The International Journal of Biostatistics

Volume 8, *Issue* 1 2012 *Article* 5

Evaluating a New Marker for Risk Prediction Using the Test Tradeoff: An Update

Stuart G. Baker, *National Cancer Institute*
Ben Van Calster, *Katholieke Universiteit Leuven and Erasmus MC*
Ewout W. Steyerberg, *Erasmus MC*

Recommended Citation:
Baker, Stuart G.; Van Calster, Ben; and Steyerberg, Ewout W. (2012) "Evaluating a New Marker for Risk Prediction Using the Test Tradeoff: An Update," *The International Journal of Biostatistics*: Vol. 8: Iss. 1, Article 5.

DOI: 10.1515/1557-4679.1395

Evaluating a New Marker for Risk Prediction Using the Test Tradeoff: An Update

Stuart G. Baker, Ben Van Calster, and Ewout W. Steyerberg

Abstract

Most of the methodological literature on evaluating an additional marker for risk prediction involves purely statistical measures of classification performance. A disadvantage of a purely statistical measure is the difficulty in deciding the improvement in the measure that would make inclusion of the additional marker worthwhile. In contrast, a medical decision making approach can weigh the cost or harm of ascertaining an additional marker against the benefit of a higher true positive rate for a given false positive rate that may be associated with risk prediction involving the additional marker. An appealing form of the medical decision making approach involves the risk threshold, which is the risk at which the expected utility of treatment and no treatment is the same. In this framework, a readily interpretable evaluation of the net benefit of an additional marker is the test tradeoff corresponding to the risk threshold. The test tradeoff is the minimum number of tests for a new marker that need to be traded for a true positive to yield an increase in the net benefit of risk prediction with the additional marker. For a sensitivity analysis the test tradeoff is computed over multiple risk thresholds. This article updates the theory and estimation of the test tradeoff. An example is provided.

KEYWORDS: decision curves, relative utility curves, receiver-operating characteristic curve, risk threshold, test tradeoff

Author Notes: SGB was supported by the National Institutes of Health. BVC is supported by a postdoctoral fellowship from the Research Foundation - Flanders (FWO) (grant 1.2.516.09N). ES was supported by the Center for Translational Molecular Medicine (PCMM project, grant 03O-203).

We thank Hafid Narayan, Margaret Pepe, and Andrew Vickers for helpful comments.

1 Introduction

Risk prediction provides important information for making treatment decisions. For example, predictions of the risk of residual tumor inform decisions about surgical resection in patients with testicular cancer (Steyerberg et al., 1995), and predictions of the risk of breast cancer inform decisions about chemoprevention (Mealiffe et al., 2010). An important question is whether or not it is worthwhile to include an additional marker in the risk prediction model. Here we take a decision analytic perspective in weighing benefits and harms in a population of an additional marker for risk prediction. Testing for or otherwise ascertaining the additional marker could have costs or harms that need to be weighed against benefits and harms of risk prediction with the additional marker. We consider a situation in which, in the absence of risk prediction, there are only two decisions: treat or not treat.

For the sake of brevity, "develops disease" will mean "develops disease in the absence of the treatment under consideration," and "risk" will refer to "risk of developing disease." Also "additional marker" refers to either a single marker or set of markers. We define the risk prediction model as a mathematical model for the risk of developing disease as a function of a set of predictors. We consider two risk prediction models: *Model 1*, a baseline risk prediction model, and *Model 2*, a risk prediction model that uses the predictors in Model 1and an additional marker. We also define the following samples and population. The *development sample* is the sample used to formulate the risk prediction model. The *development population* is the population from which the development sample is implicitly a random draw. The *validation sample* is the sample used to evaluate risk prediction model (and sometimes also used to refine the risk prediction model). The *target population* is a population from which the validation sample is implicitly a random draw or a random draw separately for persons who develop disease and persons who do not develop disease (thus allowing for oversampling of persons who developed a rare disease). Usually the target population corresponds to a different geographic location than the development population (Justice, et al. 1999).

There is a large literature on purely statistical methods to evaluate an additional marker for risk prediction. One approach is to compare classification performance curves from Models 1 and 2 in the validation samples. Examples include receiver operating characteristic (ROC) curves and predictiveness curves (Gu and Pepe, 2009). A second approach is to construct a summary statistical measure of classification performance in the validation sample. Examples include a change in the area under the receiver operating characteristic curve (AUC) (Tzoulaki, et al., 2009), a difference in the maximum Youden indices, integrated discriminant improvement, and net reclassification improvement (Pencina, et al

2008, Gu and Pepe, 2009, Whittemore, 2010). A fundamental limitation with these approaches is deciding how large a difference between performance measures for Models 1 and Model 2 is sufficient to deem the new marker worthwhile. Also, when these measures can be framed in terms of costs and benefits, they can assume an unrealistic cost-benefit tradeoff as with the Youden index (Baker and Kramer, 2007) or the area under the ROC curve (Hand, 2010).

A medical decision-making approach to evaluating an additional marker for risk prediction circumvents the aforementioned drawbacks of a purely statistical measure. The main challenge with using costs and benefits is to determine their relative values. A sensitivity analysis over a range of costs and benefits should therefore be considered. In a landmark paper, Vickers and Elkin (2006) introduced decision curves as a method to simplify the sensitivity analysis for the net benefit of a risk prediction model. Vickers and Elkin (2006) considered an individual risk threshold, which is the risk at which a person would be indifferent between treatment and no treatment (Pauker and Kassirer, 1980). The plot of net benefit of risk prediction versus risk threshold is called a decision curve. Baker et al. (2009) and Baker (2009) extended decision curves to relative utility curves. The relative utility is the maximum expected utility of risk prediction divided by the expected utility of perfect prediction. The relative utility curve is a plot relative utility versus a population risk threshold. The population risk threshold (called risk threshold subsequently) is the risk at which the expected utility of treatment and no treatment is the same in the population.

Baker et al. (2009) also introduced the test threshold (called here the test tradeoff to better distinguish it from the risk threshold) to evaluate a new marker at a specified risk threshold. The test tradeoff corresponding to a risk threshold is the minimum number of tests for a new marker that need to be traded for a true positive to yield an increase in the net benefit of risk prediction with the additional marker. The test tradeoff can then be considered in light of any detrimental side effects or monetary costs of testing for the marker. For example a test tradeoff of 100 could be reasonable if testing for the marker involved only a simple blood test but unreasonable if testing for the marker required an invasive biopsy. Computation of the test tradeoff at various risk thresholds provides a sensitivity analysis.

Section 2 updates the theory underlying the medical decision making approach to evaluating risk prediction; it links various formulations in terms of the Fundamental Rule of cost-benefit analysis, improves two-stage risk prediction, and distinguishes population and individual risk thresholds. Section 3 expands on graphical insights concerning ROC and relative utility curves. Section 4 presents new results in the estimation of relative utility curves and test tradeoff, introduces the risk mapping plot, and provides a new comparison of estimates.

Section 5 is a simulation; Section 6 provides an application; and Section 7 is a discussion.

2 Theory

The goal is to compare the use of risk prediction Model 1 versus risk prediction Model 2 (with an additional marker) for making a decision of treatment versus no treatment when applied to a target population. Under the benefit-cost approach, this comparison is based on the net benefit, which is the total expected benefit minus the total expected harm measured in the same units as benefit. Stokey and Zeckhauser (1978) define the *Fundamental Rule* in any choice situation as selecting the alternative that produces the greatest net benefit. We discuss various versions of the Fundamental Rule that involve different simplifications in the benefit-cost inputs. To help the reader with notation, a summary of symbols is provided in Table 1.

An important quantity in the computation of the net benefit of risk prediction is the risk score. An individual's *risk score* is the risk of developing disease computed from applying the risk prediction model in the development sample to the predictors for that individual. The risk score is considered here as a *preliminary* measure of risk, rather than as a definitive probability of an event. A value of the risk score greater than a cutpoint is an indicator for treatment and a value of the risk score less than a cutpoint is an indicator for no treatment. In this respect the risk score is no different from a measure of serum cholesterol level used to predict the risk of cardiovascular disease. Let J denote a risk score computed for an individual in the target population. Let D=1 if the individual in the target population develops disease and 0 otherwise. The probability of developing disease for an individual in the target population with risk score j is written

R_j = pr(D =1| J =j, target population)
= probability of developing disease if the risk score is j.

An important requirement of the risk score j is

Risk Score Assumption: R_j increases as j increases.

Consider a cutpoint s on the risk score such that $J \geq s$ indicates a positive classification followed by treatment, and $J < s$ indicates a negative classification followed by no treatment. The following probabilities for the target population are basic to the analysis:

P = pr(D=1 | target population)=probability of developing disease,
FPR_s = pr($J \geq s$ | D =0, target population) = false positive rate at cutpoint s,
TPR_s = pr($J \geq s$ | D =1, target population) = true positive rate at cutpoint s.

Symbol	Definition (applies to target population)
J	risk score
D	indicator of developing disease
R_j	probability of developing disease if the risk score is j
P	probability of developing disease
FPR_s	false positive rate at cutpoint s
TPR_s	true positive rate at cutpoint s
$U_{(treatdis)}$	utility of treating a person who would develop disease in the absence of treatment
$U_{(no\ treat,\ dis)}$	utility of not treating a person who would develop disease in the absence of treatment
$U_{(treat,\ no\ dis)}$	utility of treating a person who would not develop disease
$U_{(no\ treat,\ no\ dis)}$	utility of not treating a person who would not develop disease
U_{Test}	utility (harm or cost) associated with testing for or otherwise ascertaining all predictors in a model
$U_{Pred(s)}$	expected utility of risk prediction at cutpoint s
$U_{NoTreat}$	expected utility of no treatment
U_{Treat}	expected utility of treatment
$U_{Pred*(s)}$	$U_{Pred(s)} - U_{NoTreat}$
B	$U_{(treat,dis)} - U_{(no\ treat,\ dis)}$ = overall gain from treating a person who would develop disease
C	$U_{(no\ treat,\ no\ dis)} - U_{(treat,\ no\ dis)}$ =overall cost from treating a person who would *not* develop disease
T	$1 / (1 + B/C)$ = risk threshold
$ROCSLOPE_t$	slope of ROC curve at cutpoint t
NB_t	net benefit for decision curves
$U_{PerfPred}$	expected utility of perfect prediction
RU_t	relative utility at cutpoint t
ΔNB_t	$NB_{t[\text{Model 2}]} - NB_{t[\text{Model1}]}$ = increase in net benefit for decision curves
ΔU_{Test}	$U_{Test[\text{Model 1}]} - U_{Test\ [\text{Model 2}]}$ = cost of the additional test used with Model 2
$\Delta NB_{t\|UTest=0}$	maximum acceptable testing harm
$1/\ \Delta NB_{t\|UTest=0}$	test tradeoff

Table 1. Summary of symbols used in Section 2.

A utility is a numerical value for a health benefit, harm, or monetary cost, measured on a common scale. A positive value for utility indicates a benefit and a negative value indicates a cost or harm. Five utilities are associated with risk prediction: $U_{(treat,dis)}$, the utility of treating a person who would develop disease in the absence of treatment, $U_{(no\ treat,\ dis)}$, the utility of not treating a person who would develop disease in the absence of treatment, $U_{(treat,\ no\ dis)}$, the utility of treating a person who would not develop disease, $U_{(no\ treat,\ no\ dis)}$, the utility of not treating a person who would not develop disease, and U_{Test} is the utility (harm or cost) associated with testing for or otherwise ascertaining all predictors in a model. It is reasonable to assume that $U_{(treat,dis)} > U_{(no\ treat,\ dis)}$ and $U_{(no\ treat,\ no\ dis)} > U_{(treat,\ no\ dis)}$. Also $U_{Test} < 0$ because any test has some burden, harmful side effect, or monetary cost. Each of these utilities is an average of the utilities of individuals in the target population.

Using the aforementioned utilities and probabilities, we discuss various formulations for comparing the net benefit of Model 2 versus Model 1starting with the most basic formulation and then discussing various simplifications and extensions.

2.1 Comparing maximum expected utilities of risk prediction

Fundamentally, the comparison of the net benefit of Model 2 versus Model 1 is a comparison of the maximum expected utility of risk prediction under Model 2 versus Model 1. The maximum expected utility of risk prediction is the maximum, over the cutpoints, of the expected utility of risk prediction. For cutpoint s the expected utility of risk prediction is the average of the aforementioned utilities weighted by their probabilities of occurrence,

$$\begin{aligned} U_{Pred(s)} = {} & P \times TPR_s \times U_{(treat,dis)} \\ & + P \times (1 - TPR_s) \times U_{(no\ treat,\ dis)} \\ & + (1 - P) \times FPR_s \times U_{(treat,\ no\ dis)} \\ & + (1 - P) \times (1 - FPR_s) \times U_{(no\ treat,\ no\ dis)} \\ & + U_{Test}. \end{aligned} \quad (1)$$

Let subscripts “Model 1” and “Model 2” denote that the risk score based on risk prediction Models 1 and 2, respectively. In this framework, the Fundamental Rule is

Fundamental Rule Version 1:
Benefit-cost inputs are $U_{(\text{treat,dis})}$, $U_{(\text{no treat, dis})}$, $U_{(\text{treat, no dis})}$, $U_{(\text{no treat, no dis})}$, and U_{Test}.
Select Model 2 if $Max_s\{U_{Pred(s)[\text{Model 2}]}\} > Max_s\{U_{Pred(s)[\text{Model 1}]}\}$.

Typically $U_{(no\ treat,\ no\ dis)}$ is set to 0, which does not affect the Fundamental Rule. Sometimes $U_{(treat,\ dis)}$ is also set to 0 (Briggs and Zaretski, 2008; Cai et al, 2011), which is a strong assumption that is not necessary for simplification, as will be seen in Section 2.2.

2.2 Comparing maximum expected utilities of risk prediction: a simplification involving the no treatment option

Fundamental Rule Version 1 can be greatly simplified by considering the expected utility of risk prediction in excess of the expected utility of no treatment. The expected utility of no treatment is

$$U_{NoTreat} = P \times U_{(no\ treat,\ dis)} + (1 - P) \times U_{(no\ treat,\ no\ dis)}. \quad (2)$$

For later discussion it also helpful to define the expected utility of treatment,

$$U_{Treat} = P \times U_{(treat,dis)} + (1 - P) \times U_{(treat,\ no\ dis)}. \quad (3)$$

The expected utility of risk prediction in excess of the expected utility of no treatment is

$$\begin{aligned} U_{Pred*(s)} &= U_{Pred(s)} - U_{NoTreat} \\ &= P \times TPR_s \times B - (1 - P) \times FPR_s \times C + U_{Test}, \text{ where} \\ B &= U_{(treat,dis)} - U_{(no\ treat,\ dis)}, \\ C &= U_{(no\ treat,\ no\ dis)} - U_{(treat,\ no\ dis)}. \end{aligned} \quad (4)$$

The quantity B is the overall benefit of treating a person who would develop disease. The quantity C is the overall cost of treating a person who would *not* develop disease. Equation (4) with $U_{Test} = 0$ was proposed by Peirce (1884). Based on equation (4), the Fundamental Rule can be expressed as

Fundamental Rule Version 2:
Benefit-cost inputs are B, C, and U_{Test}.
Select Model 2 if $Max_s\{U_{Pred*(s)}\text{[Model 2]}\} > Max_s\{U_{Pred*(s)}\text{[Model 1]}\}$.

2.3 Comparing maximum expected utilities of risk prediction: a simplification using the risk threshold

Fundamental Rule Version 2 can be simplified even further by using the risk threshold to compute the optimal cutpoint that maximizes the expected utility of risk prediction. The risk threshold, denoted T, is the probability of developing

disease in the population at which the expected utility of treatment and no treatment is the same. Substituting T for P when setting $U_{NoTreat} = U_{Treat}$ in equations (2) and (3) gives the following formula for the risk threshold,

$$T = 1 / (1 + B/C) = (C/B) / (1+C/B), \quad (5)$$

which implies $C/B = T / (1 - T)$. The quantity C/B is sometimes called the relative utility of true- and false- positve results (Halpern et al. 1996), which should not be confused with the terminology "relative utility" discussed later. The receiver operating characteristic (ROC) curve in the target population is a plot of TPR_t versus FPR_t . As derived in Appendix A, the slope of this ROC curve at cutpoint t (more precisely between cutpoint t and cutpoint t+1) is

$$\begin{aligned} ROCSLOPE_t &= (TPR_t - TPR_{(t+1)})/ (FPR_t - FPR_{(t+1)}). \\ &= \{(1 - P) / P\} \times R_t / (1 - R_t). \end{aligned} \quad (6)$$

A fundamental result (Pauker and Kassier, 1975, Metz, 1978, Gail and Pfeiffer, 2005) is the following. Under the Risk Score Assumption, for a risk threshold of T, the maximum expected utility of risk prediction occurs at cutpoint t of the risk score that satisfies what we call the

Optimization Requirement, $R_t = T$,
which is equivalent to
$ROCSLOPE_t = \{(1 - P) / P\} \times T / (1 - T)$.

A proof of the optimization requirement is given in Appendix B. Using the optimization requirement, the Fundamental Rule can be simplified as

Fundamental Rule Version 3:
Benefit-cost inputs are B, C, and U_{Test}.
Select Model 2 if $U_{Pred*(t)[\text{Model 2}]} > U_{Pred*(t)[\text{Model 1}]}$,

where the optimal cutpoint t solves $R_t = T$.

2.4 Comparing decision curves

The Fundamental Rule Version 3 can be simplified by eliminating the separate contributions of B and C and using only the risk threshold T. This is the method of decision curves (Vickers and Elkin, 2006) who use the term "net benefit" (in a more specific manner than previously discussed) to define the following quantity,

$$NB_t = (U_{Pred(t)} - U_{NoTreat}) / B$$
$$= P \times TPR_t - (1-P) \times \{T / (1 - T)\} \times FPR_t + U_{Test} / B. \quad (7)$$

This net benefit for a decision curve, NB_t, is the maximum benefit of risk prediction (in excess of the benefit of no treatment) in units of the benefit of treating a true positive. It equals the benefit of treating a true positive after subtracting the cost of treating a false positive at an "exchange rate" based on the risk threshold. In this framework, the Fundamental Rule is

Fundamental Rule Version 4:
Benefit-cost inputs are T and U_{Test}.
Select Model 2 if $NB_{t[\text{Model 2}]} > NB_{t[\text{Model 1}]}$,

where the optimal cutpoint t solves $R_t = T$. The original form of the decision curve plotted net benefit versus individual risk threshold (Vickers and Elkin, 2006) when $U_{Test} = 0$.

2.5 Comparing relative utility curves

Additional perspective in comparing the net benefits of Models 1 and 2 can be obtained by considering the expected utility of perfect prediction. The expected utility of perfect prediction is a weighted average of the utility of treatment if disease develops and the utility of no treatment if disease does not develop,

$$U_{PerfPred} = P \times U_{(treat,dis)} + (1 - P) \times U_{(no\ treat,no\ dis)}. \quad (8)$$

The expected utility of no risk prediction is the larger of (i) the expected utility of always selecting no treatment and (ii) the expected utility of always selecting treatment. *Relative utility* (Baker, et al, 2009; Baker, 2009) is the ratio of the maximum expected utility of risk prediction (in excess of the expected utility of no risk prediction) to the expected utility of perfect prediction (in excess of the expected utility of no risk prediction), which can be written as

$$RU_t = \begin{cases} (U_{Pred(t)} - U_{Treat}) / (U_{PerfPred} - U_{Treat}), & \text{if } U_{NoTreat} < U_{Treat}, \\ (U_{Pred(t)} - U_{NoTreat}) / (U_{PerfPred} - U_{NoTreat}), & \text{if } U_{NoTreat} \geq U_{Treat}. \end{cases} \quad (9)$$

In the first case in equation (9), the expected utility of no risk prediction is U_{Treat} because $U_{NoTreat} < U_{Treat}$. In the second case in equation (9), the expected utility of no risk prediction is $U_{NoTreat}$ because $U_{NoTreat} \geq U_{Treat}$. As derived in Appendix C, equation (9) can be simplified to

$$RU_t = \begin{cases} (1-FPR_t) - (1-TPR_t) / ROCSlope_t + U_{Test} / \{(1-P) \times C\}, & \text{if } T < P, \\ TPR_t - FPR_t \times ROCSlope_t + U_{Test} / (P \times B), & \text{if } T \geq P. \end{cases} \quad (10)$$

In terms of relative utilities, the Fundamental Rule is

Fundamental Rule Version 5:
Benefit-cost inputs are T and U_{Test}.
Select Model 2 if $RU_{t[\text{Model 2}]} > RU_{t[\text{Model 1}]}$,

where the optimal cutpoint t solves $R_t = T$.

A relative utility curve for the target population is a plot of RU_t versus T when $U_{Test} = 0$. The relative utility curve includes both cases in equation (9) for completeness and to fully link it to the ROC curve (as discussed in Section 3). The relative utility curve has a single maximum where T equals P and decreases to zero on either side of the maximum (Baker, et al., 2009). A relative utility of zero means there is no benefit of risk prediction. As discussed in Section 2.9 only one case in equation (10) is usually relevant.

2.6 Computing maximum acceptable testing harm

So far the versions of the Fundamental Rule have involved an input of U_{Test}, which can be difficult to specify. Here the Fundamental Rule is inverted to find a bound related to U_{Test}. For this purpose, it is convenient to define

$$\Delta NB_t = NB_{t[\text{Model 2}]} - NB_{t[\text{Model1}]}, \quad (11)$$

and, reversing the order of models in the subtraction,

$$\Delta U_{Test} = U_{Test[\text{Model 1}]} - U_{Test\,[\text{Model 2}]}, \quad (12)$$

so that ΔU_{TEST} is positive. We can then write

$$\Delta NB_t = \Delta NB_{t|UTest=0} - (\Delta U_{Test} / B), \quad (13)$$

where $\Delta NB_{t|UTest=0}$ is the value of ΔNB_t when $U_{Test} = 0$. Inverting Fundamental Rule Version 4 and applying equation (13) gives an upper bound on $\Delta U_{Test} / B$,

Fundamental Rule Version 6:
Benefit-cost input is T.
Select Model 2 if $(\Delta U_{Test} / B) < \Delta NB_{t|UTest=0}$,

where the optimal cutpoint t solves $R_t = T$. Baker (2009) called $\Delta NB_{t|UTest=0}$, the maximum acceptable testing harm. The *maximum acceptable testing harm* corresponding to a risk threshold is the maximum increased harm of testing for an additional marker (measured as true positives not treated) so that there is an increase in net benefit with an additional marker.

2.7 Computing the test tradeoff

Rather than considering an upper bound on ($\Delta U_{Test} / B$), it is sometimes easier to consider a lower bound on ($B / \Delta U_{Test}$). This lower bound is the number of tests for a new marker that would be traded for a true positive. For example suppose B represents the benefit equivalent to 2 true positives treated and ΔU_{TEST} represents the cost equivalent to 0.10 true positives not treated. Then (2 units) / (0.10 units) = 20 is the number of tests that would be traded for a true positive. This lower bound leads to

Fundamental Rule Version 7:
Benefit-cost input is T.
Select Model 2 if $(B / \Delta U_{Test}) > (1 / \Delta NB_{t|UTest=0})$,

where the optimal cutpoint t solves $R_t = T$. Baker et al (2009) called the quantity $1 / \Delta NB_{t|UTest=0}$ the test threshold but, to avoid confusion with risk threshold, it is called here the test tradeoff. The *test tradeoff* corresponding to a risk threshold is the minimum number of tests for a new marker that need to be traded for a true positive to yield an increase in the net benefit of risk prediction with an additional marker. For example, in evaluating the addition of breast density to a model to predict the risk of invasive breast cancer, Baker (2009) computed a test tradeoff of 3333. This test tradeoff of 3333 means that 3333 measurements of breast density would need to traded for correct identification of a woman who would develop invasive breast cancer in the absence of treatment to yield an increase in the net benefit of risk prediction. Because measuring breast density has little cost or harm, this test tradeoff would likely be considered acceptable.

2.8 Computing the test tradeoff with two-stage risk prediction

Costs of additional marker testing can be reduced at some loss of performance by testing for the additional marker in only persons in the target population with intermediate levels of risk scores under Model 1. The Model 2* risk scores are the sorted (from smallest to largest) union of Model 1 risk scores not in the intermediate category with the Model 2 risk scores for individuals whose Model 1 risk scores are in the intermediate category (Figure 1). This version of two-stage

risk prediction supersedes that in Baker et al. (2009) which implicitly required that Model 2 risk scores in the intermediate category were within the range of Model 1 risk scores in the intermediate category.

The choice of intermediate levels of risk score for Model 1 needs to be specified in advance. The smaller the range of intermediate levels, the greater the cost savings, but also the less informative the Model 2* risk scores. Generally the intermediate levels would correspond to risk scores in a range where there is considerable debate over treatment or no treatment, such as in cardiovascular disease prevention (Ridker, et al. 2007).

Model 1 risk scores	|← low →|←intermediate → |← high → |
	| relevant relevant
Model 2 risk scores	|← relevant → |
Model 2* risk scores	the sorted (smallest to largest) union of relevant Model 1 risk scores (low and high) and Model 2 risk scores (intermediate)

Figure 1 Two-stage risk prediction to compute Model 2* risk scores in the target population.

Let $\Delta NB^*_t = NB_{t[\text{Model 2*}]} - NB_{t[\text{Model1}]}$. Let f denote the fraction of persons in the intermediate category. The expected cost of testing for the additional marker in the population is $f / \Delta U_{Test}$. The version of the Fundamental Rule for two-stage risk prediction is

Fundamental Rule Version 8:
Benefit-cost input is T.
Select Model 2 if $B / \Delta U_{Test} > f / \Delta NB^*_{t|UTest=0}$,

where the optimal cutpoint t solves $R_t = T$. If $f / \Delta NB^*_{(t)|UTest=0} < 1 / \Delta NB_{t|UTest=0}$, two-stage risk prediction is preferable to single-stage risk prediction, because the test tradeoff is lower.

2.9 Sensitivity analysis for risk thresholds

As a sensitivity analysis, the test tradeoff is computed at different risk thresholds. It helps to anchor the range of risk thresholds at a level accepted in practice. If the level accepted in practice is not known, a good approximation may be obtained by modifying the accepted risk threshold for a different treatment or disease to

account for different levels of harm and benefit than in the situation of interest (Baker, 1998).

The default decision of treatment or no treatment *in the absence of risk prediction* provides information about the relevant range of risk thresholds for a sensitivity analysis, called the relevant region (Baker et al, 2009). If the default strategy is no treatment, the relevant region is the set of risk thresholds greater than the probability of developing disease, which is implied by $U_{NoTreat} > U_{Treat}$. Conversely, if the default strategy is treatment, the relevant region is the set of risk thresholds less than the probability of developing disease, which is implied by $U_{NoTreat} < U_{Treat}$. Thus, the sensitivity analysis should involve risk thresholds in a subset of the relevant region.

2.10 Population versus individual risk thresholds

The test tradeoff is a summary measure of the "value" of an additional marker in a population. It assumes that a policy decision is made that all individuals with estimated risks greater than or equal to the population risk threshold are recommended for treatment and all other individuals are not recommended for treatment. The reason that the test tradeoff (and relative utility and decision curves) refer to a population quantity is that the false and true positive rates at given risk score are population quantities. The test tradeoff (and relative utility and decision curves) could also apply to a subset of the population with a particular risk threshold and the same distribution of risk scores as in the population. Thus the sensitivity analysis for various values of risk threshold T could apply to either (i) uncertainty in specifying a single population risk threshold or (ii) a range of risk thresholds among various subsets of the population with each subset having the same distribution of risk scores as in the population.

In contrast, an individual may have different utilities than the population averages and this leads to an individual risk threshold. The individual can compare his individual risk score j^* with his individual risk threshold T^*. If $j^* > T^*$, the individual should receive treatment and otherwise not receive treatment (Appendix D). This type of decision making is implicitly made in practice.

3 Graphical insights

Traditionally ROC curves have been used to compare additional markers (Tzoulaki, et al., 2009). A comparison of ROC curves for Models 1 and 2 has two downsides from a medical decision-making perspective. First ROC curves measure only classification performance and not relative utility (or net benefit of decision curves). Second, comparing classification performance at the same *FPR*

(the usual procedure) can be misleading because the same *FPR's* on different ROC curves usually corresponds to different ROC slopes and risk thresholds.

It may thus be surprising that an ROC curve can be translated into a relative utility curve using only one piece of additional information, namely the probability of developing disease. The key is that ROC and relative utility curves are connected by equation (10) with $U_{Test} = 0$, namely,

$$RU_{t|Utest=0} = \begin{cases} (1-FPR) - (1 - TPR_t) / ROCSlope_t \ , & \text{if } T < P, \\ TPR_t - FPR_t \times ROCSlope_t, & \text{if } T \geq P. \end{cases} \quad (14)$$

Equation (14) corresponds to the graphical relationship in Figure 2, which extends a plot in Baker et al. (2009) to a comparison of two models.

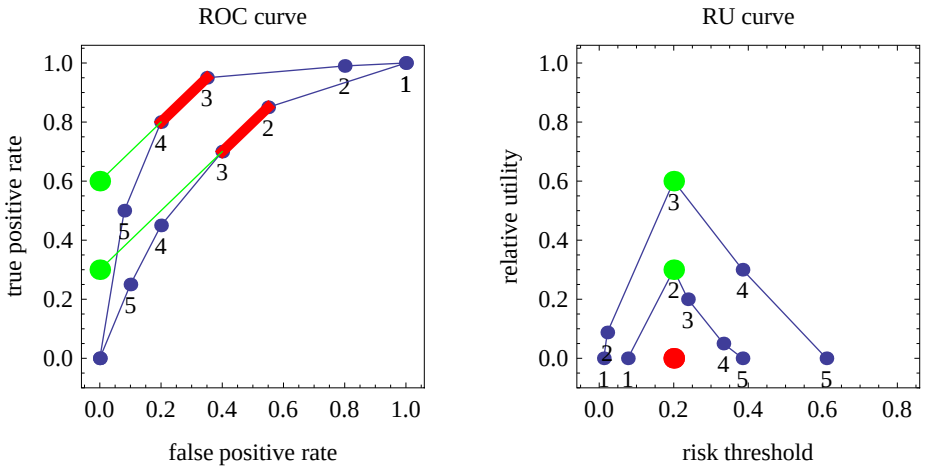

Figure 2 Translating from ROC to RU curves for Model 1 (lower curves) and Model 2 (upper curves). With $P = 0.2$, the slope of 1 of the red lines on the left side yields a risk threshold of $T = 0$ which corresponds to the red point on the right side.

Equation (14) implies that the tangent of the ROC curve at a given slope gives the relative utility on the horizontal line at $TPR=1$ (the first case) or the vertical line at $FPR=0$ (the second case).

Figure 2 illustrates the second case in equation (14). Consider the top ROC curve (for Model 2) on the left side of Figure 2 with $TPR_3 = 0.95$, $FPR_3 = 0.35$ and the $ROCSlope_3 = 1$, which corresponds to the red line segment from $t=3$ to $t=4$. The tangent of the ROC curve at $t=3$ is extended to the left vertical line at $FPR=0$ to give $RU_{3|Utest=0} = 0.6$; this value of relative utility corresponds to the top green point which is then translated horizontally to the right side of Figure 2. Because the probability of developing disease is set at $P = 0.2$, then following

equation (6), the risk threshold derived from $ROCSlope_3 = 1$ is $R_3 = 0.2$. This risk threshold is indicated by the red point on the right side of Figure 2.

Figure 2 has two main implications. First with only one additional piece of information (namely, the probability of developing disease) the relative utility curve is more informative than the ROC curve, as it allows a vertical comparison of relative utilities for each risk threshold. Second, if the ROC curves for Models 1 and 2 are similar, the relative utility curves for Models 1 and 2 will be similar. Thus small changes in ROC curves for Model 1 to Model 2 indicate little clinical utility for including an additional marker in the risk prediction model.

4 Estimation

So far we have described the theoretical underpinnings of the medical decision-making approach to evaluating an additional marker using risk thresholds in the target population. Now we turn to estimation.

The general scheme is outlined in Figure 3. The goal is to estimate the test tradeoff at various risk thresholds in the target population. A favorable test tradeoff would lead to a recommendation that individuals in the target population estimate their risk of developing disease to determine whether to receive treatment. The details of Figure 3 will be discussed.

Development sample (individual-level data): Fit risk prediction model

Validation sample (individual-level data):
- Step 1. Compute risk score (basic or adjusted) by applying risk prediction model to individual-level data.
- Step 2. Estimate risk (predicted or observed) from risk score
- Step 3. Adjust estimated risk if different probability of developing disease
- Step 4 Estimate test tradeoff (predicted, observed, or hybrid) at various risk thresholds based on estimated risks and observed outcomes. (**This is main goal of this analysis)**.

Target population (individual-level data)
(**This is a future application if the test tradeoff is favorable**).
- Step 1. Compute risk score (basic or adjusted) by applying risk prediction model to individual-level data
- Step 2 Estimate individual risk (either risk score or risk score translated to an estimate via the risk mapping plot).
- Step 3 Individuals decide on treatment based on individual risk and individual risk threshold

Figure 3 Estimation overview

4.1 Risk scores

Two types of risk scores are considered. A *basic risk score* is the risk score obtained by applying the risk prediction model from the development sample directly to predictors in the validation sample. Let $x_{VAL(hq)}$ denote predictor q for person h in the validation sample. Let a_{DEV} and b_{DEVq} denote parameter estimates from the development sample. Let expit(x)=exp(x) /{1+exp(x)}. One form for the basic risk score is

$$J_{B(h)} = \text{expit}(a_{DEV} + \Sigma_q x_{VAL(hq)} b_{DEVq}). \quad (15)$$

An *adjusted risk score* is the risk score obtained by fitting the basic risk score to the validation sample using an intercept and slope (Cox, 1958; Steyerberg et *al*. 2004). One form of the adjusted risk score is

$$J_{R(h)} = \text{expit}\{a_{VAL} + b_{VAL} \text{ logit}(J_{B(h)})\}. \quad (16)$$

where a_{VAL} and b_{VAL} are parameter estimates from the validation sample. The notation J refers to either $J_{B(h)}$ or $J_{R(h)}$, after ordering from smallest to largest risk scores.

4.2 Estimated risks

There are two estimates of risk in the validation sample:

$r_{PRED(j)} = j$ is the predicted estimate of risk,

$r_{OBS(i)}$ is the observed estimated risk for the i^{th} interval of ordered risk scores, which is based on the outcomes in the validation sample.

With a binary outcome, $r_{OBS(i)}$ equals the fraction of persons in the i^{th} interval of risk scores who develop disease. With a survival outcome, $r_{OBS(i)}$ equals 1 minus the Kaplan-Meier estimate of surviving the outcome to a pre-specified time. Unlike the observed risk, the predicted risk can be a biased in the validation sample because it is based on a model derived in the development sample. Define $r_{PREDINT(i)} = \sum_{j(i)} r_{PRED(j)} / n_i$ where $_{j(i)}$ are risk scores in the i^{th} interval of risk scores. A minimal requirement for unbiased estimation of the predicted risk is the

Calibration Condition: $r_{PREDINT(i)} \approx r_{OBS(i)}$.

Often the Calibration Condition is graphically checked by a plot of $r_{PREDINT(i)}$ versus $r_{OBS(i)}$, called a calibration plot. The necessary degree of similarity between $r_{PREDINT(i)}$ and $r_{OBS(i)}$ is difficult to specify. The Homer-Lemeshow test is sometimes used to test this similarity (Homer and Lemeshow, 2000). Alternatively the estimated parameters for intercept and slope in equation (16) to compute adjusted risk scores can provide information on the plausibility of the Calibration Condition. However, use of adjusted risk scores does not necessarily guarantee the Calibration Condition will hold.

4.3 Modified risk estimates due to sampling

For rare diseases, study costs can be reduced by creating the validation sample by taking a random sample from the target population of persons who did not develop disease and including all persons who developed disease. When using this sampling procedure, it is necessary to modify estimates of risk. In this regard, it helps to define the *validation population* as the population from which the validation sample is a random draw, as distinguished from the target population for which the validation sample is a random draw by disease status. Estimates of risk in the validation sample apply directly to the validation population but need to be modified for the target population. By definition, the validation and target populations have the same distribution of risk scores conditional on the indicator D of whether disease occurred,

$$\text{pr}(J=j \mid D=d, \text{target population}) = \text{pr}(J=j \mid D=d, \text{validation population}). \quad (17)$$

For the validation population, we define the following parameters

r_j = pr(D=1| J=j, validation population)
= risk of disease in validation population at risk score j,

w_j = pr(J=j | validation population)
= probability of risk score j in validation population,

p = pr(D=1| validation pop) $= \sum_j r_j w_j$
= probability of developing disease in validation population.

Using equation (17), the probability of risk score j in the target population is

$$\begin{aligned} W_j &= \text{pr}(J=j \mid \text{target population}) \\ &= \text{pr}(J=j \mid D=1, \text{target population})\, P + \text{pr}(J=j \mid D=0, \text{target population})\,(1-P), \\ &= \text{pr}(J=j \mid D=1, \text{validation population})\, P \\ &\quad + \text{pr}(J=j \mid D=0, \text{validation population})\,(1-P), \end{aligned} \quad (18)$$

Using equations (17) and (18), the risk in the target population is related to the risk in the validation population by the following formula (e.g., Baker, 2009; Rousson and Zumbrunn, 2011),

$$\begin{aligned} R_j &= \text{pr}(D=1 \mid J=j, \text{ target population}) \\ &= \{\text{pr}(J=j \mid D=1, \text{ target population})\, P\} \,/\, W_j, \\ &= \{\text{pr}(J=j \mid D=1, \text{ validation population})\, P\} \,/\, W_j, \\ &= (r_j\, w_j\, P\,/p) \,/\, \{r_j\, w_j\, P\,/\, p + (1 - r_j)\, w_j\, (1 - P)\,/(1 - p)\} \\ &= (r_j\, P\,/p) \,/\, \{r_j\, P\,/\, p + (1 - r_j)\, (1 - P)\,/(1 - p)\}. \\ &\equiv \mathit{RiskAdjust}(r_j, P, p) \end{aligned} \tag{19}$$

See Figure 4 for a summary of this modification.

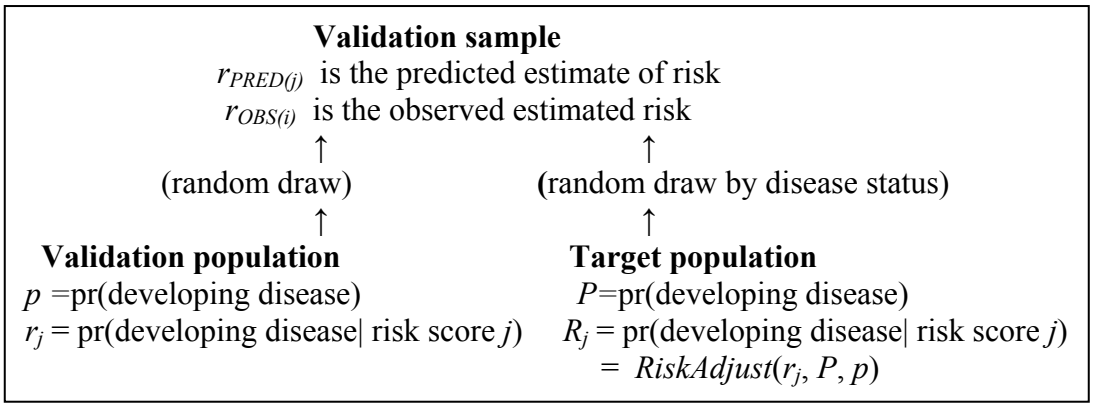

Figure 4 Risk parameters in validation and target populations.

Let n denote the number of persons in the validation sample, and let n_i denote the number of persons in the i^{th} interval of ordered risk scores in the validation sample. Based on equation (19), the estimated risks in the target population corresponding to the predicted and observed risks in the validation sample are

$$\begin{aligned} R_{PRED(j)} &= \mathit{RiskAdjust}(r_{PRED(j)}, P, p_{PRED}), \text{ where} \\ & w_{PRED(j)} = 1 / n \quad \text{and} \quad p_{\text{PRED}} = \textstyle\sum_j r_{PRED(j)} / n, \end{aligned} \tag{20}$$

$$\begin{aligned} R_{OBS(i)} &= \mathit{RiskAdjust}(r_{OBS(i)}, P, p_{\text{OBS}}), \text{ where} \\ & w_{\text{OBS(i)}} = n_i \,/\, \textstyle\sum_{\text{i}} n_i \text{ and } p_{\text{OBS}} = \sum_i r_{OBS(i)}\ w_{OBS(i)}. \end{aligned} \tag{21}$$

Based on equations (18), (A.2), and (A.3), the false and true positive rates are the same for the validation and target populations:

$$FPR_s = \sum_{s\geq j} (1-r_j)\, w_j / (1-p) = \sum_{s\geq j} (1-R_j)\, W_j / (1-P), \quad (22)$$

$$TPR_s = \sum_{s\geq j} r_j\, w_j / p = \sum_{s\geq j} R_j\, W_j / P. \quad (23)$$

4.4 Predicted estimate of test tradeoff

The predicted estimate of the relative utility curve is the estimate obtained by using only predicted risks. Let j index the ordered risk score in the validation sample. The predicted estimates of false and true positive rates and ROC slope are

$$FPR_{PRED(j)} = \sum_{j<u} (1 - r_{PRED(u)})\, w_{PRED(u)} / (1 - p_{PRED}), \quad (24)$$

$$TPR_{PRED(j)} = \sum_{j\geq u} r_{PRED(u)}\, w_{PRED(u)} / p_{PRED}, \quad (25)$$

$$ROCSLOPE_{PRED(j)} = \{p_{PRED} / (1 - p_{PRED})\}\{(1 - r_{PRED(j)}) / r_{PRED(j)}\}. \\ = \{P / (1 - P)\}\{(1 - R_{PRED(j)}) / R_{PRED(j)}\}. \quad (26)$$

Predicted estimates of relative utility curves and test tradeoff are based on $R_{PRED(j)}$, $FPR_{PRED(j)}$, and $TPR_{PRED(j)}$.

4.5 Observed estimate of test tradeoff

The observed estimate of the relative utility curve is the estimate obtained using only observed risks. To compute observed risks, a preliminary set of intervals, $i = 1, 2, \ldots, I$, is selected where, for example, each interval might correspond to a decile of predicted risk. A problem is that a preliminary estimate of the ROC curve derived from the observed risks,

$$FPR_{OBS(i)} = \sum_{i<u} (1 - r_{OBS(u)})\, w_{OBS(u)} / (1 - p_{OBS}), \quad (27)$$

$$TPR_{OBS(i)} = \sum_{i\geq u} r_{OBS(u)}\, w_{OBS(u)} / p_{OBS}, \quad (28)$$

may not be concave, violating the Risk Score Assumption.

A solution is to construct a concave ROC curve from the preliminary estimate of the ROC curve. There are at least four approaches to constructing a concave ROC curve from a non-concave preliminary ROC curve: (i) a parametric model, (ii) a semi-parametric model (e.g. Wan and Zhang, 2007), (iii) isotonic regression on the risks as a function of risk score (Lloyd, 2002), and (iv) a concave envelope (sometimes called convex hull) of the ROC points (Egan, 1975; Provost and Fawcett, 2001). We use the last approach because it gives the performance of a superior set of classification rules (as will be discussed).

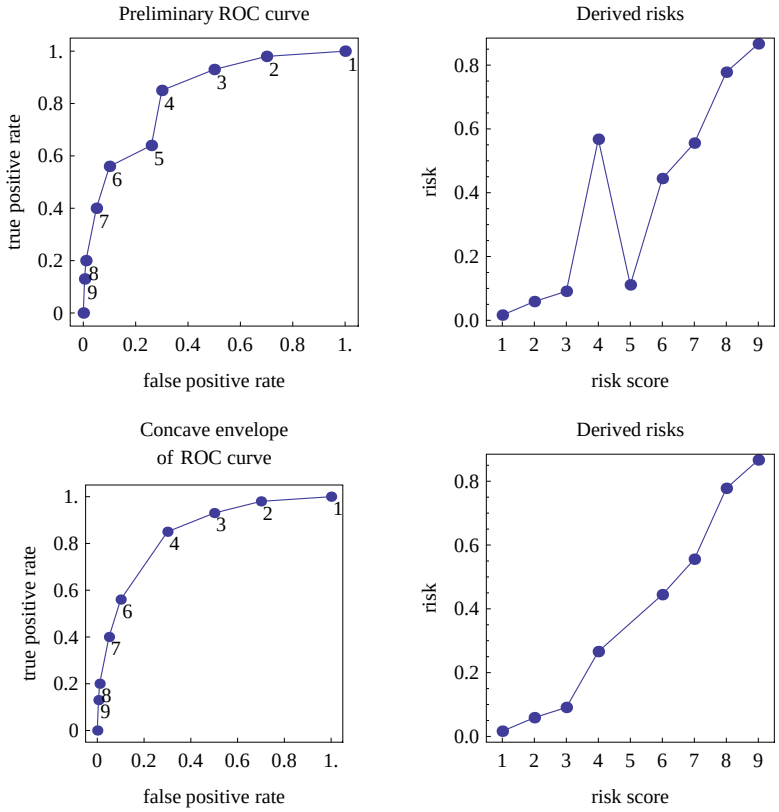

Figure 5 Construction of concave ROC curves. The plot of corresponding risks is constructed from the slopes of the ROC curve and the probability of developing disease.

The concave envelope of the ROC curve is constructed from the preliminary ROC curve by successively drawing a line from each point to the point on its right, so that the line connecting the two points has the largest slope. For example in Figure 5 (top), a preliminary non-concave ROC curve includes points 4, 5, and 6. In Figure 5 (bottom), a concave envelope of ROC points is obtained by drawing a line from point 6 to point 4, bypassing point 5 because the slope of the line from point 6 to point 4 is greater than the slope of the line from point 6 to point 5. The risk scores in Figure 5 (bottom right) are computed from the slopes of the ROC curve in Figure 5 (bottom left) and the probability of developing disease ($P = 0.20$).

The concave envelope of the ROC curve measures the performance of classification rules created by a random selection of cutpoints; this performance is better than the performance of classification rules summarized by the preliminary

ROC curve. For example, consider a classification rule based on the random selection of cutpoint j_6 (corresponding to point 6) with probability α and cutpoint j_4 (corresponding to point 4) with probability $(1-\alpha)$, for $0 \leq \alpha \leq 1$. The false and true positive rates associated with this random selection are $FPR_M = \alpha\, FPR_6 + (1-\alpha)\, FPR_4$ and $TPR_M = \alpha\, TPR_6 + (1-\alpha)\, TPR_4$. As α varies from 0 to 1, a line is created connecting points 6 and 4 on the ROC curve. This line has a higher true positive rate at the same false positive rate than classification based on the cutpoint corresponding to point 5, indicating superior classification performance.

Let $FPR_{OBSC(i)}$, $TPR_{OBSC(i)}$, and $ROCSLOPE_{OBSC(i)}$ denote estimates of false positive rate, true positive rate, and ROC slope based on the ordered cutpoint i of the concave envelope of the ROC curve. Let $R_{OBSC(i)}$ denote the observed estimate of risk derived from $ROCSLOPE_{OBSC(i)}$ and P. Observed estimates of relative utility curves and test tradeoff are based on $R_{OBSC(i)}$, $FPR_{OBSC(i)}$, and $TPR_{OBSC(i)}$.

4.6 Hybrid estimate of test tradeoff

Vickers and Elkin (2006) proposed an estimate of decision curves that can also be applied to the relative utility curve, which we call a hybrid estimate. The hybrid estimate uses outcomes in the validation sample to estimate false and true positive rates but uses the predicted estimate of risk that does not involve outcomes in the validation sample. Let $FPR_{\mathrm{OBS}(j)}$ and $TPR_{OBS(j)}$ denote the observed estimates of false and true positive rates based on ordered risk score j. With binary outcomes $FPR_{\mathrm{OBS}(j)}$ and $TPR_{OBS(j)}$ are the usual fractions. With survival data, Vickers, et al. (2008) computed $FPR_{\mathrm{OBS}(j)}$ and $TPR_{OBS(j)}$ using a Kaplan-Meier estimates for risk score j. Because this type of estimate can yield ROC curves that do not increase monotonically, Heagerty, Lumley, and Pepe (2000) introduced an estimate that avoids this drawback. Hybrid estimate of relative utility and test tradeoff are based on $R_{PRED(j)}$, $FPR_{OBS(j)}$, and $TPR_{OBS(j)}$.

4.7 Comparison of estimates of test tradeoff

We compare the three types of estimates of test tradeoff under four criteria summarized in Table 2.

Criterion 1 (necessity of Calibration Condition for unbiased estimation). The predicted and hybrid estimates of test tradeoff require the Calibration Condition for unbiased estimation because they involve predicted estimates of risk. The hybrid estimate is less sensitive to violations of the Calibration Condition than the predicted estimate because estimates of false and true positive rates do not involve predicted estimates of risk. The observed estimate of test

tradeoff does not require the Calibration Condition for unbiased estimation because it is a function of only observed estimates of risk.

	Criterion 1	Criterion 2	Criterion3	Criterion 4
Estimate of test tradeoff (estimates of risk and false and true positive rates)	Unbiased without Calibration Condition?	Do bootstrap confidence intervals from validation sample capture all variability of outcomes?	Compatibility of estimates of R with estimates of FPR_j, TPR_j	Estimated risk in target population
Predicted ($R_{PRED(j)}$, $FPR_{PRED(j)}$ $TPR_{PRED(j)}$).	no, requires Calibration Condition for $R_{PRED(j)}$, $FPR_{PRED(j)}$ $TPR_{PRED(j)}$	no, $R_{PRED(j)}$, $FPR_{PRED(j)}$ $TPR_{PRED(j)}$ do not involve outcomes	yes	risk score
Observed ($R_{OBSC(i)}$, $FPR_{OBSC(i)}$, $TPR_{OBSC(i)}$)	yes	yes	yes	risk score used with risk mapping plot
Hybrid ($R_{PRED(j)}$ $FPR_{OBS(j)}$, $TPR_{OBS(j)}$)	no, requires Calibration Condition for $R_{PRED(j)}$	no, $R_{PRED(j)}$ does not involve outcomes	no	risk score

Table 2 Comparison of estimates

Criterion 2 (role of bootstrap confidence intervals). Confidence intervals for the estimated test tradeoff are computed by bootstrapping the data (baseline variables and outcomes) in the validation sample. (Often data from the development sample are not available). For the observed estimate of test tradeoff, the confidence interval captures all the variability of outcomes in the validation sample because the observed estimate of test tradeoff is a function of only observed estimates of risks. For the predicted estimate of test tradeoff, the confidence interval does not fully capture variability because predicted estimates of R_j, FPR_j ,and TPR_j do not depend on outcomes in validation sample. For the

hybrid estimate of test tradeoff, the confidence interval does not fully capture variability because the predicted estimate of R_j does not depend on outcomes in the validation sample. Consequently, the bootstrap confidence interval is widest for the observed estimate of test tradeoff and narrowest for the predicted estimate of test tradeoff.

Criterion 3 (compatibility of estimates of risk, and false and true positive rates). A desirable property of estimates of risk and false and true positive rate is that they satisfy the fundamental relationship in equations (22) and (23). This property holds for predicted and observed estimates and ensures smooth relative utility curves. Generally this property does not hold for hybrid estimates, leading to relative utility curves that may be jagged with possibly large fluctuations in estimates of test tradeoff for small changes in the risk threshold.

Criterion 4 (risk estimates in the target population). If the test tradeoff is favorable, risk estimates would be computed for individuals in the target population in a future application. With predicted or hybrid estimates of test tradeoff, the predicted risk estimates in the target population are simply the risk scores, which is the conventional approach. With an observed estimate of test tradeoff, the risk scores need to be converted to an observed risk without the benefit of an observed outcome. This conversion may be accomplished via a *risk mapping plot* estimated from the validation sample. The horizontal axis of the risk mapping plot is the risk score and the vertical axis is the observed risk. An individual in the target population would use the plot to obtain the observed estimate of risk associated with his risk score. The risk mapping plot for the example in Section 6 is given in Figure 6.

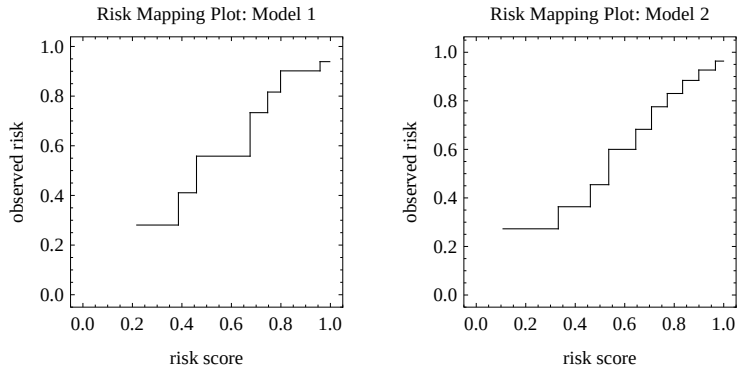

Figure 6 Mapping of risk score to observed risk estimate.

The predicted estimate of test-trade off is most appropriate if the Calibration Condition holds, smooth relative utility curves are desired, confidence intervals are not of interest, and risk scores are used to estimate risk in

future individuals . The hybrid estimate is most appropriate in the same situation as the predicted estimate but without a smooth relative utility curve and with a likely smaller bias from a violation of the Calibration Condition. The observed estimate is most appropriate when the Calibration Condition is questionable and confidence intervals are desired.

5 Simulation

To investigate the performance of different estimates, we randomly generated data under different logistic regression models in the development and validation samples. Under the simulation, the true risk in development population is

$$\text{logit}(\text{pr}(Y_i = 1|x) = 0.5 + 1.5\ x_{i1} + 1.5\ x_{i2} + 1.5\ x_{i3},$$

and the true risk in the validation population is

$$\text{logit}(\text{pr}(Y_i = 1|x) = 0.1 + 0.5\ x_{i1} + 0.5_2\ x_{i2} + 0.5\ x_{i3}.$$

The sample size was 600. Model 1 is a model fitted to x_{i1} and x_{i2}. Model 2 is a model fitted to x_{i1}, x_{i2}, and x_{i3}. As expected, the basic risk score was poorly calibrated. The adjusted risk score was also poorly calibrated for Model 1, although to a lesser degree than the basic risk score (Figure 7).

When computed from basic risk scores, the relative utility curves differed considerably among the different estimates (Figure 8). When computed from adjusted risk scores, the relative utility curves were more similar among the different estimates (Figure 9). The relative utility curve for the observed estimate is the same for the basic and adjusted risk scores because it depends only on the ranks of the risk scores, which are the same for basic and adjusted risk scores.

Estimates of test tradeoff were computed for three risk thresholds with confidence intervals computed using 2000 bootstrap replications of the validation sample. Results are shown for adjusted risk scores (Table 3). The observed or hybrid estimate of the maximum acceptable testing harm can be less than zero because these estimates involve outcomes in the validation sample. When the maximum acceptable testing harm is less than zero, the test tradeoff is labeled as "harm."

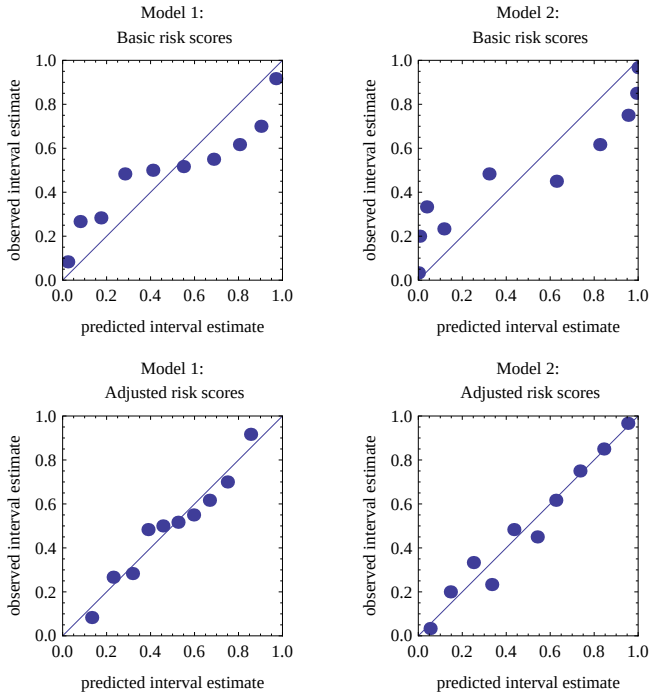

Figure 7 Calibration plots for simulated data.

	Risk thresholds		
Estimate	0.58	0.66	0.75
Predicted	15 (12, 19)	13 (11, 17)	13 (11, 16)
Observed	13 (9, 34)	12 (9, 26)	14 (9, 39)
Hybrid	11 (7, 23)	10 (7, 19)	23 (10, harm)

Table 3 Estimated test tradeoff (95% confidence intervals) in simulated data with adjusted risk scores.

For the risk threshold of 0.58 the estimated test tradeoff was similar for the three estimates, but there was a considerable difference in the widths of the bootstrap confidence intervals for reasons explained in Criterion 2 of estimation. For risk threshold 0.75, the hybrid estimate differed from the observed estimate, which may be related to Criterion 3 of estimation.

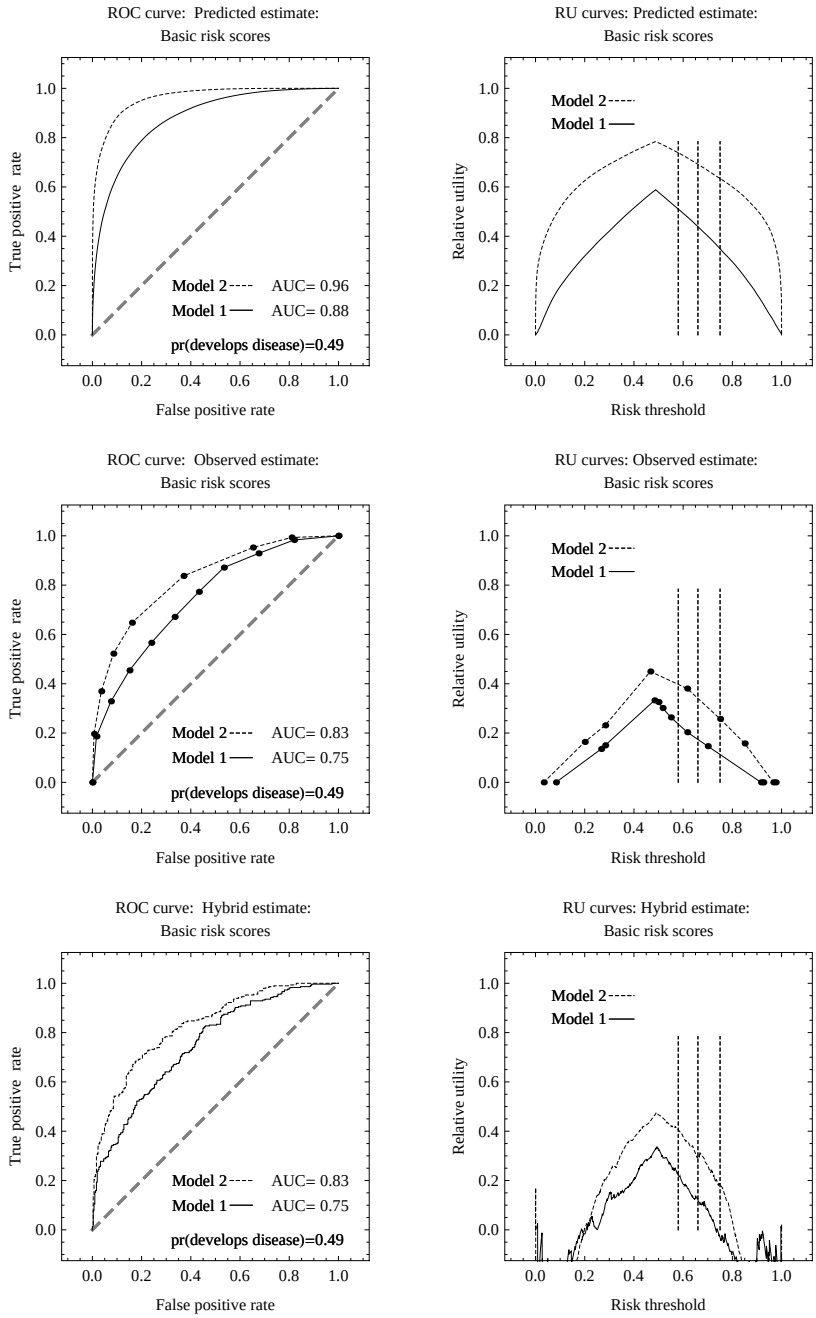

Figure 8 ROC and relative utility curves derived from basic risk scores in simulated data. Vertical lines correspond to risk thresholds in Table 3.

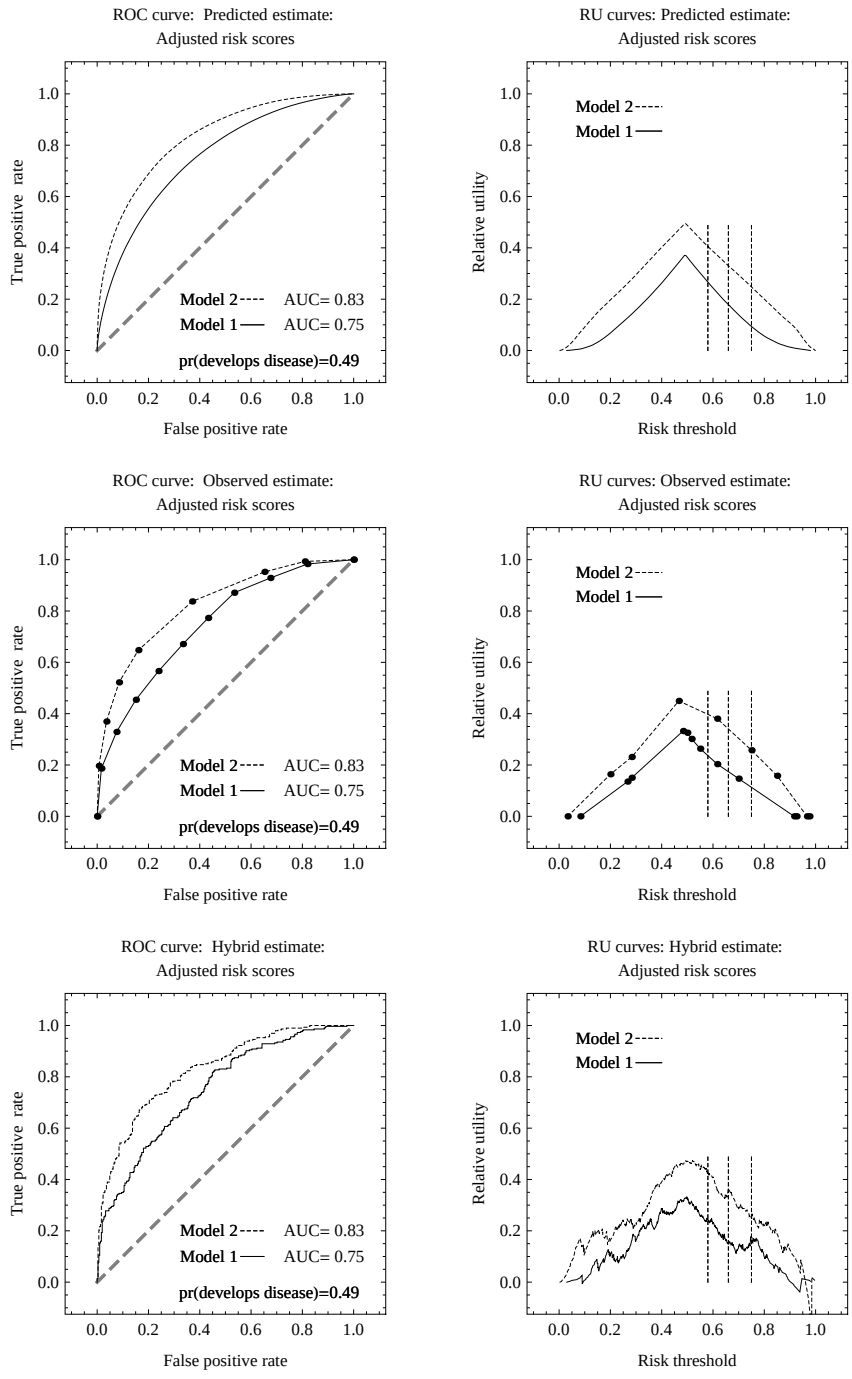

Figure 9 ROC and relative utility curves derived from adjusted risk scores in simulated data. Vertical lines correspond to the risk thresholds in Table 3.

6 Application to cancer risk prediction

We applied the methodology to a risk prediction model for the presence of residual tumor tissue in patients who received chemotherapy for testicular cancer (Steyerberg, et al., 1995; Vergouwe, et al., 2007). The development and validation samples consisted of 544 and 550 participants respectively. The fraction with disease (residual tumor) was 0.67. In the absence of risk prediction, patients receive treatment, so the relevant region for risk thresholds is less than 0.67. Treatment is removal of lymph nodes. Model 1 involves three predictors: an indicator of teratoma in the primary tumor, the size of the residual mass, and the change in mass size induced by chemotherapy. Model 2 also includes the levels of two markers in the blood, AFP and HCG. The goal is to decide if the additional markers of AFP and HCG should be included in the risk prediction model when considering risk prediction in a target population of eligible patients. The validation sample is viewed as a random draw from the target population. The calibration plot (Figure 10) shows good calibration, particularly for the adjusted risk scores.

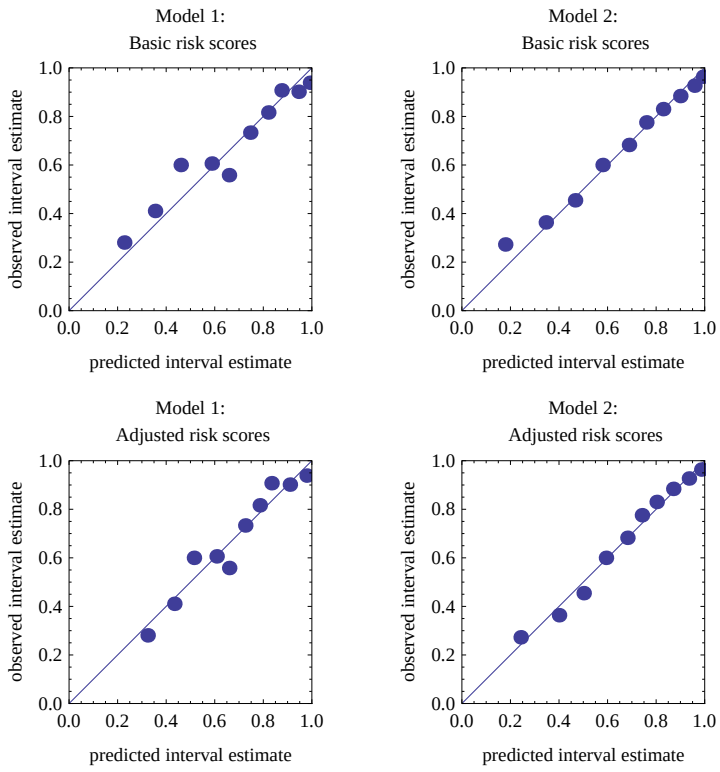

Figure 10 Calibration plots with cancer data

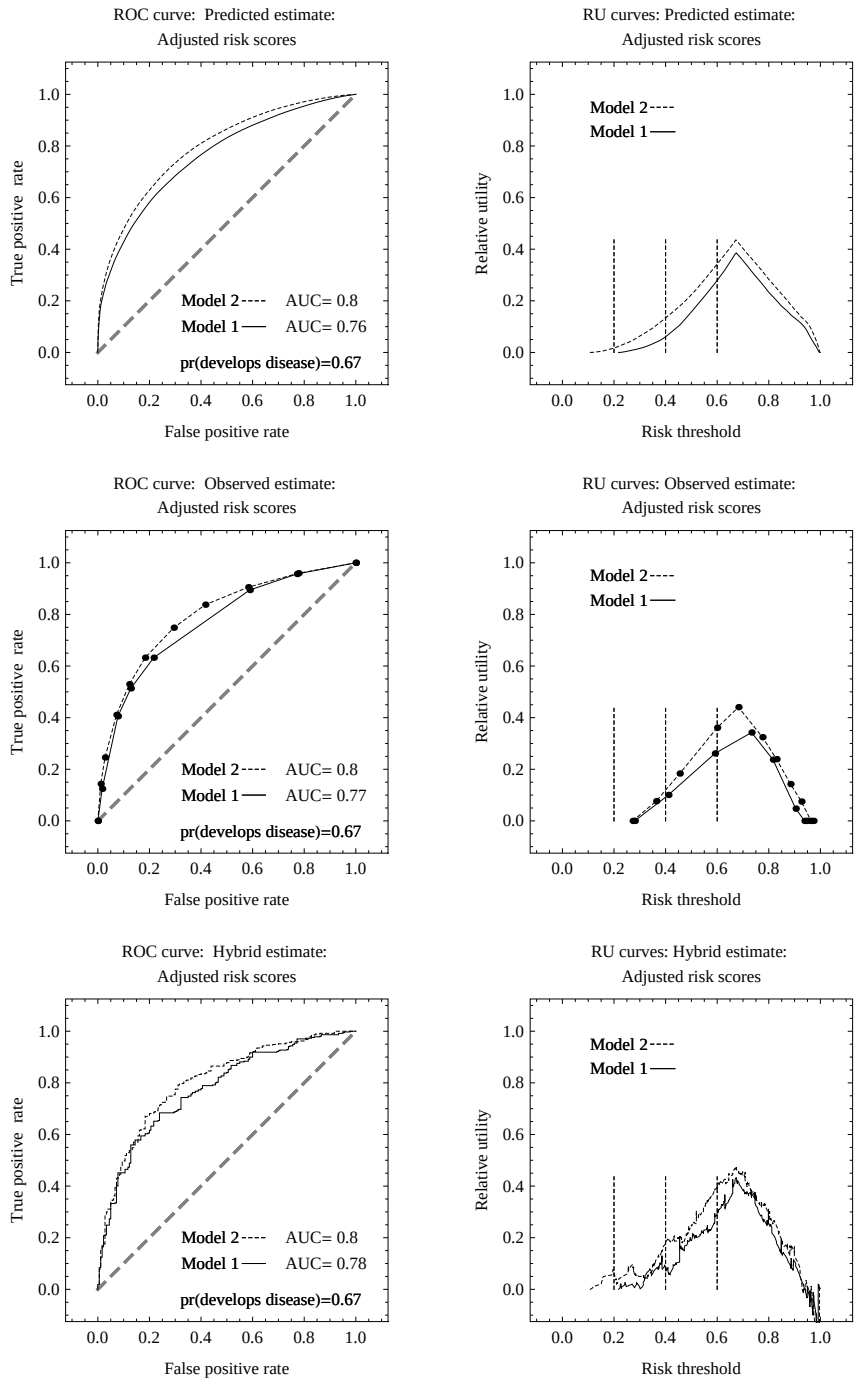

Figure 11 ROC and relative utility curves derived from adjusted risk scores for men with testicular cancer. Vertical lines correspond to the risk thresholds in Table 4

Figure 11 shows ROC and relative utility curves for adjusted risk scores. Because the ROC curves were similar for Models 1 and 2, it is not surprising, in light of results in Section 4, that the relative utility curves are similar.

The estimates of test tradeoff are shown in Table 4 for adjusted risk scores. As a sensitivity analysis, risk thresholds in the range from 0.2 to 0.4 are considered for the risk of residual tumor, in line with a formal decision analysis (Steyerberg, et al., 1999). Confidence intervals were based on 2000 bootstrap replications of outcomes in the validation sample. The predicted estimates have the narrowest confidence intervals because they are not based on outcomes of the validation sample.

As an example of the interpretation of results consider the observed estimate of test tradeoff for the risk threshold of 0.30, namely 169 with 95% confidence interval of (40, "harm"). The test tradeoff of 169 can be interpreted as follows: at least 169 marker tests need to be traded for an additional true positive to yield an increase in net benefit of risk prediction with the additional markers. Because ascertaining these markers has small monetary costs and little harm, the cost of 169 marker ascertainments seems well worth the benefit of removing lymph nodes from one patient with a residual tumor in a lymph node. However there is one major caveat, namely the upper bound on the estimated test tradeoff of "harm." Thus, there is considerable uncertainty in this estimate of test tradeoff, and researchers would need to consider the possibility that the addition of markers could make the outcomes worse for patients in the validation population.

Estimate	Risk thresholds for residual tumor		
	0.20	0.30	0.40
Predicted	712 (432, 1729)	63 (47, 91)	32 (24,49)
Observed	harm (298, harm)	169 (40, harm)	22 (13,harm)
Hybrid	188 (108,3245)	50 (24, $>10^4$)	19 (11,59)

Table 4 Estimated test tradeoff (95% confidence intervals) for testicular cancer data with adjusted risk scores.

7 Discussion

The test tradeoff at various risk thresholds is an appealing summary of the value of an additional marker in the risk prediction model. The general problem with using purely statistical measures to evaluate an additional marker is deciding how large a statistic is sufficient to declare the additional marker worthwhile. In contrast, the test tradeoff has a direct interpretation, namely the minimum number of tests for a new marker that need to be traded for a true positive to yield an increase in the net benefit of risk prediction with the additional marker. Thus the

test tradeoff accounts for the cost (monetary cost and harm) of marker ascertainment as well as the tradeoff between false and true positives summarized by the risk threshold. The more information on risk thresholds, the more focused the sensitivity analysis. The default of treatment or no treatment in the absence of risk prediction limits the range of risk thresholds to investigate. But even without any information to restrict the range of risk thresholds (which is when a purely statistical measure might be considered), a sensitivity analysis for the estimated test tradeoff could still be informative.

Somewhat remarkably, the test tradeoff is computed from relative utility curves that can be derived from ROC curves with only one piece of additional information, namely the probability of developing disease. If the ROC curves are similar the relative utility curves will be similar; However if the ROC curves differ, the relative utility curves (and the resulting test tradeoff) provide more meaningful information than the ROC curves.

Ultimately, a risk prediction model is used in a future application involving persons in a target population. The translation of risk scores into an estimated risk via a risk mapping plot (or functional equivalent) is unconventional but not difficult. In this case, the observed estimate of test tradeoff is preferred over predicted and hybrid estimates because it does not require the Calibration Condition and bootstrap confidence intervals based on the validation sample are appropriate. Generally individual-level data are needed; however, the observed estimate can be computed from tabular data (Baker, 2009). With a more conventional use of risk scores as predicted risks in a target population, the hybrid or predicted estimate of test tradeoff is appropriate if the Calibration Condition holds and there is no interest in confidence intervals. In this case, the hybrid estimate would usually be preferred over the predicted estimate because it is more robust to violations of the Calibration Condition. A general limitation with all approaches for evaluating an additional marker for risk prediction is the requirement that the validation sample be a random sample or a random sample by disease outcome from the target population.

The examples discussed here involved risk prediction models based on logistic regression but more complex multivariate approaches can be used. For example Baker (2010) created relative utility curves using a nearest-centroid analysis of microarray data (Baker, 2010). Software using Mathematica (Wolfram Research, Inc, 2010) (which also illustrates the calculations here) is available at http://dcppreview.cancer.gov/programs-resources/groups/b/software/newmarker.

Appendix A

This appendix discusses some fundamental relationships between risk and false and true positive rates. Let

W_j = pr(J =j) = probability of risk score j.

Then the following mathematical identities hold,

$$P = \sum_j R_j W_j, \tag{A.1}$$

$$TPR_s = \sum_{j \geq s} R_j W_j / P, \tag{A.2}$$

$$FPR_s = \sum_{j \geq s} (1 - R_j) W_j / (1 - P), \tag{A.3}$$

$$\begin{aligned} ROCSLOPE_s &= (TPR_s - TPR_{(s+1)}) / (FPR_s - FPR_{(s+1)}) \\ &= (R_s W_s / P) / \{(1 - R_s) W_s / (1 - P)\}. \\ &= \{(1 - P) / P\} \times R_s / (1 - R_s). \end{aligned} \tag{A.4}$$

Figure 12 shows the link between the ROC curve and the risk scores under the Risk Score Assumption.

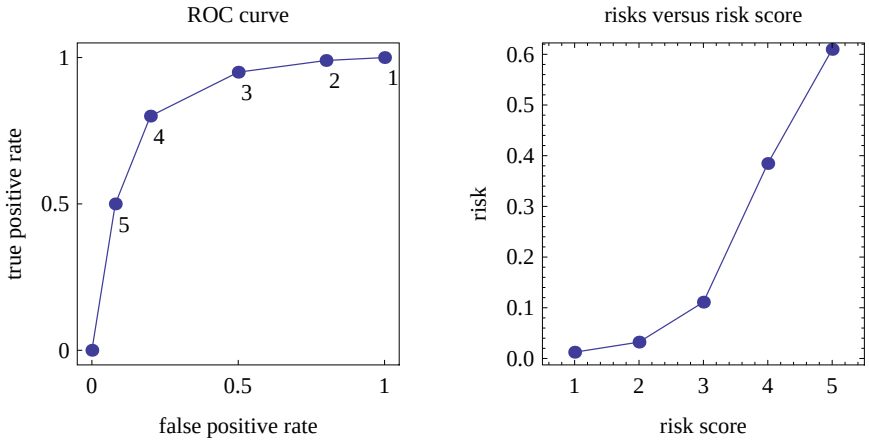

Figure 12 Relationship between the ROC curve and risk scores. Numbers next to points on the ROC curve refer to risk scores s= 1, 2, 3, 4, 5.

Appendix B

This appendix proves that the expected utility of risk prediction is maximized at $R_t = T$ when the Risk Score Assumption holds. The difference in expected utilities of prediction between risk score s and risk score (s +1) is

$$\begin{aligned} U_{Pred(s)} - U_{Pred(s+1)} &= (U_{Pred(s)} - U_{NoTreat}) - (U_{Pred(s+1)} - U_{NoTreat}) \\ &= P \times (TPR_{(s)} - TPR_{(s+1)}) \times B \\ &- (1-P) \times (FPR_{(s)} - FPR_{(s+1)}) \times C \\ &= \{P \times ROCSLOPE_s \times B - (1-P) \times C\} \\ &\quad \times \{FPR_{(s)} - FPR_{(s+1)}\}. \end{aligned} \tag{B.1}$$

If the Risk Score Assumption holds, $ROCSLOPE_s$ is increasing with s so that (B.1) equals 0 at only one point. Setting equation (B.1) equal to zero and solving for $ROCSLOPE_s$ gives the slope of the ROC curve at which the expected utility is maximized when the Risk Score Assumption holds,

$$\begin{aligned} ROCSLOPE_s &= \{(1-P)/P\} \times \{C/B\} \\ &= \{(1-P)/P\} \times \{T/(1-T)\}. \end{aligned} \tag{B.2}$$

Appendix C

Recall the formula for relative utility is

$$RU_t = \begin{cases} (U_{Pred(t)} - U_{Treat}) / (U_{PerfPred} - U_{Treat}), & \text{if } U_{NoTreat} < U_{Treat}, \\ (U_{Pred(t)} - U_{NoTreat}) / (U_{PerfPred} - U_{NoTreat}), & \text{if } U_{NoTreat} \geq U_{Treat.}, \end{cases} \tag{C.1}$$

The numerators in the formulas for the relative utility are

$$U_{Pred(t)} - U_{NoTreat} = P \times TPR_t \times B - (1-P) \times FPR_t \times C + U_{Test}, \tag{C.2}$$

$$U_{Pred(t)} - U_{Treat} = (1-P) \times (1-FPR_t) \times C - P \times (1-TPR_t) \times B + U_{Test}. \tag{C.3}$$

The denominators in the formula for the relative utility are

$$\begin{aligned} U_{PerfPred} - U_{NoTreat} &= \{P \times U_{(treat,dis)} + (1-P) \times U_{(no\ treat,no\ dis)}\} \\ &\quad - \{P \times U_{(no\ treat,\ dis)} + (1-P) \times U_{(no\ treat,\ no\ dis)}\} \\ &= P \times \{U_{(treat,dis)} - U_{(no\ treat,\ dis)}\} \\ &= P \times B, \end{aligned} \tag{C.4}$$

$$\begin{aligned} U_{PerfPred} - U_{Treat} &= \{P \times U_{(treat,dis)} + (1-P) \times U_{(no\ treat,no\ dis)}\} \\ &\quad -\{P \times U_{(treat,dis)} + (1-P) \times U_{(treat,\ no\ dis)}\} \\ &= (1-P) \times \{U_{(no\ treat,no\ dis)} - U_{(treat,\ no\ dis)}\} \\ &= (1-P) \times C. \end{aligned} \quad \text{(C.5)}$$

Therefore the relative utility when $U_{NoTreat} < U_{Treat}$ can be written

$$\begin{aligned} RU_t &= (U_{Pred(t)} - U_{Treat}) / (U_{PerfPred} - U_{Treat}) \\ &= \{(1-P) \times (1-FPR_t) \times C - P \times FNR_s \times B + U_{Test}\} / \{(1-P) \times C\} \\ &= (1-FPR_t) - (1-TPR_t) \times P/(1-P) \times B/C + U_{Test} / \{(1-P) \times C\} \\ &= (1-FPR_t) - (1-TPR_t) \times P/(1-P) \times \{(1-T)/T\} + U_{Test} / \{(1-P) \times C\} \\ &= (1-FPR_t) - (1-TPR_t) / ROCSlope_t + U_{Test} / \{(1-P) \times C\}. \end{aligned} \quad \text{(C.6)}$$

Also the relative utility when $U_{NoTreat} \geq U_{Treat}$ can be written

$$\begin{aligned} RU_t &= (U_{Pred(t)} - U_{NoTreat}) / (U_{PerfPred} - U_{NoTreat}) \\ &= \{P \times TPR_t \times B - (1-P) \times FPR_t \times C + U_{Test}\} / (P \times B) \\ &= TPR_t - FPR_t \{(1-P)/P\} \times (C/B) + U_{Test} / (P \times B) \\ &= TPR_t - FPR_t \{(1-P)/P\} \times \{T/(1-T)\} + U_{Test} / (P \times B) \\ &= TPR_t - FPR_t \ ROCSlope_t + U_{Test} / (P \times B). \end{aligned} \quad \text{(C.7)}$$

The condition for these two cases in the relative utility formula depends on

$$U_{Treat} - U_{NoTreat} = P \times B - (1-P) \times C. \quad \text{(C.8)}$$

The first case of $U_{NoTreat} < U_{Treat}$ successively implies $U_{Treat} - U_{NoTreat} > 0$, $P/(1-P) > C/B$, $P/(1-P) > T/(1-T)$, and $T < P$. Similarly $U_{NoTreat} \geq U_{Treat}$ implies $T \geq P$. Substituting the above results into the definition of relative utility in (C.0) gives the formula in the text,

$$RU_t = \begin{cases} (1-FPR_t) - (1-TPR_t) / ROCSlope_t + U_{Test} / \{(1-P) \times C\}, & \text{if } T < P, \\ TPR_t - FPR_t \times ROCSlope_t + U_{Test} / (P \times B), & \text{if } T \geq P. \end{cases} \quad \text{(C.9)}$$

Appendix D

The following derivation of an individual's risk threshold follows Pauker and Kassirer (1975) with the addition of a utility for testing costs. Let subscript "*" denote an individual's utilities. For an individual with risk score j^* the expected utility of no treatment is

$$U_{NoTreat} = \text{pr}(D=1|J=j^*) \times U^*_{(\text{no treat, dis})} + \text{pr}(D=0|J=j^*) \times U^*_{(\text{no treat, no dis})}, \quad \text{(D.1)}$$

and the expected utility of treatment is

$$U_{Treat} = \text{pr}(D=1|J=j^*) \times U^*_{(\text{treat,dis})} + \text{pr}(D=0|J=j^*) \times U^*_{(\text{treat, no dis})} + U^*_{Test}. \quad \text{(D.2)}$$

Let $B^* = U^*_{(\text{treat,dis})} - U^*_{(\text{no treat, dis})}$ and $C^* = U^*_{(\text{no treat, no dis})} - U^*_{(\text{treat, no dis})}$. The individual risk threshold T^* at which there is indifference between no treatment and treatment is obtained by setting equation (D.1) equal to equation (D.2) and solving for $\text{pr}(D=1|J=j^*)$ to obtain

$$\begin{aligned} T^* &= \{U^*_{(\text{no treat, no dis})} - U^*_{(\text{treat,dis})} + U^*_{Test}\} / \\ &\quad \{U^*_{(\text{treat,dis})} - U^*_{(\text{no treat, dis})} + U^*_{(\text{no treat, no dis})} - U^*_{(\text{treat,dis})} + U^*_{Test}\} \\ &= (C^*/B^* + U_{Test}/B^*)/ (1 + C^*/B^*). \end{aligned}$$

An individual with risk threshold T^* should receive treatment if his estimated risk is greater than T^* and no treatment if his estimated risk is less than T^*.

References

Baker, S. G. (2009). Putting risk prediction in perspective: relative utility curves. J Natl Cancer Inst 101, 1538-1542.

Baker, S. G. (2010). Simple and flexible classification via Swirls-and-Ripples. BMC Bioinformatics 11,452.

Baker, S. G., Cook, N. R., Vickers, A., Kramer, B. S. (2009). Using relative utility curves to evaluate risk prediction. J R Stat Soc [Ser A] 172, 729-748.

Briggs, W.M., and Zaretzki, R. (2008). The skill plot: a graphical technique for evaluating continuous diagnostic tests. Biometrics 63, 250–261

Cai, T., Tian, L. And Lloyd-Jones, D.M. (2011). Comparing costs associated with risk stratification rules for t-year survival. Biostatistics 12, 597-609.

Cox, D. R.. (1958) Two further applications of a model for binary regression. Biometrika 45, 62–565.

Egan, J.P. (1975). Signal detection theory and ROC analysis. New York: Academic Press.

Gail, M. H. and Pfeiffer, R. M. (2005). On criteria for evaluating models for absolute risk. Biostatistics 6, 227 -239.

Gu, W. and Pepe, M. (2009) Measures to summarize and compare the predictive capacity of markers. Int J Biostat 5, 1.

Halpern EJ, Albert M, Krieger AM, Metz CE, Maidment AD. (1996) Comparison of receiver operating characteristic curves on the basis of optimal operating points. Academic Radiology 3, 245–253

Hand, D. J. (2010). Evaluating diagnostic tests: The area under the ROC curve and the balance of errors. Stat Med. 29, 1502–1510.

Heagerty, P.J., Lumley, T., Pepe, M.S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker**.** *Biometrics*, 56, 337-344

Hosmer, D.W., Lemeshow, S. (2000). App*lied Logistic Regression*. New York: Wiley.

Justice, A.C., Covinsky, K. E., Berlin, J.A, (1999) Assessing the generalizability of prognostic information. An*nals of Internal Medicine* 130: 515-524

Khoury M. J., Mcbride, C.M., Schully S.D. (2009). The scientific foundation for personal genomics: recommendations from a national institutes of health-centers for disease control and prevention multidisciplinary workshop. Genetics in Medicine, 11, 559-567.

Mealiffe, M. E., Stokowski, R. P., Rhees, B. K., Prentice, R. L., Pettinger, M., Hinds, D. A. (2010). Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. J Natl Cancer Inst 102, 1618-1627.

Metz, C. E. (1978). Basic principles of ROC analysis. Semin Nucl Med 8, 283-98.

Pauker, S. G. and Kassirer, J. P. (1975). Therapeutic decision making: a cost-benefit analysis. N Engl J Med 293, 229-234.

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

Peirce, C .S. (1984). The numerical measure of the success of predictions. Science, 4, 453-454.

Provost, F. and Fawcett, T. (2001). Robust classification for an imprecise environment. Machine Learning, 42, 203-231.

Ridker PM, Buring JE, Rifai N, Cook N. (2007). Development and validation of improved algorithms for the assessment of global cardiovascular risk in women. *JAMA*. 297, :611-619.

Rosner, B., Colditz, G. A., Iglehart, J. D., Hankinson, S.E. (2008). Risk prediction models with incomplete data with application to prediction of estrogen receptor-positive breast cancer: prospective data from the Nurses' Health Study. Breast Cancer Res 10, R55.

Rousson, V. and Zumbrunn, T. (2011). Decision curve analysis revisited: overall net benefit, relationships to ROC curve an. *BMC Medical Informatics and Decision Making* 2011**,** 11:45.

Steyerberg, E. W., Borsboom, G. J. J. M.van Houwelingen, H.C. Eijkemans, M. J. C., Habbema, J. D. F. (2004). Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Statistics in Medicine, 23, 2567–2586.

Steyerberg, E. W., Keizer, H. J., Fosså, S. D., et al. (1995) Prediction of residual retroperitoneal mass histology following chemotherapy for metastatic nonseminomatous germ cell tumor: multivariate analysis of individual patient data from six study groups. J Clin Oncol. 13, 1177-87.

Steyerberg, E.W., Marshall, P.B., Keizer, H. J., Habbema, JDF. (1999) Resection of small, residual retroperitoneal masses after chemotherapy for nonseminomatous testicular cancer: A decision analysis Cancer 85, 1331–1341.

Stokey, E. and Zechauser R. (1978). A primer for policy analysis. W.W. Norton Company: New York.

Tzoulaki ,I., Liberopoulos, G., Ioannidis, J.P. (2009) Assessment of claims of improved prediction beyond the Framingham risk score. *Journal of the American Medical Association* 302, 2345-52.

Vergouwe, Y. ,Moons,K.G.M., and Steyerberg E.W. (2010) External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *American Journal of Epidemiology* 172, 971-980.

Vergouwe Y, Steyerberg E. W., Foster R. S., Sleijfer D. T., Fosså S. D., Gerl A., De Wit R., Roberts J. T., Habbema J. D. (2007). Predicting retroperitoneal histology in postchemotherapy testicular germ cell cancer: a model update and multicentre validation with more than 1000 patients. *Eur Urol* 51, 424- 32.

Vickers, A. J. (2011). Prediction models: revolutionary in principle, but do they do more good than harm? *Journal of Clinical Oncology* 29, 2951-2952.

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

Vickers, A. J., Cronin, A.M., Elkin, E.B., And Gonen, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC MedInform Decis Mak* 8, 53.

Wan S. and Zhang B. (2007). Smooth semiparametric receiver operating characteristic curves for continuous diagnostic tests. *Stat Med* 26, 2565-2586.

Whittemore, A. S. (2010). Evaluating health risk models. *Stat Med* 29, 2438-2452.

Wolfram Research, Inc. (2010). Mathematica, Version 8.0. Champaign, IL.