Evaluating the incremental value of a new model: Area under the ROC curve or under the PR curve

Qian M. Zhou\textsuperscript{a,1,2}, Zhe Lu\textsuperscript{b}, Russell J. Brooke\textsuperscript{c}, Melissa M Hudson\textsuperscript{c}, Yan Yuan\textsuperscript{b,2}

\textsuperscript{a} Department of Mathematics and Statistics, Mississippi State University, Starkville, Mississippi 39762, USA
\textsuperscript{b} School of Public Health, University of Alberta, Edmonton, AB T6G1C9, Canada
\textsuperscript{c} St Jude Children’s Research Hospital, Memphis, TN 38105, USA

Abstract

Incremental value (IncV) evaluates the performance improvement from an existing risk model to a new model. In this paper, we compare the IncV of the area under the receiver operating characteristic curve (IncV-AUC) and the IncV of the area under the precision-recall curve (IncV-AP). Since they are both semi-proper scoring rules, we also compare them with a strictly proper scoring rule: the IncV of the scaled Brier score (IncV-sBrS). The comparisons are demonstrated via a numerical study under various event rates. The results show that the IncV-AP and IncV-sBrS are highly consistent, but the IncV-AUC and the IncV-sBrS are negatively correlated at a low event rate. The IncV-AUC and IncV-AP are the least consistent among the three pairs, and their differences are more pronounced as the event rate decreases. To investigate this phenomenon, we derive the expression of these two metrics. Both are weighted averages of the changes (from the existing model to the new one) in the separation of the risk score distributions between events and non-events. However, the IncV-AP assigns heavier weights to the changes in the higher risk group, while the IncV-AUC weighs the entire population equally. We further illustrate this point via a data example of two risk models for predicting acute ovarian failure. The new model has a slightly lower AUC but increases the AP by 48\%. We conclude that when the objective is to identify the high-risk group, the IncV-AP is a more appropriate metric, especially when the event rate is low.

Keywords: High-risk group identification; Prediction performance; Brier score; Proper scoring rules; Rare outcome

\textsuperscript{1}Dr. Zhou is the corresponding author.
\textsuperscript{2}Dr. Yuan and Dr. Zhou are equally contributing senior authors.
1 Introduction

Risk prediction is a crucial component in managing disease prognosis. Numerous research have been dedicated to continually updating risk models for better prediction accuracy. For example, several papers have investigated the improvement in predicting the risk of cardiovascular disease by adding new biomarkers to the existing Framingham risk model, such as the C-reactive protein Cook et al. (2006); Buckley et al. (2009), or more recently, a polygenic risk score Mosley et al. (2020); Elliott et al. (2020).

In some applications, an existing marker is replaced with a new marker that provides more precise information. For example, cancer treatment such as radiation can have significant long-term health consequences for cancer survivors. Prescribed radiation doses to body regions, such as the abdomen and chest, are routinely available in medical charts. But to predict the risk of an organ-specific outcome, e.g., secondary lung cancer or ovarian failure, a more precise measurement of the radiation exposure to specific organs provides better information. Radiation oncologists developed and applied algorithms to estimate these organ-specific exposures Howell et al. (2019).

The measurement of a new marker or the more precise measurement of a known risk factor is often costly and time-consuming. Thus, it is vital to evaluate the incremental value (IncV) of the risk model that incorporates the new information. The IncV has primarily been discussed in settings where new markers are added to the existing risk profile Pencina et al. (2008); Pepe et al. (2013). In this paper, the term IncV refers to the change of the prediction performance whenever an existing risk model is compared with a new one.

For evaluating risk models in medical research, two curves have been considered: the receiver operating characteristic (ROC) curve and the precision-recall (PR) curve Pepe (2003); Yuan et al. (2015). The ROC curve is a curve of the true positive rate (TPR) versus the false positive rate (FPR), while the PR curve is a curve of the positive predicted value (PPV) versus the TPR. The ROC curve has been the most popular tool in medical research, dating back to the 1960’s when it was applied in diagnostic radiology and diagnostic imaging systems Zweig and Campbell (1993); Pepe (2003). The PR curve was first considered in the information retrieval community Raghavan et al. (1989); Manning and Schütze (1999), and is gaining more popularity in medical research Badawi et al. (2018); Chaudhury et al. (2019); Tang et al. (2019); Xiao et al. (2019).

The area under the ROC curve (AUC) captures the discriminatory ability of a model, i.e., how well a model separates the events (subjects who experience the event of interest) from the non-events (subjects who are event-free). The area under the PR curve, called the average PPV or the average precision (AP) Yuan et al. (2015, 2018), evaluates the prospective prediction performance of a risk model, because it is conditional on the risk score obtained at baseline. In contrast, the AUC is a retrospective metric that is conditional on the future disease status, unknown at baseline.

The IncV can be quantified as the change of the AUC or AP from an existing risk model to a new one. In this paper, we are interested in the comparison between the IncV in AUC (IncV-AUC) and the IncV in AP (IncV-AP). We do not consider two other popular IncV metrics: net reclassification improvement (NRI) and integrated discrimination improvement (IDI) Pencina et al. (2008), because neither is a proper scoring rule. A misspecified model may have better performance than the true model based on the NRI or IDI Pepe et al.
Both the IncV-AUC and IncV-AP are proper scoring rules, i.e., the true model has the maximum AUC and AP among all the models. However, our results of a numerical study show that these two metrics do not agree in various scenarios when neither of the two competing working risk models is the true model, especially when the event rate is low. How can we judge which one is better? The answer should depend on the purpose of the risk model. Is it for making a diagnosis, screening for subjects with some disease, or identifying the high-risk subjects that will develop an event? The purpose requires that the risk model performs well in one or more areas, such as calibration, discrimination, and identification of the high-risk group. We show that different IncV metrics focus on different aspects of the model performance. Thus, one metric is more appropriate than the other in a particular setting.

As the main contribution, we express both the IncV-AUC and IncV-AP as functions of how far the risk score’s distributions are separated between the events and non-events. They are both weighted averages of the change in the separation of these two distributions from the existing risk model to the new one. The discrepancy between these two metrics result from how they weigh the change in the separation. The IncV-AUC weighs equally over the entire population, but the IncV-AP gives more weights to the subjects with higher risk scores. Thus, when the study objective concerns the high-risk group, the IncV-AP is more appropriate than the IncV-AUC.

Additionally, since both metrics are rank-based, they are semi-proper scoring rules: a misspecified working risk model can have the same AUC and AP as the true model when they assign the same ranks to the subjects. Thus, we compare them with the Brier score (BrS), the only strictly proper scoring rule. The BrS can be rescaled to range between 0 and 1 with larger values indicating better performance [Steyerberg et al. (2010)]. In Kattan and Gerds (2018), the scaled Brier score (sBrS) is referred to as the index of prediction accuracy [Kattan and Gerds (2018)]. In our numerical study, we compare the IncV-AUC and IncV-AP with the IncV in sBrS.

The remainder of the paper is organized as follows. In Section 2 we describe the three accuracy measures: AUC, AP, and sBrS, as well as their respective IncV metrics. In Section 3 we present the numerical study in which the IncV is used to evaluate adding a new marker to an existing risk model. Our numerical study does not involve generating data or estimating the parameters, as our interest is in the IncV parameters themselves. In Section 4 we evaluate two risk models for predicting acute ovarian failure using the three IncV metrics with the data from the St. Jude Lifetime Cohort. Discussions are in Section 5.

2 AUC, AP, and Brier Score

Let $D = 0$ or 1 denote a binary outcome. For studies with an event time $T$, define $D = I(T \leq \tau)$ for a given prediction time period $\tau$, which indicates that the outcome is time-dependent. In the rest of the manuscript, we refer the subjects with $D = 1$ as the events, and those with $D = 0$ as the non-events. Let $\pi = Pr(D = 1)$ denote the event rate.
2.1 Risk model and risk score

A risk model is a function of a set of predictors $X = (X_1, \cdots, X_{k-1})$, which might include interaction terms and polynomial terms, to obtain the probability of $D = 1$. Usually, we write this model as a regression model

$$p(X) = g(\beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1}), \quad (1)$$

where $g(\cdot)$ is a smooth and monotonic link function, such as a logit link. For the censored event time outcomes, a risk model could be Cox’s proportional hazards model [Cox et al. (1972)] or the time-specific generalized linear model [Uno et al. (2007)]; both models can be expressed as equation (1).

In practice, the underlying data generating mechanism is often complicated, and the working risk model in equation (1) is usually misspecified. Let $\tilde{p}(X)$ denote the resulting estimates, and $\hat{p}(X) = g(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_{k-1} X_{k-1})$ is the estimated risk given $X$.

Remark 1. In practice, the working risk $p(X)$ is estimated from a data sample, denoted as $\{(D_i, X_i), i = 1, \cdots, n\}$. The regression coefficients $\beta = (\beta_0, \beta_1, \cdots, \beta_{k-1})'$ are estimated by solving an estimating equation: $\Psi(\beta) = \sum_{i=1}^{n} \Psi(\beta; D_i, X_i) = 0$. Let $\hat{\beta}_j$, $j = 0, 1, \cdots, k - 1$, denote the resulting estimates, and $p(X) = g(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_{k-1} X_{k-1})$ is the estimated risk given $X$.

However, in this paper, our interest is not in the estimated working risk $\hat{p}(X)$. Instead, we investigate the predictive performance of the population working risk $p(X) = g(\beta_0^* + \beta_1^* X_1 + \cdots + \beta_{k-1}^* X_{k-1})$ where $\beta_j^*$ is the limiting value of $\hat{\beta}_j$ as $n \to \infty$, obtained by solving the following estimating equation: $E_{(D,X)} [\Psi(\beta)] = 0$. Here, the expectation is taken with respect to (w.r.t.) the true joint distribution of $(D, X)$.

2.2 Accuracy measures and IncV metrics

In this section, we describe three accuracy measures: BrS, AUC, and AP. The BrS needs to be defined on the working risk $p(X)$. The AUC and AP are both rank-based, and thus, they can be defined on any risk score $r(X)$.

BrS. BrS is the expected squared difference between the binary outcome $D$ and $p(X)$, i.e., $\text{BrS} = E_{(D,X)} \{[D - p(X)]^2\}$. This measure depends on the event rate $\pi$ [Assel et al. (2017)]. The BrS is minimized at the true model, i.e., $p(X) = \pi(X)$, and thus, it is a proper scoring
rule. A non-informative model assigns the event rate to every subject, i.e., \( p(X) \equiv \pi \). This model leads to the maximum BrS value \( \pi (1 - \pi) \). A scaled Brier score (sBrS) is defined as \( \text{sBrS} = 1 - \frac{\text{BrS}}{\pi (1 - \pi)} \), ranging from 0 and 1, with larger values indicating better performance [Steyerberg et al. (2010)].

**ROC curve and AUC.** The ROC curve is a curve of the TPR versus the FPR. Given a cut-off value \( c \), the TPR and FPR are the proportions of higher risk \( r(X) > c \) among the events and non-events respectively, i.e., \( \text{TPR}(c) = \Pr [r(X) > c \mid D = 1] \) and \( \text{FPR}(c) = \Pr [r(X) > c \mid D = 0] \). The AUC can be interpreted as the conditional probability that given a pair of an event and a non-event, the event is assigned with a higher risk score, i.e., \( \text{AUC} = \Pr [p(X_i) > p(X_j) \mid D_i = 1, D_j = 0] \). Unlike the BrS, the AUC is conditional on the event status, and thus, it does not depend on \( \pi \). It is also a proper scoring rule because the true model maximizes the ROC curve at each point and has the maximum AUC.

The ROC curve is maximized at each point under the true model [McIntosh and Pepe (2002)], and thus, the true model has the maximum AUC, which makes the AUC a proper scoring rule.

**PR curve and AP.** The PR curve is a curve of the PPV versus the TPR. The PPV is defined as \( \text{PPV}(c) = \Pr [D = 1 \mid r(X) > c] \), the proportion of subjects with higher risk scores that are events. Davis and Goadrich (2006) showed that if one ROC curve is above another curve everywhere, their PR curves have the same relationship [Davis and Goadrich (2006)]. Thus, the AP is a proper scoring rule as it is maximized under the true model. It can be expressed as \( \text{AP} = \mathbb{E} [\text{PPV}(r_1(X))] \) [Yuan et al. (2018)], where \( r_1(X) \) denotes the risk score of an event, and the expectation is taken w.r.t. the distribution of \( r_1(X) \). Like the BrS, the AP is event-rate dependent [Yuan et al. (2018)].

**IncV.** Let \( \Psi_{\text{old}} \) and \( \Psi_{\text{new}} \) denote the accuracy measure \( \Psi \) (AUC, AP, or sBrS) of the existing and new risk models respectively. The IncV parameter is defined as \( \Delta \Psi = \Psi_{\text{new}} - \Psi_{\text{old}} \), which quantifies the change in \( \Psi \) when comparing the new model to the existing one.

In Section 3, we compare these three IncV parameters: \( \Delta \text{AUC} \), \( \Delta \text{AP} \), and \( \Delta \text{sBrS} \) via a numerical study in which a new marker is added to an existing model. We observe that the \( \Delta \text{AUC} \) and \( \Delta \text{AP} \) is the least consistent among all the three pairs, especially when the event rate is low. For example, when \( \pi = 0.01 \), in about 20% of the scenarios in our study, one metric is positive but the other is zero or negative. Their Pearson correlation coefficient is also negative. Thus, we next examine the \( \Delta \text{AUC} \) and \( \Delta \text{AP} \) in more depth.

### 2.3 AUC versus AP: How risk scores are separated between events and non-events

Consider the following example of three risk scores: \( r_1, r_2, \) and \( r_3 \). Assume that all the risk scores among non-events follow a standard normal distribution, i.e., \( r_j \mid D = 0 \sim N(0, 1) \), for \( j = 1, 2, 3 \). However, their distributions among events are different: (i) \( r_1 \mid D = 1 \sim N(1.8, 2) \), (ii) \( r_2 \mid D = 1 \sim N(1.5, 1.5) \), and (iii) \( r_3 \mid D = 1 \sim N(3, 1.5) \).
Figure 1 illustrates the comparisons of these three risk scores under an event rate $\pi = 0.05$. Figure 1 (a) shows the density curves of these three risk scores for events and non-events. Among them, the two density curves for $r_3$ are the most separated. Thus, the ROC and PR curves of $r_3$ dominate those of $r_1$ and $r_2$ (Figure 1 (b)), and consequently, $r_3$ has the largest AUC and AP. In contrast, the ROC and PR curves of $r_1$ and $r_2$ cross: $r_2$ has a slightly larger AUC with $\Delta \text{AUC}_{r_2} = 0.007$, but $r_1$ has a larger AP with $\Delta \text{AP}_{r_1} = 0.096$.

We next show that both the AUC and AP depend on the separation of the risk score distributions between events and non-events. Let $F_1(x)$ and $F_0(x)$ denote the cumulative distribution functions (CDFs) of a risk score $r(X)$ conditional on $D = 1$ (events) and $D = 0$ (non-events), respectively. Let $q_\alpha = F_1^{-1}(\alpha)$ denote the $\alpha$-th quantile for the distribution $F_1$, $0 \leq \alpha \leq 1$. The AUC and AP can be expressed as,

$$\text{AUC} = \int_0^1 F_0(q_\alpha) d\alpha,$$

(2)

$$\text{AP} = \int_0^1 \left\{ 1 + \frac{\pi - 1}{1 - \alpha} \left[ 1 - F_0(q_\alpha) \right] \right\}^{-1} d\alpha.$$

(3)

The derivation is provided in Appendix A.1.

As shown above, both the AUC and AP are functions of $F_0(q_\alpha)$, the proportion of non-events whose risk scores are below the $\alpha$-th quantile of the risk scores among events. The $F_0(q_\alpha)$ measures the separation of the two distributions $F_1$ and $F_0$: more non-events having lower risk scores indicates a further separation between these two distributions at a given $\alpha$. For example, the $F_0(q_\alpha)$ curve of $r_3$ dominates those of $r_1$ and $r_2$ (Figure 1 (b)), which is consistent with the fact that $r_3$ has the best separation between events and non-events (Figure 1 (a)).

Based on equations (2) and (3), we can express the $\Delta \text{AUC}$ and $\Delta \text{AP}$ as

$$\Delta \Psi = \int_0^1 w_\Psi(\alpha) \Delta(\alpha) d\alpha, \quad \Psi = \text{AUC or AP},$$

(4)

where $\Delta(\alpha) = F_{\text{new},0}(q_{\text{new},\alpha}) - F_{\text{old},0}(q_{\text{old},\alpha})$, capturing how much the new working risk model changes the separation of these two distributions at a given $\alpha$. Because it is conditional on the outcome, $\Delta(\alpha)$ is independent of $\pi$.

Based on equation (4), the $\Delta \text{AUC}$ and $\Delta \text{AP}$ are weighted averages of $\Delta(\alpha)$, but their weights are different. For the $\Delta \text{AUC}$, $w_{\text{AUC}} \equiv 1$; in other words, the $\Delta(\alpha)$ is equally weighted. In contrast, for the $\Delta \text{AP}$,

$$w_{\text{AP}}(\alpha) = \frac{\pi^{-1} - 1}{1 - \alpha} \left[ 1 + (\pi - 1) \frac{1 - F_{\text{new},0}(q_{\text{new},\alpha})}{1 - \alpha} \right]^{-1} \left[ 1 + (\pi - 1) \frac{1 - F_{\text{old},0}(q_{\text{old},\alpha})}{1 - \alpha} \right].$$

(5)

which varies with the values of $\alpha$ and $\pi$.

**Weight $w_{\text{AP}}$.** Let us compare $r_1$ and $r_2$, assuming $r_2$ is from an existing risk model and $r_1$ is from a new one. The comparison is illustrated in Figure 2 for three different event rates $\pi = 0.2, 0.05, \text{and} 0.01$. 


To visualize how $w_{\alpha}(\alpha)$ changes with $\alpha$ and $\pi$, we plot the log $w_{\alpha}(\alpha)$ against $\alpha$ for different $\pi$ in Figure 2(a). For any given $\pi$, $w_{\alpha}(\alpha)$ is an increasing function of $\alpha$. This tells us that the $\Delta AP$ assigns heavier weights to the upper-tail quantiles of the risk score, representing the higher-risk group, and lighter weights to the lower-tail quantiles, representing the lower-risk group.

This weighting mechanism demonstrates that the $\Delta AP$ emphasizes the high-risk group, whereas the $\Delta AUC$ treats the whole population equally. Thus, when the main objective is to identify the high-risk group, such as when screening for a rare outcome, the $\Delta AP$ is more appropriate than the $\Delta AUC$. On the other hand, for some classification problems, such as classifying early-stage or late-stage cancer, the whole population of cancer patients may be of interest. The $\Delta AUC$ serves this purpose better.

Back to the comparison between $r_1$ and $r_2$. The AUCs of $r_1$ and $r_2$ are similar but their APs differ: $\Delta AUC = AUC_{r_1} - AUC_{r_2} = -0.007$ and $\Delta AP = AP_{r_1} - AP_{r_2} = 0.096$ (Figure 1(b)). This phenomenon results from their different weighting mechanisms. Specifically, compared to $r_2$, $r_1$ has a better separation for the upper quantiles of the risk score, but worse for the lower quantiles. As shown in Figure 2(b), $\Delta (AP)$ of $r_1$ is an increasing function of $\alpha$, whereas the $\Delta (AP)$ treatment of the lower-risk group.

Besides $\alpha$ and $\pi$, the weight $w_{\alpha}(\alpha)$ in equation (5) also depends on $F_{new,0}(q_{new,\alpha})$ and $F_{old,0}(q_{old,\alpha})$. In general, $F_{0}(q_{\alpha}) \geq \alpha$ because the density curve for non-events is to the left of that for events. Thus, how the weight changes with $\alpha$ and $\pi$ is mainly determined by the numerator $(\pi^{1-1} - 1)/(1 - \alpha)$. However, when $\pi$ and $\alpha$ are fixed, larger values of $F_{old,0}(r_{old,\alpha})$ or $F_{new,0}(r_{new,\alpha})$ or both, i.e., better performance of at least one model, lead to larger weights.

Remark 3. Although the BrS cannot be directly expressed as a function of $F_{0}(q_{\alpha})$, it is closely related with the two distributions $F_1$ and $F_0$. Specifically, it can be written as

$$BrS = E \left\{ [1 - p(X)]^2 \mid D = 1 \right\} \pi + E \left\{ [p(X)]^2 \mid D = 0 \right\} (1 - \pi).$$

The first expectation can be expressed as $Var [p(X) \mid D = 1] + E[p(X) \mid D = 1 - 1]^2$. This quantity is determined by $F_1$, the distribution of the working risk $p(X)$ among the events, in terms of its spread and how far its mean is from 1. Similarly, the second expectation, expressed as $Var [p(X) \mid D = 0] + E[p(X) \mid D = 0]^2$, is determined by the spread of the risk score’s distribution among the non-events as well as the distance of its mean from 0.

A smaller BrS can result from one, or a combination, of the following: (i) the mean of the working risk for events closer to 1, (ii) the mean of the working risk for non-events closer
to 0, (iii) less variation in the working risk for events or non-events or both. All of these lead to a further separation of the risk score’s distributions between events and non-events.

3 Numerical Study

As mentioned in Remark 1, we are interested in the IncV parameters defined on the population working risk, not in the IncV estimates obtained from a sample. Thus, we do not use simulation studies. Here, we evaluate the IncV of adding a new marker, denoted by $Y$, to a model with an existing marker, denoted by $X$. We compare the following three IncV parameters: $\Delta \text{AUC}$, $\Delta \text{AP}$, and $\Delta \text{sBrS}$, which can be directly derived from the distributional assumptions described below.

3.1 Setting: Distributional assumptions

Let the markers $X$ and $Y$ be independent standard normal random variables. Given the values of these two markers, a binary outcome $D$ follows a Bernoulli distribution with the probability of $D = 1$ via the following model:

$$
\pi(X,Y) = \Phi(\beta_0 + \beta_1 X + \beta_2 Y + \beta_3 XY),
$$

where $\Phi(\cdot)$ is the CDF of a standard normal distribution. Given $X$ and $Y$, $\pi(X,Y)$ is the true risk. The true model in equation (6) includes an interaction between $X$ and $Y$, indicating the effect of $X$ on the risk changes with the value of $Y$, and vice versa.

Typically, in practice, none of the working models are the true model. Having this in mind, we compare the following two misspecified working models: (i) one-marker model: $p(X) = \Phi(\gamma_0 + \gamma_1 X)$, and (ii) two-marker model: $p(X,Y) = \Phi(\gamma_0 + \gamma_1 X + \gamma_2 Y)$.

Here, we consider different values of $\beta_1$, $\beta_2$, $\beta_3$ and $\pi$: $\beta_1 = 0.3, 0.4, \ldots, 0.9, 1$, $\beta_2 = 0.3, 0.4, \ldots, 0.9, 1$, $\beta_3 = -0.5, -0.4, \ldots, -0.1, 0.1, \ldots, 0.4, 0.5$ (excluding 0), and $\pi = 0.01, 0.05, 0.1, 0.2, 0.5$. Each combination of $(\beta_1, \beta_2, \beta_3, \pi)$ values is referred to as a scenario. Given a scenario, the value of $\beta_0$ can be determined. In the supplementary material, we explain how to obtain the value of $\beta_0$ and calculate the AUC, AP, and sBrS of the one-marker and two-marker models as well as the IncV parameter.

3.2 Results

We compare these three IncV parameters in two main aspects: (1) size and range, and (2) agreement. A desirable IncV metric should be sensitive to the change in the predictive performance. If a new model improves the prediction accuracy, an IncV should have a sizable positive value. It should also be able to reflect a performance deterioration with a sizable negative value. When an IncV being close to 0 occurs often, we might question its utility in decision-making. We are also interested in the agreement among the three IncV parameters, given that they focus on different aspects of the prediction performance.
3.2.1 Size and Range

Figure 3 plots the summary statistics of the three IncV metrics under different event rates: minimum, 25-th percentile, median, 75-th percentile, and maximum. The ∆AP has the widest range, followed by ∆sBrS, and ∆AUC has the narrowest range.

This difference between the ∆AUC and ∆AP is more evident for a lower event rate. For example, under $\pi = 0.01$, the inter-quartile range (IQR) and median of ∆AUC are both 0.07. It implies that the magnitude is rather small most of the time. Even if a new marker is useful, the evidence, reflected by the size of ∆AUC, is not compelling. In contrast, the IQR of ∆AP is much wider, about 0.41, with a median of 0.21.

In addition, the ∆AUC is negative in less than 1% of the scenarios (29 out of 3200). Furthermore, when it is negative, the value is very close to 0, which indicates that ∆AUC cannot distinguish between a useless model and harmful model Kattan and Gerds (2018). On the other hand, the ∆AP is negative in about 12% of the scenarios (389 out of 3200), with a much larger size.

As $\pi$ changes, the range of ∆AP varies the most among the three, whereas the quantiles of ∆AUC remain almost constant. As $\pi$ increases, the ranges of all the IncV metrics get narrower, and closer to each other. When $\pi = 0.5$, both the ∆AUC and ∆AP range from 0.015 to 0.25 with a median of 0.089, and the ∆sBrS ranges from 0.019 to 0.32 with a median of 0.12.

3.2.2 Agreement

**Correlation.** We calculate the Pearson correlation between each pair of the IncV metrics under each $\pi$ (Table 1). The ∆AP and ∆sBrS are highly correlated for all values of $\pi$. As $\pi$ increases, their correlation decreases from about 1 ($\pi = 0.01$) to 0.84 ($\pi = 0.5$), but the correlations of ∆AUC with the other two IncV metrics increases with $\pi$. When $\pi = 0.01$, ∆AUC and ∆sBrS are negatively correlated and their correlation $-0.11$ is the smallest among the three pairs; when $\pi = 0.5$, they are the highest positively correlated. We also show the scatter plots of each pair under different $\pi$ in Figure S6 (supplementary material).

**Concordance.** The sign of an IncV metric is often used to decide whether the new model is better than the existing one. Positive IncVs favor the new model, while negative or zero values favor the existing one. Here, we define a concordance measure, which quantifies the consistency of the conclusions reached by a pair of IncV metrics.

Take the ∆AP and ∆sBrS as an example. Under a scenario, we call the pair concordant if both are $> 0$ or $\leq 0$. If one is $> 0$ and the other is $\leq 0$, the pair is discordant. The measure of concordance is defined as the proportion of scenarios where the pair is concordant minus the proportion of scenarios where it is discordant. For instance, when $\pi = 0.01$, ∆AP and ∆sBrS are concordant in about 97% of the total 640 scenarios (i.e., all the combinations of $\beta_1$, $\beta_2$, and $\beta_3$ values at each $\pi$), and discordant in about 3%. Thus, the concordance measure is 0.93.

Table 1 reports the concordance for all three pairs of the IncV metrics under each $\pi$. The results are similar to those above for the Pearson correlation. When $\pi$ is small, such as 0.01, 0.05, and 0.1, the ∆AP and ∆sBrS are the most concordant; when $\pi = 0.2$ or 0.5,
\(\Delta\text{AUC}\) and \(\Delta\text{sBr}\) are the most concordant. The \(\Delta\text{AUC}\) and \(\Delta\text{AP}\) are the least concordant for all values of \(\pi\).

When \(\pi\) is close to 0.5, the three IncV metrics tend to agree, and using any of them, we would very likely reach the same conclusion about whether the new model is better. However, when the event rate is low, i.e., for a rare outcome, the \(\Delta\text{AUC}\) can be inconsistent with both the \(\Delta\text{sBr}\) and \(\Delta\text{AP}\). This disagreement creates a dilemma: one metric tells us that the new risk model is better, but another metric indicates no improvement or even a worse performance.

### 3.2.3 \(\Delta\text{AUC}\) versus \(\Delta\text{AP}\)

Next, we single out four scenarios for an in-depth comparison between \(\Delta\text{AUC}\) and \(\Delta\text{AP}\) at \(\pi = 0.01\). The first two scenarios have similar \(\Delta\text{AUC}\) but different \(\Delta\text{AP}\) (Figure 4), whereas the next two have similar \(\Delta\text{AP}\) but different \(\Delta\text{AUC}\) (Figure 5).

**Similar \(\Delta\text{AUC}\) but different \(\Delta\text{AP}\).** The two scenarios are (i) \(\beta_1 = 1, \beta_2 = 0.8, \beta_3 = 0.2\), and (ii) \(\beta_1 = 1, \beta_2 = 0.8, \beta_3 = -0.5\). In both cases, the \(\Delta\text{AUC}\) is around 0.06, but the \(\Delta\text{AP}\) is 0.33 for scenario (i) and -0.072 for scenario (ii).

In scenario (i), both the ROC and PR curves of the two-marker model dominate those of the one-marker model, respectively. This indicates that the two-marker model is better at each point, and consequently, \(\Delta(\alpha)\) is positive throughout (Figure 4 (c)). In this case, both the \(\Delta\text{AUC}\) and \(\Delta\text{AP}\) are positive. However, the size of \(\Delta\text{AP}\) 0.33 is much larger than the \(\Delta\text{AUC}\) of 0.06, due to the large weight \(w_{\text{AP}}(\alpha)\) at the upper quantiles (Figure 4 (c)).

In scenario (ii), both the two ROC curves and PR curves cross, and \(\Delta(\alpha)\) is below zero for upper quantiles and above zero for lower quantiles (Figure 4 (c)). This implies that the two-marker model can better separate between events and non-events for the lower-risk group, but not for the higher-risk group. As a result, the \(\Delta\text{AUC}\) and \(\Delta\text{AP}\) are conflicting. The \(\Delta\text{AUC}\) is positive because the area under the \(\Delta(\alpha)\) curve above zero is slightly larger than that below zero. However, the below-zero \(\Delta(\alpha)\) are weighed heavily, and thus, the \(\Delta\text{AP}\) is negative. In such a situation, is the two-marker model better? The answer should be determined by the objective of the study. If this is a classification problem, the new marker indeed improves the discrimination on average. However, if the goal is to screen for the high-risk group, the two-marker model is worse.

**Similar \(\Delta\text{AP}\) but Different \(\Delta\text{AUC}\).** The next two scenarios are (iii) \(\beta_1 = 0.7, \beta_2 = 0.3, \beta_3 = -0.3\), and (iv) \(\beta_1 = 0.6, \beta_2 = 0.7, \beta_3 = -0.4\). In both cases, the \(\Delta\text{AP}\) values are almost 0, but \(\Delta\text{AUC}\) is approximately 0 for scenario (iii) and 0.202 for scenario (iv).

In scenario (iii), the two ROC curves and the two PR curves are almost identical. This indicates that adding the new marker does not change the separation of the distributions of the risk score between events and non-events. It is also reflected in Figure 5 (c) where the entire \(\Delta(\alpha)\) curve almost overlaps with the zero line. Thus, both the \(\Delta\text{AUC}\) and \(\Delta\text{AP}\) are close to zero. This is an example of both metrics agreeing that the new marker is “useless”.

In scenario (iv), although the two-marker model makes poorer prediction for the higher-risk group, its prediction are significantly better for the rest of the population. Thus, the \(\Delta\text{AUC}\) is positive and sizable. However, since the \(\Delta\text{AP}\) weighs heavily on the high-risk group,
the improvement for the majority of the population is offset by the worse performance at the upper quantiles, which leads to a close-to-zero $\Delta$AP. If the objective is to identify the high-risk group, the new marker is not helpful because it results in a much lower PPV for subjects with high risk scores.

### 3.3 What if the two-marker model is the true model, i.e., $\beta_3 = 0$?

Figures S8 and S9 in the supplementary material examine this question and show the scatter plots and plots of the summary statistics of the $\Delta$AUC, $\Delta$AP, and $\Delta$sBrS for different $\pi$. As expected, all the IncVs are positive. For a smaller $\pi$, the $\Delta$AP ranges wider than the $\Delta$AUC does. As $\pi$ increases, these two metrics get closer to each other. When $\pi = 0.5$, the $\Delta$sBrS has the widest range.

Since all the IncVs are positive, their concordance is all 1. Table S1 (supplementary material) lists the Pearson correlation between each pair of the IncV metrics, which are all positive. When $\pi$ is small, the $\Delta$sBrS is more strongly correlated with the $\Delta$AP than with the $\Delta$AUC. As $\pi$ increases, all three IncV metrics are strongly correlated with each other.

### 4 Data Example

We illustrate our findings via comparing two recently published risk models [Clark et al. (2020)] for predicting acute ovarian failure (AOF) among female childhood cancer survivors. AOF is a complication from acute toxicity of exposure to radiation and chemotherapy. It is defined as permanent loss of ovarian function within 5 years of cancer diagnosis or no menarche after cancer treatment by age 18. About 6% female childhood cancer survivors have AOF.

Both models include the following risk factors: age at cancer diagnosis, cumulative dose of alkylating drugs measured using the cyclophosphamide-equivalent dose, haematopoietic stem-cell transplant, and radiation exposure. The difference is in the measurement of radiation exposure. One model, called the *prescribed-dose* model, uses the prescribed radiation doses to the abdominal and pelvic regions, which are routinely available in medical charts. The other model, called the *ovarian-dose* model, uses the minimum of the organ-specific radiation exposure for both ovaries estimated by radiation oncologists.

These two models were developed using 5886 survivors from the Childhood Cancer Survivors Study (CCSS) [Robison et al. (2002)]. In Clark et al. (2020), these two models were evaluated on their AUC, AP, and sBrS using the CCSS and an external data set from the St Jude Lifetime Cohort (SJLIFE) [Hudson et al. (2011)].

The equation for calculating the AOF risk using each model is given in the supplementary material of Clark et al. (2020). We estimate the IncV from the prescribed-dose model to the ovarian-dose model using the SJLIFE cohort of 875 survivors with 50 AOF events. The estimation procedure is explained in Appendix A.2.

Figure 6 (a) shows the ROC curves and PR curves of these two models. Based on the AUC, we conclude that both models perform equally well: the estimated AUC is 0.96 for the prescribed-dose model, and 0.94 for the ovarian-dose model. The ovarian model is slightly worse than the prescribed-dose model with $\Delta$AUC estimated to be $-0.02$. Thus, if using
the AUC as the evaluation criteria, there is no need to obtain the more expensive ovary
dosimetry. On the other hand, the ovarian-dose model clearly outperforms the prescribed-
dose model in terms of AP, with $\Delta AP$ estimated to be 0.22. This conclusion agrees with
the scientific intuition: the disease-specific exposure is more accurately measured in the
ovarian-dose model, and thus, this model should perform better.

Let $\Delta(\alpha) = F_{\text{ovarian},0}(q_{\text{ovarian},\alpha}) - F_{\text{prescribed},0}(q_{\text{prescribed},\alpha})$. Figure 6 (b) plots the estimated
$\Delta(\alpha)$, $w_{\text{AP}}(\alpha)$, and $w_{\text{AP}}(\alpha)\Delta(\alpha)$. Here, the event rate is estimated by
$\hat{\pi} = \sum_{i=1}^{n} D_i/n$.

Compared to the prescribed-dose model, the ovarian-dose model better separates the
events and non-events among individuals predicted to be at a higher risk. It shows that
$\Delta(\alpha) > 0$ for $\alpha > 10\%$, whereas the prescribed-dose model performs better with $\Delta(\alpha) < 0$
when $\alpha < 10\%$. Overall, under the $\Delta(\alpha)$ curve, the area below zero is slightly larger than the
area above zero. Thus, the $\Delta AUC$ is negative but close to zero. This indicates that, over
the whole population, these two models have comparable performance in terms of discrimination.

However, the ovarian-dose model demonstrates greater precision in identifying the high-
risk group. The $\Delta AP$ is positive and sizable, because it rewards the superior performance
of the ovarian-dose model at the upper quantiles with large weights.

Clark et al. (2020) created four risk groups: low (< 5%), medium-low (5% to <20%),
medium (20% to <50%), and high risk (≥ 50%). The ovarian-dose model classifies 37
individuals (out of 875) to be at a high risk, among which 30 (81%) are AOF events, while
the prescribed-dose model predicted 13 individuals at a high risk, with 6 (46%) having
developed AOF. This again confirms that the ovarian-dose model is better at identifying
high-risk individuals.

What about the sBrS? It is estimated to be 0.23 for the prescribed-dose model, and 0.50
for the ovarian-dose model, and $\Delta sBrS$ is estimated to be 0.27. Thus, similar to the $\Delta AP$,
the $\Delta sBrS$ favors the ovarian-dose model.

Specifically, for the non-AOF individuals, the mean and variance of $\hat{p}_i$ from both models
are very similar: the mean is 0.033 for the ovarian-dose model and 0.042 for the prescribed-
dose model; their variances are both about 0.0053. Thus, the sum of the squared prediction
error (SSPE) for the ovarian-dose model is 5.83, slightly higher than 5.25 for the prescribed-
dose model. However, for the AOF events, the mean of $\hat{p}_i$ for the ovarian-dose model is
0.48, much closer to 1 than 0.23 for the prescribed-dose model. The variance is 0.10 for the
ovarian dose model and 0.023 for the prescribed-dose model. As a result, the SSPE of the
ovarian-dose model is 18.4, much smaller than 30.7 of the prescribed-dose model. Combining
the events and non-events weighted by their respective rates, the overall SSPE, i.e., BrS, for
the ovarian-dose model is smaller than the prescribed-dose model.

Figure S9 in the supplementary material shows the histogram of the predicted risk from
each model among the AOF and non-AOF individuals. The distribution of the risk score for
the non-AOF individuals is similar between the two models, and thus, as shown above, their
mean and variance are similar. In contrast, for the AOF individuals, the distribution for the
ovarian model has a heavier right tail. This indicates that the ovarian-dose model pushes
more AOF events to the high-risk group, and consequently, the average $\hat{p}_i$ among events is
closer to 1. This explains the consistency between the $\Delta sBrS$ and $\Delta AP$. 
5 Discussion

We investigated the properties and relationship among the IncV metrics of three proper scoring rules: AUC, AP, and sBrS, through analytical and numerical studies. We showed that the $\Delta AUC$ and $\Delta AP$ are intrinsically connected. Both can be expressed as a function of $\Delta(\alpha)$, a quantity characterizing the change of the separation of the risk score distributions between events and non-events from an existing working risk model to a new one.

However, the $\Delta AP$ emphasizes the change in the higher-risk group, whereas the $\Delta AUC$ treats the whole population equally. Because of this difference, they do not always agree with each other. In the numerical study where both working models are misspecified, the correlation between the $\Delta AUC$ and $\Delta AP$ ranges from negative to highly positive as the event rate $\pi$ increases from 0.01 to 0.5. Their relationship with the $\Delta sBrS$ is also different. In general, the $\Delta AP$ is always strongly positively correlated with the $\Delta sBrS$. However, as $\pi$ increases, the correlation of the $\Delta AUC$ with $\Delta sBrS$ changes from negative to positive, similar to its correlation with $\Delta AP$.

Additionally, the $\Delta AP$ can vary widely, from negative to positive. Its magnitude is, in general, the largest among the three metrics. In contrast, most of the $\Delta AUC$ values have a much smaller size, which is consistent with the criticisms of the AUC being insensitive [Pepe et al. (2004)]. The $\Delta AUC$ is also rarely negative, unable to reflect the situations when the new model is worse than the existing one. The difference between the $\Delta AUC$ and $\Delta AP$ is more evident when the event rate is low.

The data example illustrates a situation where the $\Delta AUC$ is at best an inadequate measure and at worst a misleading measure, while the $\Delta AP$ is a more appropriate metric. However, in other applications, $\Delta AUC$ might be better. We suggest choosing the IncV metrics best reflecting the specific aspects of the prediction performance that align with the study objectives.

We provide a new perspective that explains the insensitivity of the AUC, especially for a rare outcome. When a new model improves the prediction performance for a subgroup, for example, the high-risk group, the AUC weakens the “local” superior performance by averaging over the entire population. In contrast, the AP augments this “local” improvement by assigning heavier weights. Likewise, if a study objective concerns the low-risk group, we could consider using the area under a curve of negative predicted values versus specificity ($1-FPR$) as the accuracy metric. Following our derivation of AP, the IncV in this area should be also a weighted average of the change in the separation of the risk score distributions between events and non-events, and the weight is larger for the lower-tail quantiles of the risk score.

A good IncV metric should be a proper scoring rule (strictly or semi). Hilden (2014) pointed out that “Under no circumstances should a risk assessor, by some clever systematic distortion of the risk assessments, be able to improve his apparent performance.” [Hilden (2014)] That is, the true model should always be the winner or among the winners. Another proper scoring rule is the decision curve and net benefit (NB) [Pepe et al. (2015)]. We did not include the NB in our analysis. The NB is designed for making treatment decisions given a patient’s risk tolerance, while all three metrics in this paper are about model evaluation. Also, the NB depends on a threshold probability $p_t$, but the metrics here are threshold-free. As a future work, we are interested in comparing the $\Delta AUC$, $\Delta AP$, and $\Delta sBrS$ with $\Delta NB$. 

13
The domain of each IncV parameter is different. Specifically, the ∆AUC ranges from -0.5 to 0.5 since the range of AUC is [0, 1]; the ∆AP ranges from π − 1 to 1 − π since the range of AP is [π, 1]; the ∆sBrS ranges from −1 to 1 since the range of sBrS is [0, 1]. It may be worthwhile to consider rescaling these IncV metrics to range from −1 and 1. Alternatively, an IncV metric can be defined as a ratio such as \( \Psi_{\text{new}}/\Psi_{\text{old}} \).

Our paper focuses on situations when neither working models is the true model. When one is the true model, Pepe et al. (2013) prove that \( H_0 : p(X,Y) = p(X) \) is equivalent to the null hypotheses concerning no improvement in prediction performance w.r.t. the accuracy measures such as the AUC, NRI, or IDI Pepe et al. (2013). In this case, the two ROC or PR curves never cross. However, when both working models are misspecified, the two curves might cross, and thus, the above equivalence among the null hypotheses does not hold.

As mentioned earlier, the AUC is conditional on the disease status. Thus, it can be estimated from either a prospective cohort study or a case-control study. In contrast, the AP is conditional on the risk score, and consequently, its estimate can only be obtained from a cohort study. The derived expression of the AP in equation (3) provides a potential solution to estimating the AP from a case-control design: one can estimate the AP using (i) an estimated or assumed event rate, and (ii) the risk score distributions of events and non-events estimated from a case-control study.

### 6 Supplementary Material

The supplementary material includes (i) the procedure of obtaining the IncV parameters under the distributional assumptions of the numerical study (Section 3.1), (ii) the results of the numerical study (Section 3.2), including plots of the values of each IncV metric for all the scenarios under different event rates, and the scatter plots of each pair of the IncV metrics, (iii) the results for the scenarios where the two-marker model is the true model (Section 3.3), including plots of the values of each IncV metric for all the scenarios under different event rates, plots of their summary statistics, and a table listing the Pearson correlation of each pair of the IncV metrics, and (iv) the histograms of the predicted AOF risk (Section 4) obtained from the prescribed-dose model and ovarian-dose model for individuals with and without AOF, respectively.

### 7 Acknowledgement

Dr Yuan’s research is supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2019-04862). The data example is from St Jude Lifetime cohort Study, supported by National Cancer Institute grant U01CA195547 (PIs Hudson MM and Robinson LL).

### References

Assel, M., Sjoberg, D. D., and Vickers, A. J. (2017). The brier score does not evaluate the clinical utility of diagnostic tests or prediction models. *Diagnostic and Prognostic*
Badawi, O., Liu, X., Hassan, E., Amelung, P. J., and Swami, S. (2018). Evaluation of icu risk models adapted for use as continuous markers of severity of illness throughout the icu stay. *Critical care medicine*, 46(3):361–367.

Buckley, D. I., Fu, R., Freeman, M., Rogers, K., and Helfand, M. (2009). C-reactive protein as a risk factor for coronary heart disease: a systematic review and meta-analyses for the us preventive services task force. *Annals of internal medicine*, 151(7):483–495.

Chaudhury, S., Brookes, K. J., Patel, T., Fallows, A., Guetta-Baranes, T., Turton, J. C., Guerreiro, R., Bras, J., Hardy, J., Francis, P. T., et al. (2019). Alzheimer’s disease polygenic risk score as a predictor of conversion from mild-cognitive impairment. *Translational psychiatry*, 9(1):1–7.

Clark, R. A., Mostoufi-Moab, S., Yasui, Y., Vu, N. K., Sklar, C. A., Motan, T., Brooke, R. J., Gibson, T. M., Oeffinger, K. C., Howell, R. M., Smith, S. A., Lu, Z., Robison, L. L., Chemaitilly, W., Hudson, M. M., Armstrong, G. T., Nathan, P. C., and Yuan, Y. (2020). Predicting acute ovarian failure in female survivors of childhood cancer: a cohort study in the childhood cancer survivor study (cess) and the st jude lifetime cohort (sjlife). *The Lancet Oncology*, 21(3):436–445.

Cook, N. R., Buring, J. E., and Ridker, P. M. (2006). The effect of including c-reactive protein in cardiovascular risk prediction models for women. *Annals of internal medicine*, 145(1):21–29.

Cox, D. R. et al. (1972). Regression models and life tables. *JR stat soc B*, 34(2):187–220.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML ’06, pages 233–240, New York, NY, USA. ACM.

Elliott, J., Bodinier, B., Bond, T. A., Chadeau-Hyam, M., Evangelou, E., Moons, K. G., Dehghan, A., Muller, D. C., Elliott, P., and Tzoulaki, I. (2020). Predictive accuracy of a polygenic risk score–enhanced prediction model vs a clinical risk score for coronary artery disease. *Jama*, 323(7):636–645.

Hilden, J. (2014). Commentary: On nri, idi, and “good-looking” statistics with nothing underneath. *Epidemiology*, 25(2):265–267.

Howell, R. M., Smith, S. A., Weathers, R. E., Kry, S. F., and Stovall, M. (2019). Adaptations to a generalized radiation dose reconstruction methodology for use in epidemiologic studies: an update from the md anderson late effect group. *Radiation research*, 192(2):169–188.

Hudson, M. M., Ness, K. K., Nolan, V. G., Armstrong, G. T., Green, D. M., Morris, E. B., Spunt, S. L., Metzger, M. L., Krull, K. R., Klosky, J. L., et al. (2011). Prospective medical assessment of adults surviving childhood cancer: study design, cohort characteristics, and feasibility of the st. jude lifetime cohort study. *Pediatric blood & cancer*, 56(5):825–836.
Kattan, M. W. and Gerds, T. A. (2018). The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagnostic and prognostic research*, 2(1):7.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

McIntosh, M. W. and Pepe, M. S. (2002). Combining several screening tests: optimality of the risk score. *Biometrics*, 58(3):657–664.

Mosley, J. D., Gupta, D. K., Tan, J., Yao, J., Wells, Q. S., Shaffer, C. M., Kundu, S., Robinson-Cohen, C., Psaty, B. M., Rich, S. S., et al. (2020). Predictive accuracy of a polygenic risk score compared with a clinical risk score for incident coronary heart disease. *Jama*, 323(7):627–635.

Pencina, M. J., D’Agostino Sr, R. B., D’Agostino Jr, 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. *Statistics in medicine*, 27(2):157–172.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA.

Pepe, M. S., Fan, J., Feng, Z., Gerds, T., and Hilden, J. (2015). The net reclassification index (nri): a misleading measure of prediction improvement even with independent test data sets. *Statistics in biosciences*, 7(2):282–295.

Pepe, M. S., Janes, H., Longton, G., Leisenring, W., and Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American journal of epidemiology*, 159(9):882–890.

Pepe, M. S., Kerr, K. F., Longton, G., and Wang, Z. (2013). Testing for improvement in prediction model performance. *Statistics in medicine*, 32(9):1467–1482.

Raghavan, V., Bollmann, P., and Jung, G. S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*, 7(3):205–229.

Robison, L. L., Mertens, A. C., Boice, J. D., Breslow, N. E., Donaldson, S. S., Green, D. M., Li, F. P., Meadows, A. T., Mulvihill, J. J., Neglia, J. P., et al. (2002). Study design and cohort characteristics of the childhood cancer survivor study: a multi-institutional collaborative project. *Medical and Pediatric Oncology: The Official Journal of SIOP—International Society of Pediatric Oncology (Societé Internationale d’Oncologie Pédiatrique)*, 38(4):229–239.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128.
Tang, M., Hu, P., Wang, C.-F., Yu, C.-Q., Sheng, J., and Ma, S.-J. (2019). Prediction model of cardiac risk for dental extraction in elderly patients with cardiovascular diseases. *Gerontology*, 65(6):591–598.

Uno, H., Cai, T., Tian, L., and Wei, L. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102:527–537.

Xiao, J., Ding, R., Xu, X., Guan, H., Feng, X., Sun, T., Zhu, S., and Ye, Z. (2019). Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *Journal of translational medicine*, 17(1):119.

Yuan, Y., Su, W., and Zhu, M. (2015). Threshold-free measures for assessing the performance of medical screening tests. *Frontiers in Public Health*, 3:57.

Yuan, Y., Zhou, Q. M., Li, B., Cai, H., Chow, E. J., and Armstrong, G. T. (2018). A threshold-free summary index of prediction accuracy for censored time to event data. *Statistics in medicine*, 37(10):1671–1681.

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577.
Figure 1: Comparison of three risk scores at event rate $\pi = 0.05$.

(a) Density curves for events ($D = 1$) and non-events ($D = 0$)

(b) ROC, PR curve, and $F_0(q_\alpha)$ versus $\alpha$. 
Figure 2: Comparison of $r_1$ and $r_2$ under different event rates.

(a) Log of $w_{AP}(\alpha)$ versus $\alpha$.

(b) $\Delta(\alpha)$, $w_{AP}(\alpha)\Delta(\alpha)$, and $w_{AP}(\alpha)$ versus $\alpha$.

(c) PR curves and $\Delta AP$. 

$\pi = 0.2$ $\Delta AP = 0.045$ $\pi = 0.05$ $\Delta AP = 0.097$ $\pi = 0.01$ $\Delta AP = 0.129$
Figure 3: Summary statistics of the three IncV metrics versus different event rates $\pi$. The statistics include minimum, first quartile, median, third quartile, and maximum.
Figure 4: Comparison of two scenarios at event rate \( \pi = 0.01 \): similar \( \Delta \text{AUC} \) but different \( \Delta \text{AP} \).
Figure 5: Comparison of two scenarios at event rate $\pi = 0.01$: similar $\Delta$AP but different $\Delta$AUC.
Figure 6: Data Example: (a) ROC with the estimated AUC and PR curves with the estimated AP of the ovarian-dose model and prescribed-dose model as well as the estimated ΔAUC and ΔAP; (b) plots of the estimated Δ(α) (ovarian − prescribed), w_{AP}(α) and w_{AP}(α)Δ(α) versus α.
Table 1: Pearson correlation and concordance of each pair of the IncV metrics for different event rates $\pi$.

|                  | Comparison     | $\pi = 0.01$ | $\pi = 0.05$ | $\pi = 0.1$ | $\pi = 0.2$ | $\pi = 0.5$ |
|------------------|----------------|--------------|--------------|--------------|--------------|--------------|
| Pearson Correlation | $\Delta s_{BrS}$ vs $\Delta AP$ | 0.995        | 0.992        | 0.986        | 0.971        | 0.837        |
|                  | $\Delta s_{BrS}$ vs $\Delta AUC$ | -0.111       | 0.262        | 0.479        | 0.718        | 0.932        |
|                  | $\Delta AUC$ vs $\Delta AP$     | -0.086       | 0.296        | 0.505        | 0.708        | 0.888        |

|                  | Comparison     | $\pi = 0.01$ | $\pi = 0.05$ | $\pi = 0.1$ | $\pi = 0.2$ | $\pi = 0.5$ |
|------------------|----------------|--------------|--------------|--------------|--------------|--------------|
| Concordance       | $\Delta s_{BrS}$ vs $\Delta AP$ | 0.931        | 0.922        | 0.897        | 0.856        | 0.922        |
|                  | $\Delta s_{BrS}$ vs $\Delta AUC$ | 0.659        | 0.750        | 0.828        | 0.928        | 1.000        |
|                  | $\Delta AUC$ vs $\Delta AP$     | 0.591        | 0.672        | 0.725        | 0.784        | 0.922        |
A Appendices

A.1 Derivation of AUC and AP

Let $\pi = Pr(D = 1)$ be the event rate, and $r(X) = r_X$ be a risk score. Let $F(c) = Pr(r_X \leq c)$ denote its cumulative distribution function (CDF) for the entire population, and $F_1(c) = Pr(r_X \leq c \mid D = 1)$ and $F_0(c) = Pr(r_X \leq c \mid D = 0)$ denote its CDFs for events and non-events, respectively.

The TPR, FPR, and PPV are

$$TPR(c) = Pr(r_X > c \mid D = 1) = 1 - F_1(r) \quad (7)$$
$$FPR(c) = Pr(r_X > c \mid D = 0) = 1 - F_0(r) \quad (8)$$
$$PPV(c) = Pr(D = 1 \mid r_X > c) = \frac{Pr(D = 1, r_X > c)}{Pr(r_X > c)} = \frac{\pi [1 - F_1(c)]}{1 - F(c)} \quad (9)$$

where $1 - F(c) = \pi [1 - F_1(c)] + (1 - \pi) [1 - F_0(c)]$.

**AUC.** AUC is the area under the ROC curve, which can be expressed as

$$AUC = \int_{-\infty}^{\infty} TPR(c)dFPR(c) = 1 - \int_{-\infty}^{\infty} FPR(c)dTPR(c) = \int_{-\infty}^{\infty} [1 - FPR(c)] dTPR(c),$$

because $\int_{-\infty}^{\infty} dTPR(c) = 1$. Using the expressions in equations (7) and (8), we have

$$AUC = \int_{-\infty}^{\infty} F_0(c)d[1 - F_1(c)] = \int_{-\infty}^{\infty} F_0(c)dF_1(c).$$

Let $q_\alpha = F_1^{-1}(\alpha)$ be the $\alpha$-th quantile of the $F_1$ distribution, i.e., $F_1(q_\alpha) = \alpha$. Thus, let $c = q_\alpha$, and we have $AUC = \int_{0}^{1} F_0(q_\alpha)d\alpha$.

**AP.** AP is the area under the PR curve, which can be expressed as

$$AP = \int_{-\infty}^{\infty} PPV(c)dTPR(c).$$

Using the expressions in equations (7) and (9), we have

$$AP = \int_{-\infty}^{\infty} \frac{\pi F_1(c)}{\pi F_1(c) + (1 - \pi) F_0(c)} d[1 - F_1(c)]$$
$$= \int_{-\infty}^{\infty} \frac{\pi [1 - F_1(c)]}{\pi [1 - F_1(c)] + (1 - \pi) [1 - F_0(c)]} dF_1(c)$$
$$= \int_{-\infty}^{\infty} \left\{ \frac{\pi [1 - F_1(c)] + (1 - \pi) [1 - F_0(c)]}{\pi [1 - F_1(c)]} \right\}^{-1} dF_1(c)$$
$$= \int_{-\infty}^{\infty} \left\{ 1 + \frac{1 - \pi}{1 - F_1(c)} \right\}^{-1} dF_1(c).$$

Again, let $c = q_\alpha$, we have

$$AP = \int_{0}^{1} \left\{ 1 + \frac{\pi^{-1} - 1}{1 - \alpha} [1 - F_0(q_\alpha)] \right\}^{-1} d\alpha.$$
A.2 Estimation of AUC, AP, and sBrS for Binary Outcomes

Suppose that the data $\mathcal{D} = \{(D_i, X_i), i = 1, \cdots, n\}$ is collected from $n$ subjects. Let $\hat{p}_i$ denoted the estimated risk, described in Remark 1. Let $\hat{r}_i$ be a risk score, which is a non-decreasing transformation of $\hat{p}_i$. The AUC and AP are estimated using $\hat{r}_i$ by the following nonparametric estimators

$$\hat{\text{AUC}} = \frac{\sum_{i=1}^n \sum_{j=1}^n I(D_i = 1)I(D_j = 0)I(\hat{r}_i > \hat{r}_j)}{\sum_{i=1}^n \sum_{j=1}^n I(D_i = 1)I(D_j = 0)},$$

and

$$\hat{\text{AP}} = \frac{\sum_{i=1}^n \left[ I(D_i = 1) \sum_{j=1}^n I(D_j = 1)I(\hat{r}_j > \hat{r}_i)/\sum_{j=1}^n I(\hat{r}_j > \hat{r}_i) \right]}{\sum_{i=1}^n I(D_i = 1)}.$$

The BrS can be estimated using $\hat{p}_i$ by $\hat{\text{BrS}} = n^{-1} \sum_{i=1}^n (D_i - \hat{p}_i)^2$. The event rate is estimated as $\hat{\pi} = n^{-1} \sum_{i=1}^n D_i$. Then the sBrS is estimated as $\hat{s\text{BrS}} = 1 - \hat{\text{BrS}}/\hat{\pi}(1 - \hat{\pi})$. 