OPEN ACCESS Freely available online

PLOS ONE

Alternative Performance Measures for Prediction Models

Yun-Chun Wu[1], **Wen-Chung Lee**[1,2]*

1 Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan, **2** Research Center for Genes, Environment and Human Health, College of Public Health, National Taiwan University, Taipei, Taiwan

Abstract

As a performance measure for a prediction model, the area under the receiver operating characteristic curve (AUC) is insensitive to the addition of strong markers. A number of measures sensitive to performance change have recently been proposed; however, these relative-performance measures may lead to self-contradictory conclusions. This paper examines alternative performance measures for prediction models: the Lorenz curve-based Gini and Pietra indices, and a standardized version of the Brier score, the scaled Brier. Computer simulations are performed in order to study the sensitivity of these measures to performance change when a new marker is added to a baseline model. When the discrimination power of the added marker is concentrated in the gray zone of the baseline model, the AUC and the Gini show minimal performance improvements. The Pietra and the scaled Brier show more significant improvements in the same situation, comparatively. The Pietra and the scaled Brier indices are therefore recommended for prediction model performance measurement, in light of their ease of interpretation, clinical relevance and sensitivity to gray-zone resolving markers.

Citation: Wu Y-C, Lee W-C (2014) Alternative Performance Measures for Prediction Models. PLoS ONE 9(3): e91249. doi:10.1371/journal.pone.0091249

Editor: Ju-Seog Lee, University of Texas MD Anderson Cancer Center, United States of America

Received September 27, 2013; **Accepted** February 10, 2014; **Published** March 7, 2014

Funding: This paper is partly supported by grants from National Science Council, Taiwan (NSC 102-2628-B-002-036-MY3) and National Taiwan University, Taiwan (NTU-CESRP-102R7622-8). No additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: wenchung@ntu.edu.tw

Introduction

Risk prediction models are important for both patients and physicians alike. A prediction model can be used to integrate an individual's socio-demographic variables, medical histories and biomarker values, etc., and to translate them into a disease risk, upon which prognostication and/or treatment decision can be based. Examples are the prediction models for cardiovascular diseases [1], hypertension [2], diabetes [3] and different forms of cancer [4–6]. Prediction model performance must be evaluated in a scientific way. There are two aspects to model performance: calibration and discrimination. Calibration is a measure of how well predicted probability agrees with actual observed risk, while discrimination is a measure of how well a model separates those who do and do not have the disease of interest [7]. This study focuses on evaluating the discrimination ability of a prediction model.

The area under the receiver operating characteristic (ROC) curve (AUC) (also referred to as the c statistic) is by far the most popular index of discrimination ability [8]. AUC is defined as the probability that the predicted probability of a randomly selected diseased subject will exceed that of a randomly selected non-diseased subject. AUC is a value between 0.5 and 1.0, with a higher value indicating better prediction performance. A prediction model with an AUC value of 0.5 is no better than tossing a coin, and at the other extreme, a model with a 1.0 AUC value is a perfect model, with 100% accurate predictions. However, AUC has been criticized as insensitive to the addition of strong marker(s), typically resulting in only small changes in value [9,10]. A small change in AUC (ΔAUC), even though it is statistically significant, can be difficult to interpret. For example, the addition of C-reactive protein to a set of standard risk factors predicting cardiovascular disease only increases the model AUC from 0.72 to 0.74 [11], and the ΔAUC is a mere 0.001 (from 0.900 to 0.901) when a genotype score (derived from a total of 18 alleles) is added into the prediction model for type 2 diabetes [3]. One cannot help wondering whether this is because the C-reactive protein and the genotype score (despite their strong associations with the disease) are actually useless in disease prediction, or whether the AUC's insensitivity to model performance change is entirely to blame.

Recently, a number of 'relative-performance' indices that are sensitive to performance change have been proposed [12]. These measures specifically compare models with and without new markers, and include net reclassification improvement (NRI), continuous NRI (cNRI) and integrated discrimination improvement (IDI) [13,14]. NRI is defined as the difference between the proportion of subjects 'moving up' (changing to higher risk categories in the model with the new marker(s)) and the proportion of subjects 'moving down' (changing to lower risk categories) for diseased subjects, and the corresponding difference in proportions for non-diseased subjects [13]. cNRI and IDI also hinge on such up and down movement. In cNRI, any increase (decrease) in predicted probability constitutes a movement up (down) [14]. In IDI, the actual amount of increase/decrease in predicted probability is counted [13]. However, a relative-performance measure can sometimes lead to self-contradictory conclusions. For example, a situation may occur in which the prediction performances of models A, B and C are rated, using a relative performance index, as A>B and B>C, yet paradoxically, A<C.

This paper describes and compares a number of alternative performance measures for prediction models. These include the Lorenz curve-based Gini and Pietra indices [15] and a

standardized version of the Brier score, the scaled Brier (sBrier) [7]. All these are absolute measures, directly reflecting the prediction performance of a specific model, and when used for model comparisons they do not produce self-contradictory results. The sensitivity of these measures to performance change when new marker(s) are added to a baseline model will also be examined.

Methods

Formulas for Various Performance Measures

Assume that there are a total of n subjects (indexed i) in a population, of which n_1 ($i=1,...,n_1$) subjects are diseased ($D_i=1$), and n_2 ($i=n_1+1,...,n$) subjects are non-diseased ($D_i=0$). Assume a prediction model which yields a predicted probability, $\hat{p}_i$, for each and every subject in the population. The prediction model is well calibrated and unbiased such that the mean predicted probability, $\bar{p}$, is equal to disease prevalence in the population, that is, $\bar{p}=n_1/n$. Figure 1 presents the computing formulas and interpretations of various performance measures, including AUC, Gini, Pietra and sBrier.

The formula for AUC is

$$\mathrm{AUC}=\frac{\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n} S\left(\hat{p}_i,\hat{p}_j\right)}{n_1\times n_2},$$

where $S\left(\hat{p}_i,\hat{p}_j\right)$ is a scoring function comparing the predicted probabilities for a pair of subjects: $S\left(\hat{p}_i,\hat{p}_j\right)=1$ if $\hat{p}_i>\hat{p}_j$, 0.5 if $\hat{p}_i=\hat{p}_j$, and 0 if otherwise. The formula clearly shows that AUC is the probability that the predicted probability of a randomly selected diseased subject exceeds that of a randomly selected non-diseased subject.

It is of interest to compare the computing formulas for Gini, Pietra and sBrier:

$$\begin{aligned}\mathrm{Gini}&=\frac{\text{mean separation for the current model}}{\text{mean separation for an error-free model}}\\&=\frac{\frac{1}{n^2}\times\sum_{i=1}^{n}\sum_{j=1}^{n}|\hat{p}_i-\hat{p}_j|}{\frac{1}{n^2}\times\sum_{i=1}^{n}\sum_{j=1}^{n}|D_i-D_j|}\\&=\frac{\sum_{j=1}^{n}\sum_{i=1}^{n}|\hat{p}_i-\hat{p}_j|}{2\times n^2\times\bar{p}\times(1-\bar{p})},\end{aligned}$$

$$\begin{aligned}\mathrm{Pietra}&=\frac{\text{mean gain for the current model}}{\text{mean gain for an error-free model}}\\&=\frac{\frac{1}{n}\times\sum_{i=1}^{n}|\hat{p}_i-\bar{p}|}{\frac{1}{n}\times\sum_{i=1}^{n}|D_i-\bar{p}|}\\&=\frac{\sum_{i=1}^{n}|\hat{p}_i-\bar{p}|}{2\times n\times\bar{p}\times(1-\bar{p})},\end{aligned}$$

and

$$\begin{aligned}\mathrm{sBrier}&=1-\frac{\text{mean squared error for the current model}}{\text{mean squared error for the null model}}\\&=1-\frac{\frac{1}{n}\times\sum_{i=1}^{n}(D_i-\hat{p}_i)^2}{\frac{1}{n}\times\sum_{i=1}^{n}(D_i-\bar{p})^2}\\&=\frac{\sum_{i=1}^{n}(\hat{p}_i-\bar{p})^2}{n\times\bar{p}\times(1-\bar{p})}\\&=\frac{\frac{1}{n}\times\sum_{i=1}^{n}(\hat{p}_i-\bar{p})^2}{\frac{1}{n}\times\sum_{i=1}^{n}(D_i-\bar{p})^2}\\&=\frac{\text{mean squared gain for the current model}}{\text{mean squared gain for an error-free model}},\end{aligned}$$

respectively. Note that initially, all subjects in the population are on the same footing - the same *a priori* probability ($\bar{p}$). When a prediction model is used, however, they diverge ($\hat{p}_i$s are different in general). Gini quantifies the "separation" (subject-to-subject variation in the *a posteriori* probability) of a model, while Pietra and sBrier quantify the "gain" (deviation of the *a posteriori* probability from the *a priori* probability).

Simulation Schemes

Three variables are assumed to be predictive of a particular disease (D): the baseline score (S) and two new markers (M_1 and M_2). It is assumed that S is a composite of traditional risk factors (age, smoking, systolic blood pressure, total and high density lipoprotein cholesterol levels, etc.) standardized to a normal distribution with a mean of 0 and a standard deviation of 1. The new markers are assumed to be binary. In order to acknowledge a correlation between S and the two new markers, let the prevalence of M_1 and M_2 be 85% when S is above average ($S>0$), and 75%, when otherwise.

It is assumed that the discrimination power of M_1 is independent of the baseline score, whereas the discrimination power of M_2 is not uniform, but is concentrated in the gray zone of the baseline model (where the predicted probability using the baseline model is close to the *a priori* probability). Specifically, the disease risk is assumed to follow a logistic model, as below:

$$\begin{aligned}&\mathrm{logit}\,\Pr(D=1|B,M_1,M_2)\\&\quad=-3+2\times B+1.5\times M_1+2.2\times \mathrm{K}(B)\times M_2,\end{aligned}$$

where $\mathrm{K}(x)$ is a Gaussian kernel function centered at 0: $\mathrm{K}(x)=\exp\left(-x^2/0.5\right)$. In this model, the disease odds ratio per unit increase in the baseline score (disease odds ratio for one standard deviation increase in the composite variable of traditional risk factors) is $\exp(2)=7.4$. To simulate new markers that are strong predictors for the disease, we let the disease odds ratio for M_1 to be $\exp(1.5)=4.5$ irrespective of the baseline score (Figure 2), and the disease odds ratio for M_2 to reach a peak [$\exp(2.2)=9.0$] when the baseline score is at its average value

	Performance Measure			
	AUC	Gini	Pietra	sBrier
Formula	$\frac{\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n} S(\hat{p}_i,\hat{p}_j)}{n_1 \times n_2}$	$\frac{\sum_{j=1}^{n}\sum_{i=1}^{n}\lvert\hat{p}_i-\hat{p}_j\rvert}{2\times n^2\times\bar{p}\times(1-\bar{p})}$	$\frac{\sum_{i=1}^{n}\lvert\hat{p}_i-\bar{p}\rvert}{2\times n\times\bar{p}\times(1-\bar{p})}$	$\frac{\sum_{i=1}^{n}(\hat{p}_i-\bar{p})^2}{n\times\bar{p}\times(1-\bar{p})}$
Interpretation	A probability that the predicted probability of a randomly selected diseased subject exceeds that of a randomly selected non-diseased subject.	A measure of "separation" (subject-to-subject variation in the *a posteriori* probability) of a model.	A measure of "gain" (deviation of the *a posteriori* probability from the *a priori* probability).	A measure of "squared gain" (deviation of the *a posteriori* probability from the *a priori* probability).

Figure 1. Computing formulas and interpretations of various performance measures.
doi:10.1371/journal.pone.0091249.g001

($S=0$) and rapidly decay when the baseline score is above or below average (Figure 2).

A total of 500 subjects were simulated as the training sample, and another 500 subjects were simulated as the validation sample. The performances of three prediction models were compared: (I) the model with the baseline score only, (II) the model with the baseline score plus M_1 and (III) the model with the baseline score plus M_2. A total of 10000 simulations were performed.

Results

In Figure 3, it can be seen that there is almost no change in the distributions of the predicted probabilities between the baseline model (A) and the model with M_1 added (B). Using the AUC

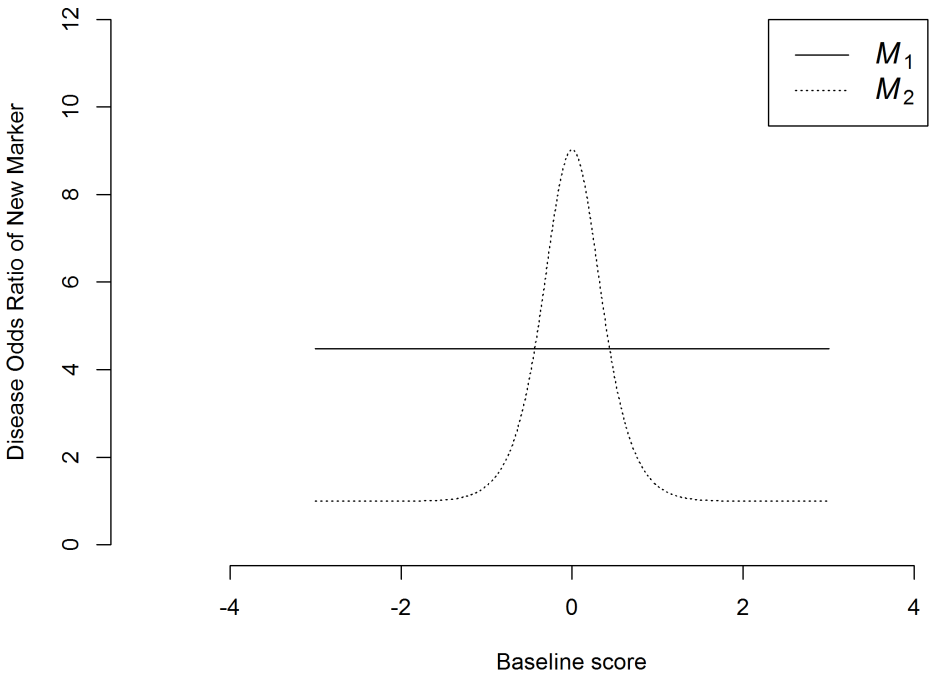

Figure 2. Disease odds ratios (discrimination powers) of the new markers (M_1 and M_2) (solid line: when the discrimination power of the new marker (M_1) is independent of the baseline score; dotted line: when the discrimination power of the new marker (M_2) is concentrated in the gray zone of the baseline model).
doi:10.1371/journal.pone.0091249.g002

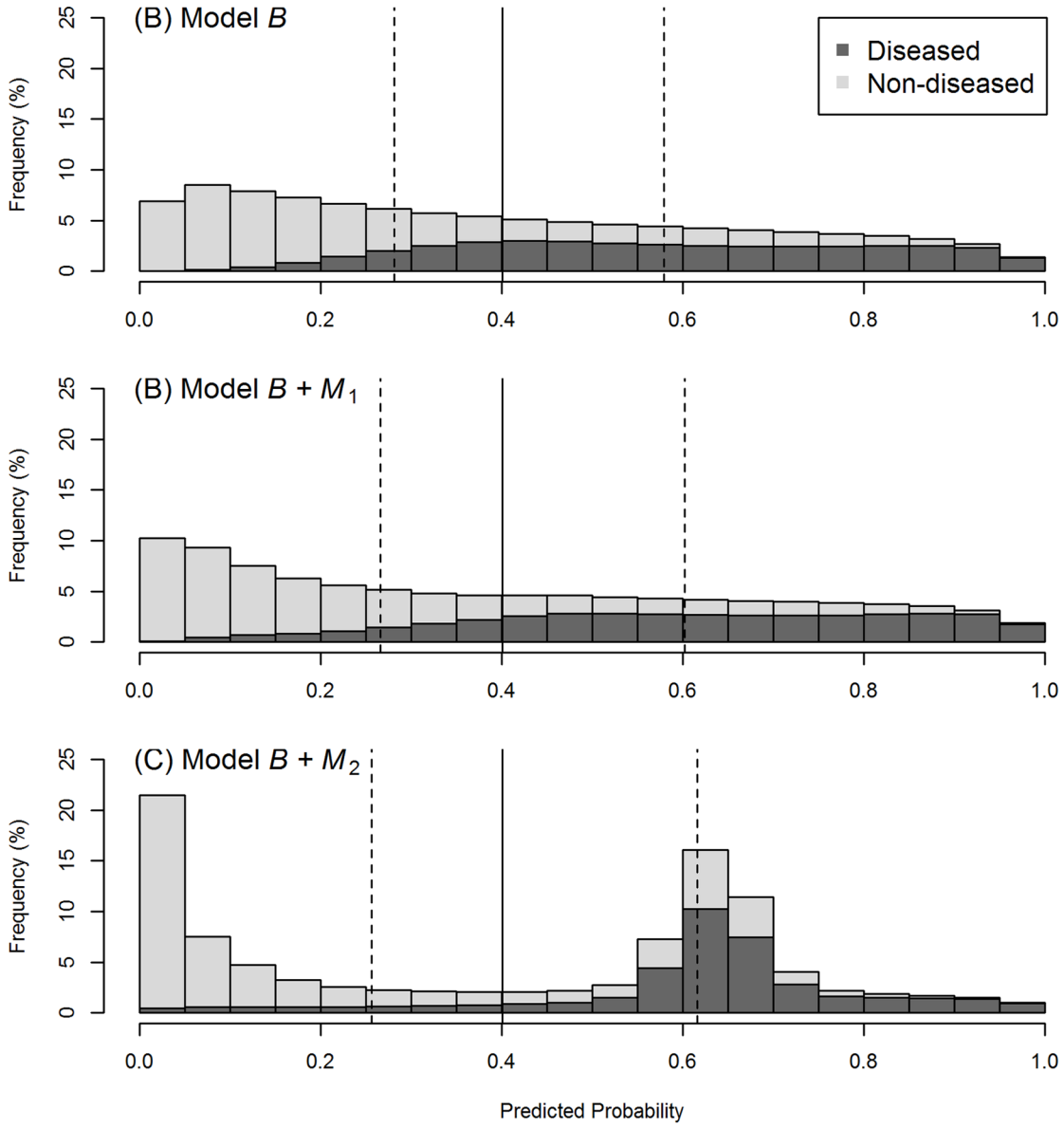

Figure 3. Distribution of the predicted probabilities for a baseline model (A), and the model with the new marker M_1 added (B), or M_2 added (C). The discrimination power of M_1 is independent of the baseline score, and that of M_2 is concentrated in the gray zone of the baseline model. The solid vertical bar indicates the grand mean of the predicted probabilities, and the two dotted vertical bars, the means of the predicted probabilities for the diseased subjects and the non-diseased subjects, respectively.
doi:10.1371/journal.pone.0091249.g003

index, it can be seen that adding M_1 increases the prediction performance of the model from 0.822 to 0.841, an absolute (relative) improvement of a mere +0.019 (+2.3%) (Table 1). Noted that the absolute improvement gauged by the Gini index (+0.039) is twice that by AUC (apart from the rounding error; in fact, $\text{Gini} = 2 \times \text{AUC} - 1$, see [15]), and the relative improvement is +6.1%. The Pietra [+0.036 (+7.4%)] and the sBrier [+0.038 (+12.4%)] also demonstrate more significant improvements than that of AUC.

By contrast, the results are much more intriguing when M_2 is added. In Figure 3, the number of people (diseased or non-diseased) in the gray zone (near the solid vertical bars) is drastically reduced when M_2 is added (C) to the baseline model (A); most diseased individuals move to the right (higher predicted probability), whereas most non-diseased individuals move to the left. An informative marker like M_2 certainly deserves a high rate; however, the AUC credits it with an absolute (relative) improvement in prediction performance of only +0.022 (+2.7%), and the Gini, twice that value, but still only +0.043 (+6.7%) (Table 1). Comparatively, the Pietra [+0.083 (+17.1%)] and the sBrier [+0.057 (+18.6%)] indices more fittingly judge the value of the marker.

It is also of interest to compare models "$B + M_1$" and "$B + M_2$" head to head. Figure 3 shows that the two models generate predicted probabilities that are quite different in distribution (B vs. C); however, AUC and Gini fail to set them apart (AUC: 0.841 vs. 0.844; Gini: 0.683 vs. 0.687). By contrast, Pietra and sBrier clearly differentiate between the two models (Pietra: 0.521 vs. 0.568; sBrier: 0.344 vs. 0.363).

Table 1. Improvements in prediction performances when new markers, M_1 and M_2, are added to a baseline model (B), respectively.

	Performance Measure			
	AUC	**Gini**	**Pietra**	**sBrier**
Model				
B	0.822	0.644	0.485	0.306
B+M_1	0.841	0.683	0.521	0.344
B+M_2	0.844	0.687	0.568	0.363
Absolute (Relative) Improvement				
from B to B+M_1	+0.019 (+2.3%)	+0.039 (+6.1%)	+0.036 (+7.4%)	+0.038 (+12.4%)
from B to B+M_2	+0.022 (+2.7%)	+0.043 (+6.7%)	+0.083 (+17.1%)	+0.057 (+18.6%)

The discrimination power of M_1 is independent of the baseline score, whereas that of M_2 is concentrated in the gray zone of the baseline model.
doi:10.1371/journal.pone.0091249.t001

In addition, this study examined situations when a strong continuous-scale marker (Exhibit S1) and multiple weak binary markers (Exhibit S2; to simulate genetic markers that are by themselves weak predictors for the disease but are strongly predictive of the disease if used collectively as a genetic score) were added to the baseline model, respectively. The conclusions regarding the comparisons of the various performance indices remain the same as when one strong binary marker is added, as shown above.

Discussion

ROC curve analysis is the most widely used method for the evaluation of diagnostic test or prediction model performance [16–19]. For any subject to be diagnosed/predicted, a diagnostic test yields a single test value which, depending on the test used, can be in binary, ordinal or continuous scale, whereas a prediction model, upon integrating the information of more than one predictor, produces a probability, which is a value between 0 and 1. Lorenz curve analysis has also enjoyed a long history of use, dating back to 1905 [20]. However, it has been primarily used by economists (demographers) to study inequality in income (population) distribution [21,22]. Lee [15] pioneered the use of Lorenz curve analysis in biomedicine (in the context of diagnostic test evaluation, although he did not consider prediction models). The interpretation of the ROC curve-based AUC index is actually rather unrealistic - subjects will not come in pairs, one being diseased and the other non-diseased, with their predicted probabilities to be compared. By contrast, Lorenz curve-based Gini and Pietra indices follow-up study subjects from their *a priori* probabilities to their *a posteriori* probabilities (after using a prediction model), and should have more relevance for actual clinical practices.

Brier score has been used to evaluate the accuracy of weather forecasting since 1950 [23]. In recent decades it has seen use in applications in biomedical fields [24–26]. Brier score depends on the disease prevalence (the *a priori* probability) of the population where the prediction model is built, and therefore it is unsuitable for making a comparison between populations. Steyerberg et al. [7] proposed a standardized version of the Brier score, the sBrier, which is an index between 0 and 1, and is prevalent-independent. Austin and Steyerberg [27] used sBrier to examine performance changes when new markers were added to a baseline model. However, they did not consider the type of markers with discrimination power concentrating in the gray zone of the baseline model, and therefore did not recognize that sBrier was sensitive to gray-zone resolving markers. Another, lesser known fact about sBrier is that the change in sBrier upon addition of new markers is equal to the IDI index itself. A proof of this is given in Exhibit S3.

It is worth noting that Gini, Pietra and sBrier indices can be expressed as ratios, comparing the resolution power (separation for Gini; gain for Pietra; squared gain for sBrier) of the current model with that of an error-free model. They are all therefore indices between 0 and 1, and can be neatly interpreted as a per cent maximum resolution power of the current model. In Table 1, the prediction performances of the baseline model are 0.644 (Gini), 0.485 (Pietra), and 0.306 (sBrier), respectively. This means that the baseline model still has a great deal of room for improvement; currently, it only achieves 64.4% separation/48.5% gain/30.6% squared gain of a sure-fire prediction model.

In this study, it is felt that patients (and their physicians) should be more interested in the gain (or squared gain) of a model (this tells how much their disease probability could be expected to be revised if they use that model), than in the separation (this compares two randomly chosen people). This study found that the two indices that quantify gains (Pietra and sBrier) are also those that are most sensitive to gray-zone resolving markers.

Taken together, Pietra and sBrier are promising alternative prediction model performance measures, in light of their ease of interpretation, clinical relevance and sensitivity to gray-zone resolving markers. Further work is needed to fully develop the statistical inference procedures (hypothesis tests and confidence intervals etc.) regarding these two indices.

Supporting Information

Exhibit S1 Simulation when a strong continuous-scale marker is added to the prediction model.
(PDF)

Exhibit S2 Simulation when multiple weak binary markers are added to the prediction model.
(PDF)

Exhibit S3 A proof that the change in sBrier upon addition of new marker(s) is equal to the IDI index.
(PDF)

Author Contributions

Conceived and designed the experiments: WCL. Performed the experiments: YCW. Analyzed the data: YCW. Contributed reagents/materials/analysis tools: WCL. Wrote the paper: YCW WCL.

References

1. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, et al. (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97: 1837–1847.
2. Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, et al. (2008) A risk score for predicting near-term incidence of hypertension: the Framingham Heart Study. Ann Intern Med 148: 102–110.
3. Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, et al. (2008) Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med 359: 2208–2219.
4. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, et al. (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 81: 1879–1886.
5. AlHilli MM, Tran CW, Langstraat CL, Martin JR, Weaver AL, et al. (2013) Risk-scoring model for prediction of non-home discharge in epithelial ovarian cancer patients. J Am Coll Surg 217: 507–515.
6. Cohen RJ, Chan WC, Edgar SG, Robinson E, Dodd N, et al. (1998) Prediction of pathological stage and clinical outcome in prostate cancer: an improved pre-operative model incorporating biopsy-determined intraductal carcinoma. Br J Urol 81: 413–418.
7. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, et al. (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21: 128–138.
8. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143: 29–36.
9. Cook NR (2010) Assessing the incremental role of novel and emerging risk factors. Curr Cardiovasc Risk Rep 4: 112–119.
10. Cook NR (2008) Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem 54: 17–23.
11. Rutter MK, Meigs JB, Sullivan LM, D'Agostino RB, Wilson PW (2004) C-reactive protein, the metabolic syndrome, and prediction of cardiovascular events in the Framingham Offspring Study. Circulation 110: 380–385.
12. Pencina MJ, D'Agostino RB, Pencina KM, Janssens AC, Greenland P (2012) Interpreting incremental value of markers added to risk prediction models. Am J Epidemiol 176: 473–481.
13. Pencina MJ, D'Agostino RB, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 27: 157–172; discussion 207-112.
14. Pencina MJ, D'Agostino RB, Steyerberg EW (2011) Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med 30: 11–21.
15. Lee WC (1999) Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. Stat Med 18: 455–471.
16. Zou KH, O'Malley AJ, Mauri L (2007) Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 115: 654–657.
17. Oates J, Casikar I, Campain A, Muller S, Yang J, et al. (2013) A prediction model for viability at the end of the first trimester after a single early pregnancy evaluation. Aust N Z J Obstet Gynaecol 53: 51–57.
18. Beukers W, Kandimalla R, van Houwelingen D, Kovacic H, Chin JF, et al. (2013) The use of molecular analyses in voided urine for the assessment of patients with hematuria. PLoS One 8: e77657.
19. Scott IC, Seegobin SD, Steer S, Tan R, Forabosco P, et al. (2013) Predicting the risk of rheumatoid arthritis and its age of onset through modelling genetic risk variants with smoking. PLoS Genet 9: e1003808.
20. Lorenz MO (1905) Methods of measuring the concentration of wealth. Publ Am Stat Asso 9: 209–219.
21. Ekelund RB, Tollison RD (1986) Economics. Boston: Little, Brown, and Company.
22. Shryock HS, Siegel JS (1975) The Methods and Materials of Demography. Washington, DC: U.S. Government Printing Office.
23. Brier GW (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev: 1–3.
24. Braun J, Bopp M, Faeh D (2013) Blood glucose may be an alternative to cholesterol in CVD risk prediction charts. Cardiovasc Diabetol 12: 24.
25. Held U, Bove DS, Steurer J, Held L (2012) Validating and updating a risk model for pneumonia - a case study. BMC Med Res Methodol 12: 99.
26. Meyfroidt G, Guiza F, Cottem D, De Becker W, Van Loon K, et al. (2011) Computerized prediction of intensive care unit discharge after cardiac surgery: development and validation of a Gaussian processes model. BMC Med Inform Decis Mak 11: 64.
27. Austin PC, Steyerberg EW (2013) Predictive accuracy of risk factors and markers: a simulation study of the effect of novel markers on different performance measures for logistic regression models. Stat Med 32: 661–672.