A Modified Net Reclassification Improvement Statistic

Glenn Heller

Department of Epidemiology and Biostatistics,

Memorial Sloan Kettering, New York, NY 10017, U.S.A.

*email: *hellerg@mskcc.org*

ABSTRACT

The continuous net reclassification improvement (NRI) statistic is a popular model change measure that was developed to assess the incremental value of new factors in a risk prediction model. Two prominent statistical issues identified in the literature call the utility of this measure into question: (1) it is not a proper scoring function and (2) it has a high false positive rate when testing whether new factors contribute to the risk model. For binary response regression models, these subjects are interrogated and a modification of the continuous NRI, guided by the likelihood-based score residual, is proposed to address these issues. Within a nested model framework, the modified NRI may be viewed as a distance measure between two risk models. An application of the modified NRI is illustrated using prostate cancer data.

KEYWORDS

Binary response model, L_1 distance, Nested models, Proper score, Valid test

1. INTRODUCTION

In the clinical setting, individual risk assessment is often derived through a regression model, which incorporates a combination of risk factors due to biological complexity. These risk models are used in forecasting future health outcomes of an individual such as treatment response or survival. The quality of the risk model, evaluated using statistical measures such as calibration, discrimination, explained variation, and likelihood based, reflects the level of confidence in the forecast (Gerds and Kattan 2021). When the objective is to incorporate a new set of factors to an existing risk model, assessing the impact of these new factors on the forecast is critical. For binary response regression, a discrimination measure, the net reclassification improvement (NRI), is one statistic used for this evaluative process. The NRI, also referred to as the net reclassification index, was developed to ascertain whether the introduction of new risk factors move a model derived forecast in a direction consonant with the binary response outcome (Pencina et al. 2008).

The NRI statistic has been criticized on numerous grounds. Two prevailing points of contention are: (1) it is not a proper scoring function and (2) it has a high false positive rate when testing whether the new factors contribute to the risk model, even in situations that include independent training and test datasets (Kerr et al. 2014 and Pepe et al. 2014, 2015). Despite of these critiques, the NRI is a popular statistic, and in the three-year time period 2019-2021, it was cited in PubMed over 800 times. The purpose of this work is to elucidate the methodology underlying these two concerns and to propose a likelihood guided modification to the NRI to rectify these issues.

The NRI is defined through a series of nested regression models, where it is assumed that the existing factors alone ($\boldsymbol{x}$), which includes a constant for the intercept

term, or combined with new factors ($\boldsymbol{z}$) are modeled as

$$\begin{aligned} \Pr(Y = 1) &= G(\beta^{\bullet}) \\ \Pr(Y = 1|\boldsymbol{x}) &= G(\boldsymbol{\beta}^{0^T}\boldsymbol{x}) \\ \Pr(Y = 1|\boldsymbol{x}, \boldsymbol{z}) &= G(\boldsymbol{\beta}_0^T\boldsymbol{x} + \boldsymbol{\gamma}_0^T\boldsymbol{z}) \end{aligned} \tag{1}$$

where Y is a binary outcome denoted as event ($Y = 1$) or non-event ($Y = 0$), G is a monotone function representing the probability of an event, the base model risk score is $\boldsymbol{\beta}^{0^T}\boldsymbol{x}$, the expanded model risk score is $\boldsymbol{\beta}_0^T\boldsymbol{x} + \boldsymbol{\gamma}_0^T\boldsymbol{z}$, and for the constant model, $\pi_0 = G(\beta^{\bullet})$. Throughout this work, random variables are represented with upper case, their observed copies are written in lower case, and vectors are indicated in bold.

The log-likelihood used to estimate the model parameters is

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_i \left[y_i \log G(\boldsymbol{\beta}^T\boldsymbol{x}_i + \boldsymbol{\gamma}^T\boldsymbol{z}_i) + (1 - y_i) \log(1 - G(\boldsymbol{\beta}^T\boldsymbol{x}_i + \boldsymbol{\gamma}^T\boldsymbol{z}_i)) \right],$$

where $\{(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)\}, i = 1, \ldots, n$ are independent identically distributed copies of $(Y, \boldsymbol{X}, \boldsymbol{Z})$. The maximum likelihood estimates from the three models are represented as: $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$, $\hat{\boldsymbol{\theta}}^0 = (\hat{\boldsymbol{\beta}}^0, \mathbf{0})$, and $\hat{\pi} = \bar{y}$, the observed proportion of events.

Historically, the NRI was developed under the assumption that the base model risk score could be placed in risk classification categories. It was a measure of whether the expanded model risk score, due to the addition of new factors, would move into higher risk categories for subjects with an event and into lower risk categories for subjects without an event. This framework, however, requires apriori clinically meaningful risk categories, which are often not apparent at the time of analysis, particularly in the early stage of model development. As a result, the continuous NRI was developed (Pencina et al. 2011) and it is this measure that is the focus of this work.

The population NRI is defined as

$$\rho(\boldsymbol{\theta}_0; \boldsymbol{\theta}^0; \pi_0) =$$
$$2\left\{\Pr(\boldsymbol{\beta}_0^T \boldsymbol{X} + \boldsymbol{\gamma}_0^T \boldsymbol{Z} \geq \boldsymbol{\beta}^{0T} \boldsymbol{X} | Y = 1) - \Pr(\boldsymbol{\beta}_0^T \boldsymbol{X} + \boldsymbol{\gamma}_0^T \boldsymbol{Z} \geq \boldsymbol{\beta}^{0T} \boldsymbol{X} | Y = 0)\right\}$$

where $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$ and $\boldsymbol{\theta}^0 = (\boldsymbol{\beta}^0, \mathbf{0})$. When multiplied by $1/2$, the population NRI is estimated as

$$R_n(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0; \hat{\pi}) = [n\bar{y}(1-\bar{y})]^{-1} \sum_i [y_i - \bar{y}] \left[I(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i + \hat{\boldsymbol{\gamma}}^T \boldsymbol{z}_i - \hat{\boldsymbol{\beta}}^{0T} \boldsymbol{x}_i > 0) - \frac{1}{2} \right]. \quad (2)$$

Assuming at least one component of $\boldsymbol{x}$ is continuous, it can be asserted without loss of generality, that the indicator function can be extended as

$$\begin{aligned} I(u > 0) &= 1 \quad \text{if} \quad u > 0 \\ I(u > 0) &= \frac{1}{2} \quad \text{if} \quad u = 0 \\ I(u > 0) &= 0 \quad \text{if} \quad u < 0. \end{aligned} \quad (3)$$

Although the net reclassification improvement statistic is a frequently applied model change measure, its lack of propriety and high false positive rate are problematic. In Section 2, a modified NRI (mNRI) is developed that satisfies the concept of a proper change score, which adapts the proper scoring principle to model change measures (Pepe et al. 2015). Section 3 demonstrates that a smooth version of the mNRI provides a valid test procedure when the population NRI is zero. This result is established in the single sample and the independent training and test data case. In Section 4, a prostate cancer data example is used to illustrate these concepts and Section 5 contains a discussion.

2. THE mNRI IS A PROPER CHANGE SCORE

For a correctly specified parametric risk model, a performance measure is a proper score if its expected value is minimized/maximized at the true model parameter value (Gneiting and Raftery, 2007). For example, the expected value of the Brier score applied to the expanded model

$$E[Y - G(\boldsymbol{\beta}^T \boldsymbol{X} + \boldsymbol{\gamma}^T \boldsymbol{Z})]^2,$$

is minimized at $(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$. If a performance measure is not a proper score, then the analyst may find inconsistent parameter estimates that make the measure look better. Population performance measures such as the expected value of the area under the curve (AUC), the Brier score (BS), and Kullback-Leibler divergence (KL), are maximized/minimized at their true parameter values and therefore are proper scores.

Proper scoring is more difficult to achieve for model change measures. Consider the case where a performance measure M is applied separately to the expanded model and the base model, and the change measure is

$$\Delta M(\boldsymbol{b}, \boldsymbol{g}; \boldsymbol{b}^0) = M(\boldsymbol{b}, \boldsymbol{g}) - M(\boldsymbol{b}^0).$$

If the performance measure (M) is convex,

$$(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = \underset{(\boldsymbol{b}, \boldsymbol{g})}{\arg\min} \quad E[M(\boldsymbol{b}, \boldsymbol{g})]$$

$$\boldsymbol{\beta}^0 = \underset{\boldsymbol{b}^0}{\arg\min} \quad E[M(\boldsymbol{b}^0)],$$

but the difference of two convex functions is not necessarily convex, and in general,

$$(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{\beta}^0) \neq \underset{(\boldsymbol{b}, \boldsymbol{g}, \boldsymbol{b}^0)}{\arg\min} \quad E[\Delta M(\boldsymbol{b}, \boldsymbol{g}; \boldsymbol{b}^0)].$$

To adapt proper scoring to change measures, Pepe et al. (2015) orient the model parameter space so that the base model is evaluated at the true parameter $\boldsymbol{\beta}^0$. In this setting, ΔM is termed a proper change score, since

$$(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = \underset{(\boldsymbol{b},\boldsymbol{g})}{\arg\min} \quad E\left[\Delta M(\boldsymbol{b}, \boldsymbol{g}; \boldsymbol{\beta}^0)\right],$$

recreating the single model evaluation. The term proper change score is used here to acknowledge the adaptation of the proper scoring principle to change measures. Under this definition, ΔAUC, ΔBS, and ΔKL are proper change scores.

The NRI differs from other change measures because it is a statistic based on within subject change and not between model change as above. In addition, the statistic is composed of parameter estimates from three nested models. As a result, it is not covered under the previous argument. To satisfy the proper change score criterion, the NRI is modified

$$T_n(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0; \hat{\pi}) = [n\bar{y}(1-\bar{y})]^{-1} \sum_i r(\hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_i) \left[I(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i + \hat{\boldsymbol{\gamma}}^T \boldsymbol{z}_i - \hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_i > 0) - \frac{1}{2} \right],$$

which is constructed by replacing the constant model score residual $y - \bar{y}$ in (2) with the base model score residual $r(\hat{\boldsymbol{\beta}^0}^T \boldsymbol{x})$, where

$$r(\boldsymbol{\beta}^{0T} \boldsymbol{x}) = \left[\frac{\partial G(\boldsymbol{\beta}^{0T} \boldsymbol{x})}{\partial (\boldsymbol{\beta}^{0T} \boldsymbol{x})} \right] \left[G(\boldsymbol{\beta}^{0T} \boldsymbol{x})(1 - G(\boldsymbol{\beta}^{0T} \boldsymbol{x})) \right]^{-1} \left[y_i - G(\boldsymbol{\beta}^{0T} \boldsymbol{x}) \right]. \qquad (4)$$

The modified NRI (mNRI) is closely akin to the maximum score statistic and the least absolute deviation statistic (Manski 1985, Horowitz 1998), which provide the framework for the derivation in Theorem 1.

Theorem 1.

Consider the mNRI scoring function derived from a single random variable, with the base and constant model parameters given

$$T_1(\boldsymbol{\theta};\boldsymbol{\theta}^0;\pi_0) = [\pi_0(1-\pi_0)]^{-1} r(\boldsymbol{\beta}^{0T}\boldsymbol{X}) \left[I(\boldsymbol{\beta}^T\boldsymbol{X} + \boldsymbol{\gamma}^T\boldsymbol{Z} - \boldsymbol{\beta}^{0T}\boldsymbol{X} > 0) - \frac{1}{2} \right].$$

The mNRI scoring function is a proper change score,

$$E[T_1(\boldsymbol{\theta}_0;\boldsymbol{\theta}^0;\pi_0)] \geq E[T_1(\boldsymbol{\theta};\boldsymbol{\theta}^0;\pi_0)] \quad \text{for any } \boldsymbol{\theta} = (\boldsymbol{\beta},\boldsymbol{\gamma}).$$

The theorem is proved in the appendix.

An interpretation of the mNRI statistic is obtained by rewriting it as

$$T_n(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}}^0,\hat{\pi}) = [2\bar{y}(1-\bar{y})]^{-1} \frac{[\boldsymbol{s}(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}^0})]^T[\boldsymbol{r}(\hat{\boldsymbol{\theta}^0})]}{[\boldsymbol{s}(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}^0})]^T[\boldsymbol{s}(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}^0})]}$$

where $\boldsymbol{r}(\hat{\boldsymbol{\theta}^0}) = [r(\hat{\boldsymbol{\beta}^0}^T\boldsymbol{x}_1),\ldots,r(\hat{\boldsymbol{\beta}^0}^T\boldsymbol{x}_n)]$ is the base model score residual vector and $\boldsymbol{s}(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}^0})$ is a sign vector with subject components $s_i(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}^0}) = 2I(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i + \hat{\boldsymbol{\gamma}}^T\boldsymbol{z}_i - \hat{\boldsymbol{\beta}^0}^T\boldsymbol{x}_i > 0) - 1$. The mNRI is a function of the propensity of the event outcome ($\bar{y}$) and a regression coefficient representing the association between the direction of the risk score due to adding $\boldsymbol{z}$ and the event outcome after taking into account $\boldsymbol{x}$. This perspective is analogous to a partial residual plot, where a model covariate of interest $\boldsymbol{z}$ is replaced by a between model directional covariate $\boldsymbol{s}(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}^0})$.

An alternative interpretation of the mNRI may be considered from the viewpoint of its limiting value

$$\lim_{n\to\infty} T_n(\boldsymbol{\theta}_0;\boldsymbol{\theta}^0;\pi_0) = [2\pi_0(1-\pi_0)]^{-1} E_{X,Z}\left\{ h(\boldsymbol{\beta}^{0T}\boldsymbol{X}) \left| G(\boldsymbol{\beta}_0{}^T\boldsymbol{X} + \boldsymbol{\gamma}_0^T\boldsymbol{Z}) - G(\boldsymbol{\beta}^{0T}\boldsymbol{X}) \right| \right\}$$

where the weight $h(\boldsymbol{\beta}^{0T}\boldsymbol{X})$ stems from the base model score residual (4),

$$h(\boldsymbol{\beta}^{0T}\boldsymbol{X}) = \left[\frac{\partial G(\boldsymbol{\beta}^{0T}\boldsymbol{x})}{\partial(\boldsymbol{\beta}^{0T}\boldsymbol{x})} \right] \left[G(\boldsymbol{\beta}^{0T}\boldsymbol{x})(1 - G(\boldsymbol{\beta}^{0T}\boldsymbol{x})) \right]^{-1}$$

$$r(\boldsymbol{\beta}^{0T}\boldsymbol{X}) = h(\boldsymbol{\beta}^{0T}\boldsymbol{X})[y_i - G(\boldsymbol{\beta}^{0T}\boldsymbol{X})].$$

Thus, the population mNRI is a weighted L_1 distance measure between the nested event probabilities. An important special case occurs when G is logistic and

$$\lim_{n\to\infty} T_n(\boldsymbol{\theta}_0; \boldsymbol{\theta}^0; \pi_0) = [2\pi_0(1-\pi_0)]^{-1} E_{X,Z}\left|G(\boldsymbol{\beta}_0^T\boldsymbol{X} + \boldsymbol{\gamma}_0^T\boldsymbol{Z}) - G(\boldsymbol{\beta}^{0T}\boldsymbol{X})\right|,$$

which results in an unweighted L_1 distance measure. Here, the population mNRI is proportional to the mean absolute deviation (MAD) of the nested event probabilities. In addition to using the MAD as a summary measure, this result suggests that graphical insight into the mNRI may be obtained by plotting the base model event probability estimates by the expanded model event probability estimates.

3. THE NRI FALSE POSITIVE RATE

Empirical research on the utility of the NRI has raised questions as to whether it has an unacceptably high false positive rate, signifying a larger than anticipated value when the new factors have no effect on the binary response (Kerr et al. 2014 and Pepe et al. 2014, 2015). As a practical matter, measures with high false positive rates lead to the introduction of irrelevant factors into the model development process. In this section, this issue is investigated, and a valid test procedure is developed, both in the case of a single sample and when independent training and test samples are included.

Pencina et al (2008) state that under the null $\rho(\boldsymbol{\theta}_0; \boldsymbol{\theta}^0; \pi_0) = 0$, the asymptotic distribution of the estimated NRI in (2) is

$$n^{1/2} R_n(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0; \hat{\pi}) \quad \xrightarrow{D} \quad N\left[0, A\right] \tag{5}$$

where accounting for the multiplication by 1/2 to produce (3), the asymptotic variance is estimated as $\hat{A} = [(4n_1)^{-1} + (4n_0)^{-1}]$. Further work by Pencina et al. (2011, 2012) modified the asymptotic variance calculation. In a series of simulation experiments, Kerr et al. (2014) and Pepe et al. (2014, 2015) evaluated the adequacy of this result, using a conditional binormal model to produce nested logistic regression models. They found that on average, under the null, the NRI estimate was positive and that the type 1 error rate using the asymptotic normal reference distribution was as high as 0.63. Additional simulations that incorporated independent training and test datasets produced similar conclusions. Taken in total, these results represent a critical indictment against the test procedure in (5). A problem, recognized by these authors, and Demler et al. (2017), is that the asymptotic normal reference distribution is incorrect.

Consider a smooth NRI

$$R_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0; \hat{\pi}) = [n\bar{y}(1-\bar{y})]^{-1} \sum_i [y_i - \bar{y}] \left[\Phi(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i + \hat{\boldsymbol{\gamma}}^T \boldsymbol{z}_i - \hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_i) - \frac{1}{2} \right], \quad (6)$$

where the extended indicator function, which is discontinuous in $\boldsymbol{\theta}$, is replaced by the continuous standard normal distribution function $\Phi(\cdot)$. A heuristic for this substitution is that when $\rho(\boldsymbol{\theta}_0; \boldsymbol{\theta}^0; \pi_0) = 0$, as n gets large, $\hat{\boldsymbol{\gamma}} \xrightarrow{p} 0$, $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}^0} \xrightarrow{p} 0$ (Pepe et al. 2013), and therefore (Horowitz 1998)

$$I(\hat{\boldsymbol{\beta}}^T \boldsymbol{x} + \hat{\boldsymbol{\gamma}}^T \boldsymbol{z} - \hat{\boldsymbol{\beta}^0}^T \boldsymbol{x} > 0) \approx \Phi\left(\hat{\boldsymbol{\beta}}^T \boldsymbol{x} + \hat{\boldsymbol{\gamma}}^T \boldsymbol{z} - \hat{\boldsymbol{\beta}^0}^T \boldsymbol{x} \right) \approx \frac{1}{2}.$$

The purpose of this local smoothing is to facilitate the derivation of the asymptotic null reference distribution.

Theorem 2. Assume the binary response regression models in (1) are properly

specified and the covariate vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ have dimension p and q, respectively. If $\rho(\boldsymbol{\theta}_0; \boldsymbol{\theta}^0; \pi_0) = 0$, then

$$nR_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0; \hat{\pi}) \xrightarrow{D} [\boldsymbol{N}_q(0, V_1)]^T [\boldsymbol{N}_q(0, V_2)] + \frac{1}{2} \boldsymbol{D}_p^T \mathcal{I}_{\beta\beta}^{-1} \boldsymbol{c}_p.$$

The first term is the inner product of two positively correlated, q-dimensional, mean zero normal random vectors, and the second term is bilinear, where $\boldsymbol{D}_p$ is a p dimensional random vector with quadratic components, and $\boldsymbol{c}_p$ is a p dimensional constant vector. This result demonstrates that the null distribution of the NRI is not normal and is not symmetric about zero, which explains the anomalous findings in Kerr et al. (2014) and Pepe et al. (2014, 2015).

The reference distribution for the NRI test statistic $R_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0; \hat{\pi})$ is complex and difficult to apply. In contrast, the mNRI test statistic

$$T_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0; \hat{\pi}) = [n\bar{y}(1-\bar{y})]^{-1} \sum_i r(\hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_i) \left[\Phi(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i + \hat{\boldsymbol{\gamma}}^T \boldsymbol{z}_i - \hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_i) - \frac{1}{2} \right]$$

has a straightforward null reference distribution.

Theorem 3. Assume the binary regression models in (1) are properly specified and the covariate vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ have dimension p and q, respectively. If $\rho(\boldsymbol{\theta}_0; \boldsymbol{\theta}^0; \pi_0) = 0$, then

$$nT_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0; \hat{\pi}) \xrightarrow{D} k\chi_q^2,$$

where $k = \phi(0)[\pi_0(1-\pi_0)]^{-1}$, $\phi(0)$ is the standard normal density function evaluated at 0, and χ_q^2 is a chi-square random variable with q degrees of freedom. A proof of this result is found in the appendix.

Theorems 2 and 3 reorient one's understanding of what constitutes meaningful

NRI and mNRI statistics and Theorem 3 provides an uncomplicated metric to test the mNRI distance from zero. If the new clinical factors ($\boldsymbol{z}$) are noise, then small positive values are simply random variation under the null, and only large positive values, as determined by the scaled chi-square reference distribution, are considered meaningful. A precursor to this result is found in Kerr et al. (2011).

Theorem 3 covers the single sample case. Alternatively, the test statistic may be constructed from two independent data sets from the same population, where the regression coefficients are estimated from the training data ($\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^0$) and the test data ($\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}^0$), and the data for the test statistic ($y_i, \boldsymbol{x}_i, \boldsymbol{z}_i$) are drawn from the independent test data. Under these conditions, the reference distribution for the smooth mNRI test statistic

$$T_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0, \tilde{\boldsymbol{\theta}}^0; \tilde{\pi}) = [n\bar{y}(1-\bar{y})]^{-1} \sum_i r(\tilde{\boldsymbol{\beta}^0}^T \boldsymbol{x}_i) \left[\Phi(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i + \hat{\boldsymbol{\gamma}}^T \boldsymbol{z}_i - \hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_i) - \frac{1}{2} \right].$$

is provided in Theorem 4.

Theorem 4. Assume the binary regression models for the training and test data have the same specification and are given in (1), where the covariate vector $\boldsymbol{x}$ has dimension p and the covariate vector $\boldsymbol{z}$ has dimension q. If $\rho(\boldsymbol{\theta}_0; \boldsymbol{\theta}^0; \pi_0) = 0$,

$$nT_n^s(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0, \tilde{\boldsymbol{\theta}}^0; \tilde{\pi}) \xrightarrow{D} \frac{k}{2} \sum_{j=1}^{2q} \lambda_j \chi_j^2,$$

where k is defined in Theorem 3, $\{\chi_j^2\}$ are independent chi-square random variables each with one degree of freedom, and $\{\lambda_j\}$ represent eigenvalues determined from the product matrix VC (Baldessari 1967), where

$$V = \begin{pmatrix} \text{var}(\tilde{\boldsymbol{\gamma}}) & 0 \\ 0 & \text{var}(\hat{\boldsymbol{\gamma}}) \end{pmatrix} \qquad C = \begin{pmatrix} 0 & D \\ D & 0 \end{pmatrix}.$$

and $D = [\text{var}(\tilde{\boldsymbol{\gamma}})]^{-1}$. The details are provided in the appendix.

A simulation study was performed to assess the false positive rate using the reference distributions in Theorems 3 and 4, and the normal reference distribution in (5). A conditional bivariate normal covariate distribution was used to generate nested logistic risk models. The conditioning variable was the event status with $\Pr(Y = 1) = \{0.25, 0.50, 0.75\}$. The bivariate normal had a common variance-covariance matrix across event status, with correlation parameters 0 or 0.5. The mean of Z was 0 for $Y = 0$ or $Y = 1$, and the mean of X was set equal to 0 for $Y = 0$ and took on values $\{0.25, 0.50, 0.75, 1.0\}$ for $Y = 1$. Simulations with 200 and 500 observations per replicate were conducted. Five thousand replicates were run for each simulation. Tables 1 and 2 compare the size estimates for the mNRI reference distributions in Theorems 3 and 4 with the NRI normal reference distribution. The nominal type 1 error in all simulations was 0.05.

For the single sample simulations in Table 1, using Theorem 3, the average type 1 error was 0.048 (n=200) and 0.050 (n=500). In contrast, applying the normal reference distribution in (5), produced average type 1 errors equal to 0.079 (n=200) and 0.129 (n=500). Similar results were found for the independent training-test sample simulations in Table 2. From Theorem 4, the average type 1 error was 0.051 (n=200) and 0.052 (n=500), whereas when using the normal reference distribution it was 0.079 (n=200) and 0.124 (n=500). These simulation results confirm that the modified NRI test statistics, with their associated reference distributions, are valid test procedures, and they confirm the poor operating characteristics of the asymptotic normal reference distribution, with divergence increasing with sample size.

5. PROSTATE CANCER DATA

Patients with metastatic prostate cancer are by definition high risk. Nevertheless, there is significant variability in the survival times of these patients (Sayegh, Swami, and Agarwal, 2021). Given this heterogeneity, there is a pressing need to identify new biomarkers that can accurately assess patient risk. Historically, the use of prostate specific antigen (PSA) and other blood based biomarkers have produced risk models with only moderate calibration and discrimination in the metastatic prostate cancer setting (Gafita et al. 2021). As a result, exploring informative new biomarkers continues, with a recent focus around circulating tumor cells and serum testosterone (Cieslikowski et al. 2021; Ryan et al. 2019).

An application of the net reclassification improvement (NRI), based on the addition of circulating tumor cells and serum testosterone, was undertaken for metastatic prostate cancer patients treated on the control arm of a multicenter phase 3 randomized clinical trial (Saad et al. 2015). The control arm of the randomized trial, patients treated with steroids alone, is useful to assess the added prognostic utility of new biomarkers, because it approximates the natural history of the disease.

Four hundred and eighteen patients with a complete set of biomarkers and sufficient follow-up were used in the analysis. The binary endpoint was survival 24 months after the start of treatment. In this cohort, forty seven percent of the patients survived longer than two years. In addition to circulating tumor cells and serum testosterone, traditional biomarkers for metastatic prostate cancer were incorporated into the risk model. The complete set of eight biomarkers included in the analysis were: albumin, alkaline phosphatase, circulating tumor cells, Gleason score, hemoglobin, lactate dehydrogenase, prostate specific antigen, and serum testosterone.

Nested logistic regression models were fit for the binary 24 month survival endpoint; the expanded model incorporated all eight biomarkers and the base model represented a subset of seven biomarkers. All biomarkers except Gleason score were continuous. To create greater flexibility in the models, a restricted cubic spline with four knots was fit to each continuous biomarker. Gleason score, an ordinal variable ranging from 2-10, representing tumor complexity as determined by pathology, and was dichotomized as 1-7 and 8-10.

Table 3 summarizes the results of the NRI, mNRI, and the p-values generated from their respective test procedures described in Section 3. For the logistic models, the mNRI equates to a scaled mean absolute difference (MAD) between the estimated event probabilities

$$[2n\bar{y}(1-\bar{y})]^{-1}\sum_{i}|G(\hat{\boldsymbol{\beta}}^{T}\boldsymbol{x}_i+\hat{\boldsymbol{\gamma}}^{T}\boldsymbol{z}_i)-G(\hat{\boldsymbol{\beta}}^{0^{T}}\boldsymbol{x}_i)|.$$

For the prostate data, the observed proportion of events was 0.47, and so the mNRI $\approx$ $2 \times$ MAD.

With the addition of serum testosterone, the mean absolute distance was only 0.022, and using the smooth mNRI, a test of whether the population NRI differed from zero generated a p-value equal to 0.490. Figure 1 provides corroborating evidence that adding serum testosterone does not meaningfully change the predicted event probabilities. An application of the NRI with a normal reference distribution (5), however, produced a p-value equal to 0.046, which mirrors the high false positive rate for the NRI found in the simulations. When the circulating tumor cell (CTC) biomarker was added to the risk model, the mean absolute difference between the estimated event probabilities was large and equal to 0.095, with an attending p-value

less than 0.001. The addition of circulating tumor cells had a marked effect on the predicted probability of death within 24 months. This result is confirmed visually in Figure 2, where the estimated event probabilities change significantly from the base model to the expanded model due to the addition of CTC. Thus, the addition of CTC but not serum testosterone would consequentially change the predicted probabilities of surviving greater than 24 months. Furthermore, for other single variable deletions, only the addition of alkaline phosphatase and hemoglobin appreciably change the expanded model probabilities.

6. DISCUSSION

The net reclassification improvement (NRI) statistic is a measure of change for a model based risk score due to the addition of new factors. Although the NRI is frequently applied, identified weaknesses of the statistic include that it is not a proper scoring function (or proper change score) and it does not produce a valid test procedure. A modification of this statistic (mNRI) corrects these deficiencies. The mNRI can be interpreted as a measure of association between the directional change in the risk score and the base model score residual. In the special but frequently applied case of logistic regression, an asymptotic analysis demonstrates that the mNRI is proportional to a mean absolute deviation measure, putting the mNRI on an easily interpretable difference in probability scale.

There remain, however, some concerns with the NRI that are not resolved through the mNRI (Kerr et al. 2014). The mNRI does not include risk thresholds for the purpose of intervention strategies, and therefore does not include the costs and benefits

of a risk threshold guided intervention. As a result, its application should be directed to the model development stage. On this topic, there has been significant discussion surrounding the utility of the NRI, and even with the modification proposed here, the debate will almost surely continue. The contribution of this work is to put the statistic on a stronger statistical foundation and to clear away some of the arguments that obscure its properties, perhaps shedding more light and less heat on this measure.

ACKNOWLEDGEMENTS.

This work was supported by NIH Grants R01CA207220 and P30CA008748.

REFERENCES

Baldessari, B. (1967), "The Distribution of a Quadratic Form of Normal Random Variables," *Annals of Mathematical Statistics*, 38, 1700-1704.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, The Johns Hopkins University Press.

Cieslikowski, W. A., Antczak, A., Nowicki, M., Zabel, M., Budna-Tukan, J. (2021), "Clinical Relevance of Circulating Tumor Cells in Prostate Cancer Management," *Biomedicines*, 9, 1179.

Demler, O. V., Pencina, M. J., Cook, N. R., and D'Agostino Sr, R. B. (2017), "Asymptotic distribution of ΔAUC, NRIs, and IDI based on theory of U-statistics, "*Statistics in Medicine*, 36, 3334-3360.

Gafita, A., Calais, J., Grogan, T. R., Hadaschik, B., Wang, H., Weber, M., Sandhu, S., Kratochwil, C., Esfandiari, R., Tauber, R., Zeldin, A., Rathke, H., Armstrong, W. R., Robertson, A., Thin, P., D'Alessandria, C., Rettig, M. B., Delpassand, E. S., Haberkorn, U., Elashoff, D., Herrmann, K., Czernin, J., Hofman, M. S., Fendler, W. P., Eiber, M. (2021), "Nomograms to predict outcomes after 177 Lu-PSMA therapy in men with metastatic castration-resistant prostate cancer: an international, multicentre, retrospective study," *Lancet Oncology*, 22, 1115–25.

Gerds, T. A. and Kattan, M. W. (2021), *Medical Risk Prediction Models With Ties to Machine Learning.* CRC Press.

Gneiting, T. and Raftery, A. E. (2007), "Strictly proper scoring rules, prediction, and estimation," *Journal of The American Statistical Association*, 102, 359-378.

Horowitz, J. L. (1998), *Semiparametric Methods in Econometrics*. Springer-Verlag.

Kerr, K. F., McClelland, R. L., Brown, E. R., and Lumley, T. (2011), "Evaluating the Incremental Value of New Biomarkers With Integrated Discrimination Improvement," *A*merican Journal of Epidemiology, 174, 364-374.

Kerr, K. F., Wang, Z., Janes, H., McClelland, R. L., Psaty, B. M., and Pepe, M. S. (2014), "Net reclassification indices for evaluating risk-prediction instruments: A critical review," *Epidemiology*, 25, 114-121.

Manski, C. F. (1985), "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27, 313-333.

Pencina, M. J., D'Agostino Sr, R. B., D'Agostino Jr, R. D., and Vasan, R. (2008), "Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond," *Statistics in Medicine*, 27, 157-172.

Pencina, M. J., D'Agostino Sr, R. B., and Steyerberg, E. W. (2011), "Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers," *Statistics in Medicine*, 30, 11-21.

Pencina, M. J., D'Agostino Sr, R. B., and Demler O. V. (2012), "Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models," *Statistics in Medicine*, 31, 101-113.

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, 282-295.

Pepe, M. S., Janes, H., and Li, C. I. (2014), Net risk reclassification p values: Valid or misleading? *Journal of the National Cancer Institute*, 106, 1-6.

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

Ryan, C. J., Dutta, S., Kelly, W. K., Russell, C., Small, E. J., Morris, M. J., Taplin, M. E., Halabi, S. (2020), "Androgen Decline and Survival During Docetaxel Therapy in Metastatic Castration Resistant Prostate Cancer (mCRPC)," *Prostate Cancer and Prostatic Disease*, 23, 66-73.

Saad, F., Fizazi, K., Jinga, V., Efstathiou, E., Fong, P. C., Hart, L. L., Jones, R., McDermott, R., Wirth, M., Suzuki, K., MacLean, D. B., Wang, L., Akaza, H., Nelson, J., Scher, H. I., Dreicer, R., Webb, I. J., de Wit, R. ELM-PC 4 investigators. (2015), "Orteronel plus prednisone in patients with chemotherapy naive metastatic castration-resistant prostate cancer (ELM-PC 4): a double-blind, multicentre, phase 3, randomised, placebo-controlled trial," *Lancet Oncology*, 16, 338–348.

Sayegh, N., Swami, U., and Agarwal, N. (2021), "Recent Advances in the Management of Metastatic Prostate Cancer," *JCO Oncology Practice*, 18, 45-55.

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*. Springer-Verlag.

Appendix: Proof of theorems

The following conditions and notation will be used in the appendix.

(C1) The set of binary response nested models

$$\Pr(Y = 1) = G(\beta^\bullet)$$

$$\Pr(Y = 1|\boldsymbol{x}) = G(\boldsymbol{\beta}^{0T}\boldsymbol{x})$$

$$\Pr(Y = 1|\boldsymbol{x}, \boldsymbol{z}) = G(\boldsymbol{\beta}_0^T\boldsymbol{x} + \boldsymbol{\gamma}_0^T\boldsymbol{z})$$

specify the relationship between the p-dimensional existing factors $\boldsymbol{x}$, the q-dimensional new factors $\boldsymbol{z}$, and the binary event outcome y. The model with no covariates is the constant model, $\boldsymbol{x}$ alone is the base model and $(\boldsymbol{x}, \boldsymbol{z})$ is the expanded model. The inverse link function G is known. Throughout this work, random variables are represented with upper case, their observed copies are written in lower case, and vectors are indicated in bold.

(C2) The log-likelihood used to estimate the regression coefficients is

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_i \left[y_i \log G(\boldsymbol{\beta}^T\boldsymbol{x}_i + \boldsymbol{\gamma}^T\boldsymbol{z}_i) + (1 - y_i) \log(1 - G(\boldsymbol{\beta}^T\boldsymbol{x}_i + \boldsymbol{\gamma}^T\boldsymbol{z}_i)) \right],$$

where $\{(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)\}, i = 1, \ldots, n$ are independent identically distributed copies of $(Y, \boldsymbol{X}, \boldsymbol{Z})$. For $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$, the expanded model maximum likelihood estimate is denoted by $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$, and the two sets of restricted maximum likelihood estimates are $\hat{\boldsymbol{\theta}}^0 = (\hat{\boldsymbol{\beta}}^0, \mathbf{0})$ for the base model, and $\hat{\pi} = G(\hat{\boldsymbol{\beta}}^\bullet)$, which is equal to the mean number of events $\bar{y}$, for the constant model.

(C3) The score vector, observed information matrix, and expected information matrix

for $L(\boldsymbol{\theta})$ are partitioned as

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} U_\beta \\ U_\gamma \end{pmatrix}; \qquad \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \begin{pmatrix} U_{\beta\beta} & U_{\beta\gamma} \\ U_{\gamma\beta} & U_{\gamma\gamma} \end{pmatrix}; \qquad -\mathrm{E}\left[n^{-1} U_{\theta\theta}\right] = \begin{pmatrix} \mathcal{I}_{\beta\beta} & \mathcal{I}_{\beta\gamma} \\ \mathcal{I}_{\gamma\beta} & \mathcal{I}_{\gamma\gamma} \end{pmatrix}$$

(C4) The likelihood parameterization $L(\boldsymbol{\eta})$ will be utilized, where $\eta_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + \boldsymbol{\gamma}^T \boldsymbol{z}_i$ is the risk score and the corresponding score residual $r(\eta_i)$ is

$$\frac{\partial L(\boldsymbol{\eta})}{\partial \eta_i} = \left(\frac{dG(\eta_i)}{d\eta_i}\right) \left[G(\eta_i)(1 - G(\eta_i))\right]^{-1} \left[y_i - G(\eta_i)\right],$$

which will be useful to rewrite as

$$r(\eta_i) = h(\eta_i)[y_i - G(\eta_i)].$$

Proof of Theorem 1: The modified NRI (mNRI) is a proper change score

For a single random variable, the modified NRI with the base and constant model parameters evaluated at their true value is

$$T_1(\boldsymbol{\theta}; \boldsymbol{\theta}^0; \pi_0) = \left[\pi_0(1 - \pi_0)\right]^{-1} r(\boldsymbol{\beta}^{0T} \boldsymbol{X}) \left[I(\boldsymbol{\beta}^T \boldsymbol{X} + \boldsymbol{\gamma}^T \boldsymbol{Z} - \boldsymbol{\beta}^{0T} \boldsymbol{X} > 0) - \frac{1}{2}\right].$$

Its expected value is equal to

$$\mathrm{E}_{X,Z} \Big\{ \left[\pi_0(1 - \pi_0)\right]^{-1} h(\boldsymbol{\beta}^{0T} \boldsymbol{X}) \times \\ \left[G(\boldsymbol{\beta}_0^T \boldsymbol{X} + \boldsymbol{\gamma}_0^T \boldsymbol{Z}) - G(\boldsymbol{\beta}^{0T} \boldsymbol{X})\right] \left[I(\boldsymbol{\beta}^T \boldsymbol{X} + \boldsymbol{\gamma}^T \boldsymbol{Z} - \boldsymbol{\beta}^{0T} \boldsymbol{X} > 0) - \frac{1}{2}\right] \Big\}$$

where $h(\boldsymbol{\beta}^{0T} \boldsymbol{X})$ is a component of the score residual in (C4) evaluated under the base model.

To show $\mathrm{E}[T_1(\boldsymbol{\theta};\boldsymbol{\theta}^0;\pi_0)]$ is maximized at $\boldsymbol{\theta}=\boldsymbol{\theta}_0$, and therefore the modified NRI is a proper change score, consider

$$\begin{aligned}&\mathrm{E}[T_1(\boldsymbol{\theta}_0;\boldsymbol{\theta}^0;\pi_0)-T_1(\boldsymbol{\theta};\boldsymbol{\theta}^0;\pi_0)]=\\&\quad\mathrm{E}_{X,Z}\Big\{[\pi_0(1-\pi_0)]^{-1}\,h(\boldsymbol{\beta}^{0T}\boldsymbol{X})\left[G(\boldsymbol{\beta}_0{}^T\boldsymbol{X}+\boldsymbol{\gamma}_0{}^T\boldsymbol{Z})-G(\boldsymbol{\beta}^{0T}\boldsymbol{X})\right]\;\times\\&\quad\left[I(\boldsymbol{\beta}_0{}^T\boldsymbol{X}+\boldsymbol{\gamma}_0{}^T\boldsymbol{Z}-\boldsymbol{\beta}^{0T}\boldsymbol{X}>0)-I(\boldsymbol{\beta}^T\boldsymbol{X}+\boldsymbol{\gamma}^T\boldsymbol{Z}-\boldsymbol{\beta}^{0T}\boldsymbol{X}>0)\right]\Big\}\end{aligned}$$

This expectation is evaluated under two cases:

Case (i): $\boldsymbol{\beta}_0^T\boldsymbol{X}+\boldsymbol{\gamma}_0^T\boldsymbol{Z}\geq\boldsymbol{\beta}^{0T}\boldsymbol{X}$

The first term in square brackets, $G(\boldsymbol{\beta}_0{}^T\boldsymbol{X}+\boldsymbol{\gamma}_0{}^T\boldsymbol{Z})-G(\boldsymbol{\beta}^{0T}\boldsymbol{X})$, is non-negative due to the monotonicity of G, and the second term in square brackets, the difference in indicator functions, is either 0 or 1. Therefore, since the weight function $h(\boldsymbol{\beta}^{0T}\boldsymbol{X})$ is positive, the expectation is non-negative for any $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\gamma})$.

Case (ii): $\boldsymbol{\beta}_0^T\boldsymbol{X}+\boldsymbol{\gamma}_0^T\boldsymbol{Z}<\boldsymbol{\beta}^{0T}\boldsymbol{X}$.

Under this constraint, the first term in square brackets is negative and the second term in square brackets is either 0 or -1. It follows that the expectation is again non-negative for any $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\gamma})$.

Combining these two cases, $\mathrm{E}[T_1(\boldsymbol{\theta};\boldsymbol{\theta}^0;\pi_0)]$ is maximized at $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ and therefore, the modified NRI is a proper change score.

Theorem 2. Assume the covariate vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ have dimension p and q, respectively. If $\rho(\boldsymbol{\theta}_0;\boldsymbol{\theta}^0;\pi_0)=0$, then the smooth NRI test statistic

$$nR_n^S(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}}^0;\hat{\pi})\;\xrightarrow{D}\;[\boldsymbol{N}_q(0,V_1)]^T[\boldsymbol{N}_q(0,V_2)]+\frac{1}{2}\boldsymbol{D}_p^T\mathcal{I}_{\beta\beta}^{-1}\boldsymbol{c}_p.$$

The first term is the inner product of two positively correlated, q-dimensional, mean zero normal random vectors, and the second term is bilinear, where $\boldsymbol{D}_p$ is a p dimensional random vector with quadratic components, and $\boldsymbol{c}_p$ is a p dimensional constant vector.

Proof of Theorem 2:

The smooth NRI test statistic is

$$R_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}^0}; \hat{\pi}) = \left[n\bar{y}(1-\bar{y})\right]^{-1} \sum_i [y_i - \bar{y}] \left[\Phi(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i + \hat{\boldsymbol{\gamma}}^T \boldsymbol{z}_i - \hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_i) - \frac{1}{2}, \right],$$

where $\Phi(\cdot)$ is the standard normal distribution function.

To determine its null reference distribution, Pepe et al. (2013) demonstrate that for correctly specified nested models (C1), $\rho(\boldsymbol{\theta}_0; \boldsymbol{\theta}^0; \pi_0) = 0$ *iff* $\boldsymbol{\gamma}_0 = 0$. This allows consideration of a second order Taylor expansion of $R_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0; \hat{\pi})$ around $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^0$,

$$nR_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}^0}; \hat{\pi}) = \left[\frac{\phi(0)}{\bar{y}(1-\bar{y})}\right] \sum_i \left[(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^0)^T \boldsymbol{x}_i + (\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^0)^T \boldsymbol{z}_i\right] [y_i - \bar{y}] + o_p(1), \quad \text{(A.1)}$$

where $\phi(0)$ represents the standard normal density function evaluated at 0, and since its derivative evaluated at zero, $\phi'(0) = 0$, each element of the matrix in the quadratic term of the expansion is equal to 0.

To further simplify, note that

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^0) = -\mathcal{I}_{\beta\beta}^{-1}\mathcal{I}_{\beta\gamma}[n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)] + (4n)^{-1/2}\mathcal{I}_{\beta\beta}^{-1}\boldsymbol{d}(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0) + o_p(n^{-1/2}) \quad \text{(A.2)}$$

which follows from a second order Taylor series approximation of the score statistic

(C3), $U_\beta(\boldsymbol{\theta})$ around $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^0$, with

$$\boldsymbol{d}(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0) = \begin{bmatrix} n^{1/2}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^0)^T[n^{-1}H^{(1)}(\hat{\boldsymbol{\theta}}^0)]n^{1/2}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^0) \\ \vdots \\ n^{1/2}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^0)^T[n^{-1}H^{(p)}(\hat{\boldsymbol{\theta}}^0)]n^{1/2}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^0) \end{bmatrix} \quad \text{and} \quad H^{(j)}(\boldsymbol{\theta}) = \frac{\partial^2 U_{\beta_j}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} .$$

Substituting (A.2) into (A.1),

$$\begin{aligned} nR_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0; \hat{\pi}) = & \left[\frac{\phi(0)}{\bar{y}(1-\bar{y})}\right] \times \\ & \left\{ \left[n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)\right]^T \left[n^{-1/2} \textstyle\sum_i \left(\boldsymbol{z}_i - \mathcal{I}_{\gamma\beta}\mathcal{I}_{\beta\beta}^{-1}\boldsymbol{x}_i\right)(y_i - \bar{y})\right] + \right. \\ & \left. \tfrac{1}{2}\left[\boldsymbol{d}(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0)\right]^T \mathcal{I}_{\beta\beta}^{-1}\left[n^{-1}\textstyle\sum_i \boldsymbol{x}_i(y_i - \bar{y})\right] \right\} + o_p(1). \end{aligned} \tag{A.3}$$

To obtain the result in Theorem 2, consider the elements in (A.3),

$$\frac{\phi(0)}{\bar{y}(1-\bar{y})} \xrightarrow{p} \frac{\phi(0)}{\pi_0(1-\pi_0)}$$

$$n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \xrightarrow{D} \boldsymbol{N}_q(0, V_\gamma)$$

$$n^{-1}\sum_i \boldsymbol{x}_i(y_i - \bar{y}) \xrightarrow{p} \boldsymbol{c}_p$$

$$\boldsymbol{d}(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0) \xrightarrow{D} \boldsymbol{D}_p$$

The remaining element is

$$n^{-1/2}\sum_i \left(\boldsymbol{z}_i - \mathcal{I}_{\gamma\beta}\mathcal{I}_{\beta\beta}^{-1}\boldsymbol{x}_i\right)(y_i - \bar{y}) .$$

First, under the null

$$n^{-1}\sum_i \left(\boldsymbol{z}_i - \mathcal{I}_{\gamma\beta}\mathcal{I}_{\beta\beta}^{-1}\boldsymbol{x}_i\right)(y_i - \bar{y}) \xrightarrow{p} E_{X,Z}\left\{(\boldsymbol{Z} - \mathcal{I}_{\gamma\beta}\mathcal{I}_{\beta\beta}^{-1}\boldsymbol{X})(G(\boldsymbol{\beta}^{0T}\boldsymbol{X}) - \pi_0)\right\}$$

which is rewritten as

$$E_{X,Z}\left\{(\boldsymbol{Z}_* - \mathcal{I}_{\gamma\beta}\mathcal{I}_{\beta\beta}^{-1}\boldsymbol{X}_*)(W_X^{-1/2}[G(\boldsymbol{\beta}^{0T}\boldsymbol{X}) - \pi_0])\right\} \tag{A.4}$$

where $\boldsymbol{Z}_* = \boldsymbol{Z}W_X^{1/2},\ \boldsymbol{X}_* = \boldsymbol{X}W_X^{1/2},$ and $W_X = \mathrm{var}[r(\boldsymbol{\beta}^{0T}\boldsymbol{X})|\boldsymbol{X}]$.

The motivation for the weight W_X comes from the Bernoulli loglikelihood (C2, C3)

$$\mathcal{I}_{\beta\beta} = E[\boldsymbol{X}_*\boldsymbol{X}_*^T] \qquad \mathcal{I}_{\gamma\beta} = E[\boldsymbol{Z}_*\boldsymbol{X}_*^T]$$

and the recognition that

$$E_{X,Z}\left\{(\boldsymbol{Z}_* - \mathcal{I}_{\gamma\beta}\mathcal{I}_{\beta\beta}^{-1}\boldsymbol{X}_*)\boldsymbol{X}_*^T\right\} = 0,$$

a $q \times p$ matrix of zeros.

Therefore by projection theory (Tsiatis 2006),

$$E[\boldsymbol{Z}_*|\boldsymbol{X}_*] = \mathcal{I}_{\gamma\beta}\mathcal{I}_{\beta\beta}^{-1}\boldsymbol{X}_*$$

and so the expectation in (A.4) is equal to zero.

It now follows from the central limit theorem,

$$n^{-1/2}\sum_i \left(\boldsymbol{z}_i - \mathcal{I}_{\gamma\beta}\mathcal{I}_{\beta\beta}^{-1}\boldsymbol{x}_i\right)(y_i - \bar{y}) \ \xrightarrow{D}\ \boldsymbol{N}_q(\boldsymbol{0}, V_2).$$

Theorem 2 is the result of Slutsky's theorem applied to the elements in (A.3).

Theorem 3. Assume the binary regression models in (C1) are properly specified and the covariate vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ have dimension p and q, respectively. If $\rho(\boldsymbol{\theta}_0; \boldsymbol{\theta}^0; \pi_0) = 0$, then

$$nT_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0; \hat{\pi}) \ \xrightarrow{D}\ k\chi_q^2,$$

where $k = \phi(0)[\pi_0(1-\pi_0)]^{-1}$, $\phi(0)$ is the standard normal density function evaluated at 0, and χ^2_q is a chi-square random variable with q degrees of freedom.

Proof of Theorem 3:

The mNRI test statistic is,

$$nT_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}^0}; \hat{\pi}) = [\bar{y}(1-\bar{y})]^{-1} \sum_i r(\hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_i) \left[\Phi(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i + \hat{\boldsymbol{\gamma}}^T \boldsymbol{z}_i - \hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_i) - \frac{1}{2} \right],$$

where $r(\cdot)$ is the score residual defined in (C4).

A second order Taylor expansion of T_n^S around $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^0$ results in

$nT_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}^0}; \hat{\pi}) =$

$$\frac{\phi(0)}{\bar{y}(1-\bar{y})} \left[n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \right]^T \left[n^{-1/2} \sum_i r(\hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_i) \left(\boldsymbol{z}_i - \mathcal{I}_{\gamma\beta} \mathcal{I}_{\beta\beta}^{-1} \boldsymbol{x}_i \right) \right] + o_p(1).$$

This approximation may be further simplified through the recognition that $\sum_i r(\hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_i) \left(\boldsymbol{z}_i - \mathcal{I}_{\gamma\beta} \mathcal{I}_{\beta\beta}^{-1} \boldsymbol{x}_i \right)$ is the efficient score statistic for estimating $\boldsymbol{\gamma}$ in the presence of $\boldsymbol{\beta}$ and evaluated under the constraint $\boldsymbol{\gamma} = 0$. It follows that (Bickel, Klassen, Ritov, and Wellner, 1993)

$$n^{-1/2} \sum_i r(\hat{\boldsymbol{\beta}^0}^T \boldsymbol{x}_i) \left(\boldsymbol{z}_i - \mathcal{I}_{\gamma\beta} \mathcal{I}_{\beta\beta}^{-1} \boldsymbol{x}_i \right) = [\mathcal{I}^{\gamma\gamma}]^{-1} [n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)] + o_p(1),$$

and therefore,

$$nT_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0; \hat{\pi}) = \frac{\phi(0)}{\pi_0(1-\pi_0)} \left[n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \right]^T [\mathcal{I}^{\gamma\gamma}]^{-1} [n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)] + o_p(1).$$

That is,

$$\Pr\left(nT_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0; \hat{\pi}) \leq u \right) = \Pr\left(k\chi^2_q \leq u \right)$$

where $k = \phi(0)[\pi_0(1-\pi_0)]^{-1}$ and χ^2_q is a chi-square random variable with q degrees of freedom.

Theorem 4. Assume the binary regression models for the training and test data have the same specification and are given in (C1), where the covariate vector $\boldsymbol{x}$ has dimension p and the covariate vector $\boldsymbol{z}$ has dimension q. Denote the estimated regression coefficients from the training data by $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^0, \hat{\pi})$, the coefficients from the test data by $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}^0, \tilde{\pi})$, and the data $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ are drawn from the test sample. If $\rho(\boldsymbol{\theta}_0; \boldsymbol{\theta}^0; \pi_0) = 0$,

$$nT_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0, \tilde{\boldsymbol{\theta}}^0; \tilde{\pi}) \ \xrightarrow{D} \ \frac{k}{2} \sum_{j=1}^{2q} \lambda_j \chi_j^2,$$

where k is defined in Theorem 3, $\{\chi_j^2\}$ are independent chi-square random variables each with one degree of freedom, and $\{\lambda_j\}$ represent eigenvalues determined from the product matrix VC, where

$$V = \begin{pmatrix} \text{var}(\tilde{\boldsymbol{\gamma}}) & 0 \\ 0 & \text{var}(\hat{\boldsymbol{\gamma}}) \end{pmatrix} \qquad\qquad C = \begin{pmatrix} 0 & D \\ D & 0 \end{pmatrix}.$$

and $D = [\text{var}(\tilde{\boldsymbol{\gamma}})]^{-1}$.

Proof of Theorem 4:

The test statistic for the NRI derived from training and test data are

$$nT_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0, \tilde{\boldsymbol{\theta}}^0; \tilde{\pi}) = [\bar{y}(1-\bar{y})]^{-1} \sum_i r(\tilde{\boldsymbol{\beta}}^{0^T} \boldsymbol{x}_i) \left[\Phi(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i + \hat{\boldsymbol{\gamma}}^T \boldsymbol{z}_i - \hat{\boldsymbol{\beta}}^{0^T} \boldsymbol{x}_i) - \frac{1}{2} \right].$$

Employing the arguments provided in the proof of Theorem 3, the smooth mNRI may be asymptotically approximated by

$$nT_n^S(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}^0, \tilde{\boldsymbol{\theta}}^0; \tilde{\pi}) = \left[\frac{\phi(0)}{\pi_0(1-\pi_0)} \right] [n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)]^T [I^{\gamma\gamma}]^{-1} [n^{1/2}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)] + o_p(1).$$

The test statistic T_n^S is bilinear, due to the different coefficient estimates $(\hat{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\gamma}})$ from

the training and test data. This statistic may be transformed to the quadratic

$$\frac{k}{2}\begin{bmatrix}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)\\(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)\end{bmatrix}^T\begin{pmatrix}0 & D\\D & 0\end{pmatrix}\begin{bmatrix}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)\\(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)\end{bmatrix}.$$

It follows from Baldessari (1967) that as $n \to \infty$,

$$\Pr\left(nT_n^S(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}}^0,\tilde{\boldsymbol{\theta}}^0;\tilde{\pi}) \le t\right) = \Pr\left(\frac{k}{2}\sum_{j=1}^{2q}\lambda_j\chi_j^2 \le t\right).$$

TABLE 1. Type 1 error for the NRI and the modified NRI test procedures using a single sample

			$\rho = 0$			$\rho = 0.5$	
n	π_0	μ_X	mNRI test	NRI test		mNRI test	NRI test
200	0.25	0.25	0.0494	0.0468		0.0496	0.0504
		0.50	0.0504	0.0578		0.0466	0.0582
		0.75	0.0484	0.0776		0.0470	0.0724
		1.00	0.0452	0.1046		0.0494	0.1034
	0.50	0.25	0.0508	0.0574		0.0516	0.0678
		0.50	0.0488	0.0820		0.0546	0.0804
		0.75	0.0538	0.1028		0.0500	0.1008
		1.00	0.0510	0.1242		0.0444	0.1276
	0.75	0.25	0.0454	0.0466		0.0444	0.0466
		0.50	0.0432	0.0588		0.0462	0.0568
		0.75	0.0464	0.0756		0.0474	0.0832
		1.00	0.0426	0.1032		0.0462	0.1036
500	0.25	0.25	0.0522	0.0630		0.0456	0.0604
		0.50	0.0468	0.0926		0.0510	0.1040
		0.75	0.0496	0.1468		0.0480	0.1392
		1.00	0.0572	0.2012		0.0502	0.1910
	0.50	0.25	0.0596	0.0726		0.0494	0.0646
		0.50	0.0494	0.1128		0.0478	0.1116
		0.75	0.0462	0.1532		0.0482	0.1590
		1.00	0.0582	0.2152		0.0506	0.2076
	0.75	0.25	0.0470	0.0624		0.0480	0.0650
		0.50	0.0480	0.0976		0.0504	0.0940
		0.75	0.0470	0.1488		0.0506	0.1472
		1.00	0.0474	0.1946		0.0490	0.1864

mNRI test = Modified NRI test with Theorem 3 reference distribution

NRI test = NRI test with normal reference distribution

n = Sample size within each simulation; ρ = Correlation between covariates (X, Z);

$\pi_0 = \Pr(Y = 1)$; μ_X = Population mean for X when $Y = 1$

TABLE 2. Type 1 error for the NRI and the modified NRI test procedures using a training and an independent test sample

			$\rho = 0$			$\rho = 0.5$	
n	π_0	μ_X	mNRI test	NRI test		mNRI test	NRI test
200	0.25	0.25	0.0518	0.0492		0.0490	0.0454
		0.50	0.0500	0.0586		0.0528	0.0590
		0.75	0.0506	0.0750		0.0498	0.0718
		1.00	0.0524	0.1128		0.0468	0.1058
	0.50	0.25	0.0500	0.0602		0.0508	0.0556
		0.50	0.0456	0.0722		0.0534	0.0818
		0.75	0.0496	0.0966		0.0486	0.0984
		1.00	0.0568	0.1210		0.0484	0.1288
	0.75	0.25	0.0516	0.0486		0.0532	0.0520
		0.50	0.0496	0.0582		0.0478	0.0640
		0.75	0.0492	0.0834		0.0516	0.0834
		1.00	0.0560	0.1076		0.0484	0.1070
500	0.25	0.25	0.0510	0.0620		0.0528	0.0634
		0.50	0.0528	0.1060		0.0494	0.0932
		0.75	0.0542	0.1340		0.0532	0.1448
		1.00	0.0536	0.2012		0.0554	0.1862
	0.50	0.25	0.0594	0.0684		0.0480	0.0654
		0.50	0.0518	0.1068		0.0560	0.1140
		0.75	0.0516	0.1510		0.0488	0.1528
		1.00	0.0524	0.1924		0.0464	0.1922
	0.75	0.25	0.0526	0.0628		0.0564	0.0610
		0.50	0.0530	0.0980		0.0498	0.0998
		0.75	0.0514	0.1476		0.0500	0.1302
		1.00	0.0504	0.1894		0.0486	0.1862

mNRI test = Modified NRI test with Theorem 4 reference distribution

NRI test = NRI test with normal reference distribution

n = Sample size within each simulation; ρ = Correlation between covariates (X, Z);

$\pi_0 = \Pr(Y = 1)$; μ_X = Population mean for X when $Y = 1$

TABLE 3. NRI and modified NRI for the prostate data

Omitted factor	NRI	P-value NRI test	mNRI	P-value mNRI test
Albumin	0.116	0.236	0.018	0.920
Alkaline phosphatase	0.336	< 0.001	0.106	0.014
Circulating tumor cells	0.627	< 0.001	0.190	< 0.001
Gleason score	0.086	0.381	0.034	0.849
Hemoglobin	0.351	< 0.001	0.088	0.020
Lactate dehydrogenase	0.027	0.787	0.056	0.322
Prostate specific antigen	0.359	< 0.001	0.080	0.138
Serum testosterone	0.195	0.046	0.044	0.490

P-value NRI test = P-value generated from the NRI test procedure with a normal reference distribution

P-value mNRI test = P-value generated from the mNRI test procedure with the reference distribution specified in Theorem 3.

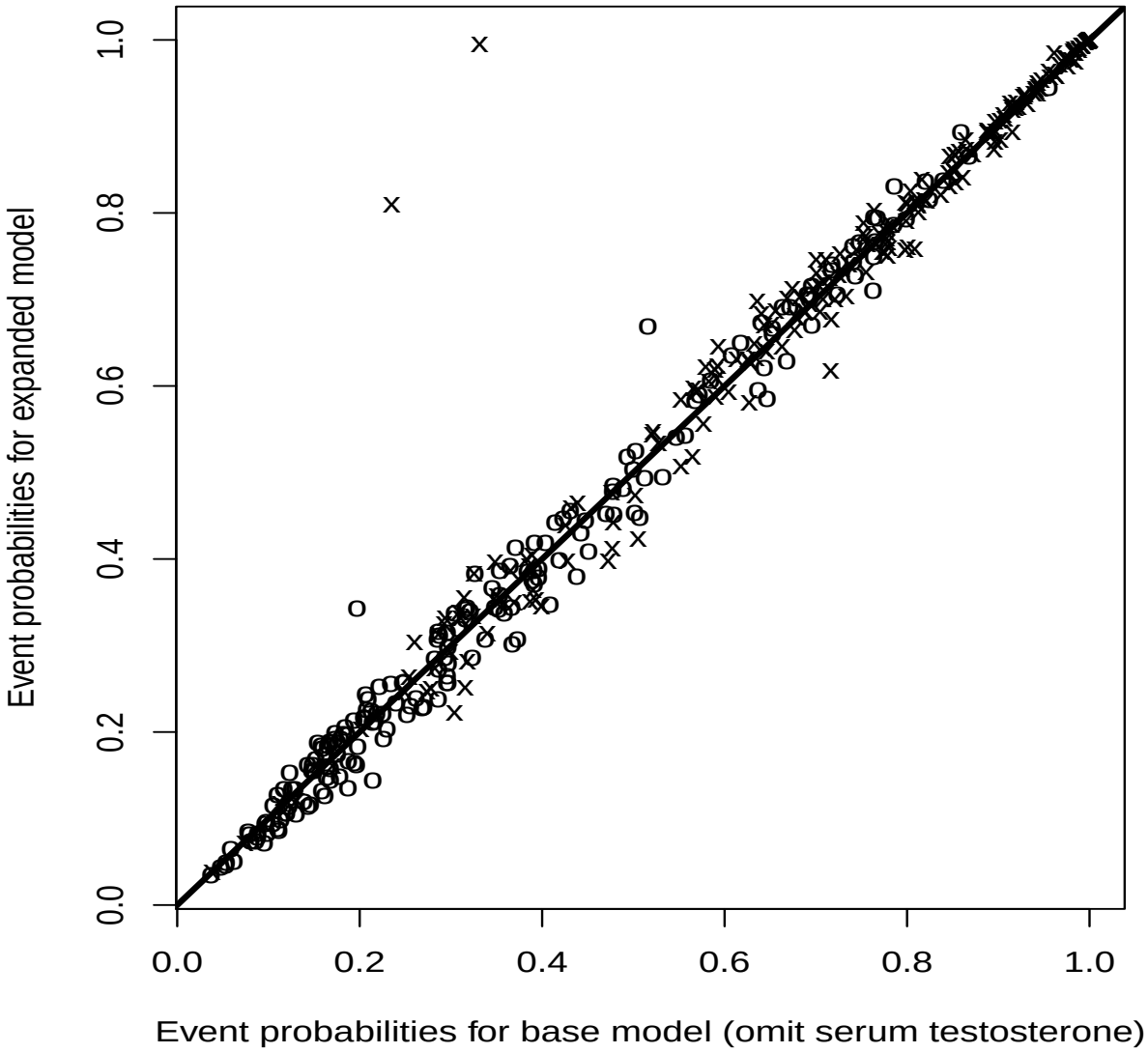

FIGURE 1 Event probabilities for each individual estimated from the base model and the expanded model. The expanded model includes all eight biomarkers and the base model omits the biomarker serum testosterone. The symbols 'o' and 'x' represent individuals that survived 24 months from the start of treatment and those who did not.

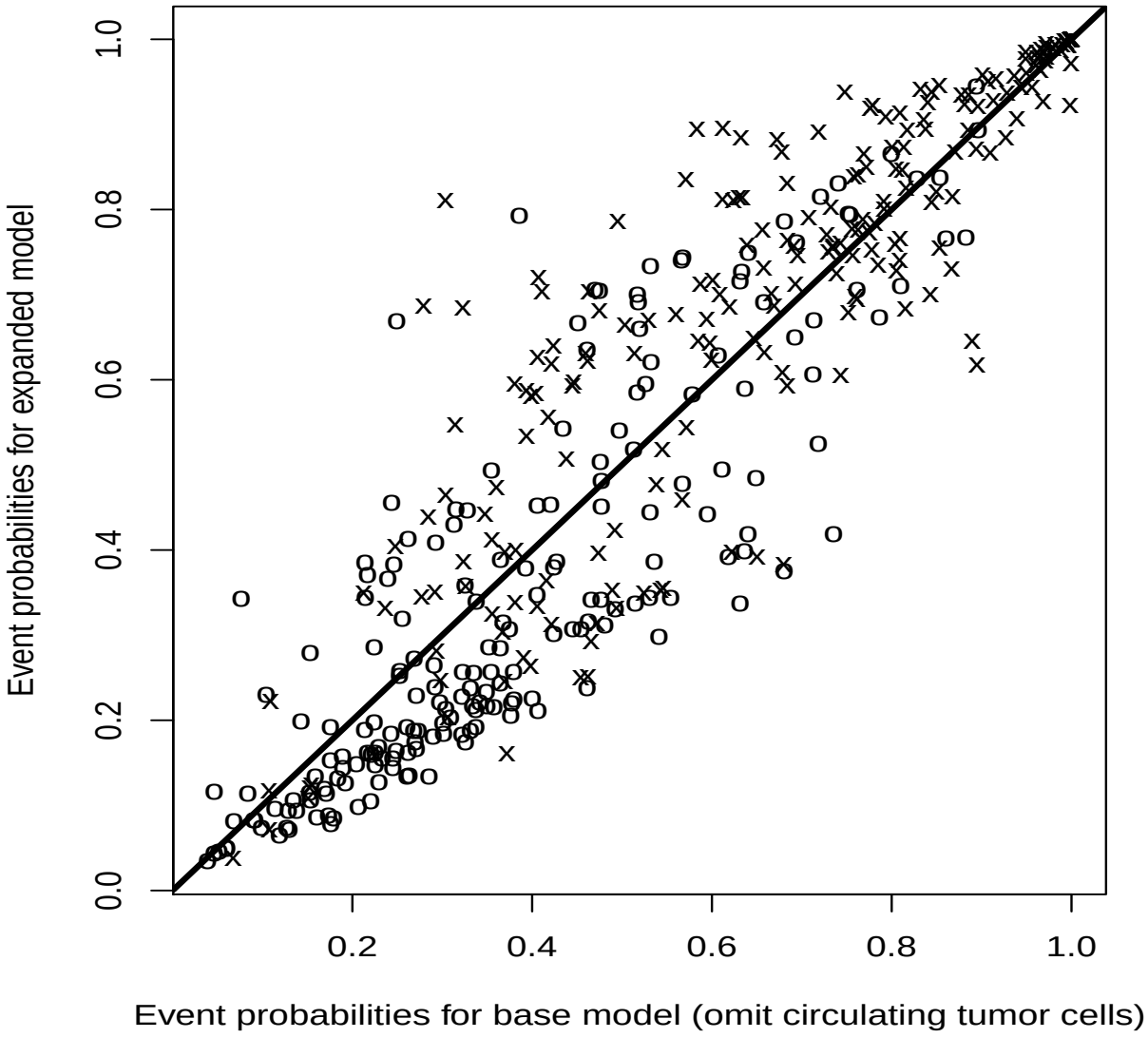

FIGURE 2 Event probabilities for each individual estimated from the base model and the expanded model. The expanded model includes all eight biomarkers, and the base model omits the biomarker circulating tumor cells. The symbols 'o' and 'x' represent individuals that survived 24 months from the start of treatment and those who did not.