Relationship between

Brier score and area under the binormal ROC curve

Mitsuru Ikeda[1], Takeo Ishigaki[2], and Kazunobu Yamauchi[1]

[1]Department of Medical Information and Medical Records,

Nagoya University Hospital, 65, Tsurumai-cho, Showa-ku, Nagoya 466-8560, Japan.

[2]Department of Radiology, Nagoya University, School of Medicine,

65, Tsurumai-cho, Showa-ku, Nagoya 466-8560, Japan.

(corresponding author)

Mitsuru Ikeda,

Department of Medical Information and Medical Records,

Nagoya University Hospital, 65, Tsurumai-cho, Showa-ku,

Nagoya 466-8560, Japan.

Phone: +81-52-744-2666; Fax: +81-52-744-2973

Email: a40495a@nucc.cc.nagoya-u.ac.jp

Key words: Brier Score;

Receiver Operating Characteristic (ROC) Curve;

Probabilistic Judgments;

Medical Decision Making.

Abstract

If we consider the Brier Score (*B*) in the context of the signal detection theory and assume that it makes sense to consider the existence of *B* as a parameter for the population (let $\overline{B}$ be this *B*), and if we assume that the calibration in the observer's probability estimate is perfect, we find that there is a theoretical relationship between $\overline{B}$ and the area under the binormal receiver operating characteristic (ROC) curve, $\mathrm{A_Z}$. We have derived this theoretical functional relationship between *B* and $\mathrm{A_Z}$, by using the parameter a and b in the binormal ROC model and the prior probability of signal events (α); here, the two underlying normal distributions are $N(\mu_s, \sigma_s)$ and $N(\mu_n, \sigma_n)$; and, $a = (\mu_s - \mu_n)/\sigma_s$ and $b = \sigma_n/\sigma_s$. We empirically found that, if parameters b and α are constant, $\overline{B}$ values in relation to given $\mathrm{A_Z}$ values monotonically decrease as $\mathrm{A_Z}$ values increase, and these relationship curves have monotonically decreasing slopes.

1. Introduction

In several situations, physicians are required to express probabilistic judgments in numerical terms, and there is some evidence that such judgments have operational meaning to physicians [1]. Thus, it is very important to evaluate these judgments properly. Today, there are several quantitative methods to assess the quality of physicians' probabilistic judgments [1]. Scoring rules are one of the methods for the assessment of such probabilistic judgments, and the Brier score (*B*) is one of these best-known rules [1].

On the other hand, receiver operating characteristic (ROC) analysis is the most common and sophisticated method for evaluating the signal-detection capability of observers or imaging modalities [2]. ROC curve shows the ability of probability estimates to separate patients into groups ordered by the prevalence of disease [1], and the area under the ROC curve represents the probability that a randomly chosen diseased subject is correctly rated or ranked with greater suspicion than a randomly chosen nondiseased subject [3]. Various methods for estimating the ROC curve from test results have been developed. In particular, the methods for estimating the ROC curve based on normal (Gaussian) probability distributions [4] are well known; this ROC curve is called the binormal ROC curve, and the area under the binormal ROC curve is termed $\mathrm{A_Z}$. Further, one can estimate a binormal ROC curve from continuously-distributed test results by using Metz LABROC4 algorithm [5].

So, one can now use at least two indexes, to evaluate probabiltistic judgments quantitatively: indexes *B* and $\mathrm{A_Z}$. Then, under the conditions in which both *B* and $\mathrm{A_Z}$ can be calculated, which index of the two should be used to evaluate probabilistic judgments? For example, Gurney suggested that ROC analysis may not be the ideal method to judge predictive accuracy, and that a true test of predictive accuracy such as *B* should be used [2].

Therefore, we believe that it will be very important to study the relationship between *B* and A_Z, and we have investigated the theoretical relationship between them. Here, we must first realize that there is no general theoretical relationship between *B* and A_Z; the reason is that ROC curves and A_Z's are invariant under order-preserving transformations [6], although *B*'s change by these order-preserving transformations. However, if we make several assumptions in the context of signal detection theory (SDT), we find a theoretical relationship between *B* and A_Z, and we then derive this functional relationship. One of these strong assumptions is that the calibration in the observer's probability estimate is perfect. The purpose of the present study is to report this functional relationship and its application to the assessment of probabilistic judgments.

2. Theoretical development

ROC analysis is based on true-positive probability, $P(S \mid s)$, and false-positive probability, $P(S \mid n)$, in fundamental detection problems with only two events and two responses [4, 7, 8]. According to SDT, we have assumed that there are two probability distributions of the random variable X, one associated with the signal event s [$f(x \mid s)$] and the other with the nonsignal event n [$f(x \mid n)$] [8]; these probability (or density) distributions of a given observation x are conditional upon the occurrence of s and n [8]. In the medical context, the signal event corresponds to the abnormal (diseased) group, and the nonsignal event to the normal (nondiseased) group [3]. If the cutoff value is c, corresponding to a particular likelihood ratio, the true- and false-positive probabilities are given by the following expressions [8]:

$$P(S|s)=\int_c^\infty f(x|s)dx \tag{1}$$

$$P(S|n)=\int_c^\infty f(x|n)dx \tag{2}$$

When we consider the conventional binormal model, the probability distributions associated with the signal event [$f(x|s)$] and those associated with the nonsignal event [$f(x|n)$] are assumed to be represented by two overlapping normal distributions [4, 9]; that is,

$$f(x|s)=\frac{1}{\sqrt{2\pi}\sigma_s}\exp\left[-\frac{(x-\mu_s)^2}{2\sigma_s^2}\right] \tag{3}$$

$$f(x|n)=\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{(x-\mu_n)^2}{2\sigma_n^2}\right] \tag{4}$$

We designate the normal-deviate values of the true-positive probability [$P(S|s)$] and the false-positive probability [$P(S|n)$] as $z(S|s)$ and $z(S|n)$, respectively [4, 9]. Thus, from equations (1), (2), (3), and (4), $P(S|s)$ and $P(S|n)$ can be described as

$$P(S|s)=\Phi\left(\frac{-c+\mu_s}{\sigma_s}\right) \tag{5}$$

and

$$P(S|n)=\Phi\left(\frac{-c+\mu_n}{\sigma_n}\right) \tag{6}$$

where,

$$\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp\left(-\frac{x^2}{2}\right)dx \tag{7}$$

and the relationship of $z(S|s)$ and $z(S|n)$ is given by

$$z(S|s)=bz(S|n)+a \tag{8}$$

where, $a = (\mu_s - \mu_n)/\sigma_s$ and $b = \sigma_n/\sigma_s$. These relations were shown by Green [7] and Simpson [9].

Now we consider a set of M signal-detection tasks with αM signal events and $(1-\alpha)M$ nonsignal events ($0 \le \alpha \le 1$), where the subscript i indexes the individual event, and we postulate that αM and $(1-\alpha)M$ are natural numbers (therefore, α denotes the *a priori* probability of signal events). Let y_i indicate the true state of the event, such that $y_i = 0$ if the event is nonsignal and $y_i = 1$ if the event is signal [10]. Let p_i denote this observer's (or the physician's) probability estimate that the *i*th event will be the signal one [10]. The definition of *B* is thus as follows [10, 11]:

$$B = \frac{1}{M}\sum_{i=1}^{M}(y_i - p_i)^2 \tag{9}$$

In the following, we consider *B* in the context of SDT. We assume that the observer (or physician) makes a probability estimate upon the evidence of the decision variable x, and we denote x for the *i*th event by x_i. That is, we assume that p_i is estimated upon x, and that p_i is a function of x_i, $p(x_i)$. In the above-mentioned conditions, when one knows x_i for the *i*th event, the probability of the signal event's occurrence for this *i*th event, $\Pr(s \mid x_i)$, can be calculated from the Bayes theorem. Therefore, when the calibration in the observer's probability estimate is perfect [that is, $p_i = \Pr(s \mid x_i)$], p_i in the equation (9) will be given as

$$p_i = p(x_i) = \Pr(s \mid x_i) = \frac{\alpha f(x_i \mid s)}{(1-\alpha) f(x_i \mid n) + \alpha f(x_i \mid s)} \tag{10}$$

Now we consider $\mathbf{x}_i = (p_i, y_i)$ $(i = 1, 2, \ldots, M)$ as a random sample of size M extracted from the population, and assume that it makes sense to consider the existence of the Brier score as a parameter for the population. Further, we assume that the convergence in probability of B given by the law of large numbers as M tends to infinity is the Brier score as a parameter for the population, and make the assumption that the calibration in the observer's probability estimate is perfect. Let $\overline{B}$ denote this Brier score as a parameter for the population under the assumption that the calibration in the observer's probability estimate is perfect.

From Appendix 1, $\overline{B}$ is given by

$$\overline{B} = \int_0^1 \frac{(1-\alpha)\alpha}{(1-\alpha) + \alpha \dfrac{dP(S\,|\,s)}{dP(S\,|\,n)}} dP(S\,|\,s) \tag{11}$$

or

$$\overline{B} = \int_0^1 \frac{(1-\alpha)\alpha}{\alpha + (1-\alpha) \dfrac{dP(S\,|\,n)}{dP(S\,|\,s)}} dP(S\,|\,n) \tag{12}$$

Now, we can describe $\overline{B}$ and $\mathrm{A_Z}$ as functions of a, b, and α. $\mathrm{A_Z}$ is given by the following equation [9]:

$$\mathrm{A_Z} = \int_0^1 \Phi\left[b\Phi^{-1}(x) + a\right] dx = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \tag{13}$$

From the equation (11) or (12), $\overline{B}$ can be expressed as

$$\overline{B} = \int_0^1 \frac{(1-\alpha)\alpha b}{\alpha b + (1-\alpha)\exp\left(\dfrac{(b^2-1)\left[\Phi^{-1}(x)\right]^2 + 2ab\Phi^{-1}(x) + a^2}{2}\right)} dx, \tag{14}$$

because,

$$\frac{dP(S\,|\,n)}{dP(S\,|\,s)} = \frac{1}{b}\exp\left(\frac{(b^2-1)\left[\Phi^{-1}(x)\right]^2 + 2ab\Phi^{-1}(x) + a^2}{2}\right), \tag{15}$$

etc. Therefore, the theoretical relationship between $\overline{B}$ and $\mathrm{A_Z}$ can be described as these functions [equations (13) and (14)] by using the three parameters a, b, and α.

From the above discussions, $\overline{B}$ can also be calculated in the following way: $\mathrm{A_Z}$

is estimated to continuously-distributed p_i by using the Metz LABROC4 algorithm [5], and, then, $\overline{B}$ is calculated from our derived theoretical relationship between $\overline{B}$ and $\mathrm{A_Z}$.

Here, we must also draw attention to the following: from Appendix 2, $\overline{B}$ is equal to the expected *B* as given by Spiegelhalter [11].

3. Graph of functional relationship between $\overline{B}$ and $\mathrm{A_Z}$

In our derived functional relation, there are three parameters for the respective functions describing $\overline{B}$ and $\mathrm{A_Z}$. Therefore, our derived functional relation is not one-to-one. However, if we determine two of a, b, and α, we can obtain the one-to-one functional relation between $\overline{B}$ and $\mathrm{A_Z}$. Therefore, if a and α are fixed, we can obtain the one-to-one functional relation between $\overline{B}$ and $\mathrm{A_Z}$ from the equations (13) and (14); if b and α are fixed, we can also obtain the one-to-one functional relation between $\overline{B}$ and $\mathrm{A_Z}$; and, if a and b are fixed, $\mathrm{A_Z}$ is constant for a given α value. Fig. 1 and Fig. 2 show the graphs of these one-to-one functions; Fig. 1 illustrates the graph of $\overline{B}$ as a function of $\mathrm{A_Z}$, where parameters a and α are fixed ($a = 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0$, $\alpha = 0.5$); and Fig. 2 illustrates the graph of $\overline{B}$ as a function of $\mathrm{A_Z}$, where parameters b and α are fixed ($b = 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0$, $\alpha = 0.5$).

4. Functional relationship between $\overline{B}$ and the area under the ROC curve for another distribution function

In the above, we discussed the relationship between $\overline{B}$ and the area under the ROC curve (AUC) for the conventional binormal model (that is, $\overline{B}$ and $\mathrm{A_Z}$). Now, for comparison, we have derived the functional relationship between $\overline{B}$ and AUC for another distribution function.

Egan [8] gave various ROCs based upon assumed probability distributions that have the same probability law for each member of a given pair of distributions. He also treated a theoretical ROC derived from normal probability distributions as a normal-normal ROC (N-N ROC) and as a standard for comparison with the ROCs based upon other probability laws [8]. A power-law ROC is one such ROC; its mathematical treatment is easy, and this family of ROCs represents a useful contrast to N-N ROCs [8]. Thus, we have derived the functional relationship between $\overline{B}$ and AUC for a power-law ROC.

Again, according to Egan [8], for a power-law ROC, $f(x \mid s)$ and $f(x \mid n)$ in the equations (1) and (2) are given by

$$f(x \mid s) = K_s \exp(-K_s x) \tag{16}$$

$$f(x \mid n) = K_n \exp(-K_n x) \tag{17}$$

with $0 \le x < \infty$, and with $K_n \ge K_s > 0$. Thus, $dP(S|n)/dP(S \mid s)$ in the equation (12) is given by

$$\frac{dP(S|n)}{dP(S \mid s)} = \frac{1}{k}\left[P(S \mid n)\right]^{1-k} \tag{18}$$

where, $k = K_s / K_n$. Therefore, from equation (12), $\overline{B}$ is

$$\overline{B} = \int_0^1 \frac{(1-\alpha)\alpha}{\alpha + \frac{(1-\alpha)}{k}\left[P(S \mid n)\right]^{1-k}} dP(S \mid n) \tag{19}$$

On the other hand, from $P(S \mid s) = \left[P(S \mid n)\right]^k$, AUC for this model can be expressed as

$$\text{AUC} = \int_0^1 P(S \mid s) dP(S \mid n) = \frac{1}{1+k} \tag{20}$$

Therefore, just as with the binormal model, the theoretical relationship between $\overline{B}$ and

AUC can be described as these functions [the equations (19) and (20)] by using the two parameters k and α. Further, if we determine α, we can obtain the one-to-one functional relation between $\overline{B}$ and AUC.

Fig. 3 illustrates the graph of $\overline{B}$ as a function of AUC for Power-ROCs and the graph for N-N ROCs, where parameters b and α are fixed with $b=1$ and $\alpha = 0.1(0.1)0.5$. From Fig. 3, one can see that the theoretical relationship between $\overline{B}$ and AUC for a power-law ROC was similar to that between $\overline{B}$ and $\mathrm{A_Z}$ with $b=1$ for an N-N ROC.

5. Discussion

Somoza [12] notes that most diagnostic tests are described by ROCs where $0.2 \le b \le 5$; ROCs where b falls outside this range are "eccentric" and require some rethinking of the binormal model [13]. It is for this reason that we calculated $\overline{B}$ values in relation to given $\mathrm{A_Z}$ values with b fixed for $0.2 \le b \le 5$.

From these results, we empirically found that, if parameters b and α are constant, $\overline{B}$ values in relation to given $\mathrm{A_Z}$ values monotonically decreases as $\mathrm{A_Z}$ values increase, and that this relationship curve of $\overline{B}$ values in relation to given $\mathrm{A_Z}$ values approaches the point, $\overline{B} = 0$ and $\mathrm{A_Z} = 1$; in addition, these curves have monotonically decreasing slopes (Fig. 2). Especially, for $b=1$ and $\alpha = 0.5$, the relationship curve of $\overline{B}$ values to given $\mathrm{A_Z}$ values starts from the point, $\overline{B} = 0.25$ and $\mathrm{A_Z} = 0.5$, corresponding to the values indicating valueless predictions for each index, and approaches the point, $\overline{B} = 0$ and $\mathrm{A_Z} = 1$, corresponding to the values indicating errorless predictions for each index; here, the $\overline{B}$ values for $0.7 \le b \le 1.5$ are in rather good agreement with the ones for $b=1$ (Fig. 2).

We think this empirically determined fact is very important for two reasons.

First, when the calibration in the observer's probability estimate is approximately perfect, in a comparison of binormal ROCs with similar slopes, there will be no inconsistencies in the relations between $\overline{B}$ and A_Z as accuracy indices. Green has shown that the percentage correct in the two-alternative forced-choice situation is equal to AUC (Green's theorem) [7, 9]; yet from our derived relationship, A_Z has another meaning. Further, practically, in the case in which slopes in binormal ROC models for comparing data sets are similar, and in which the calibration in the observer's probability estimate is good, *B* could be used to replace A_Z as an accuracy index. Because the use of A_Z as an accuracy index is properly restricted to a comparison of binormal ROCs with slopes that are similar or not materially different [4], one cound use *B* reasonably in place of A_Z, when the calibration in the observer's probability estimate could be considered to be good.

Secondly, when parameters b and α are constant, around $A_Z = 1$, A_Z value changes slowly relative to $\overline{B}$, and vice versa around $A_Z = 0.5$. From this functional relationship, *B* calculated from large data sets with good calibration in the observer's probability estimate may be more appropriate than A_Z in comparative evaluations of highly accurate probabilistic judgments, for assessing their discriminating power. On the contrary, *B*, even if calculated from large data sets with very good calibration, may not be as appropriate as A_Z for comparative evaluations of lower accurate probabilistic judgments.

Although, in the derivation of the theoretical relationship of $\overline{B}$ and A_Z, we postulated that underlying distributions would be normal (that is, binormal), the definition of *B* does not require the assumption of underlying distributions, and the

ROC curve and AUC in themselves are independent of the form of the underlying signal and nonsignal distributions [9]. Therefore, our theory is a restricted one in the sense that it holds only in parametric situations, although *B* and A_Z in themselves are nonparametric. However, the relationship between $\overline{B}$ and AUC for a power-law ROC with $\alpha = 0.5$ was similar to the one between $\overline{B}$ and A_Z with $b = 1$ and $\alpha = 0.5$ for an N-N ROC. This fact suggests that the relationship between $\overline{B}$ and AUC for "proper" ROCs would be similar; if this is true, this limitation of our theory would not be significant for "proper" ROCs.

From our derived relationship between $\overline{B}$ and A_Z, if one compares ROCs with slopes that are materially different, the cases in which there are inconsistencies in the relations between *B* and A_Z do exist. Now, we make the reasons for these inconsistencies clear.

Before we present this fuller discussion, however, we must mention once again, that the relationship between *B* and A_Z, derived in this study, is valid only under the various above-mentioned assumptions. Especially, the assumption of perfect calibration in the observer's probability estimate is strong. Here, what is important is that, if the assumptions that we have made for deriving the relation between *B* and A_Z do not hold, a theoretical relationship between *B* and A_Z does not exit.

Let us now return to the previous discussion. If the calibration in the observer's probability estimate is perfect or can be considered to be approximately perfect, our derived relationship between $\overline{B}$ and A_Z will account for the relations between *B* and A_Z, and the inconsistency in the relation between *B* and A_Z is only superficial. Here, the extent to which the observer's probability estimate coincides with the true probability of the signal event's occurrence can be measured by the Sanders decomposition of B into the calibration (or reliability) component and the discrimination (or resolution) component [14-16], or by our derived relationship between $\overline{B}$ and A_Z; that is, if $\overline{B}$ calculated from the corresponding A_Z value by using our derived theoretical relationship is close to B directly calculated from the observer's probability estimates, the calibration in the observer's probability estimate will be good.

On the other hand, if the calibration in the observer's probability estimate is not good, the inconsistency in the relation between B and A_Z is both due to the imperfect calibration in the observer's probability estimate and due to the theoretical relationship between $\overline{B}$ and A_Z. Therefore, in that case, one must consider very carefully which index of B and A_Z should be used to evaluate probabilistic judgments.

ROC curves are invariant under order-preserving transformations [6]. Thus, any monotonic transformation of the decision variable changes the form of the decision-variable distributions but does not change the ROC. Therefore, the ROC analysis and its index A_Z are independent of the accuracy of calibration in probabilistic judgments [1], and the ROC analysis is ineffective in the evaluation of the calibration problem of physicians' probabilistic judgments, which is the problem of evaluating the degree to which an appraiser's probabilities correspond to the actual frequencies of outcome [1]. However, the calibration is usually considered to be important in "external correspondence" [14]. Therefore, to evaluate probabilistic judgments completely, not only should ROC analysis be performed, but also B, especially the calibration component of B, should be evaluated.

Further, if the calibration in the observer's probability estimate is perfect, the observer's judgments can be said to be perfectly "internally consistent" [14]. In that case, the estimated ROC parameters for the observer's judgments can be expected to be accurate from the standpoint of the "internal sampling error." Thus, when ROC analysis is performed to evaluate probabilistic judgments, it will be important to evaluate the calibration component of *B* in combination with the ROC analysis.

6. Conclusions

If we consider *B* in the context of SDT and assume that it makes sense to consider the existence of *B* as a parameter for the population (that is, $\overline{B}$), and if we assume that the calibration in the observer's probability estimate is perfect, we have found that there is a theoretical relationship between $\overline{B}$ and $\mathrm{A_Z}$. Here, we empirically found that, if parameters b and α are constant, $\overline{B}$ values in relation to given $\mathrm{A_Z}$ values monotonically decrease as $\mathrm{A_Z}$ values increase, and that this relationship curve of $\overline{B}$ values to given $\mathrm{A_Z}$ values approaches the point, $\overline{B}=0$ and $\mathrm{A_Z}=1$; in addition, these curves have monotonically decreasing slopes. Thus, in the case in which the slopes in binormal ROC models for comparing data sets are similar, and where the calibration in the observer's probability estimate is good, *B* could be used in place of $\mathrm{A_Z}$ as an accuracy index.

Appendix 1

As mentioned in the text, let us consider a set of M signal-detection tasks with only two events (s and n), using the same notations and assumptions as in the text. Now we consider the expected value of $(y_i - p_i)^2$ in the equation (9), when the calibration in the observer's probability estimate is perfect. Since, in that case, p_i in the equation (9) is given by the equation (10) from the Bayes theorem, the expected value of $(y_i - p_i)^2$, $\mathrm{E}\left[(y_i - p_i)^2\right]$, is given by

$$\begin{aligned}
\mathrm{E}\left[(y_i - p_i)^2\right] &= \int_{-\infty}^{\infty}[1 - p(x)]^2 \alpha f(x \mid s)dx + \int_{-\infty}^{\infty} p^2(x)(1-\alpha) f(x \mid n)dx \\
&= \int_{-\infty}^{\infty}\left[1 - \frac{\alpha f(x \mid s)}{(1-\alpha) f(x \mid n) + \alpha f(x \mid s)}\right]^2 \alpha f(x \mid s)dx \\
&\quad + \int_{-\infty}^{\infty}\left[\frac{\alpha f(x \mid s)}{(1-\alpha) f(x \mid n) + \alpha f(x \mid s)}\right]^2 (1-\alpha) f(x \mid n)dx \\
&= \int_{-\infty}^{\infty} \frac{(1-\alpha)\alpha f(x \mid n) f(x \mid s)}{(1-\alpha) f(x \mid n) + \alpha f(x \mid s)} dx
\end{aligned} \tag{A1}$$

From the law of large numbers, the convergence in probability of B as M tends to infinity is given by $\int_{-\infty}^{\infty} \frac{(1-\alpha)\alpha f(x \mid n) f(x \mid s)}{(1-\alpha) f(x \mid n) + \alpha f(x \mid s)} dx$. Therefore, when the Brier score as a parameter for the population under the assumption that the calibration in the observer's probability estimate is perfect, $\overline{B}$ can be given by

$$\begin{aligned}
\overline{B} &= \int_{-\infty}^{\infty} \frac{(1-\alpha)\alpha f(x \mid n) f(x \mid s)}{(1-\alpha) f(x \mid n) + \alpha f(x \mid s)} dx \\
&= \int_{-\infty}^{\infty} \frac{(1-\alpha)\alpha f(x \mid s)}{(1-\alpha) + \alpha \dfrac{f(x \mid s)}{f(x \mid n)}} dx \\
&= \int_{-\infty}^{\infty} \frac{(1-\alpha)\alpha f(x \mid n)}{\alpha + (1-\alpha) \dfrac{f(x \mid n)}{f(x \mid s)}} dx
\end{aligned} \tag{A2}$$

Thus, $\overline{B}$ is given by

$$\overline{B} = \int_0^1 \frac{(1-\alpha)\alpha}{(1-\alpha)+\alpha\dfrac{dP(S\,|\,s)}{dP(S\,|\,n)}} dP(S\,|\,s), \tag{A3}$$

or,

$$\overline{B} = \int_0^1 \frac{(1-\alpha)\alpha}{\alpha+(1-\alpha)\dfrac{dP(S\,|\,n)}{dP(S\,|\,s)}} dP(S\,|\,n). \tag{A4}$$

Appendix 2

Spiegelhalter has shown that the expected B ($EBrier$) is given as

$$EBrier = \frac{1}{M}\sum_{i=1}^{M} p_i(1-p_i), \tag{A5}$$

under the null hypothesis of perfect calibration [11]. Here, when the calibration in the observer's probability estimate is perfect, and p_i is given by the equation (10), the expected value of $p_i(1-p_i)$, $\mathrm{E}[p_i(1-p_i)]$, is expressed as,

$$\begin{aligned}\mathrm{E}[p_i(1-p_i)] &= \int_{-\infty}^{\infty} p(x)[1-p(x)]\alpha f(x\,|\,s)dx \\ &\quad + \int_{-\infty}^{\infty} p(x)[1-p(x)](1-\alpha)f(x\,|\,n)dx. \\ &= \int_{-\infty}^{\infty} \frac{(1-\alpha)\alpha f(x\,|\,n)f(x\,|\,s)}{(1-\alpha)f(x\,|\,n)+\alpha f(x\,|\,s)}dx\end{aligned} \tag{A6}$$

Thus, from the law of large numbers, the convergence in probability of $EBrier$ as M tends to infinity is given by $\int_{-\infty}^{\infty} \frac{(1-\alpha)\alpha f(x\,|\,n)f(x\,|\,s)}{(1-\alpha)f(x\,|\,n)+\alpha f(x\,|\,s)}dx$. That is, $EBrier$ is equal to $\overline{B}$.

References

[1] R.M. Poses, R.D. Cebul, R.M. Centor, Evaluating physicians' probabilistic judgments, Med. Decis. Making. 8 (1988) 233-240.

[2] J.W. Gurney, Neural networks at the crossroads: caution ahead, Radiology 193 (1994) 27-30.

[3] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29-36.

[4] J.A. Swets, R.M. Pickett, Evaluation of Diagnostic Systems: Methods from Signal Detection Theory, Academic Press, New York, 1982.

[5] C.E. Metz, B.A. Herman, J.H. Shen, Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data, Stat. Med. 17 (1998) 1033-53.

[6] D. Bamber, The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, J. Math. Psych. 12 (1975) 387-415.

[7] D.M. Green, J.A. Swets, Signal Detection Theory and Psychophysics, Wiley, New York, 1966.

[8] J.P. Egan, Signal Detection Theory and ROC Analysis, Academic Press, New York, 1975.

[9] A.J. Simpson, M.J. Fitter, What is the best index of detectability?, Psychol. Bull. 80 (1973) 481-488.

[10] D.A. Redelmeier, D.A. Bloch, D.H. Hickam, Assessing predictive accuracy: how to compare Brier scores, J. Clin. Epidemiol. 44 (1991) 1141-1146.

[11] D.J. Spiegelhalter, Probabilistic prediction in patient management and clinical trials, Stat. Med. 5 (1986) 421-433.

[12] E. Somoza, Classifying binormal diagnostic tests using separation-asymmetry diagrams with constant-performance curves, Med. Decis. Making. 14 (1994) 157-168.

[13] D. Mossman, Resampling techniques in the analysis of non-binormal ROC data, Med. Decis. Making. 15 (1995) 358-366.

[14] J.F. Yates, External correspondence: decompositions of the mean probability score, Org. Behav. Human. Perf. 30 (1982) 132-156.

[15] D.K. McClish, S.H. Powell, How well can physicians estimate mortality in a medical intensive care unit?, Med. Decis. Making. 9 (1989) 125-132.

[16] J.W. Gurney, S.J. Swensen, Solitary pulmonary nodules: determining the likelihood of malignancy with neural network analysis, Radiology 196 (1995) 823-829.

Figure Legends

Fig. 1.

Our derived theoretical curve of $\overline{B}$ as a function of A_Z, where parameters a and α are fixed with $a = 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0$ and $\alpha = 0.5$. $\overline{B}$ = the Brier score as a parameter for the population under the assumption that the calibration in the observer's probability estimate is perfect; A_Z = the area under the receiver operating characteristic (ROC) curve for the binormal model; a = the parameter in the binormal ROC model; α = the prior probability of signal events.

Fig. 2.

Our derived theoretical curve of $\overline{B}$ as a function of A_Z, where parameters b and α are fixed with $b = 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0$ and $\alpha = 0.5$. $\overline{B}$ = the Brier score as a parameter for the population under the assumption that the calibration in the observer's probability estimate is perfect; A_Z = the area under the receiver operating characteristic (ROC) curve for the binormal model; b = the parameter in the binormal ROC model; α = the prior probability of signal events.

Fig. 3.

Our derived theoretical curve of $\overline{B}$ as a function of AUC, where parameters b and α are fixed with $b = 1.0$ and $\alpha = 0.1(0.1)0.5$. (A) the one for the Power-ROC model. (B) the one for the binormal ROC model. $\overline{B}$ = the Brier score as a parameter for the population under the assumption that the calibration in the observer's probability estimate is perfect; AUC = the area under the receiver operating characteristic (ROC) curve; b = the parameter in the binormal ROC model; α = the prior probability of signal events.

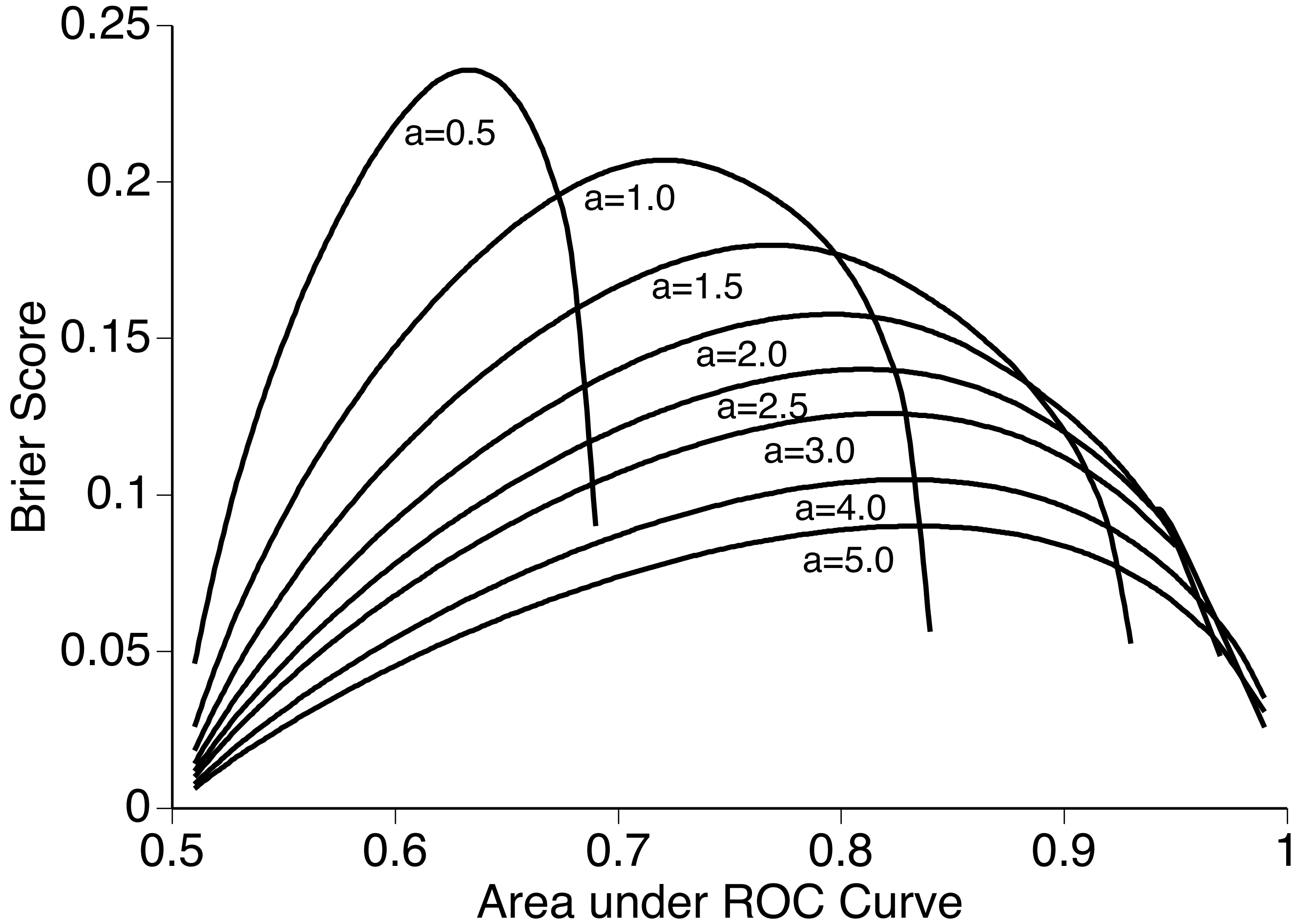
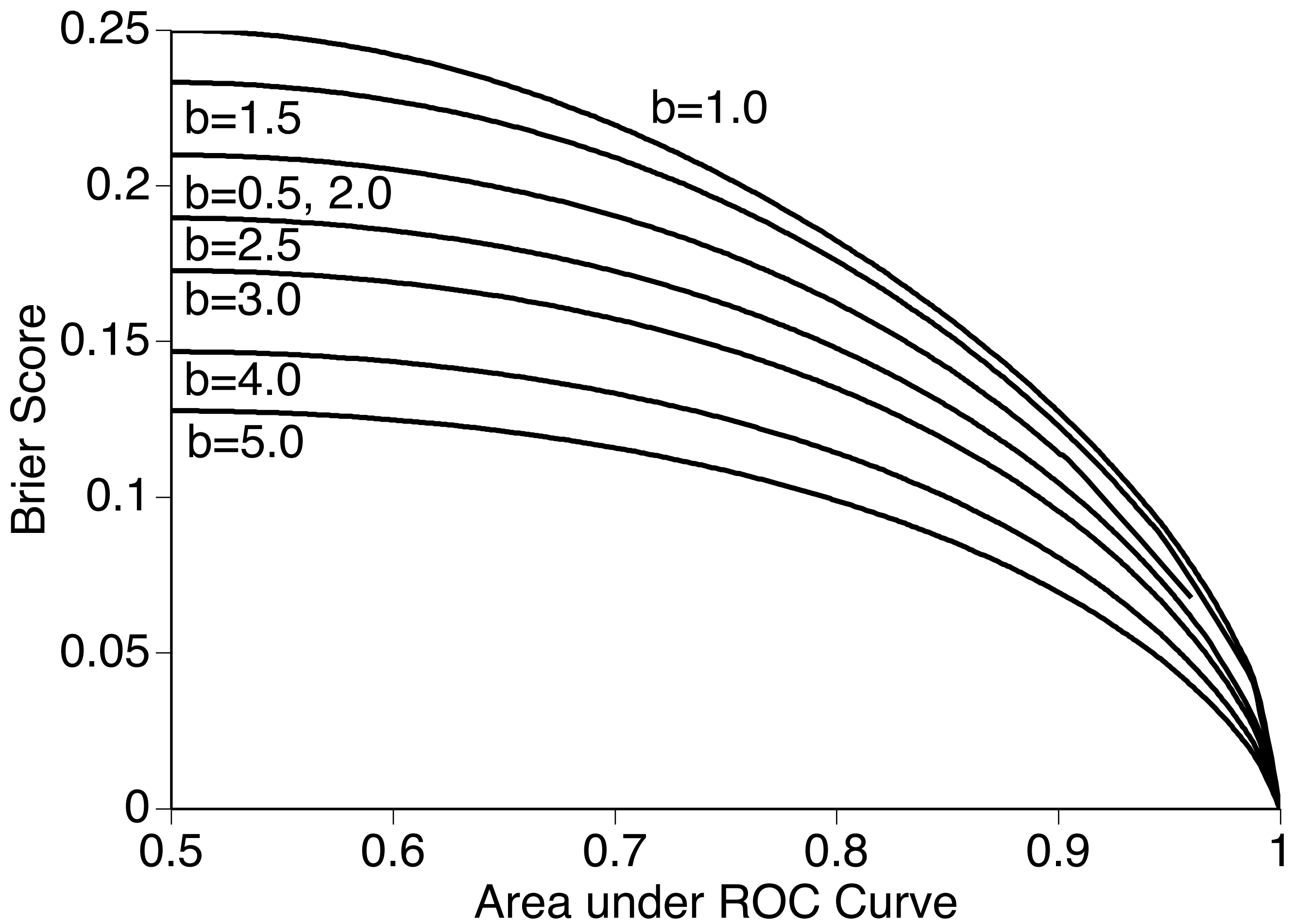
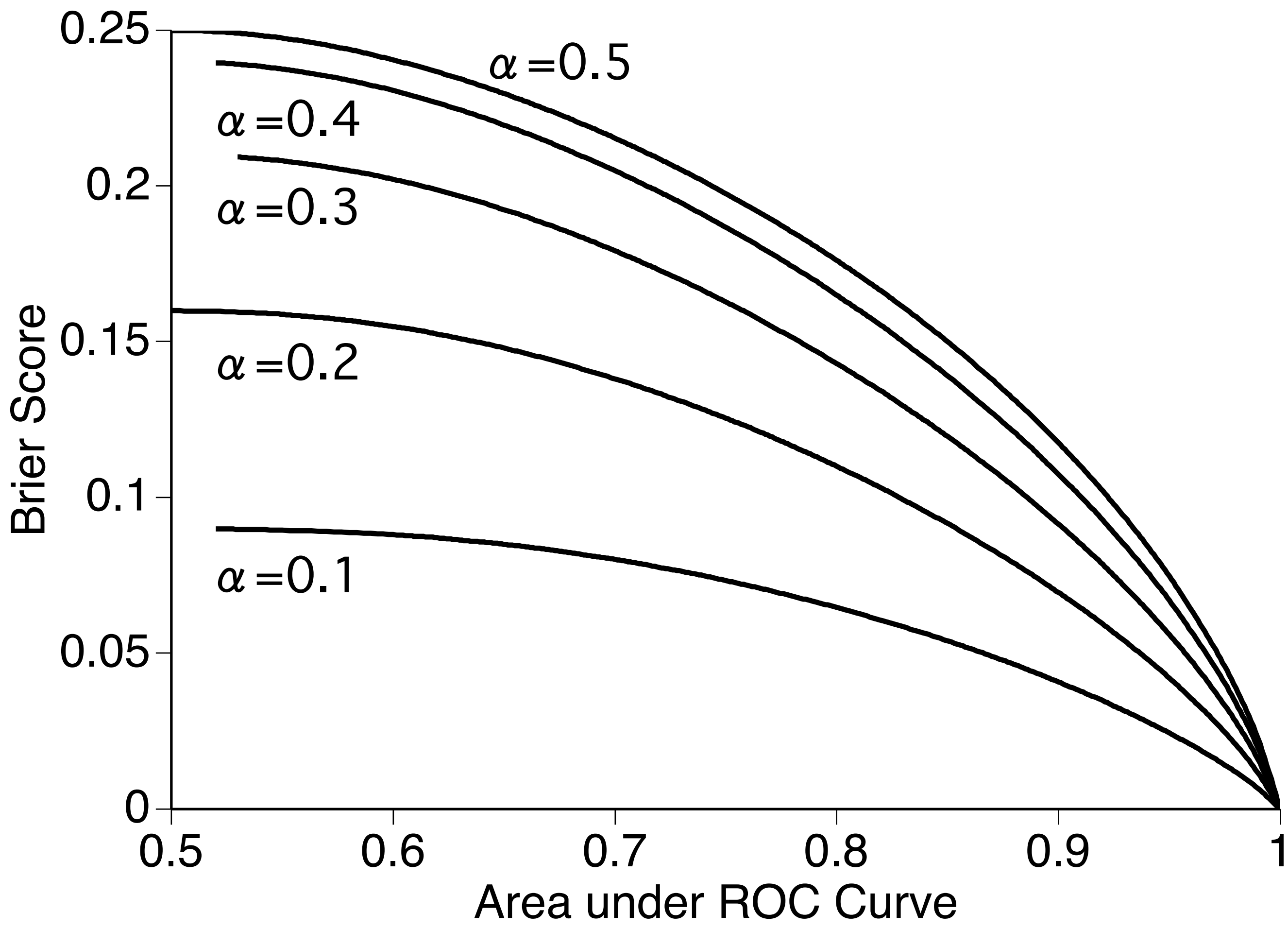

0.25
0.2
0.15
0.1
0.05
0
0.5
0.6
0.7
0.8
0.9
1
Brier Score
Area under ROC Curve
a=0.5
a=1.0
a=1.5
a=2.0
a=2.5
a=3.0
a=4.0
a=5.0

Fig. 1.



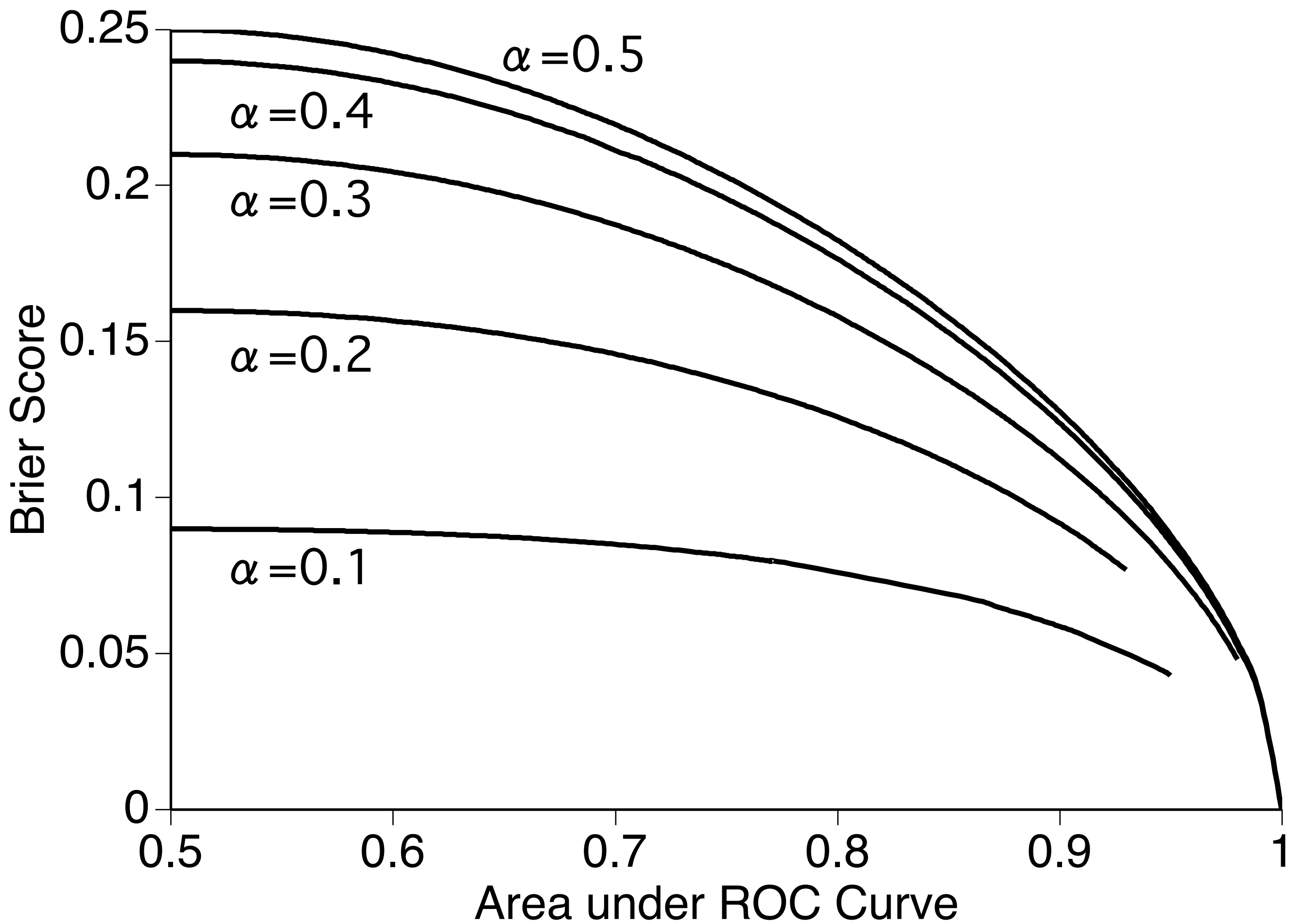
0.25
0.2
0.15
0.1
0.05
0
Brier Score
0.5
0.6
0.7
0.8
0.9
1
Area under ROC Curve
b=1.0
b=1.5
b=0.5, 2.0
b=2.5
b=3.0
b=4.0
b=5.0

Fig. 2.



0.25
0.2
0.15
0.1
0.05
0
0.5
0.6
0.7
0.8
0.9
1
Brier Score
Area under ROC Curve
$\alpha=0.5$
$\alpha=0.4$
$\alpha=0.3$
$\alpha=0.2$
$\alpha=0.1$

Fig. 3 (A).



0.25
0.2
0.15
0.1
0.05
0
0.5
0.6
0.7
0.8
0.9
1
Brier Score
Area under ROC Curve
$\alpha = 0.5$
$\alpha = 0.4$
$\alpha = 0.3$
$\alpha = 0.2$
$\alpha = 0.1$

Fig. 3 (B).