# Evaluating Probability Forecasts

Shulamith T. Gross*,[1] and Catherine Huber-Carol[2,3]

[1] The City University of New York *Bernard M. Baruch College, Department of Statistics and CIS, One Baruch way, NY, NY 10010, USA.*

[2,3] Université René Descartes CNRS 2428, MAP 5, UFR de Mathématiques, 45, rue des Saints-Pères, 75 270 Paris Cedex 06, and U 1018 INSERM, 16 bis avenue Paul Vaillant-Couturier, 94 804 Villejuif, France.

**Abstract**

Mostly employed in medicine and epidemiology research, the topic we discuss is of interest in any area that is concerned with covariate-based risk evaluation. A variety of indices, such as IDI (Integrated Discrimination Improvement), NRI (Net Reclassification Improvement), the area under the ROC curve (AUC) and difference in AUC and PEV (Proportion Explained Variation) as well as predictiveness curves that compare two models' predictive capacity are routinely used. They are often used however without adequate inferential tools. We provide such tools for the IDI and the Brier Improvement (BRI) when model parameters are estimated and indices are computed on the same data, i.e., when predictiveness is evaluated within the sample. We chose these two indices as the first measures discrimination difference and the latter mostly evaluates calibration differences between the two models. We show that both sample indices are consistent and asymptotically normal as long as the true indices are not null. We also provide consistent estimates for the standard errors, leading to confidence intervals for the true indices. We evaluate our confidence intervals with percentile Bootstrap intervals through simulation. We discuss our results in the context of existing work in epidemiology, and pay special attention to the zero-index case when two models cannot be distinguished by the index under discussion.

**Key Words:** Probability Forecast; IDI; Brier Calibration Improvement; Within the sampe; Confidence Interval; Confidence Intervals; Asymptotic properties

## 1. Introduction

As long as the objective of an epidemiological study is explanatory and the monetary and availability costs of collecting information on patients is not an issue, or is not taken into account, it is most appropriate to select the best fitting model for the data. Model selection procedures for fit experienced many advances in recent years, most notably through regularization. For recent references see Friedman and Hastie (2010) and Meishausen and Bühlmann (2010). Arlot and Celisse (2010) provide an interesting survey of model selection developments in the context of cross validation, with a large bibliography. Final models are often selected using Akaike's AIC (1973) or Schwartz's BIC (1978) criteria. But if the study objective is risk prediction and the explanation of observed events and relationships are not considered, one is led to compare models by their predictive rather than fit qualities.

We point out that Akaike's AIC is an approximation defined for "within the sample" evaluation of model fit as measured by Kullback Leibler information distance and maximum likelihood estimation. "Within the sample" simply refers to the evaluation of fit (or prediction) quality of a model on the same data that had been used for model estimation.

Zheng et al (2013) have recently derived the asymptotic distribution of a modified Net Reclassification Improvement (NRI) measure of model discrimination when the same survival data are used for model estimation and modified NRI evaluation. In this paper we present similar results for the sample IDI, $\widehat{IDI}$, and the sample $\widehat{BRI}$ (the Brier improvement measure), when these are computed within the sample.

The risk predicting models considered in this paper are parametric linear logistic models, although other linear parametric models such as probit models may be similarly analyzed. Several measures of the effectiveness of risk predicting models, or markers, have been discussed in the literature. Among important contributors are Pencina et al (2008), Gu and Pepe (2009) and Uno et al (2010) where additional references may be found. More recent contributions include Pepe et al (2013) and Muhlenbruch et al (2015) where some fundamental results are obtained.

We selected two widely used measures of prediction improvement: the Integrated Discrimination Improvement (IDI), and the Brier's Improvement measure (BRI) between two models. The latter is simply the difference in Brier Score of the two models under consideration: the smaller mode's minus the larger model's Brier score. These indices of discrimination and calibration respectively have not undergone a careful study of their asymptotic properties when the indices are evaluated on the same data on which model estimation was performed. When within-the-sample evaluation is carried out, the asymptotic approximation to the sampling distribution of the index must take into account the prior estimation. The main contribution of the present paper is the derivation of this distribution for the sample $\widehat{\text{IDI}}$ and $\widehat{\text{BRI}}$. A concise discussion of earlier related contributions to the analysis of these and other discrimination or calibration measures is found in Lai et al (2011) and in the last section.

The paper is organized as follows: In section 2 basic model and index definitions are presented. Section 3 is devoted to the presentation of the models under consideration and to issues associated with the fact that the 'true' model generating the data cannot coincide with both models being compared, and therefore estimation under 'false models' must be taken into account. Since we consider maximum likelihood estimation, we refer to Hjort's (1992) seminal paper on the subject. The fundamental problem is that we consider two possible model for our data. Simple estimates of the asymptotic standard errors and confidence intervals for the IDI and the BRI are proposed (section 4) and compared to competing nonparametric bootstrap estimates and confidence intervals in a fairly large simulation study (section 5). We provide only one consistency proof, which is a prototype for all our proofs, and is presented in section 4. Proofs for theorems 2-4 will appear elsewhere. We do not report the application of our methods to a Dementia study where the efficacy of the addition of genetic Markers to the usual predictors of dementia is considered, for lack of space. In section 6 we discuss our results within the framework of much statistical and epidemiological work in this area.

## 2. Framework

## 2.1 Description of the data set and the models to be compared

The data consists of a sample of $n$ observations $X_n = (X_1, \dots, X_n)$ where $X = (Y, Z)$ : the binary response variable Y and a p-dimensional real vector of covariates Z. The true model M has risk $R(z)$ of occurrence of the event based on the covariates $z$

$$R(z) \coloneqq P(Y = 1 | Z = z, M) \tag{1}$$

The true distribution of $(Y, Z)$ referred to as M is unknown, while two competing models M1 and M2 for the data are defined for the joint distribution of $(Y, Z)$ leading to two risk models to be compared using the data :

$$R_1(z) \coloneqq P(Y = 1 | Z = z, M_1) \tag{2}$$

$$R_2(z) \coloneqq P(Y = 1 | Z = z, M_2) \tag{3}$$

The risk models for $R_1(Z)$ and $R_2(Z)$ are often taken to be logistic or probit and $M_2$ is nested in $M_1$ so that one or more covariates in $M_1$ are not included in $M_2$. Neither model $M_1$ nor the smaller model $M_2$ are assumed to coincide with the *true model* M *which remains unspecified in the sequel.* We remark that some previous works (e.g., Uno et al (2011) and Kerr et al (2011)) considered maximum likelihood estimates when the assumed model is not necessarily the true model in a variety of contexts, but none derived the asymptotic properties of the sample IDI and BRI. Hjort (1992) showed that when maximum likelihood estimated are derived under a false model, these MLE's do converge in probability to some parameter value that is the *least false value* in the sense that it minimizes the Kullback Leibler distance between of corresponding model and the true model. These MLE'1 have been also shown by Hjort (op. cit.) to be root-n asymptotically normal with an asymptotic variance provided by the author.

For ease of presentation, we assume the two models to be logistic, but any parametric model that is linear in the covariates may be similarly treated. Let $u = \theta^T z$ be the scalar product of two $(p + 1)$ dimensional real column vectors $(\theta = (\theta_0, \theta_1, \dots, \theta_p), z = (1, z_1, \dots, z_p))$, for $p \geq 1$ and $g$ the logistic function

$$g(u) = \frac{e^u}{1 + e^u} \tag{4}$$

We have a data set $X_n = ((Y_i, Z_i), i = 1, \dots, n)$ of i.i.d. random variables, and two logistic models:

$$g_1(z) = g(\theta_1^T z)$$

$$g_2(z) = g(\theta_2^T z)$$

Where some components of $\theta_1$ and of $\theta_1$ are predefined. In general the nesting model 1 may have several parameters that are not included in the nested model 1:

$$\theta_1 = (\theta_0, \theta_1, \dots, \theta_k, \dots, \theta_{k+m}) ; k + m \leq p ; m \geq 1$$

$$\theta_2 = (\theta'_0, \theta'_1, \dots, \theta'_k)$$

## 2.2 Definition of the prediction performance criteria: IDI and BRI

Where some components of $\theta_1$ and of $\theta_1$ are predefined. If we refer to the motivating example of the introduction, we have then

From Pencina et al (2008), the IDI (Integrated Discrimination Improvement) of model 2 with respect to model 1 is

$$IDI_{2/1} = E[R_2(Z) - R_1(Z)|Y = 1] - E[R_2(Z) - R_1(Z)|Y = 0] \quad (5)$$

where $E$ always denotes the expectation with respect to the true distribution of $X$. Denoting by the true $P[Y = 1]$, i.e. the population prevalence of the event under study

$$\rho \;:=\; P(Y = 1) \;=\; E[R(Z)]$$

we obtain a simpler expression for $IDI_{2/1}$. In order to derive the first expression below from (5), we note that the expectation of $R_1(Z) - R_2(Z)$ conditional on $Y = 1$ is obtained by integrating the function $R_1(z) - R_2(z)$ with respect to $P(Z = z|Y = 1)$ which equals $P(Y = 1|Z = z) * P(Z = z)/P(Y = 1) = R(z) * P(Z = z)/\rho$. The remaining expressions are similarly obtained:

$$IDI_{2/1} \;=\; \frac{1}{\rho}\, E\big[\big(R_2(Z) - R_1(Z)\big)R(Z)\big] - \frac{1}{(1-\rho)} E[\big(R_2(Z) - R_1(Z)\big)\big(1 - R(Z)\big)] \quad (6)$$

$$= \; E\{[\,\big(R_2(Z) - R_1(Z)\big)][(1/\rho + 1/(1-\rho))R(Z) - 1/(1-\rho)]\} \quad (7)$$

$$= \; E[\big(R_2(Z) - R_1(Z)\big)\left(\frac{R(Z)-\rho}{\rho(1-\rho)}\right)] \quad (8)$$

$$= \; E[\big(R_2(Z) - R_1(Z)\big)\left(\frac{Y-\rho}{\rho(1-\rho)}\right)] \quad (9)$$

The equality of (8) and (9) is proved by taking in equation (9) the expectation with respect to Z so that equation (9) may be written as:

$$E\left[\big(R_2(Z) - R_1(Z)\big)\frac{(E(Y|Z)-\rho)}{\rho(1-\rho)}\right] := E[\big(R_2(Z) - R_1(Z)\big)\left(\frac{R(Z)-\rho}{\rho(1-\rho)}\right)] \quad (10)$$

which is equal to (8).

For a single model $M_1$, the Brier's score BR is defined as

$$BR(M_1) = E[\big(Y - R_1(Z)\big)^2] \quad (11)$$

It measures the difference between the observed $(Y)$ and the predicted $(R_1(Z))$ risk of occurrence so that the bigger the Brier's score the worst is the model. Thus we define the Brier's score Improvement provided by model 2 with respect to model 1, denoted by $BRI_{2/1}$, as

$$BRI_{2/1} = BR(M_1) - BR(M_2) \quad (12)$$

$$= 2E\left[\big(R_2(Z) - R_1(Z)\big)\left(Y - \frac{(R_1(Z)+R_2(Z))}{2}\right)\right] \quad (13)$$

A positive $IDI_{2/1}$, as well as a positive $BRI_{2/1}$, means that the discrimination or calibration properties respectively of model $M_2$ are better than those of model $M_1$.

## 3. Estimation of IDI and BRI

We now assume that we have a sample of size $n$ of $X$, and the two prediction models are two nested logistic models $g(\theta_j^T\, z)$, defined for $j = 1,2$ through the respective parameters $\theta_j; j = 1,2$ as:

Model $M_1$ : $R_1(Z) = g(\theta_1^T z)$ ; $\theta_1 = (\theta_0, \theta_1, \dots, \theta_{k+m})$, $k+m \leq p$

Model $M_2$ : $R_2(Z) = g(\theta_2^T z)$ ; $\theta_2 = (\theta_0, \theta_1, \dots, \theta_k)$, $k \geq 1$

while under the true prediction model the risk is $R(z)$. Let $i = 1, \dots, n, j = 1,2$ and

$$g_{ji} := g\left(\theta_j^T z_i\right) \; for\, j = 1,2 \; and \; i = 1,2, \dots, n \tag{14}$$

Then, using (9) and (13), and a notation analogous to (14), we write

$$\widehat{g_{ji}} := g\left(\widehat{\theta}_j^T z_i\right) for j = 1,2 \; and \; i = 1,2, \dots n. \;\; 1,2, \dots, n. \tag{15}$$

where $\widehat{\theta_j}, j = 1,2$, are the maximum likelihood estimators of the parameters of models $M_1$ and $M_2$. Natural estimators of $IDI_{2/1}$ and $BRI_{2/1}$ are then respectively

$$\widehat{IDI}_{2/1} = \frac{1}{\bar{y}(1-\bar{y})}\left(\frac{1}{n}\sum_{i=1}^{n}(\widehat{g_{2i}} - \widehat{g_{1i}})(y_i - \bar{y})\right) \tag{16}$$

$$\widehat{BRI}_{2/1} = \frac{2}{n}\sum_{i=1}^{n}\left[(\widehat{g_{1i}} - \widehat{g_{2i}})\left(\frac{(\widehat{g_{1i}} + \widehat{g_{2i}})}{2} - y_i\right)\right] \tag{17}$$

Under usual regularity conditions on models $M_1$ and $M_2$, the maximum likelihood estimators $\widehat{\theta_j}, j = 1,2$ of their respective parameters $\theta_j$ converge to the values $\theta_j^*$ of $\theta_j$ that minimize the Kullback-Leibler distance of $M_j$ to the true model $M$. We shall refer to these limits as the 'least false' parameters, rather than the usual 'true' parameters. Moreover, $\widehat{\theta_j}, j = 1,2$, are asymptotically normal, with an information matrix $I_j$ that is defined below in (18). Let $L_j$ be the log-likelihood, $L'_j$ the vector of first derivatives with respect to $\theta_j$ and $L''_j$ the matrix of second derivatives. Then, the matrices $J_j = -E(L''_j)$ and $K_j = E[(L'_j)^t(L'_j)]$, which are two representations of the information matrix when the assumed model is the true model, are not equal, due to the fact that the expectation is taken with respect to the true model $M$, and the information matrix $I_j$ (see Hjort (1992)) is then :

$$I_j = J_j K_j^{-1} J_j \tag{18}$$

We now drop the index $j$ for simplicity. In our case, the likelihood, and $1/n$ the log-likelihood are respectively

$$L(\theta|y,z) := \prod_{i=1}^{n}\left(g(\theta^T z_i)^{y_i}\left(1 - g(\theta^T z_i)\right)^{1-y_i}\right)$$

$$\ln L(\theta|y,z) := \frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\left(g(\theta^T z_i)\right) + (1-y_i)\log\left(1 - g(\theta^T z_i)\right)\right]$$

and, since the logistic function $g$ defined in (4) has the following properties:

$$g'(u) = g(1-g)u' \;\; ; \;\; g(-u) = 1 - g(u)$$

$$g^{-1}(u) = \log\left(\frac{u}{1-u}\right) \;\; ; \;\; \left(\log(1-g(u))\right)' = -g(u)u'$$

The gradient $L'$ (the score vector of first derivatives), and the Hessian $L''$ (the second derivatives) of $L$ with respect to $\theta$ are respectively equal to

$$L' = \frac{1}{n}\sum_{i=1}^{n}(y_i - g_i)z_i$$

$$L'' = -\frac{1}{n}\sum_{i=1}^{n} g_i(1-g_i)z_i({}^t z_i)$$

**Lemma 3.1** (Asymptotic behavior of $\hat{\theta}$)

*Under the assumption that Z has a moment of order two, if we denote $I = JK^{-1}J$ the information matrix, we have*

$$\sqrt{n}\left(\widehat{\theta_n} - \theta^*\right) \xrightarrow{\varphi, n\to\infty} N(0, J^{-1}KJ^{-1}) := N(0, I^{-1}) \tag{19}$$

*and*

$$J := E(-L'') = E[-g(1-g)Z({}^t Z)] = \lim J_n := -lim\frac{1}{n}\sum_{i=1}^{n}\widehat{g_i}\,(1-\widehat{g_i})z_i({}^t z_i)$$

$$K := E((L')({}^t(L'))) = E[(y-g)^2 Z({}^t Z)] = \lim K_n := lim\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(y_i - \widehat{g_i})^2 z_i({}^t z_i)$$

This lemma is a direct consequence of Hjort result (1992, page 358) as the logistic model meets the usual regularity conditions that imply the consistency of $\hat{\theta}$ for the "least false" parameter $\theta^*$ and also the consistency of $J_n$ and $K_n$ as respective estimators of $J$ and $K$. Result (19) is a direct consequence of

$$\sqrt{n}\left(\widehat{\theta_n} - \theta^*\right) = \sqrt{n}((J(\theta^*)^{-1})L'(\theta^*) + O_P(1) \tag{20}$$

We now apply these results to the two models $M_j, j = 1,2$.

## 4 Asymptotic properties of $\widehat{IDI_{2/1}}$ *and* $\widehat{BRI_{2/1}}$

In order to use $\widehat{IDI}$ and $\widehat{BRI}$ in tests of significance and confidence intervals, we need to establish the consistency and asymptotic normality of these estimates of the IDI and BRI respectively. The following four theorems accomplish the task. Most proofs will appear elsewhere. In the sequel $I_j$ is the information matrix of model $j, j = 1,2$ as defined in (18).

**Theorem 4.1** (Consistency of $\widehat{IDI}$) *As* $n \to \infty, \widehat{IDI} \xrightarrow{a.s.\ n\to\infty} IDI$.

**Proof of consistency of** $\widehat{IDI}$

We now drop, for simplicity, the index 2/1. Define $\widehat{IDI}$, which would be the estimator of IDI if the two models were perfectly known to be

$$\widehat{IDI} = \frac{1}{\bar{y}(1-\bar{y})}\left[\frac{1}{n}\sum_{i=1}^{n}(g_{2i} - g_{1i})(y_i - \bar{y})\right] \tag{21}$$

Using (16), we can write

$$\widehat{IDI} - IDI = \left(\widehat{IDI} - \widetilde{IDI}\right) + \left(\widetilde{IDI} - IDI\right) := T_{1n} + T_{2n}$$

Consider $T_{1n}$ first:

$$T_{1n} = \frac{1}{\bar{y}(1-\bar{y})} \times \frac{1}{n}\sum_{i=1}^{n}[(\widehat{g_{2i}} - g_{2i}) - (\widehat{g_{1i}} - g_{1i})][y_i - \bar{y}] \tag{22}$$

$$:= \quad A_n \times B_n \tag{23}$$

By the delta method applied to the function $\frac{1}{\rho(1-\rho)}$ where $u = \bar{y}$, we get that

$$A_n = \frac{1}{\rho(1-\rho)}\left[1 + \frac{2\rho-1}{\sqrt{\rho(1-\rho)}}\frac{\varepsilon}{\sqrt{n}}\right] + O_P(\frac{1}{\sqrt{n}}) \tag{24}$$

where $\varepsilon$ is a standard normal variable, $\varepsilon \sim N(0,1)$, so that $A_n = A + O_P(\frac{1}{\sqrt{n}})$ with $A = \frac{1}{\rho(1-\rho)}$. To take care now of $B_n$ we use (20) and get

$$\widehat{\theta_j} - \theta_j^* = J^{-1}(\widetilde{\theta_j})L'(\widetilde{\theta_j})(\widetilde{\theta_j} - \theta_j^*)\,;\; j = 1,2 \tag{25}$$

where $\widetilde{\theta_j}$ lies between $\widehat{\theta_j}$ and $\theta_j^*$

Dropping the index $j$ in $g_{ji}$ in $B_n$, we look at the behavior of $\frac{1}{n}\sum_{i=1}^{n}[(\widehat{g_i} - g_i)(y_i - \bar{y})]$. The delta method applied to function $g$ of (4) yields $\widehat{g_i} - g_i = g_i(1-g_i)z_i(\hat{\theta} - \theta^*) + O_P(\frac{1}{\sqrt{n}})$ so that $B_n$ may be written as:

$$B_n = \quad \frac{1}{n}\sum_{i=1}^{n}[(\widehat{g_{2i}} - g_{2i}) - (\widehat{g_{1i}} - g_{1i})][y_i - \bar{y}] \tag{26}$$

$$= \quad \frac{1}{n}\sum_{i=1}^{n}\left[\left(g_{2i}(1-g_{2i})z_{2i}(\widetilde{\theta_2} - \theta_2^*)\right) - \left(g_{1i}(1-g_{1i})z_{1i}(\widetilde{\theta_1} - \theta_1^*)\right)\right][y_i - \bar{y}] \tag{27}$$

As $|g_{ji}|, |(1-g_{ji})|$ and $|y_i - \bar{y}|$ are bounded by 1 and $|Z|$ is integrable, and using (19), we find that $B_n = O_P(\frac{1}{\sqrt{n}})$ and $T_{1n} = O_P(\frac{1}{\sqrt{n}})$.

Finally, $T_{2n} := \widehat{IDI} - IDI$

$$T_{2n} = \frac{1}{\bar{y}(1-\bar{y})} \times \left[\frac{1}{n}\sum_{i=1}^{n}(g_{2i} - g_{1i})(y_i - \bar{y})\right] - \frac{1}{\rho(1-\rho)} \times E[(g_2 - g_1)(Y - \rho)] \tag{28}$$

$$= A_n \quad \times \quad C_n \quad - \quad A \quad \times \quad C \tag{29}$$

$$= A_n \quad \times \quad (C_{n1} - C_{n2}) \quad - \quad A \quad \times \quad \mathrm{C} \tag{30}$$

$$where \quad A \times C = IDI$$

$$C_{n1} := \quad \frac{1}{n}\sum_{i=1}^{n}(g_{2i} - g_{1i})(y_i - \pi) \tag{31}$$

$$C_{n2} := \quad (\bar{y} - \rho)\frac{1}{n}\sum_{i=1}^{n}(g_{2i} - g_{1i}) \tag{32}$$

Since we have seen that $A_n = A + O_P(\frac{1}{\sqrt{n}})$, $C_{n1} = C + O_P(\frac{1}{\sqrt{n}})$ and $C_{n2}$ is $O_P(\frac{1}{\sqrt{n}})$, we obtain that $T_{2n} = O_P(\frac{1}{\sqrt{n}})$. This terminates the proof of the consistency of $\widehat{IDI}$.

**Theorem 4.2** (CLT of $\widehat{IDI}$) as $n \to \infty$.

$$\sqrt{n}(\widehat{IDI} - IDI) \xrightarrow{L,\ n\to\infty} N(0, \sigma^2). \tag{33}$$

*provided* $\sigma^2 \neq 0, where\ \sigma^2 = \left(\frac{1}{(1-\rho)\rho}\right)^2\ var(V)\ where\ V\ is\ defined\ as$

$$V = (g(\theta_2^{*T} Z_2) - g(\theta_1^{*T} Z_1) - E_\Delta)(Y - \rho)$$
$$+ \big(Y - g(\theta_2^{*T} Z_2)\big) Z_2^T (I_2^{-1}(\theta_2^*)) E_2$$
$$- \big(Y - g(\theta_1^{*T} Z_1)\big) Z_1^T (I_1^{-1}(\theta_1^*)) E_1$$
$$+ IDI\ (2\rho - 1)(Y - \rho) - IDI\rho(1 - \rho)$$

*where $I_1$ and $I_2$ are the matrices defined in lemma* (3.1) *for models $M_1$ and $M_2$ respectively and*

$$E_\Delta = E(g_2 - g_1) \tag{34}$$

$$E_j = E\big[g_j(1 - g_j)(Y - \rho) Z_j\big],\ j = 1{,}2 \tag{35}$$

*where we recall that $g_j := g\big(\theta_j^{*T} Z_j\big), for\ j = 1{,}2$. A consistent estimator of $\sigma^2 = (\frac{1}{(1-\rho)\rho})^2\ var(V)$ is given by*

$$\widehat{\sigma^2} = [\frac{1}{\bar{y}(1 - \bar{y})}]^2 \frac{1}{n-1} \sum_{i=1}^{n} (\hat{V}_i - \hat{\bar{V}})^2.$$

Proofs of theorems 4.2 - 4.4 will appear elsewhere.

**Theorem 4.3** (Consistency of the estimated $BRI_{2/1}$)

$$\widehat{BRI}_{2/1} \xrightarrow{a.s.\ n\to\infty} BRI_{2/1} \tag{36}$$

**Theorem 4.4** (CLT for $\widehat{BRI_{2/1}}$) *As* $n \to \infty$, $\sqrt{n}\big(\widehat{BRI_{2/1}} - BRI_{2/1}\big) \xrightarrow{L\ n\to\infty} N(0, \sigma_B^2)$ (37)

*provided* $\sigma_B^2 = var(W) \neq 0, where\ \sigma_B^2 = var(W)\ with\ W\ defined\ by$

$$W_i = 2\left[\big(g(\theta_1^{*T} Z_{1i}) - g(\theta_2^{*T} Z_{2i})\big)\left(\frac{\big(g(\theta_2^{*T} Z_{2i}) + g(\theta_1^{*T} Z_{1i})\big)}{2} - Y_i\right)\right]$$

$$- \big(Y_i - g(\theta_2^{*T} Z_{2i})\big) Z_{2i}^T I_2^{-1}(\theta_2^*)(2E_2 - E_4)$$

$$+ \big(Y_i - g(\theta_1^{*T} Z_{1i})\big) Z_{1i} I_1^{-1}(\theta_1^*)(2E_1 - E_3)$$

*and*

$$E_1 = E\left[g(\theta_1^{*T} Z_{1i})\big(1 - g(\theta_1^{*T} Z_{1i})\big) Z_1\left(\frac{g(\theta_2^{*T} Z_{2i}) + g(\theta_1^{*T} Z_{1i})}{2} - Y\right)\right]$$

$$E_2 = E\left[g(\theta_2^{*T} Z_{2i})\big(1 - g(\theta_2^{*T} Z_{2i})\big) Z_2\left(\frac{g(\theta_2^{*T} Z_{2i}) + g(\theta_1^{*T} Z_{1i})}{2} - Y\right)\right]$$

$$E_3 = E\big[g(\theta_1^{*T} Z_{1i})\big(1 - g(\theta_1^{*T} Z_{1i})\big) Z_1\big(g(\theta_2^{*T} Z_{2i}) - g(\theta_1^{*T} Z_{1i})\big)\big]$$

$$E_4 = E\big[g(\theta_2^{*T} Z_{2i})\big(1 - g(\theta_2^{*T} Z_{2i})\big) Z_2\big(g(\theta_2^{*T} Z_{2i}) - g(\theta_1^{*T} Z_{1i})\big)\big]$$

*A consistent estimator of* $\sigma_B^2$ *is given by*

$\widehat{\sigma_B^2} = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{W_i} - \overline{W})^2$.

**Remark**: For a pair of nested logistic models it is easy to verify that when $Z_{k+1}, \ldots, Z_{k+m}$, the covariates in model $M_1$ that are not in model $M_2$, are all independent of the response $Y$ and of the covariates $Z_1, \ldots, Z_k$ common to the two models, $IDI_{2/1} = 0$ and $var(V) = 0$ as well. Similarly, under these conditions, $BRI_{2/1} = 0$ and $var(W) = 0$. Therefore in these cases of independent additional covariates root-n normal convergence does not hold for either the sample IDI or BRI. Our simulations in section 5 that address the case of null IDI suggest a constant multiple of a Chisquare asymptotic distribution at rate n, rather than root-n. A theoretical treatment of the null IDI or null BRI case will require Taylor expansions of second order and will be undertaken elsewhere.

## 5 Simulation studies

We ran several simulation studies based on $p = 4$ covariates:

- $Z_1$ is trinomial $\in \{-1, 0, 1\}$ with respective probabilities $(.2, .4, .4)$.
- $Z_2$ is Bernoulli with respective probabilities $(.2, .8)$.
- $Z_3 \sim \delta(1)$, exponential with parameter 1.
- $Z_4 \sim N(0.5, 1)$, normal, independent of $Y$ and of all $Z_j, j = 1,2,3$.

Based on some of these covariates, two logistic models were used to generate the data, with $\theta = (\theta_0, \ldots, \theta_4)$ defined by

- $M_1$: $\theta = (0,2,1,0,0)$,
- $M_2$: $\theta = (0,2,0,1,0)$,

so that $M_2$ is a mixed model featuring both categorical and continuous covariates, while $M_1$ is based on categorical covariates only. With data generated by these two models, several alternative models, $M_3$ to$M_6$, defined below, are evaluated relative to $M_1$ and $M_2$, via IDI and BRI. Some of them, $M_3, M_4$ *and* $M_6$, are embedded in $M_1$ or $M_2$, and one, $M_5$, contains $M_2$.
Models

- $M_3$: $\theta = (\theta_0^*, \theta_1^*, 0,0,0)$ was compared to $M_1$and $M_2$.
- $M_4$: $\theta = (\theta_0^*, 0, \theta_2^*, 0,0)$ was compared to $M_1$.
- $M_5$: $\theta = (\theta_0^*, \theta_1^*, 0, \theta_3^*, \theta_4^*)$ Was compared to $M_2$.
- $M_6$: $\theta = (\theta_0^*, 0,0, \theta_3^*, 0)$ Was compared to $M_2$.

**Remarks**

1. The value of $\theta^*$ for model $M_3$ depends on the true underlying model so that its value of $\theta^* = \theta_{3,1}^*$ when model $M_1$ is true is different from its value $\theta^* = \theta_{3,2}^*$ when model $M_2$ is the true underlying model.
2. The true values of $\theta^*$of an alternative model $M_m$ with respect to the true generating model $M_1$ that contains only categorical covariates, were computed by minimizing the Kullback-Leibler distance of model $M_m$ with respect to model $M_1$, as well as by simulating a large sample of size 1,000,000 and computing $\theta^*$ in that sample. The difference between the $\theta^*$ and the corresponding IDI and BRI of the two methods is very small as shown in Table 1. For the model $M_2$ only simulation was used because the direct computation would imply numerical integration.
3. The IDI of $M_5$ with respect to the true model $M_2$ is equal to 0, so that the estimator $\widehat{IDI}$ is estimating 0.
4. In table 2 we compare the standard error of $\widehat{IDI}$ *and* $\widehat{BRI}$ as computed using our formulae on samples of size 200 and 1000 when 5000 samples were simulated from the generating models $M_1$ or $M_2$, to the standard deviation of these statistics computed from 5000 samples

of size 200 or 1000. The results lend support to our estimated standard errors for $\widehat{IDI}$ *and* $\widehat{BRI}$.

5. In Tables 3 we consider two different confidence intervals (CI) for IDI and BRI. Both are based on $nsim$ = 5000 samples, each of size $n$ = 1000 (or *n=200*) of data generated from model $M_1$ or model $M_2$. The first confidence interval, based on theorem (4.2) is computed, for each sample, from the variance of $V_i$, as defined in theorem (4.2) as a normal 95% CI. The coverage reported is the % of these normal CI's that contain the true parameter, IDI, or BRI. The second confidence interval is the usual percentile confidence interval with 5000 replication obtained for each of 500 independent samples of size 1000 (or 200) from the generating model. The coverage probability reported for the Bootstrap is then the average coverage of the 500 samples.
6. We have also compared (table not included) our variance estimate to that of Pencina et al (2008) and found, as expected that their estimate is appropriate only when IDI=0.

**Table 1** True values of $\theta^*_{M_3/M_1}, \theta^*_{M_4/M_1}$, IDI and BRI of $M_3$ and $M_4$ with respect to $M_1$ evaluated from minimum Kullback-Leibler and a sample of size 1 million.

| $\theta^*_{M_3/M_1}$ | $\theta_0^* = 0.7781;\ \theta_1^* = 1.9386$ | $\theta_0^* = 0.7781;\ \theta_1^* = 1.9386$ |
|---|---|---|
| $\theta^*_{M_4/M_1}$ | $\theta_0^* = 0.3070;\ \theta_2^* = 0.6736$ | $\theta_0^* = 0.3107;\ \theta_2^* = 0.6697$ |
| $IDI_{M_3/M_1}$ | -0.0211 | -0.0215 |
| $IDI_{M_4/M_1}$ | -0.3130 | -0.3133 |
| $BRI_{M_3/M_1}$ | -0.0044 | -0.0044 |
| $BRI_{M_4/M_1}$ | -0.0661 | -0.0663 |

**Table 2** Standard Errors of $\widehat{IDI}$ *and* $\widehat{BRI}$ as estimated from nsim=5000 samples of size n=1000 using our formulae, and via simulations of nsim=5000 samples of size n=1000, and samples of size n=200.

| **True Model** | **Model Pair** | **Estimated S.E. of $\widehat{IDI}$** | **Simulation S.E. of $\widehat{IDI}$ nsim=5000** | **Estimated S.E. of $\widehat{BRI}$** | **Simulation S.E. of $\widehat{BRI}$ nsim=5000** |
|---|---|---|---|---|---|
| **n=1000 nsim=5000** | | | | | |
| M1 | M4:M1 | 0.028 | 0.028 | 0.0650 | 0.0620 |
| M1 | M3:M1 | 0.008 | 0.009 | 0.0200 | 0.0190 |
| M2 | M3:M2 | 0.014 | 0.015 | 0.0032 | 0.0032 |
| M2 | M5:M2 | 0.0013 | 0.0013 | 0.0003 | 0.0003 |
| M2 | M6:M2 | 0.026 | 0.027 | 0.0061 | 0.0063 |
| **n=200 nsim=5000** | | | | | |
| M1 | M4:M1 | 0.061 | 0.064 | 0.0147 | 0.0150 |
| M1 | M3:M1 | 0.018 | 0.021 | 0.0047 | 0.0048 |
| M2 | M3:M2 | 0.032 | 0.034 | 0.0074 | 0.0075 |
| M2 | M5:M2 | 0.0069 | 0.0067 | 0.0015 | 0.0016 |
| M2 | M6:M2 | 0.057 | 0.061 | 0.0137 | 0.0139 |

**Table 3** Actual coverage for nominal 95% Confidence Intervals: Asymptotic normal CI's and average of 500 percentile bootstrap confidence intervals' coverages for IDI and BRI for sample sizes n=1000 and n=200. For each of 500 samples, 5000 Bootstrap replications were taken.

**n=200 nsim=5000 B=5000**

| **Models ompared** | **Data Model** | ***True IDI*** | **IDI Estimated var normal CI** | **IDI Bootstrap CI: Average Coverage** | ***True BRI*** | **BRI Estima - ted var normal CI Coverage** | **BRI Bootstrap CI: Average Coverage** |
|---|---|---|---|---|---|---|---|
| M4:M1 | M1 | 0.3133070 | 0.930 | 0.962 | 0.0662868 | 0.935 | 0.964 |
| M3:M1 | M1 | 0.0205001 | 0.869 | 0.978 | 0.0044574 | 0.843 | 0.974 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| M3:M2 | M2 | 0.0740809 | 0.917 | 0.938 | 0.0155497 | .926 | 0.938 |
| M5:M2 | M2 | | 0.998 | 0.958 | 0.0000013 | 0.999 | 0.962 |
| M6:M2 | M2 | 0.2958171 | 0.933 | 0.954 | 0.0608998 | 0.942 | 0.940 |

**n=1000 nsim=5000 B=5000**

| **Models compared** | **Data Model** | **True IDI** | **IDI Estimated var normal CI** | **IDI Bootstrap CI: Average Coverage** | **True BRI** | **BRI Estimated var normal CI** | **BRI Bootstrap CI: Average Coverage** |
|---|---|---|---|---|---|---|---|
| M4:M1 | M1 | 0.3133070 | 0.941 | 0.960 | 0.0662868 | 0.950 | 0.956 |
| M3:M1 | M1 | 0.0205001 | 0.914 | 0.954 | 0.0044574 | 0.928 | 0.954 |
| M3:M2 | M2 | 0.0740809 | 0.935 | 0.944 | 0.0155497 | 0.945 | 0.942 |
| M5:M2 | M2 | -0.000142 | 0.979 | 0.982 | 0.0000013 | 1.000 | 0.968 |
| M6:M2 | M2 | 0.2958171 | 0.936 | 0.954 | 0.0608998 | 0.951 | 0.962 |

7. In Tables 1 and 3, we have the true IDI and true BRI for the five model pairs in our study. Note that $IDI_{M_3/M_1}$ is particularly small and $IDI_{M_5/M_2} = 0$. Despite the lack of normality of the estimated IDI, clearly portrayed in our QQ plots (not shown here) for samples even as large as 1000, our 95% confidence intervals (CLT with var(V) in Table 3) display % coverage, of 5000 samples, that are rather close to the nominal 95%. The only intervals with unexplained coverage are those for $IDI_{\mathrm{M3}/M_1}$ and for $IDI_{M_6/M_2}$ both of which have rather low values. The Bootstrap percentile confidence intervals, based on a nonparametric bootstrap with 5000 repetitions on 500 samples, are in general somewhat closer to the nominal value of 95% but are mostly on the high side. Table 3 also displays the corresponding results of the BRI. Here the coverage of the normal confidence intervals based on our estimate of the standard error of $\widehat{BRI}$ have mostly a coverage probability that is closer to the nominal value of 9. The most extreme is the interval for $BRI_{M_5/M_2}$ with true BRI equal to 0.

A final word about the nonparametric Bootstrap we used to obtain our alternative confidence intervals and their coverage. Babu et al (1989) report that under appropriate smoothness conditions on the population distribution F, the sup distance between the sampling distribution of a statistic T and its Bootstrap distribution is $o\left(\frac{1}{\sqrt{n}}\right)$. Thus, in case IDI or BRI are very small, it is safer to report the nonparametric percentile Bootstrap confidence intervals that do not rely on symmetry.

## 6 Concluding remarks.

We have presented results that enable researchers do inference on two important indices for measuring the relative effectiveness of two models in predicting the probabilities of future events. Most importantly, we allowed for model estimation prior to index computation on the same data by providing new standard errors for both the IDI and the Brier score when the indices are computed on the same data that provided model parameter estimates. One referee inquired whether our contribution is meant to replace cross-validation that is often proposed for model validation after model selection. Our contribution is in fact not meant to replace cross validation, but simply to provide correct inferential tools for model selection when prediction rather than fit is the desired criterion. We point out that the AIC (Akaike 1973), which is often employed for model selection for fit, does indeed take into account the maximum likelihood estimation prior to fit evaluation by Kullback Leibler distance between the data and the model under consideration. For a recent illuminating discussion of the AIC, see Shuhua Hu (http://www4.ncsu.edu/ shu3/Presentation/AIC2012.pdf)

There are additional indices for comparison of models' calibration and discrimination. We mention in particular the difference in the area under the ROC curve of two models and Pencina et al's (2008) Net Reclassification Improvement NRI. The recent paper by Zheng et al (2013) and the work of Uno et al (2010) and the work of Muhlenbruch (2015) are relevant in this context. The first tackles the problem of within the sample estimation for the NRI as we have done for the IDI. Uno's C-statistic applied to uncensored duration data that is converted to a binary outcome by replacing the time to event by one if event occurred before some predetermined time t∗, and zero otherwise, reduces exactly to the area under the ROC curve resulting from a model for binary outcomes. However, Uno's methods must be modified to apply to binary outcomes following logistic or probit models. In addition they do not present explicit normal convergence results. Muhlenbruch et al (op. cit.) adopt a very elementary approach to the asymptotic distribution of the NRI based on the multinomial distribution, and appears to cover the within the sample setting of Zheng et al. (Op. Cit.)

One of the referees has pointed out the paper by Kerr et al (2011) which was published after we first submitted our paper for publication. They point out the popularity of the IDI which warrants further investigation of its behavior. They treat testing the hypothesis of null IDI whereas we estimate the IDI and its standard error and provide confidence intervals, precisely when IDI is not zero. Their theory for the assumed linear model M for the risk, $P[Y = 1|Z, M]$, is a short cut for studying the behavior of the $\widehat{IDI}$ when IDI = 0. They suggest however that when IDI = 0, the asymptotic distribution is not normal at rate$\sqrt{n}$, but rather a multiple of Chi-square at rate n. A significant contribution is their simulation of $\widehat{IDI}$ and the Pencina et al proposed estimate of its standard error for very large samples of sizes 1,000 to 10,000, particularly for the zero-IDI case. To test their null hypothesis of zero-IDI, the authors use a parametric bootstrap in which model parameters are replaced by their samples estimates, instead of using hypotheses testing bootstrap that, in their case, would assume a zero-IDI within their logistic model. In their simulations, the only possible value of $\gamma$ that would yield a zero-IDI is $\gamma = 0$. Since sample estimates of zero coefficients will always yield non-zero results their bootstrap yields a biased estimate of the $\widehat{IDI}$ distribution. We in contrast used a non-parametric Bootstrap which is the one recommended by Hjort (1992) as the only asymptotically correct Bootstrap when the model being estimated is a false model. As Hjort (Op. Cit.) proves, parametric bootstraps under a false model are, even asymptotically, biased.

In this connection we mention a more recent important paper by Pepe, Kerr et al (2013). In this paper they prove the very plausible claim that the null hypothesis of equal risk $R_{M1}(\mathbf{Z}) = R_{M2}(\mathbf{Z})$ for two nested models $M_1$ and $M_2$ is *equivalent* to a whole slew of null hypotheses of equality of the two ROC curves as well as $IDI_{2/1}=0$ and $NRI_{2/1} = 0$. They conclude that testing risk equality is more powerful and should precede any presentation of indices such as IDI. Upon rejection of the null hypothesis one can use our results to produce confidence intervals for the non-null IDI or BRI.

In this paper we assume nested logistic models and treat non-zero IDI and BRI model-pair estimation: both their asymptotic distribution and estimation of their standard error. In addition, we find, via simulations, that the zero IDI/BRI cases are a special case in which the normal convergence we prove in the non-null case does not appear to hold. We intend to treat the null case in a future paper.

In this paper we have tested our asymptotic results and standard error formulae in simulation studies and found them to hold even in medium size samples. Elsewhere we applied these results to Dementia data from the French Three Cities study. The analysis included a Bootstrap confidence interval for IDI between the models with and without genetic marker. The IDI analysis did not provide any evidence to suggest the effectiveness of the genetic marker APOE4 in predicting Dementia beyond that achieved by standard non-genetic predictor variables such as age, education, and additional health variables. This in spite of the fact that the model including the genetic marker turned out to have a lower AIC than the model without it, and the marker coefficient was

significantly different from zero when included in the model. This suggests that IDI may be very close to zero even when the two nested models considered are significantly different from each other. This suggests that a covariate that contributes to model fit may not contribute significantly to prediction in finite samples. In infinite samples Pepe Kerr et al (2013) have proved that this cannot happen.

## Acknowledgements

We wish to thank our student Yang Liu for the simulations and preparations of the tables in this paper.

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Proc. 2nd International Symposium on Information Theory. (eds. B.N. Petrov and F. Csaki), Akademiai Kiado, Budapest, 267–281. (Reproduced in: Breakthroughs in Statistics, I, Foundations and Basic Theory, (eds S. Kotz and N. L. Johnson), Springer-Verlag, New York, (1992), 610–624).

Arlot, S. and Celisse, A. (2010).A survey of cross-validation procedures for model selection. Statistical Surveys, **4**, 40–79.

Babu, G.J., Bose, A. (1989). Bootstrap Confidence Intervals. Statistics and Probability Letters 7, 151-160.

Empana, J.P., Tafflet, M., Escolano, S., Vergnaux, A.C., Bineau, S., Ruidavets, J.B., Montaye, M., Haas, B., Czernichow, S., Balkau, B. and Ducimetiere, P. (2011). Predicting CHD risk in France: a pooled analysis of the D.E.S.I.R., Three City PRIME and SU.VI.MAX studies. Eur J Cardiovasc Prev Rehabil. Epub, 18:2, 175–85.

Friedman, J. and Hastie, T.(2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of statistical software, 33:1, 1–22.

Gu, W. and Pepe, M. (2009). Measures to Summarize and Compare the Predictive Capacity of Biomarkers. International Journal of Biostatistics, **5**.

Hjort, N.L. (1992). On Inference in Parametric Survival Data Models. International Statistical Review, 60:3, 355–387.

Kerr, K.F., McClelland, R.L., Brown,E.R. and Lumley,T. (2011). Evaluating the Incremental Value of New Biomarkers With Integrated Discrimination Improvement. American Journal of Epidemiology,174:3, 364–374.

Lai, T.L., Gross, S.T. and Shen D.B. (2011). Evaluating probability forecasts. Annals of statistics 39:5 2356–2382.

Meishausen, N. and Bühlmann, P. (2010). Stability selection. Journal of the Royal statistical society B, 72:4, 417–473.

Muhlenbruch, K., Kuxhaus, O., Pencina, M. J., Boeing, H., Hannelore, L., Schulze, M. B., (2015). A confidence ellipse for the net reclassification improvement. Eur J Epidemiol 30:299-304.

Pencina, M.J., D'Agostino Sr, R.B., D'Agostino Jr, R.B., 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 ; 27 157–172.

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

Schwartz, G. (1978). Estimating the dimension of a model, Annals of Statistics, 6, 461–464.

Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B., Wei, L.J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medecine, 30(10), 1105–17.

Zheng, Y., Parast, L., Cai, T., Brown, M. (2013) Evaluating incremental values from new predictors with net reclassification improvement in survival analysis. Lifetime Data Anal. 19(3): 350-370.