfect prediction and provides a benchmark for evaluating risk prediction models. An attractive aspect of relative utility curves, which is also shared by decision curves (10), is that they show a range of values for \( R \), rather than focusing a single value for \( R \), so that precise specification of \( R \) is not required.

**Comparison of Relative Utility Curves**

Another attractive aspect of relative utility curves is that two relative utility curves are aligned at the same risk threshold, which is ideal for making comparisons for a person with a given risk threshold. In contrast, two ROC curves are aligned at the same false-positive rates, which generally correspond to different slopes of the ROC curves and hence are associated with optimal cut points for different risk thresholds, making visual comparison at the same risk threshold difficult (2).

For purposes of illustration, the relative utility curves for the data on invasive breast cancer were computed for a target population with overall probability \( p^* = 0.010 \) of developing invasive breast cancer in the absence of treatment (Figure 2). The relative utility curves in the relevant region were similar for the two models, “Baseline Factors” and “Baseline Factors and Breast Density.” For example, when \( R \) equals 0.013, the relative utility ignoring testing harms was 0.15 for “Baseline Factors” and 0.18 for “Baseline Factors and Breast Density.” Thus, adding breast density to the risk prediction model increased the relative utility ignoring testing harms by 0.03 (with standard error of 0.003). To put this increase in relative utility in perspective, the largest possible increase would be from 0.15 to 1.00, a difference of 0.85.

An additional risk factor will generally increase the performance of the risk prediction model but it comes at a “price” of the harm of testing for the additional risk factor. An additional test is deemed worthwhile if its harm is less than the maximum acceptable testing harm, which is the increased harm of testing for or measuring an additional risk factor that would exactly counterbalance the increased utility from better risk prediction with the inclusion of the additional risk factor.

The harm of an additional test or measurement, which could also include monetary costs, is measured as equivalent to a fraction of persons not treated among persons who would develop disease in the absence of treatment. When there is no treatment in the
absence of risk prediction, so that the relevant region of the relative utility curve is $R > p^*$, the maximum acceptable testing harm equals $p^*$ multiplied by the difference in relative utilities when the testing harm is ignored.

For the breast cancer example at a risk threshold $R$ of 0.013, the maximum acceptable testing harm equals 0.0003, which is computed as a 0.03 increase in the relative utility ignoring testing harm multiplied by 0.01, the overall probability of developing disease in the target population, $p^*$. Therefore, measuring breast density is worthwhile if the harm of measuring breast density is less than 0.0003—the equivalent of three in 10,000 women not being treated with chemoprevention among those who would have developed invasive breast cancer in the absence of treatment for invasive breast cancer. Alternatively, one can say that breast density is worth measuring if one would be willing to exchange at least $1/0.0003 = 3333$ measurements of breast density to treat with chemoprevention one woman who would develop invasive breast cancer in the absence of treatment for invasive breast cancer (9).

Additional Considerations

To avoid using the same data to estimate the risk prediction model and to compute the relative utility curve, the data should be randomly split into what are called training and test samples, the risk prediction model should be fit to the training sample, and the relative utility curve should be computed from data in the test sample. Standard errors can be computed by randomly repeating this procedure (13). Although the total sample for the risk prediction model for invasive breast cancer had been split into separate training and test samples, the published counts for the risk groups were from the total sample and not the test sample, as would ideally have been preferred for this analysis.

Relative utility curves should be computed from a concave ROC curve because concavity in the ROC curve is a prerequisite for optimal medical decision making (14). An ROC curve can be made concave by combining categories, smoothing, or creating a concave envelope of points. Alternatively, a concave ROC curve can be created by computing the estimated risk of developing disease in risk group $j$ in the absence of treatment, $r_j$, as an average of predicted risks of developing disease among persons in risk group $j$. This latter approach requires the validity of the model, which can be partly checked by comparing predicted and observed estimates, a process called calibration. Smoother relative utility curves could be created by computing false-positive rate and true-positive rate for each individual’s predicted risk, as is done with decision curves (10).

The application that was presented in this commentary involved evaluating the addition of breast density to a risk prediction model for invasive breast cancer. A previous application involved evaluating the addition of a test for high-density lipoprotein to a risk prediction model for cardiovascular disease (9). The comparison of relative utility curves could also be used to evaluate the addition of a genomic test to a risk prediction model that is based on conventional risk factors either in a cancer-prevention setting, such as the value of adding information on information on single nucleotide polymorphisms to known risk factors for breast cancer risk (8), or a cancer treatment setting, such as the value of adding genetic information to established clinical prognostic factor in models to predict survival of lymphoma patients, survival of breast cancer patients, or diagnosis of lymph node metastases in head and neck cancer (15).

References


Funding

Division of Cancer Prevention in the National Cancer Institute and the National Institutes of Health.

Notes

The data were previously published (1).

The author had full responsibility for the analysis and interpretation of the data, the decision to submit the manuscript for publication, and the writing of the manuscript.

Manuscript received February 10, 2009; revised May 28, 2009; accepted September 4, 2009.