Assessing the skill of yes/no predictions

William Briggs
General Internal Medicine, Weill Cornell Medical College
525 E. 68th, Box 46, New York, NY 10021
email: wib2004@med.cornell.edu

and

David Ruppert
School of Operations Research & Industrial Engineering,
Rhodes Hall, Cornell University, Ithaca, NY 14853
email: dr24@cornell.edu

November 29, 2004

SUMMARY: Should healthy, middle-aged women receive precautionary mammograms? Should trauma surgeons use the popular TRISS score to predict the likelihood of patient survival? These are examples of questions confronting us when we decide whether to use a Yes/No prediction. In order to trust a prediction we must show that it is more valuable than would be our best guess of the future in absence of the prediction.

Calculating value means identifying our loss should the prediction err and examining the past performance of the prediction with respect to that loss. A statistical test to do this is developed. Predictions that pass this test are said to have skill.

Only skillful predictions should be used. Graphical and numerical methods to identify skill will be demonstrated. The usefulness of mammograms is explored.

KEY WORDS: Brier score; Expected loss; Mammogram testing; Skill score; Sensitivity; Specificity; Test of dependence; TRISS.

1. INTRODUCTION

We are concerned with testing the hypothesis of skill of Yes/No predictions for dichotomous events. The terms *prediction*, *forecast*, *diagnosis*, and *prognosis* are used interchangeably in this paper. Skill can be defined solely as *accuracy* or instead as *value*. Accurate predictions have a small probability of error while valuable predictions minimize risk (expected loss). Thus, value depends upon both accuracy and the costs of prediction errors.

Skill testing has many potential uses and, in particular, is important for deciding when to use diagnostic procedures in medicine. As Gigerenzer (2002, pp. 55-86) discusses, medical diagnostics such as mammogram screening of asymptomatic women may have less value and possibly less accuracy than the naive forecast (always predicting no disease), depending on the costs of false positives and negatives and their probabilities. This is true even if the sensitivity (probability of a positive given disease) and specificity (probability of a negative given no disease) are very high (Berry, 1998, 2002; Bluman et al., 1999). The asymmetry of the error loss and the genuine (and often ignored) importance of the false positive costs are returned to latter in Section 5 (Thorne et al., 1999). Asymmetric losses are often used, however, in classification problems (e.g. Hastie et al., 2001) where their motivation has traditionally been stronger.

Tests have been developed for comparing competing predictions in specific applications. Parker and Davis (1999) specifically address testing the efficacy of predictions in medical contexts, a topic pursued here. Diaconis (1978), Diaconis and Graham (1981), and Diaconis and Mosteller (1989) give a formal hypothesis test of skill for card guessing experiments with feedback. Hamill (1998) showed how to use simple statistical tests to evaluate quantitative precipitation predictions. Diebold and Mariano (1995) and Diebold (2001) develop a test based on the mean squared error (MSE) score for continuous-valued observations. Kolb and Stekler (1993) have a test that is similar to the MSE test. Briggs and Levine (1998) use the bootstrap to test for a difference in prediction accuracy. But, overall, relatively little has been done to develop formal tests of skill. Mozer and Briggs (2003) first introduced a simplified presentation of the skill score given below; a full theoretical treatment is given here.

Much more work has been done on prediction verification, most notably by Murphy (Murphy, 1991; Murphy, 1997; Murphy and Winkler, 1987; Murphy and Ehrendorfer, 1987; Ferrell, 1994; to name only a few). Many others have also contributed to this field (for example, Meeden, 1979; Solow and Broadus, 1988), but the key concept we use here is due to Murphy and Winkler (1987), which is that the joint distribution of the observations and predictions provide the framework for a complete verification analysis. Specifically, for an i.i.d. sequence of dichotomous predictions (X) and observations (Y), we model this joint distribution using a calibration-refinement factorization of the probability density (namely: $P(Y,X) = P(Y|X)P(X)$, where Y represents the observation and X the forecast; see Murphy, 1994). We assume that prediction and observation pairs are independent and therefore non-dynamic in the sense that future predictions do not depend on previous events or previous predictions. In a separate paper, we relax the independence assumption and allow the observations to be a first-order Markov chain (Briggs and Ruppert, 2004). Work in modeling precipitation series as Markovian was done, for example, by Wilks (1991).

Two concepts should be distinguished: prediction accuracy and prediction value. Analysts have long recognized that predictions can be accurate (having a high proportion of correct predictions) but without value. Likewise, a prediction might have low accuracy but still possess value. The distinctions between these two concepts are laid out in, among other places, Murphy (1991), Thompson and Zucchini (1990), Murphy and Winkler (1987), and Murphy and Ehrendorfer (1987). We shall show how the the concepts of value and accuracy combine to define skill.

In Section 2 we start with the simplest case and define the optimal naive prediction, X^N, of a binary Y to be the optimal prediction knowing only the unconditional probability of the event, Y. By *expert* we mean any prediction not the optimal naive prediction. Skill of an expert prediction, X^E, over X^N is defined in two ways. We assume no loss for accurate forecasts: loss (or cost) for accurate forecast is easily added to our framework, though we don't do so here (see Briggs, 2004; Katz, 1993).

Skill scores of these tests are formed in Section 3; their relation to the popular Brier score is also shown there. The new skill score is simple to calculate, which should ease its acceptance. We compare the skill test to the usual test for independence in Section 4; we also show how these two tests relate to the test developed in Parker and Davis (1999).

Examples of each test of skill and of the corresponding skill scores are given in Section 5 where medical diagnoses data of Parker and Davis (1999), mammogram screening data of Gigerenzer (2002), and data on the TRISS model (Boyd et al, 1987) used in trauma research are analyzed. Section 6 presents concluding remarks and suggests new lines of research.

2. Basic Test

We are concerned with binary Y. Predictions for Y can be either dichotomous ($\widetilde{X} \in \{0,1\}$) or probabilistic ($\widetilde{X} \in [0,1]$), but we only consider decisions based on dichotomous predictions.

We follow the notation developed in Schervish (1989). Let $Y_i \in \{0,1\}$ designate the ith observation of a dichotomous event. Using the function loss k_{yx}, let the losses k_{11} and k_{00} of correct decisions equal 0. The finite loss k for making an error can always be quantified such that the total loss is normalized to 1, so that with $Y = 0$ and decision d_1 the loss can be written as some $k_{01} = \theta < 1$, which implies that with $Y = 1$ and decision d_0 the loss is $k_{10} = 1 - \theta$.

The decision maker minimizes the expected loss and makes decision d_1 whenever $\tilde{X} \geq \theta$ or makes decision d_0 if $\tilde{X} < \theta$, where $\tilde{X}$ indicates a (possibly probabilistic) expert prediction. The loss for an expert prediction can be now written as

$$k = \theta I(\tilde{X} \geq \theta, Y = 0) + (1 - \theta) I(\tilde{X} < \theta, Y = 1). \tag{2.1}$$

(A user might not choose this decision rule and instead pick $X = I(\widetilde{X} > h)$ for some $h \in (0, 1)$, h not necessarily θ. Users may do this if they believe, for example, that the forecast is not calibrated (McClelland and Bolger, 1994). The work below does not depend on the particular decision rule, though the θ rule is used as a reference.)

Skill will be framed in terms of expected loss (risk). A collection of predictions has skill if its expected loss is less than that of the optimal naive predictions. (Typical definitions of skill, e.g., Wilks (1995), refer to skill as relative accuracy of an expert to a naive prediction, a distinction used when developing skill scores below.) The naive information we have about Y is that we know $P(Y = 1)$, the unconditional probability of occurrence, so that skill, if it exists, is known as *simple*, *base rate*, or *climate skill* to reflect the idea that the expert prediction can beat the simple prediction. The expected loss for a decision maker for a collection of expert predictions is

$$\begin{aligned} E(k^E) = {} & E\left[\theta I(\tilde{X} \geq \theta, Y = 0) + (1 - \theta) I(\tilde{X} < \theta, Y = 1)\right] \\ = {} & \theta P(\tilde{X} \geq \theta, Y = 0) + (1 - \theta) P(\tilde{X} < \theta, Y = 1) \\ = {} & \theta P(X = 1, Y = 0) + (1 - \theta) P(X = 0, Y = 1). \end{aligned} \tag{2.2}$$

The last step uses the definition that $X = I(\tilde{X} \geq \theta)$ is the transformation of the probabilistic prediction to a dichotomous one. Unless otherwise indicated, it shall be assumed that the (transformed if necessary) dichotomous predictions are used (this transformation differs from that of Mason (1979) whose goal was to maximize the expected value of various scores that differed from loss functions).

Thus, let $X_i \in \{0, 1\}$ designate the ith (possibly transformed) prediction. We assume that the observations (Y_i, X_i) are i.i.d. In particular, Y_i and X_j are independent for $i \neq j$, so the prediction and observation process is not dynamic and future observations do not depend on past predictions. Using standard notation for 2 × 2 tables (Bishop et al., 1975; Fienberg, 1980), (2.2) becomes

$$E(k^E) = \theta p_{01} + (1 - \theta)p_{10}. \tag{2.3}$$

The expected loss for the optimal naive prediction depends both on p_{1+} and on the value of θ. If $p_{1+} \leq \theta$ the optimal naive prediction is $X^N = 0$, where the superscript N denotes "optimal naive". This gives an expected loss of $p_{1+}(1 - \theta)$. If $p_{1+} > \theta$ the optimal naive prediction is $X^N \equiv 1$ and the expected loss is $(1 - p_{1+})\theta$. It is convenient, though not necessary, to transform both the observations and the loss so that the optimal naive prediction is always $X^N = 0$ (so that $p_{1+} \leq \theta$). With the assumption that $X^N \equiv 0$, we have

$$E(k^N) = p_{1+}(1 - \theta). \tag{2.4}$$

The null hypothesis for the skill test is

$$H_0 : \quad E(k^E) \geq E(k^N) \tag{2.5}$$

where k^E and k^N correspond to the loss of the expert and optimal naive predictions, respectively, and expectation is taken over both predictions and observations. Substituting for the expected loss, and noting that $X^N \equiv 0$, (2.3)–(2.5) give

$$\theta p_{01} + (1 - \theta)p_{10} \geq p_{1+}(1 - \theta) \;\Rightarrow\; \theta p_{01} \geq p_{11}(1 - \theta) \;\Rightarrow\; \theta \geq \frac{p_{11}}{p_{+1}} \;\Rightarrow\; p_{1|1} \leq \theta,$$

so (2.5) is equivalent to

$$H_0 : p_{1|1} \leq \theta. \tag{2.6}$$

The alternative is that $p_{1|1} > \theta$.

Theorem 2.1. *Skill defined in terms of expected loss is the same as skill defined in terms of accuracy when the loss is symmetric, that is, when $\theta = 1/2$.*

Proof. Let $\theta = 1/2$. The null hypothesis is

$$\begin{aligned} H_0: \quad P(Y = X^E) &\leq P(Y = X^N) \\ \Rightarrow p_{1|1}p_{+1} + (1 - p_{1|0})(1 - p_{+1}) &\leq (1 - p_{1|1})p_{+1} + (1 - p_{1|0})(1 - p_{+1}) \\ \Rightarrow p_{1|1} &\leq 1/2. \end{aligned}$$

This is identical to the original definition of skill with $\theta = 1/2$, as was to be proved. □

The likelihood of the model for Y, X, written in terms of $p_{1|1}$, $p_{1|0}$, and p_{+1}, is

$$L(p_{1|1}, p_{1|0}, p_{+1}|Y, X) =$$

$$\prod_{i=1}^{n} p_{+1}^{X_i}(1 - p_{+1})^{1-X_i} p_{1|1}^{X_iY_i}(1 - p_{1|1})^{X_i(1-Y_i)} p_{1|0}^{(1-X_i)Y_i}(1 - p_{1|0})^{(1-X_i)(1-Y_i)}$$

(2.7)

The estimates for these parameters are found from the cell counts of a 2 × 2 observation and prediction contingency table in Table 1. The unrestricted maximum likelihood estimates (MLEs) are the same as in ordinary 2 × 2 tables:

$$\widehat{p}_{+1} = \frac{n_{11} + n_{01}}{n_{++}}, \qquad \widehat{p}_{1|1} = \frac{n_{11}}{n_{11} + n_{01}}, \qquad \widehat{p}_{1|0} = \frac{n_{10}}{n_{10} + n_{00}}.$$

Under the null hypothesis that $p_{1|1} \leq \theta$, the MLE for p_{+1} and the estimate for $p_{1|0}$ remain unchanged and the MLE of $p_{1|1}$ is $\widetilde{p}_{1|1} = \min\{\frac{n_{11}}{n_{11}+n_{01}}, \theta\}$. Calculation of the likelihood ratio statistic (LRS or G) is simple as the terms involving p_{+1} and $p_{1|0}$ drop out and

$$G = -2\log\left[\left(\frac{\widetilde{p}_{1|1}}{\widehat{p}_{1|1}}\right)^{n_{11}}\left(\frac{1 - \widetilde{p}_{1|1}}{1 - \widehat{p}_{1|1}}\right)^{n_{01}}\right] = 2n_{11}\log\left[\frac{\widehat{p}_{1|1}}{\widetilde{p}_{1|1}}\right] + 2n_{01}\log\left[\frac{1 - \widehat{p}_{1|1}}{1 - \widetilde{p}_{1|1}}\right].$$

When $\frac{n_{11}}{n_{+1}} > \theta$ then $\widetilde{p}_{1|1} = \theta$ and $G > 0$, and when $\frac{n_{11}}{n_{+1}} \leq \theta$ then $\widetilde{p}_{1|1} = \widehat{p}_{1|1}$ and $G = 0$. This allows us to rewrite G as

$$G = \left(2n_{11} \log \left[\frac{n_{11}}{n_{+1}\theta} \right] + 2n_{01} \log \left[\frac{n_{01}}{n_{+1}(1-\theta)} \right] \right) I \left(\frac{n_{11}}{n_{+1}} > \theta \right) \tag{2.8}$$

where $n_{+1} = n_{11} + n_{01}$ and $0 \log(0) = 0$.

G has asymptotic distribution $1/2\chi_0^2 + 1/2\chi_1^2$ (Self and Liang, 1987; their case 5). Tests are carried out similarly to a standard χ_1^2 test, except the chosen test level is doubled, or equivalently one divides the p-value by 2. Examples are given in Section 5.

We conducted simulations (not shown) on G and found that the approximation to the distribution of G improves as p_{1+} approaches θ and becomes worse as p_{1+} approaches 0. As p_{1+} nears 0, the approximation becomes worse in that more observations are needed to approach the asymptotic distribution of G. But as p_{1+} approaches θ the approximation is fairly good for sample sizes n over 20, becoming excellent for $n \geq 50$. For small n and p_{1+} the test is conservative (rejects too infrequently) because $P(G = 0)$ is larger than the asymptotic value: therefore, users of the test should not give undue weight to the test with small n.

2.1. **Calibration.** A collection of predictions need not be calibrated to be skillful. Calibration is defined as having $P(Y = 1|X = x) = x$ for all x (McClelland and Bolger, 1994). In the two decision problem X can take only two values, 0 and 1, so that if a collection of predictions was empirically calibrated then $\widehat{P}(Y = 1|X = 1) = \widehat{p}_{1|1} = 1$ and $\widehat{P}(Y = 1|X = 0) = \widehat{p}_{1|0} = 0$. This, of course, means that calibrated dichotomous predictions are perfect ($Y_i \equiv X_i$) and hence skillful. The set of probability predictions (if any) that are transformed to dichotomous predictions need not be calibrated to be skillful. Calibrated probability predictions are not a guarantee of skill, e.g., $\widetilde{X} \equiv p_{1+}$ is calibrated, but is just the optimal naive prediction.

3. Skill Score

Typically one wants both a test of skill and a score that measures this skill. Such a score is useful, for example, in tracking skill for a system of predictions over time or for comparing predictions for similar events. Skill scores are in widespread use in the meteorological community (Wilks, 1995; Kryzysztofowicz, 1992). Normally, skill scores K take the form

$$K(y, x^E) = \frac{S(y, x^N) - S(y, x^E)}{S(y, x^N)}, \tag{3.1}$$

where $S(y, x^N)$ and $S(y, x^E)$ are some error scores (usually a distance or loss: square error is common) for collections of naive and expert predictions, respectively. The divisor normalizes the error scores so that rough comparisons can be made between skill scores across different situations. Scores of the type (3.1) are not proper (Winkler, 1996). A proper score is one in which $E_p(K(Y, p)) \leq E_p(K(Y, x))$ for all $x, p \in [0, 1]$, and reflects the idea that the forecaster can only minimize the loss by ensuring that the forecast x is the same as his true feeling p (Hendrickson and Buehler, 1971). Winkler shows a proper score related to (3.1) is $K(y, x^E) = S(y, x^N) - S(y, x^E)$; however, this loses the desirable normalization of (3.1). Scores in the form of (3.1) are in widespread use, and the departure from properness does little harm in that a forecaster will find it difficult if not impossible to manipulate the skill score advantagously; moreover, skill scores of this form are approximately proper for large samples (Murphy, 1973)).

The difficulty with skill scores has been that their sampling distributions were unknown making hypothesis testing impossible. However, in our case testing the significance of a skill score is the same as the climate skill test if the following skill score is taken (using (2.3) and (2.4))

$$K_\theta = \frac{E(k^N) - E(k^E)}{E(k^N)} = \frac{p_{+1}(p_{1|1} - \theta)}{p_{1+}(1 - \theta)} \tag{3.2}$$

where the expected prediction loss defined in (2.2) is taken as the error score. A collection of perfect expert predictions will have a loss of 0, so, for us, a perfect skill score will be $K_\theta \equiv 1$. A collection with "negative" skill, as defined in (2.5), will have either an expected loss the same as the naive predictions or even greater so that the skill score will be 0 or less. The null hypothesis is

$$H_0: \quad K_\theta \leq 0. \tag{3.3}$$

It can be easily seen that this translates exactly to the hypothesis and test used before defined in (2.6).

An estimate for the skill score is

$$\widehat{K}_\theta = \frac{(\widehat{p}_{1|1} - \theta)\widehat{p}_{+1}}{\widehat{p}_{1+}(1-\theta)} = \frac{n_{11}(1-\theta) - n_{01}\theta}{(n_{11} + n_{10})(1-\theta)}. \tag{3.4}$$

For general verification purposes a plausible loss is symmetric loss, that is $\theta = 1/2$. Symmetric loss is further justified below. Symmetric loss gives

$$\widehat{K}_{1/2} = \frac{n_{11} - n_{01}}{n_{11} + n_{10}}. \tag{3.5}$$

This has a particularly simple form which shows easily whether predictions have skill: this is when $n_{11} > n_{01}$, which makes $\widehat{K}_{1/2} > 0$. Our score for symmetric loss is also similar in form to other skill scores which are summarized in, among other places, Wilks (1995).

3.1. **Brier score.** The most popular score is the Brier score, (Brier, 1950) which is $B(X) = E\{(Y - X)^2\}$. The Brier score is symmetric, predictions that are wrong in either direction receive the same weight. A Brier score $B = 0$ indicates a correct prediction, while a $B = 1$ indicates an incorrect prediction. Note again that the optimal naive prediction when $P(Y = 0) > 1/2$ is $X^N = 0$. We have

$$E\{B(X^N)\} = p_{1+}, \quad \text{and} \quad E\{B(X^E)\} = p_{10} + p_{01}. \tag{3.6}$$

Our climate skill score and the Brier score have an interesting relationship.

Theorem 3.1. *Testing for skill using the Brier score, where skill is defined as when a collection of expert predictions have a lower Brier score than the Brier score for the optimal naive predictions, is equivalent to the climate skill test with symmetric loss.*

Proof. We have by (3.6) that

$$H_0: \ E\{B(X^E)\} \geq E\{B(X^N)\} \Rightarrow p_{10} + p_{01} \geq p_{11} + p_{10} \Rightarrow p_{01} \geq p_{11} \Rightarrow p_{1|1} \leq 1/2,$$

which is identical to the climate skill test with $\theta = 1/2$, that is, symmetric loss. Note: The Brier score itself is generated by integrating equation (2.2) (the loss) over all possible values of θ and assigning equal weight to each possible value. This technique of generating general scores is developed in Schervish (1989); it should also be noted that “scores” are not necessarily skill scores. □

That testing the Brier score is equivalent to testing for climate skill under symmetric loss suggests using (3.5) for general scoring. It is also useful to say something more about when a Brier score is skillful.

Theorem 3.2. *A collection of predictions has skill (with symmetric loss) when $\widehat{B} < \widehat{p}_{1+}$.*

Proof. We know that

$$\widehat{B} = \frac{\sum(y-x)^2}{n} = \frac{\sum y}{n} - \frac{\sum y}{\sum y}\frac{(2\sum yx - \sum x)}{n} = \widehat{p}_{1+}\left(1 - \frac{(2\sum yx - \sum x)}{\sum y}\right)$$

$$= \widehat{p}_{1+}\left(1 - \frac{2n_{11} - (n_{11} + n_{01})}{n_{11} + n_{10}}\right) = \widehat{p}_{1+}(1 - \widehat{K}_{1/2}).$$

And a collection of prediction has skill (with symmetric loss) when $\widehat{K}_{1/2} > 0$, so that $\widehat{B} < \widehat{p}_{1+}$ as was to be proved. □

4. Skill and Dependence

It is helpful to note how the test for climate skill relates to the more common test of independence in 2 × 2 contingency tables. A collection of predictions that have no association satisfy the null hypothesis

$$H_0: \quad p_{1|1} = p_{1|0} = p_{1+}. \tag{4.1}$$

The alternative is that $p_{1|1} \neq p_{1|0}$, which states that dependence between the predictions and observations exists. The estimate for p_{+1} is the same for under the null and alternative and the LRS is

$$\begin{aligned} LRS &= -2\log\left[\left(\frac{\widehat{p}_{1+}}{\widehat{p}_{1|1}}\right)^{n_{11}}\left(\frac{1-\widehat{p}_{1+}}{1-\widehat{p}_{1|1}}\right)^{n_{01}}\left(\frac{\widehat{p}_{1+}}{\widehat{p}_{1|0}}\right)^{n_{10}}\left(\frac{1-\widehat{p}_{1+}}{1-\widehat{p}_{1|0}}\right)^{n_{00}}\right] \\ &= 2\sum_i\sum_j n_{ij}\log\left(\frac{n_{ij}}{\pi_{ij}}\right), \end{aligned}$$

where $\pi_{ij} = n_{i+}n_{+j}/n_{++}$. This is the familiar LRS for independence in 2 × 2 tables (Agresti, 1990).

Parker and Davis (1999) developed a test that is similar to the climate skill test. The *sensitivity* $= P(X = 1|Y = 1) = p_{11}/p_{1+}$ and *specificity* $= P(X = 0|Y = 0) = p_{00}/p_{0+}$ are commonly used to describe the efficacy of a diagnostic procedure. Parker and Davis define accuracy as sensitivity plus specificity, a sum that must be greater than one for the observed accuracy to be larger than chance. This implies a null hypothesis that $p_{11}/p_{1+} + p_{00}/p_{0+} \leq 1$ or equivalently that

$$H_0: \frac{p_{11}}{p_{1+}} \leq \frac{p_{01}}{p_{0+}}.$$

This is slightly weaker than the base-rate skill test for symmetric loss in the following way. Assume that we are in the familiar case where $p_{1+} \leq 1/2$. The null hypothesis of no accuracy according to Parker and Davis can be restated as $p_{1|1} \leq p_{0|1}$ since when $p_{1+} \leq 1/2$

$$H_0 : \frac{p_{11}}{p_{1+}} \leq \frac{p_{01}}{p_{0+}} \Rightarrow p_{11} \leq p_{01} \Leftrightarrow p_{1|1} \leq p_{0|1}. \tag{4.2}$$

This is a restatement of the null hypothesis of independence, except that it is one-sided, not two-sided.

Sensitivity and specificity are also used in ROC curve analysis for diagnoses where a threshold of a diagnostic test is picked as the forecast (Zhou et al. 2002). In the standard 2 × 2 case (only one level of a diagnostic test) the area under the ROC curve is commonly used. This is equivalent to a test on the difference between two proportions (Tsimakas et al., 2002) which, of course, is different than the skill test.

5. Examples

This section gives examples and graphical techniques that allow one to assess the value of predictions of dichotomous events. The skill score developed in this work lets the decision maker specify a loss for prediction error (both false positives and false negatives) and shows whether a set of expert predictions has value or skill. Two decision makers with different losses could come to different conclusions: the prediction may be valuable for one decision maker but useless to the other. Several examples will show the wide applicability of these skill scoring methods.

5.1. **Data.**

Simple Diagnoses: Parker and Davis (1999) present data from two alternative medical practitioners who predict the presence or absence of a disease.

Mammograms: Gigerenzer (2002) summarized data from the first mammogram screening of 26,000 American women over 30 years old to detect breast cancer (Gigerenzer, p. 45 & p. 261). The data are in the form of a Performance

Table for a typical 1000 women in this study. The predictions say a woman does or does not have breast cancer.

TRISS: Boyd et al. (1987). The TRISS model, widely used in trauma research, is a logistic regression predicting probability of death. The model has three inputs: a Revised Trauma Score (RTS), an Injury Severity Score (ISS), and an indicator of Age greater than 54 years. The TRISS model will be compared a revised TRISS model to see whether the revised model is better.

5.2. **Simple Diagnoses.** As an example of a simple test, Parker and Davis (1999) evaluated the accuracy of two alternative medicine practitioners (Q1 and Q2) who were to diagnose the presence ($Y = 1$) or absence ($Y = 0$) of disease in 36 patients. From Tables 2 and 3, we see that Q2 has used the optimal naive diagnoses: the probability of disease in this sample was less than 1/2, so that the optimal naive diagnosis predicts “no disease” in each patient. Q2 obviously has no skill.

Q1 has attempted more than just parroting the optimal naive prediction, but has still not performed well enough to be skillful. We have $\widehat{K}_{1/2} = -2.75 < 0$ which indicates no skill. There isn't even dependence between Q1's diagnoses and the truth (dependence LRS $= 1.44$, $p = 0.23$). This is strong evidence that the methods used by Q1 and Q2 to diagnosis disease are wanting. Parker and Davis came to the same conclusion using their test.

5.3. **Eliciting Loss.** There are typically two situations for prediction verification, the general and specific. The general verification case is when the predictions were issued to a wide number of users whose loss functions were unknown and most likely widely different. The forecaster must consider these varied losses in ascertaining the value of the predictions. The specific case is when an exact loss function is known.

The decision maker only need consider a specific θ. A healthy woman considering a precautionary mammogram is an example of the specific case.

Both the skill score K_θ and the test statistic G show explicit dependence on θ, which is the loss. This highlights the fact that a loss must first be chosen in order to estimate and assess the statistical significance of skill.

The costs associated with prediction error are not always immediately obvious. The costs need not be monetary and can instead be emotional costs, or whatever else is relevant to the prediction user. We will use the mammography example to illustrate the key point: the user must not be influenced by the fact that the error probability is low; she must not let the small chance of error lead her to underestimate the true cost of that error when it occurs. (Harder to factor for any particular woman is that the error rate, particularly the false positive error rate, is widely variable both within and across radiologists; Elmore et al., 2002; this variability is ignored in this analysis.)

A healthy 40 year-old woman's costs for a false positive mammogram might be stress, depression, unnecessary treatment, the mislabelling of non-harmful growths such as *ductal carcinoma in situ* (Berry 1998; Berry 2002), the possible increased risk of subsequent cancer (McCann et al., 2002) and so on. False negative costs might be delayed treatment, or the ignoring of true symptoms because of false reassurance ("the test said I was fine"). Imagine for a young, healthy woman a false positive cost of 70 (in arbitrary units) and a false negatives cost of 210. This gives a $\theta = 1/4$ and shows that, as is likely for young, healthy, and asymptomatic women, false negatives incur more costs than do false positives, but the costs are not overwhelmingly higher.

Another woman, say older and more at risk, might have a different set of costs and might use a different scale in assessing her scores. This means that costs cannot be compared from woman to woman, but skill scores can be compared as the next subsection will show.

Now imagine a physician who was asked by our healthy woman whether she should receive a precautionary mammogram. The doctor's costs are similar to the woman's, but the doctor also worries about malpractice suits (and associated increase in insurance rates and loss of reputation), concerns the woman does not have. The doctor's costs for a false positive, say, are 50, and for false negatives 1050. This gives $\theta = 1/22$ and shows that, for this imaginary doctor, false negatives incur much more costs than do false positives. Again, the individual scores of the doctor cannot be directly compared to the woman's.

5.4. **Mammograms.** The Mammogram Performance Data are given in Table 4. (These data represent an accurate reduction from the full set of 26,000 women as presented in Gigerenzer (2002). Having data from all tests would not effect the calculation of the skill score, as it would involve multiplying the numerator and the denominator of K_θ by the same constant. But it would effect the calculation of the test statistic G because the sample size would increase. Because of this, our results make it appear that p-values are larger than they would ordinary be; however this section is only for illustration.)

The accuracy, A, of the mammograms at detecting breast cancer is the probability of a correct forecast and is estimated as $A = (7 + 922)/1000 = 0.929$ Since $\widehat{p}_{1+} = 0.008 < \theta_W = 1/4$, the optimal naive prediction is to say "no cancer." The accuracy of this optimal naive prediction, which might be called a *naive-o-gram*, is $A_{ONF} = (70 + 922)/1000 = 0.992$ which is much larger than the actual mammograms. This, as mentioned before, does not take into account the loss and shows why accuracy is not necessarily the best measure of performance.

Using the woman's example from Section 5.3 gives $\widehat{K}_{1/4} = -2.04$. This is, of course, less than 0 and indicates that the mammogram's performance is not skillful. That means the woman would have been better off had she opted for the naive-o-gram.

But, the skill score for the doctor is $\widehat{K}_{1/22} = 0.415$, which is greater than 0 and indicates skill. The doctor would then calculate the test statistic G and formally check for skill. The calculations is $G = 3.87$ which has a p-value of 0.025 and would lead the doctor to recommend the test.

This example highlights the important fact that a prediction may be valuable to one person (the doctor) but not valuable to another (the woman). The doctor and the woman would make different choices in this example and conflict would arise; therefore, both sides should make explicit their costs so that an understanding could be reached. A single mammogram is never taken as an authoritative diagnostic test; further tests are always required. But this is the point. If the test is so inaccurate that the expected costs of its false positives, including further tests, outweigh the expected costs of its false negatives, women would be better off not receiving the test. See Gigerenzer (2002) for an extended analysis of the problems of mammogram screenings.

5.5. **Guidance.** A more common problem is the general verification situation, where the individual loss functions are not known. One approach is to plot the prediction's skill score through a range of values for θ, as shown in Figure 1. Plots like these were first suggested in Wilks (2001) and Schervish (1989). Plotting of the skill score shows when the prediction has potential value to different decision makers.

Here, we plot $\log\theta$ in the range of 0 to 1/2 to show fine detail for small θ. We could have plotted K_θ for $\theta > 1/2$, though it is unlikely many people's losses would be in this range. Moreover, we can exactly calculate, using (3.4), the range where $\widehat{K}_\theta > 0$, which for this data is in $\theta \in [0.0011, 0.0909]$.

5.6. **TRISS and model building.** The Trauma Injury Severity Score estimates the probability of death (or, of survival) of patients who sustain trauma (Boyd et al., 1987). TRISS has been in use for more than a decade for research purposes to allow comparison of survival probabilities from various practice settings. It is the primary predicative variable used in the National Trauma Data Bank of the American College of Surgeons (Gabbe et al., 2004).

TRISS is a multifactorial construct, which incorporates the type of injury, blunt or penetrating; severity of injury, as measured by the Revised Trauma Score (RTS, a physiological component) and the Injury Severity Score (ISS, a anatomical component); and an age index. Several components were employed to make TRISS robust and generalizable to many populations of trauma patients. In this paper we only consider blunt injuries.

The Revised Trauma Score (RTS) was introduced the late 1980's and is a physiological scoring system, with high inter-rater reliability (e.g. Eichelberger et al., 1989). It is scored from the vital information on the patient at the time of presentation, and consists of a linear weighting of the Glasgow Coma Scale (which is most heavily weighted), Systolic Blood Pressure, and Respiratory Rate.

The Injury Severity Score (ISS) is based on the anatomical distribution of the injuries and provides an overall score for patients with multiple injuries (e.g. Rutledge et al., 1997). Each injury is assigned a score from the Abbreviated Injury Scale (AIS) and allocated to one of six body regions (Head, Face, Chest, Abdomen/Pelvis, Extremities, and External). Only the highest AIS score in each body region is used and the 3 most highly scored anatomical regions have their scores squared and the squared values are summed to produce the ISS score. The ISS score has values from 0 to 75, however, if an injury is assigned an AIS of 6 for a lethal injury, the ISS score of 75 is automatically assigned. The ISS score correlates well with mortality, morbidity, and length of stay in hospital.

As mentioned above, the last component of TRISS, the age index, is the indicator of age greater than 54.

We sought to build a better TRISS model by incorporating a measure of comorbidity and by possibly changing the age index. Full details on the importance of these factors, and their meaning to trauma research, is contained in Melniker et al. (2004). We do not concentrate on the revised's model development here, except to note its performance in comparison to the original TRISS model.

Data on about 46,000 patients was provided by the National Trauma Data Bank. The elements of the data were death indicator, injury type, age, RTS, ISS, and Charlson comorbidty index (Charlson et al., 1994), which has been found to correlate well with mortality. We used only the blunt trauma injury patients and TRISS model. The widely-used Charlson comorbidty index is a scale, starting a 0, indicating no medical comorbidity, to 16, indicating dire straits. The idea is that patients who arrive at the emergency unit with significant prior medical comorbidity would have less chance for survival than other patients with similar injuries and no comorbidity.

We fit a logistic regression model using a random sample of half the patients with RTS, ISS, and comorbidty index as predictors, and an indicator of age greater or equal to 70 (this cut off was better indicated by the data, though the improvement over the original cut off is minimal). We validated the model using skill with the other random half of the data for both the original and revised TRISS model.

Figure 2 shows the skill score for both the revised and the original TRISS regression models (on the validation data only). The 95% point-wise confidence intervals are also plotted for the revised TRISS model. The skill score for the revised model is larger than the skill score for the original model over most of the range of user losses. In particular, the original TRISS model at $\theta = 1/2$ has no skill, which says that users of the original model would have been better off had they used the optimal naive prediction, which is "no death" for everyone. This is despite the fact that the regression coefficients were "significant".

This informal analysis shows that the comorbidity index certainly improves the model (and hints at how skill scores may be used for model selection).

6. Conclusions

Formal tests for simple skill for predictions for dichotomous events have been developed. Skill was defined in two ways. A collection of predictions were said to be skillful if the expected loss incurred while using them was less than the expected loss that would have been incurred had the optimal naive predictions been used instead. This definition allows the loss for prediction error to be asymmetric. A second definition of skill, and equivalent to the first definition under symmetric loss, requires that a collection of predictions have a higher probability of being correct than a collection of optimal naive predictions.

Skill scores for these situations have also been given, and it was shown that testing for skill using these scores is equivalent to formal tests of skill previously developed. Skill scores are a handy measure of prediction performance and here have the interpretation of the relative difference in loss between the expert and optimal naive collection of predictions.

Skill, for a 2×2 contingency table, was shown to be a stronger association between the observations and predictions Y and X than that of dependence (see, for example, Agresti, 1990).

The costs of false positives and the manner of handling them in the decision making process are much discussed in the medical literature. The costs of false positives are not negligible. Thorne et al. (1999) report that women waiting for a subsequent diagnosis after an initial positive mammogram find it an "intense and often agonizing experience." Barton et al. (2001) report that women who have had false positives seek (and pay for) more health care over the subsequent year than do other women. Burman et al. (1999) and Pisano et al. (1998) found that women with initial false positive screenings were more likely than other women to return for future mammograms. Berry (1998) carried out a systematic statistical study of mammograms for women in their forties and argued that the mammogram's benefits were small, and could even be negative; he urged that women better understand their own potential risks. Skill scores, computed individually, could help in this goal.

There is much future work to be done. Extensions of the skill score test to the case where the observed series is multivariate would be enormously useful, particularly when the observed variables are spatially related (such as a precipitation prediction map). Extensions to categorical and continuous-valued observations should be explored. Comparisons of our skill score curves with receiver operating characteristic (ROC) curves should also prove useful (see Wilks, 2001, and Venkatraman, 2000).

In this paper it was assumed that the parameters of the observed series were constant in a given collection of observations. In practice these parameters may vary, for example, seasonally or over subsets of patients for diagnostic tests. Probability of precipitation is a prime example of varying parameters. The skill scores could still be used to compute a rough score, but the effect on the skill test remains to be investigated.

Acknowledgements

Thanks to Dan Wilks for his comments and clarifications and to Persi Diaconis for his encouragement and many helpful comments during the early stages of this research. Also thanks to the three anonymous reviewers whose comments greatly improved this manuscript.

References

1. Agresti, A. (1990). *Categorical Data Analysis.* New York: Wiley.
2. Barton, M.B., Moore, S., Polk, S., Shtatland, E., Elmore, J.G., and Fletcher, S.W. (2001). Increased patient concern after false-positive mammograms. *J. General Internal Medicine* **16**, 150–156.
3. Berry, D.A. (2002). The utility of mammography for women 40 to 50 years of age (con). In *Progress in Oncology*, Devita, V.T., Hellman, S., and Rosenberg, S.A. (eds), 233–259. Sudbury: Jones and Bartlett.
4. Berry, D.A. (1998). Benefits and risks of screening mammography for women in their forties: a statistical approach. *J. of the National Cancer Institue* **90 (19)**, 1431–1439.
5. Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete multivariate analyses: Theory and practice.* Cambridge: MIT Press.
6. Bluman, L.G., Rimer, B.K., Berry, D.A., Borstelmann, N., Regan, K., Schildkraut, J., and Winter, E. (1999). Attitudes, knowledge and risk perceptions of women with breast cancer considering testing for BRCA1 and BRCA2. *Journal of Clinical Oncology* **17**, 1040–1046.
7. Boyd, C.R., Tolson, M.A., and Copes, W.S. (1987). Evaluating Trauma Care: The TRISS Method. *J. Trauma* **27**, 370–378.
8. Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3.
9. Briggs, W.M. (2004). Incorporating Cost in the new Skill Score. *Technical Report*, wm-briggs.com/public/skillcost.pdf.
10. Briggs, W.M., and Levine, R.A. (1998). Comparison of forecasts using the bootstrap. *14th Conf. on Probability and Statistics in the Atmospheric Sciences*, Phoenix, AZ, Amer. Meteor. Soc., 1–4.
11. Briggs, W.M., and Ruppert, D. (2004). Assessing the skill of yes/no forecasts for Markov observations. *17th Conf. on Probability and Statistics in the Atmospheric Sciences*, Seattle, WA, Amer. Meteor. Soc.

12. Charlson, M.E., Szatrowski, T.P., Peterson, J., and Gold, J. (1994). Validation of a combined comorbidity index. *J. Clinical Epidemology* **47(11)**, 1245–51.
13. Diaconis, P. (1978). Statistical problems in ESP research. *Science* **201**, 131–136.
14. Diaconis, P., and Mosteller, F. (1989). Methods for studying coincidences. *JASA* **84**, 853–861.
15. Diaconis, P., and Graham, R. (1981). The analysis of sequential experiements with feedback to subjects. *Annals of Statistics* **9**, 3–23.
16. Diebold, F.X. (2001). *Elements of Forecasting* (second edition). Florence: South-Western Publishing.
17. Diebold, F.X., and Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics* **13**, 253–263.
18. Eichelberger, M.R., Bowman, L.M., Sacco, W.J., Mangubat, E.A., Lowenstein, A.D., and Gotschall, C.S. (1989). Trauma score versus revised trauma score in TRISS to predict outcome in children with blunt trauma. *Annals of Emergency Medicine* **18(9)**, 939–942.
19. Elmore, J.G., Miglioretti, D.L., Reisch, L.M., Barton, M.B., Kreuter, W., Christiansen, C.L., and S.W. Flectcher (2002). Screening mammograms by community radiologists: variability in false-positive rates. *J. of the National Cancer Institute* **94 (18)**, 1373–1380.
20. Ferrell, W.R. (1994). Discrete subjective probabilities and decision analysis: elicitation, calibration, and combination. In *Subjective Probability*, Wright, G., and P. Ayton, (eds), 411–451. New York: Wiley.
21. Fienberg, S.E. (1980). *The analysis of cross-classified categorical data.* Cambridge: MIT Press.
22. Gabbe B.J., Cameron, P.A., and Wolfe, R. (2004). TRISS: Does It Get Better than This? *Academic Emergency Medicine* **11(2)**, 181–186.
23. Gigerenzer, G. (2002). *Calculated Risks: How to know when numbers deceive you.* New York: Simon and Schuster.
24. Hamill, T. M. (1998). Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Predictioning* **14**, 155–167.
25. Hastie, T., Tibshirani, R., and Friedman, J.H. (2001). *The Elements of Statistical Learning.* New York: Springer.
26. Hendrickson, A.D., and Buehler, R.J. (1971). Proper scores for probability forecasters. *Annals of Mathematical Statistics* **42**, 1916–1920.
27. Katz, R.W. (1993). Dynamic cost-loss ratio decision-making model with an autocorrelated climate variable. *Journal of Climate* **6**, 151–160.
28. Kolb, R.A., and Stekler, H.O. (1993). Are economic predictions significantly better than naive forecasts? An appropriate test. *Int. J. Predictioning* **9**, 117–120.
29. Kryzysztofowicz, R. (1992). Bayesian correlation score: a utilitarian measure of forecast skill. *Monthly Weather Review* **120**, 208–219.
30. Mason, I. (1979). On reducing probability forecasts to Yes/No forecasts. *Monthly Weather Review* **107**, 207-211.
31. McClelland, A.G.R., and Bolger, F. (1994). The calibration of subjective probabilities: theories and models 1980–94. In *Subjective Probability*, Wright, G., and P. Ayton, (eds), 453–484. New York: Wiley.

32. Meeden, G. (1979). Comparing two probability appraisers. *JASA* **74**, 299–302.

33. Melniker, L., Briggs, W.M., and Mancuso, C. (2003). An improved TRISS model. In preparation.

34. Mozer, J.B., and Briggs, W.M. (2003). Skill in real-time solar wind shock predictions. *J. Geophysical Research: Space Physics* **108 (A6)**, SSH 9 p. 1–9, (DOI 10.1029/2003JA009827).

35. Murphy, A.H. (1973). Hedging and skill scores for probability forecasts. *J. Applied Meteorology* **12**, 215–223.

36. Murphy, A.H. (1991). Forecast verification: its complexity and dimensionality. *Monthly Weather Review* **119**, 1590–1601.

37. Murphy, A.H. (1994). A coherent method of stratification within a general framework for forecast verification. *Monthly Weather Review* **123**, 1582–1588.

38. Murphy, A.H. (1996). The Finely affair: a signal event in the history of forecast verification. *Weather and Predictioning* **11**, 3–20.

39. Murphy, A.H. (1997). Forecast verification. In *Economic Value of Weather and Climate Predictions*, Katz, R.W., and A.H. Murphy (eds), 19–74. London: Cambridge.

40. Murphy, A.H., and Ehrendorfer, A. (1987). One the relationship between the accuracy and value of forecasts in the cost-loss ratio situation. *Weather and Forecasting* **2**, 243–251.

41. Murphy, A.H., and Winkler, R.L. (1987). A general framework for forecast verification. *Monthly Weather Review* **115**, 1330–1338.

42. Murphy, A.H., and Ye, Q. (1990). Comparison of objective and subjective precipitation probability forecasts: the sufficiency relation. *Monthly Weather Review* **115**, 1330–1338.

43. Parker, R.A., and Davis, R.B. (1999). Evaluating whether a binary decision rule operates better than chance. *Biometrical Journal* **41**, 25–31.

44. Pisano, E.D., Earp, J.A., and Gallant, T.L. (1998) Screening mammography behavior after a false positive mammogram. *Cancer Detection and Prevention* **22**, 161–167.

45. Rutledge, R., Hoyt, D.B., Eastman, A.B., Sise, M.J., Vleky, T., Canty, T., Watchel, T., and Osler, T.M. (1997). Comparison of the Injury Severity Score and ICD-9 diagnosis codes as predictors of outcome in injury: analysis of 44,032 patients. *J. Trauma* **42(3)**, 477–87.

46. Schervish, M.J. (1989). A general method for comparing probability assessors. *Annals of Statistics* **17**, 1856–1879.

47. Self, S.G., and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. American Statistical Association* **82**, 605–610.

48. Solow, A.R., and Broadus, J.M. (1988). A simple model of overforecasting. *Monthly Weather Review* **116**, 1371–1373.

49. Thompson, M.L., and Zucchini, W. (1990). Assessing the value of probability forecasters. *Monthly Weather Review* **118**, 2696–2706.

50. Thorne, S.E., Harris, S.R., Hislop, T.G. and Vestrup, J.A. (1999). The experience of waiting for diagnosis after an abnormal mammogram. *The Breast Journal* **5**, 42–51.

51. Tsimikas, J.V., Bosch, R., Coull, B.A., and El Barmi, H. (2002). A profile likelihood approach for comparing highly accurate diagnostic tests. *Biometrics* **58**, 946–956.

52. Venkatraman, E.S. (2000). A permutation test to compare receiver operating characteristic curves. *Biometrics* **56**, 1134–1138.

53. Wilks, D.S. (1991). Representing serial correlation of meteorological events and forecasts in dynamic decision-analytic models. *Monthly Weather Review* **119**, 1640–1662.

54. Wilks, D.S. (1995). *Statistical Methods in the Atmospheric Sciences.* New York: Academic Press.

55. Wilks, D.S. (2001). A skill score based on economic value for probability forecasts. *Meteorological Applications* **8**, 209–219.

56. Winkler, R.L. (1994). Evaluating probabilities: asymmetric scoring rules. *Management Science* **40**, 1395–1405.

57. Winkler, R.L. (1996). Scoring rules and the evaluation of probabilities (with comments). *Test* **5**, 1–60.

58. Zhou, X.H., Obuchowsjki, N.A., and McClish, D.K. (2002). *Statistical Methods in Diagnostic Medicine.* New York: Wiley.

TABLE 1. *Standard* 2×2 *contingency table.*

	$Y = 1$	$Y = 0$
$X = 1$	n_{11}	n_{01}
$X = 0$	n_{10}	n_{00}

TABLE 2. *Diagnoses for Q1.*

	$Y = 1$	$Y = 0$
$X = 1$	3	14
$X = 0$	1	18

TABLE 3. *Diagnoses for Q2.*

	$Y = 1$	$Y = 0$
$X = 1$	0	0
$X = 0$	4	32

TABLE 4. *Mammogram Performance table. Data is presented as an average over 1000 women but the estimates are from a larger sample.*

	$Y = 1$	$Y = 0$
$X = 1$	7	70
$X = 0$	1	922

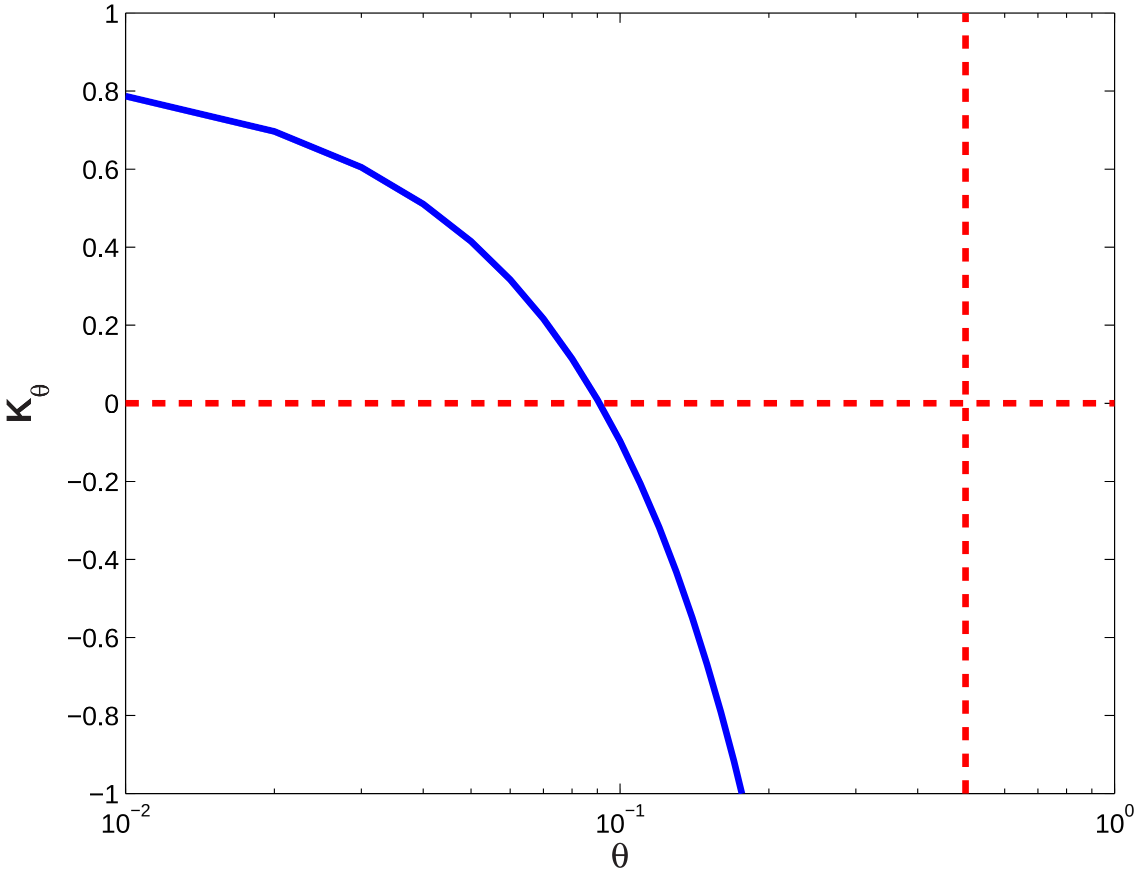

FIGURE 1. Skill score range plot for $\theta \in (0, 1/2]$, presented for $\log \theta$ for the mammogram data. The dashed horizontal line shows 0 and predictions below this line have no skill; the dashed vertical line indicates $\theta = 1/2$.

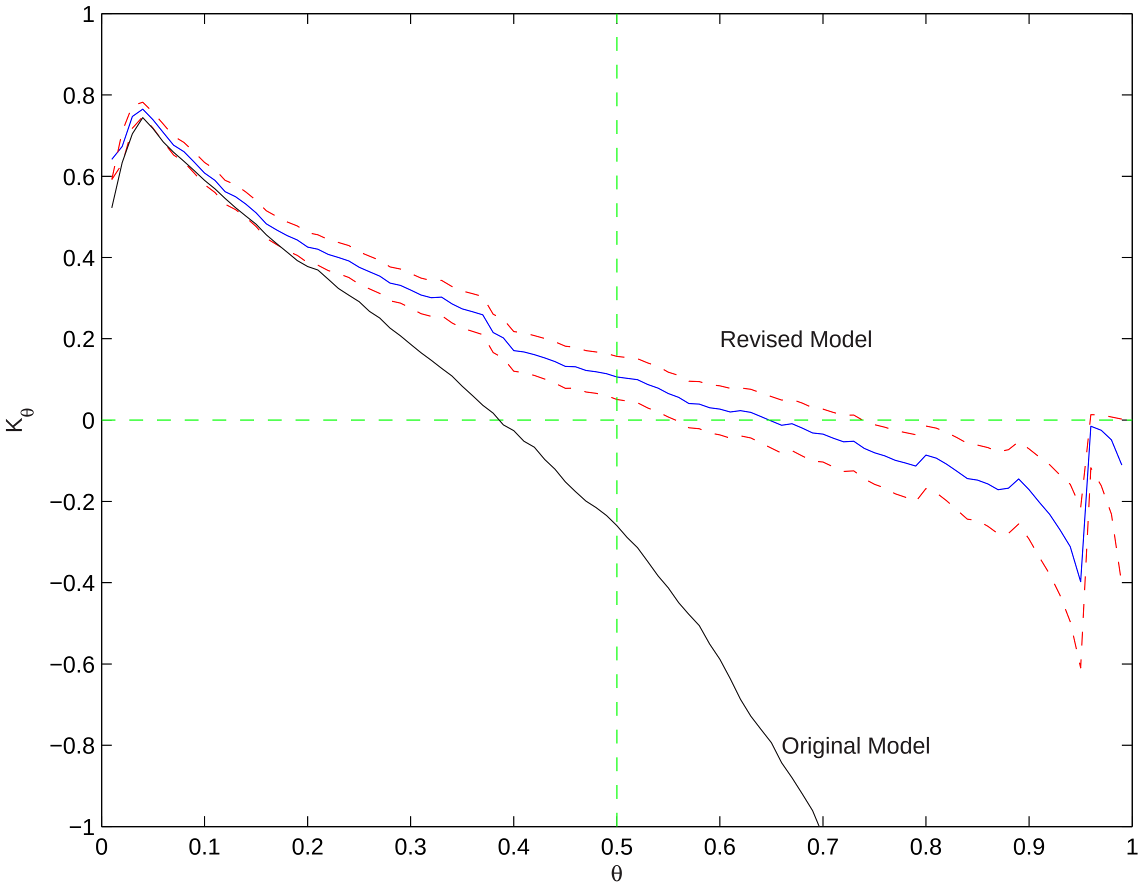

FIGURE 2. Skill score range plot for all θ for the TRISS original and revised model data. 95% confidence bounds, dashed lines, for the revised model are also plotted.