Published in final edited form as:

Stat Med. 2014 May 20; 33(11): 1914–1927. doi:10.1002/sim.6077.

Reclassification of Predictions for Uncovering Sub-Group Specific Improvement

Swati Biswas¹, Banu Arun², and Giovanni Parmigiani^{3,4}

¹Department of Mathematical Sciences, University of Texas at Dallas

²Department of Breast Medical Oncology and Clinical Cancer Genetics, University of Texas M.D. Anderson Cancer Center

³Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute

⁴Department of Biostatistics, Harvard School of Public Health

Abstract

Risk prediction models play an important role in prevention and treatment of several diseases. Models that are in clinical use are often refined and improved. In many instances, the most efficient way to improve a successful model is to identify subgroups for which there is a specific biological rationale for improvement, and tailor the improved model to individuals in these subgroups, an approach especially in line with personalized medicine. At present, we lack statistical tools to evaluate improvements targeted to specific sub-groups. Here we propose simple tools to fill this gap. First, we extend a recently proposed measure, the Integrated Discrimination Improvement, using a linear model with covariates representing the sub-groups. Next, we develop graphical and numerical tools that compare reclassification of two models, focusing only on those subjects for whom the two models reclassify differently. We apply these approaches to BRCA1/2, a genetic risk prediction model for breast and ovarian cancer, using data from MD Anderson Cancer Center. We also conduct a simulation study to investigate properties of the new reclassification measure and compare it with currently used measures. Our results show that the proposed tools can successfully uncover sub-group specific model improvements.

Introduction

Risk prediction plays an important role in prevention, management, and treatment of several diseases such as various cancers, cardiovascular disease, and diabetes [1–5]. Most statistical models used for this purpose still have varying misclassification rates, and their improvement is an active area of research. Improving risk prediction models that have been successful in the clinic is generally more effective than developing new models. In many instances, the most efficient way to improve a successful model is to identify subgroups of individuals for which there is a biological rationale for improvement and modify an existing model to better capture subjects in these subgroups. There are several ways by which an existing model may be modified to improve estimates for a sub-group, including using a

new prediction variable, different combination of current and/or new prediction variables, or different categories of a categorical variable. Even with the same variables, a model can be modified by using different estimates of one or more parameters, or using a different type of model (e.g., a non-linear model in place of a linear model) to better explain the relationships between variables.

A case in point is the genetic risk prediction model BRCAPRO, which estimates the probability that a person carries mutations of breast/ovarian cancer genes BRCA1 and BRCA2. BRCAPRO is a Mendelian model and calculates this estimate based on family history of breast and ovarian cancers [1] and other information. This model is widely used in genetic counseling. If the probability of carrying a mutation of BRCA1/2 is found to be high, the counselee is referred for genetic testing for mutations in these genes. BRCAPRO has been modified frequently to provide more accurate estimates [6–10]. In particular, BRCAPRO was recently extended to utilize information on breast tumor markers estrogen receptor (ER) and progesterone receptor (PR) status considered jointly [8]. To further improve BRCAPRO, our recent investigation revealed that information on another routinely collected tumor marker — the human epidermal growth factor receptor 2 (Her-2/neu), can help [10]. We found that the joint information on ER/PR and Her-2/neu status is more informative than ER/PR status alone, especially in the sub-group of subjects with ER/PR negative and Her-2/neu positive status. Even though negative ER/PR status makes a person more likely to be a carrier of a BRCA mutation, if the Her-2/neu status is positive, a person is much more likely to be a non-carrier, a fact with important clinical implications. To account for this, we recently updated BRCAPRO to utilize information on Her-2/neu status [10].

We expect that this approach to improving risk prediction models will become common in personalized medicine. From a statistical standpoint, we lack tools to evaluate improvements targeted to specific sub-groups. The Area Under the ROC Curve (AUC) has been the standard for evaluating the discrimination ability of a risk prediction model. However, in the past few years, it has been increasingly recognized that changes in AUC are not sensitive when few potentially useful factors (such as biomarkers) are added to a model that already comprises standard risk factors [11, 12]. Similarly, the AUC is likely to show little change when such an improvement is focused on specific sub-groups. For example, the standard BRCAPRO model uses extensive family history information including the relationships between members, affection ages for affected members, and current ages for unaffected members. This information is most important for genetic risk prediction. When BRCAPRO was run on a cohort of MD Anderson Cancer Center patients with this information only, the AUC was found to be 0.76 in [10]. With the addition of ER/PR, the AUC moved to 0.78, and further addition of Her-2/neu made the AUC to move up to 0.79; the confidence intervals (CIs) for AUCs in all three scenarios overlap. Thus, even though the markers ER/PR and Her-2/neu are well-justified biologically, and useful from a clinical perspective in discriminating between carriers and non-carriers of BRCA mutations in relevant subgroups, the overall AUC increases very little with their addition.

Recognizing the limitation of AUC, new measures such as the Integrated Discrimination Improvement (IDI) and the Net Reclassification Improvement (NRI) have been proposed to

investigate whether a new model provides better discrimination than an existing model [12, 13]. The AUC as well as these measures are global in nature, that is, they gauge if the newer model leads to an improvement in the risk prediction for the population under consideration as a whole. However, it is also useful to know if the newer model performs better for some specific sub-groups even when the two models perform similarly in the majority of cases. Thus, there is a need for measures that can help answer these questions and are more directly aimed at determining when and why two models disagree. The goal of this article is to fill this gap by proposing some simple tools.

The basis for the IDI are individual-level differences in the models' scores, with sign reversed in cases compared to controls. A simple way to identify sub-groups whose prediction is improved with the new model is to fit a linear model using these signed differences as the response and with covariates representing the sub-groups. We explore this extension. Another way to assess the usefulness of a new model could be to compare reclassifications of the two models. This is the essence of NRI, however, NRI involves all observations and finds whether a subject moves from one pre-determined risk category to another when the new model is used. If the goal is to determine when and why two models disagree, it may be helpful to focus only on those subjects for whom the two models actually disagree, i.e., those whom the two models reclassify differently. This idea has been used informally in comparing discrimination ability of risk prediction models [10, 14, 15]; here we aim to formalize it. In addition, the reclassification calibration (RC) statistic [16] can be used to assess model calibration. IDI, NRI, and RC have been evaluated through simulations in [17].

For the models that give continuous scores, such as BRCAPRO, we have to explicitly consider cutoffs of risk to decide whether a model reclassifies a subject as high or low risk. From a clinical standpoint, cutoffs are routinely used for risk counseling. For example, a subject with BRCA mutation carrier probability of 10% or higher is typically considered to be at high risk of carrying these mutations for the purpose of referral for genetic testing and sometimes insurance reimbursement. While cutoffs are clinically and practically relevant, they are not always chosen based on optimality criteria. So, we propose to investigate reclassification through a range of cutoffs. Using this information, we provide a simple graphical display that can be useful in clinical settings to decide if a model performs better at the cutoffs of interest and/or in some specific sub-groups. We also combine the information at all cutoffs and propose an overall measure of correct reclassification. We illustrate the approaches using BRCAPRO as applied to clinical data from the MD Anderson Cancer Center. Further, we conduct a simulation study to compare the properties of IDI and a continuous (or category-less) version of NRI — cNRI to those of the proposed overall reclassification statistic.

Methods

In this section, we first review the IDI and cNRI statistics. We then introduce an extension of IDI that incorporates covariates. Next we present a new reclassification-based approach that explicitly takes cutoffs into account.

We consider risk prediction models that provide probabilities (or risk scores) of a binary event (e.g., disease/no disease) as in [12]. Suppose there are a total of n individuals of which the first n_e experienced the event of interest. For individual i , let p_i and q_i denote the event probabilities from an “old” and a “new” model being compared, and let E_i be +1 if the event occurred, and -1 if not. The quantity $E_i(q_i - p_i)$ measures the improvement due to the new model, for individual i . The IDI is defined as

$$\text{IDI} = \frac{\sum_{i=1}^{n_e} E_i(q_i - p_i)}{n_e} + \frac{\sum_{i=n_e+1}^n E_i(q_i - p_i)}{n - n_e}.$$

The IDI captures the fact that it is desirable that the change in score be positive for subjects with events and negative for subjects without events. The range of IDI is from -1 to 1 . A value of 0 at the population level corresponds to the null hypothesis of no average difference in probabilities from the two models.

The cNRI is similar in motivation to IDI but it only uses information on whether the probabilities from the new model increase or decrease in the correct direction as compared to those from the old model. So, cNRI is defined as

$$\text{cNRI} = \frac{\sum_{i=1}^{n_e} E_i \text{sign}(q_i - p_i)}{n_e} + \frac{\sum_{i=n_e+1}^n E_i \text{sign}(q_i - p_i)}{n - n_e},$$

where $\text{sign}(q_i - p_i)$ is 1 , -1 , or 0 for $(q_i - p_i) > 0$, < 0 , and $=0$, respectively. Similar to IDI, the range of cNRI is $(-1, 1)$ with 0 being the null value.

The IDI can be easily extended to incorporate covariates. Here we propose a simple linear model with $E_i(q_i - p_i)$ (the improvement) as the response variable. An improvement can be positive or negative and its range is $(-1, 1)$, so we apply Fisher's Z-transformation to better match the assumptions of a linear model. Specifically, we define the response variable as

$$y_i = \frac{1}{2} \log \left(\frac{1 + E_i(q_i - p_i)}{1 - E_i(q_i - p_i)} \right).$$

Thus, the range of y is $(-\infty, \infty)$. Larger values of y_i are evidence that the new model is better than the old model. Let $\mathbf{X}$ denote the matrix of covariates including an intercept and interaction terms. The linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ with $\boldsymbol{\beta}$ as the column vector of regression coefficients and the column vector $\mathbf{y} = (y_1, \dots, y_n)'$ can be used to investigate the relationship between covariates and model improvement. We refer to this regression-based extension of IDI as IDIr. After fitting this linear model, we can test for significance of $\boldsymbol{\beta}$ coefficients using standard methods, e.g., t -test or likelihood ratio test. Significant coefficient indicates that the corresponding covariate is a significant predictor of improvement by the new model. When multiple covariates are considered, it may also be helpful to visually examine fitted values for specific subgroups defined by covariate combinations, to identify subgroups for which improvement is more pronounced.

Alternatively, we also explore a reclassification-based approach. Consider a specific cutoff c for classifying a subject as high or low risk. Applying the cutoff c on p 's and q 's, we get two sets of subjects for whom the two models classify differently, one satisfying $\{p_i < c, q_i \geq c\}$, and the other satisfying $\{p_i \geq c, q_i < c\}$. Here we assume that scores from the two models are expressed in the same scale (e.g., the probability scale) so that a common cutoff c can be used for dichotomizing both sets of scores. Suppose the total number of (reclassified)

subjects in these two sets is $n_R^c = \sum_{i=1}^n \{I(p_i < c, q_i \geq c) + I(p_i \geq c, q_i < c)\}$, where I is the indicator function. We focus on these n_R^c subjects and we find the Proportion of Correct Reclassification at cutoff c (PCR^c) by the new model. Here “Correct” means the reclassification by the new model matches the true status given by E_i . This is given by

$$\text{PCR}^c = \frac{\sum_{i=1}^n \{I(E_i=1)I(p_i < c, q_i \geq c) + I(E_i=-1)I(p_i \geq c, q_i < c)\}}{n_R^c}.$$

A tabular representation of this calculation is shown in Table 1. In simpler terms, PCR^c considers all subjects for whom the two models do not agree in reclassification. Out of these (n_R^c) subjects, it finds the proportion of those for whom the reclassification by the new model is correct (numerator). The numerator has two parts — one for subjects with events and another for those without the event. The larger the value of PCR^c , the greater the improvement of the new model at cutoff c . A plot of PCR^c versus c can serve as a visual aid in understanding how the proportion of correct reclassification varies with cutoff. Pointwise exact binomial confidence interval (CI) for PCR^c can be also added to the plot.

Alternatively, for large n_R^c , a large sample approximate confidence interval may be used (see Appendix). This plot is primarily intended as an exploratory graphical tool to examine how the proportion of correct reclassification by the new model varies with cutoff. However, if the analysis is to be used for formal statistical inference, a bootstrap method can be used to get CIs as described in the last paragraph of this section, and significance statements about single and multiple hypothesis on the comparisons considered can be derived.

We can use similar plots for sub-groups that have different values of categorical covariates to help in understanding if there are sub-groups where the new model performs better or worse than the old. We will provide examples with our real data analysis. In principle, a sample can be stratified using several covariates, however, one must be cautious as the net reclassified sample size n_R^c may become small or zero when many covariates are considered.

To get an overall summary of the Proportion of Correct Reclassification (PCR) across all cutoffs ($c_1, \dots, c_K$) values, we define

$$\text{PCR} = \frac{\sum_{k=1}^K \sum_{i=1}^n [I(E_i=1)I(p_i < c_k, q_i \geq c_k) + I(E_i=-1)I(p_i \geq c_k, q_i < c_k)]}{\sum_{k=1}^K n_R^{c_k}},$$

where K is the total number of cutoffs considered. PCR considers subjects for whom the two models do not agree on at least one cutoff and so there may be overlap of subjects across

different cutoffs. Out of these $(\sum_{k=1}^K n_R^{c_k})$ instances, PCR finds the proportion of instances for which the reclassification by the new model is correct. A large PCR value reflects the fact that the new model is better in reclassification on average across cutoffs. Even though n_R^c may be zero for some values of c , it is unlikely that $\sum_k n_R^{c_k}$ will be zero unless the two models are so similar that their classifications match at all cutoffs. The range of PCR is (0, 1) with 0.5 representing the null value of no improvement in reclassification by the new model or the old model.

For all three statistics — IDI, cNRI, and PCR, we can obtain CIs using the bootstrap method. As the calculation of PCR involves repetition of some subjects over different cutoffs, for generating a bootstrap sample, once a particular subject is selected, all occurrences of that subject over all cutoffs in the original data are included in the bootstrap sample. From this sample, PCR^c values at various c values as well as the overall PCR can be calculated. After repeating the process a large number of times, we will obtain distributions of these statistics, from which CIs can be calculated.

Application to familial risk prediction

As mentioned in the Introduction, we have recently extended the BRCAPRO model to utilize information on Her-2/neu [10]. We used a training dataset to develop this extension and an independent dataset for its validation. Thus, it is of interest to compare the old model that uses information on ER/PR only (“ER/PR” model) with the new model that uses the joint information on ER/PR and Her-2/neu (“ER/PR, Her-2” model). In familial risk prediction, a data unit is a proband (counselee) and each proband’s data consists of information about the proband and his/her family members. In BRCAPRO, this includes their relationships, ages, breast and ovarian cancer status, affection ages, and tumor marker status. The genetic test results for probands serve as proxy for the true BRCA carrier status. Here we analyze the same MD Anderson cohort that was used to validate the new model in [10]. Details of these data are reported in Table 2 of [10]. In particular, there are 412 probands with ER or PR and Her-2/neu information available. Of these, 77 (19%) are BRCA mutation carriers. We ran two versions of the BRCAPRO model on all probands — one version using ER/PR status only and the other using both ER/PR and Her-2/neu status — to obtain a pair of probabilities for each proband.

We plotted the probabilities from the two models for each of the four sub-groups of probands: (ER/PR+, Her-2/neu+), (ER/PR+, Her-2/neu−), (ER/PR−, Her-2/neu+), and (ER/PR−, Her-2/neu−), in Figure 2 of [10] and found that the “ER/PR, Her-2” model is most helpful for the probands who are ER/PR− and Her-2/neu+. These probands were not carriers of BRCA mutations and with inclusion of Her-2/neu in the new model, their carrier probabilities were reduced markedly. Also, we had reported estimates of IDI, cNRI, and their 95% CI for comparing the two versions as follows: IDI = 0.03 (0.02, 0.04) and cNRI = 0.12 (0.05, 0.18). Neither CI contains the null value of 0 and hence the improvement from the “ER/PR, Her-2” model is statistically significant [10].

Here, to investigate the proportion of correct reclassification by the new model, we plot PCR^c against c in Figure 1, for the entire data. The proportion remains at or above 0.6 for all cutoffs except for $c = 0.9$ (the CIs are wide as n_R^c values are mostly small). We also calculate the proportion of correct reclassification statistic, PCR (and its 95% CI) = 0.75 (0.63, 0.84). This CI excludes the null value of 0.5, and hence also gives evidence that the new model provides better discrimination.

Next, we ask if the improvement from the new model is focused on some sub-groups of probands as our earlier investigation [10] had suggested. For a visual insight, we plot PCR^c versus c for the four sub-groups in Figure 2. We did not plot the CIs here as the n_R^c values are too small for many cutoffs. The sub-group (ER/PR–, Her-2/neu+) shows substantial improvement, as the correct reclassification proportion is 1 for all cutoffs where $n_R^c > 0$. In other subgroups $n_R^c = 0$ for many values of c , suggesting that the two models perform similarly. Because of the small number of individuals being reclassified, it is difficult to assess whether there are improvements.

These plots help us gain a better understanding of the behavior of the new model. For a more formal investigation, we use the regression model for the (transformed) improvement, IDIr, as described in the Methods section. The explanatory variables are ER/PR status, Her-2/neu status, and their interaction. The results are in Table 2. We see that both ER/PR and Her-2/neu are significant predictors of the improvement obtained from the “ER/PR, Her-2” model. To specifically see whether one sub-group has better improvement than others, we can consider the fitted values for various sub-groups. For the four sub-groups, these estimates are: (1) (ER/PR+, Her-2/neu+): 0.031, (2) (ER/PR+, Her-2/neu–): –0.001, (3) (ER/PR–, Her-2/neu+): 0.109 (4) (ER/PR–, Her-2/neu–): 0.003. Thus, the new model gives the largest improvement for the sub-group (ER/PR–, Her-2/neu+). On the other hand, for the sub-group (ER/PR+, Her-2/neu–), the new model performs slightly worse, with the estimated y being negative. These results are consistent with the reclassification plots in Figure 2.

Simulation Study

Next, we carry out simulations to investigate the properties of IDI, cNRI, and PCR. To generate the data, we mimic the training data that was used in [10] to derive the new model. In particular, we use the same proportions of events and covariates status. We also vary these parameters to get different simulation scenarios. In general terms, for each subject, E represents the (true) event status, and X_1 and X_2 represent two binary covariates that mimic ER/PR and Her-2/neu, respectively.

In the first simulation scenario, for each subject, we first generate X_1 and X_2 . In the real data, 38% and 83% of the probands are ER/PR– and Her-2/neu–. So we generate $X_1 \sim \text{Bernoulli}(0.38)$ and $X_2 \sim \text{Bernoulli}(0.83)$. Then conditional on X_1 and X_2 values, we simulate E (true status), p and q (probabilities from the old and new models). E is generated from a Bernoulli distribution while (p, q) are jointly simulated as the expit transformation (inverse-logit function) of a bivariate normal (BVN) distribution with parameters listed in Table 3. The parameters for the distributions of E reflect the proportion of BRCA carriers in

each sub-group of joint ER/PR and Her-2/neu status. For example, in the (ER/PR–, Her-2/neu+) sub-group, 8% probands are carriers, and thus a Bernoulli(0.08) distribution is used. The same information is used for setting the mean parameter for q in the BVN distribution, e.g., $\text{logit}(0.08) = -2.44$. For setting the mean parameter for p , we used the proportion of carriers in each ER/PR status subgroup without regard to Her-2/neu status. There are 46% carriers in the ER/PR– group and so $\text{logit}(0.46) = -0.16$ is the mean of $\text{logit}(p)$ for the first two sub-groups with $X_1 = \text{“Negative”}$. In the ER/PR+ group, 25% are carriers and thus $\text{logit}(0.25) = -1.10$ is the mean $\text{logit}(p)$ for $X_1 = \text{“Positive”}$ sub-groups. So, we mimic a situation where we want to compare an old model that uses X_1 only with a new model that uses X_2 in addition to X_1 . However, for this choice of parameters, the new model here closely reflects the true status — a very favorable scenario. We refer to this setting as scenario 0.

For more realistic scenarios, we vary the mean and correlation parameters of the BVN distribution as shown in the first six columns of Table 4. For example, in scenario 1, (p, q) for group 4 is generated from the expit of a $\text{BVN}(-1.10, -1.32, 0.5, 0.5, 0.95)$ distribution. In scenarios 8 – 13, the mean parameters for p and q are the same in each group, so the two models will give similar probabilities; these scenarios are chosen to help us gauge the type I error rate for the tests based on the corresponding 95% CIs, while the rest help evaluate the power. The correlation coefficient is set to either 0.95 or 0.7 to explore high and moderate correlation scenarios. We generated 500 samples for each simulation scenario and each sample is of size 500. The distributions of (p, q) depend on the values of X_1 and/or X_2 as combinations of their values define the four groups in Table 4. The (p, q) values are then used to test if the two models differ in terms of reclassification. Thus, X_1 and X_2 are utilized in the test via (p, q) values.

The last three columns of Table 4 show the proportion of times the 95% CI for each statistic excluded its corresponding null value. So, this proportion is power for scenarios 0 – 7 and is type I error rate for scenarios 8 – 13. With higher correlation ($r = 0.95$), the powers are higher. The power of the test using cNRI is, in general, much lower than the ones using IDI and PCR. Also, the test using PCR has higher power than the test using IDI in all scenarios except scenario 5. However, the type I errors for IDI and PCR are different with IDI being more conservative.

As the type I errors corresponding to PCR and IDI are not the same and there is a continuum of null scenarios (any scenario with same means for p and q), it is challenging to fully compare the powers of these tests. Nonetheless, to get a better insight into this issue, we plot powers for scenarios 0 to 7 as function of type I errors for scenarios 9 and 10 in Figures 3 and 4. For most of the scenarios, the power of the test using PCR is higher than that using IDI, especially when scenario 10 (Figure 4) is used for the type I error. In scenario 5, IDI outperforms PCR substantially while in scenarios 3 and 7 in Figure 3, the two curves cross each other with the test using IDI having higher powers at small type I errors. A similar trend is seen when the type I errors from other scenarios are used (plots not shown). Thus, we conclude that both IDI and PCR are useful tools for comparing discrimination of two models.

Scenario 5 shows strikingly different results from rest of the scenarios — the power of the test using PCR is substantially lower than those using IDI and even cNRI. Thus, it is worth further investigation. In Figure 5, we plot q vs. p for a sample simulated under this scenario. For the new model to perform better, the red points (events) and blue points (no event) should fall below and above the diagonal line, respectively. Although it is clear that most of the red points do fall below the diagonal line, the improvement is not that obvious for the blue points. Overall, it seems that the new model does provide improvement, and this explains high powers of the tests using IDI and cNRI. To understand the behavior of PCR, let us first focus at cutoff $c = 0.2$. The pair of orange lines at this point forms four quadrants, and the points in the second and fourth quadrant are the ones reclassified ($n_R^{0.2}$). Of these, the ones correctly reclassified by the new model are the red points in the fourth quadrant ($q > 0.2$ for $E = 1$) and the blue points in the second quadrant ($p < 0.2$ for $E = -1$). The proportion of correct reclassification at this cutoff appears to be slightly greater than 0.5. Next, if we consider $c = 0.7$ (pair of green lines) in a similar manner, we see that the proportion of correct reclassification is around 0.7. Combining over all cutoffs, $\text{PCR} = 0.55$, and its 95% CI does not exclude 0.5. Now if we look more carefully into the parameters for this scenario in Table 4 and compare them with those for other scenarios, we see that all groups except Group 2 have some differences in the means for the two models but these differences are modest compared to the differences in Group 2 means in other scenarios. Moreover, Group 2 is a small group as $P(X_1 = \text{“Negative”}) = 0.38$ and $P(X_2 = \text{“Positive”}) = 0.17$. Thus, in this scenario, there is small to moderate improvement spread out in a large proportion of samples and PCR did not perform well. However, this and other types of scenarios need to be investigated in more depth before a generalization can be made.

Discussion

We have proposed new measures for comparing discrimination of two risk prediction models with the goal of discovering when and why the two models disagree. In the framework of IDI, we proposed a linear regression model, IDIr, whose response variable is the signed difference of probabilities from the two models, and whose covariates are subject characteristics. If the covariates are not known beforehand or there are many candidate covariates, one can use standard model selection techniques to select useful covariates. IDIr can be used as a follow-up tool if IDI (or PCR or NRI) shows one model is a significant improvement over the other model. IDIr will help uncover if the improvement is larger in some specific sub-group(s). Another regression based method was proposed earlier in literature for evaluating multiple risk prediction models [18]. As a visual tool, we have proposed a reclassification-based statistic, which explicitly shows the proportion of subjects whose risk assessment is directly impacted by the new model as compared to the old model. Further, by stratifying by sub-groups, we can see whether this impact is more pronounced in some specific sub-group(s). Thus, this measure has a direct clinical relevance. Moreover, by plotting the statistic PCR^c as a function of cutoff c , the graph explicitly shows correct reclassification rate of the new model at different values of c . If there are any specific cutoffs of clinical interest, then these plots will highlight reclassification performance of the new model at those cutoffs. IDIr and PCR^c plots used together can aid investigators in understanding the nature of improvement, if any, by the new model.

For an overall comparison of two models' discrimination, we have proposed the reclassification statistic PCR by averaging the the correct reclassification proportion PCR^c across all cutoffs. In most of the simulation scenarios that we considered, the test using PCR has higher power than the test using cNRI. The comparison with IDI depends on the specifics of the scenario considered: each is preferable in some cases but not in others. Following our BRCAPro example, we mostly focused on scenarios where the improvement from the new model was substantial only in one sub-group, which is also relatively small in size. In these scenarios, our results show that PCR and IDI should be used while the test using cNRI lacks power to detect this type of sub-group-specific improvement. On the other hand, results from scenario 5 suggest that IDI and cNRI may be preferable if there is a modest improvement from the new model that is spread over a larger proportion of samples; although a more in-depth investigation is needed. Thus, we conclude that these statistics complement each other.

Being based directly on reclassification, PCR is similar to cNRI and to categorical NRI based on two classes of risk [12]. However, an important distinction is that cNRI and NRI use the whole sample while PCR focuses only on the sub-sample where the two models differ in reclassification. In particular, the NRI with two risk categories summarizes whether reclassification of each subject by the new model moves to the correct category (as defined by the true status) or not. Similar to PCR, it computes the proportions of correct reclassification but uses as denominators the numbers of subjects with and without events and thereby considering the whole sample. NRI calculates the proportions of incorrect reclassification also in a similar manner. On the other hand, PCR^c and PCR calculate the proportion of correct reclassification out of reclassified subjects only. The overall NRI statistic (as well as cNRI and IDI) is defined as difference of proportions (for events and non-events) while PCR^c and PCR are proportions.

A strength of PCR^c and PCR is that from a clinical point of view, these statistics are highly relevant as they only look at those subjects for whom using one model instead of the other in clinic will make a difference in decision-making (n_R^c). Although PCR^c is not well defined when n_R^c is small, from a statistical point of view, a small value of n_R^c is itself an indication that the two models agree well in their reclassification at that c value. Thus, it is a useful piece of information, as we illustrated in our BRCAPro application, where the two models seem to be performing comparably in the sub-group ER/PR+, Her-2/neu- in Figure 2.

An important caveat is that the analyses proposed here are not a replacement for the investigation of whether new models overfit compared to older models. Our goal is to investigate whether two models give clinically equivalent predictions or not, and if not, in what subgroups the differences are more pronounced. Overfitting is a separate question, and should be addressed by different methodologies. To illustrate this point, we fitted two regression models to a simulated dataset — the first one containing a covariate X_1 that is associated with response, and the second one containing X_1 and X_2 , where X_2 is unassociated with response. Then we computed the discrimination statistics IDI, NRI, and PCR to compare the predicted values from the two models. We found these statistics to be significant a large proportion of times (much more than 5% times; results not shown). This

is because, even though the true $\beta_2 = 0$, in the fitted model, $\hat{\beta}_2 \neq 0$. Likely, the model including X_2 is an overfit. Thus, the predictive distributions of the two models are different and this scenario does not correspond to the null distribution for these statistics, which is, in general terms, that the two models give clinically equivalent risk scores.

Our approach is cast within the framework of reclassification indices, and it inherits strengths and limitations of this philosophy. Strengths are well articulated in several works such as [11–13, 16, 17]. Recent literature has also highlighted some limitations from a decision theoretic standpoint, including the fact that reclassification metrics may not be proper scoring rules [19, 20].

The motivation of this article came from our recent work with improving BRCAPRO [10]. Although the proportion of subjects with ER/PR– and Her-2/neu+ is small, the new BRCAPRO model with Her-2/neu clearly improves reclassification of these subjects. As standard risk prediction models such as BRCAPRO are further improved, it is unlikely that the additional factors incorporated into these models will provide a large improvement to a large proportion of population and so one can expect only a small change in AUC value. Thus, it is important to have statistical tools that can help uncover modest to small improvement in large proportion or large improvements in a small proportion of population. The latter is the one that will be especially valuable for personalized medicine, and we hope that the simple tools proposed here can be useful in this regard.

Acknowledgments

This work was supported in part by Susan G Komen for the Cure (KG081303), the National Cancer Institute (R03CA173834-01 and P30CA006516-47), and the Dana Farber Cancer Institute. We thank Dr. Pankaj Choudhary for his insightful discussion on the asymptotic properties. We are also thankful to the two anonymous reviewers and the Associate Editor for their constructive comments that led to improvement and clearer presentation of the manuscript.

Appendix

The estimator PCR^c can be expressed as a function of three averages as follows:

$$\text{PCR}^c = \frac{\frac{1}{n} \sum_{i=1}^n I(E_i=1)I(p_i < c, q_i \geq c) + \frac{1}{n} \sum_{i=1}^n I(E_i=-1)I(p_i \geq c, q_i < c)}{\frac{1}{n} \sum_{i=1}^n \{I(p_i < c, q_i \geq c) + I(p_i \geq c, q_i < c)\}} = \frac{\bar{X}_1 + \bar{X}_2}{\bar{X}_3} = f(\bar{X}_1, \bar{X}_2, \bar{X}_3), \text{ where}$$

$$\bar{X}_1 = \frac{1}{n} \sum_{i=1}^n I(E_i=1)I(p_i < c, q_i \geq c),$$

$$\bar{X}_2 = \frac{1}{n} \sum_{i=1}^n I(E_i=-1)I(p_i \geq c, q_i < c), \text{ and}$$

$$\bar{X}_3 = \frac{1}{n} \sum_{i=1}^n \{I(p_i < c, q_i \geq c) + I(p_i \geq c, q_i < c)\}$$

Next, we obtain the expected values, variances, and covariances of $\bar{X}_1$, $\bar{X}_2$, and $\bar{X}_3$. Let $\theta_1 = P(p < c, q \geq c)$ and $\theta_2 = P(p \geq c, q < c)$. Recall that out of n individuals, we assume that the first n_e are with events ($E_i = 1$) while the rest $n - n_e = n_{\bar{e}}$ are without events ($E_i = -1$). So

$$\begin{aligned}\mu_1 &= E[\bar{X}_1] = \frac{n_e}{n} P(p < c, q \geq c) = \frac{n_e}{n} \theta_1 \\ \mu_2 &= E[\bar{X}_2] = \frac{n_{\bar{e}}}{n} P(p \geq c, q < c) = \frac{n_{\bar{e}}}{n} \theta_2 \\ \mu_3 &= E[\bar{X}_3] = P(p < c, q \geq c) + P(p \geq c, q < c) = \theta_1 + \theta_2\end{aligned}$$

$$\begin{aligned}Var(\bar{X}_1) &= \frac{1}{n^2} Var\left(\sum_{i=1}^n I(E_i=1)I(p_i < c, q_i \geq c)\right) \\ &= \frac{1}{n^2} \sum_{i=1}^{n_e} P(p_i < c, q_i \geq c)(1 - P(p_i < c, q_i \geq c)) \\ &= \frac{n_e}{n^2} \theta_1(1 - \theta_1)\end{aligned}$$

Similarly,

$$Var(\bar{X}_2) = \frac{1}{n^2} Var\left(\sum_{i=1}^n I(E_i = -1)I(p_i \geq c, q_i < c)\right) = \frac{n_{\bar{e}}}{n^2} \theta_2(1 - \theta_2).$$

$$\begin{aligned}Var(\bar{X}_3) &= \frac{1}{n^2} [Var\left(\sum_{i=1}^n I(p_i < c, q_i \geq c)\right) + Var\left(\sum_{i=1}^n I(p_i \geq c, q_i < c)\right) + 2Cov\left(\sum_{i=1}^{n_e} I(p_i < c, q_i \geq c), \sum_{i=1}^n I(p_i \geq c, q_i < c)\right)] \\ &= \frac{1}{n} [\theta_1(1 - \theta_1) + \theta_2(1 - \theta_2) - 2\theta_1\theta_2] \\ &= \frac{1}{n} [\theta_1 + \theta_2 - (\theta_1 + \theta_2)^2]\end{aligned}$$

$$Cov(\bar{X}_1, \bar{X}_2) = Cov\left(\frac{1}{n} \sum_{i=1}^{n_e} I(p_i < c, q_i \geq c), \frac{1}{n} \sum_{i=n_e+1}^n I(p_i \geq c, q_i < c)\right) = 0,$$

as the first n_e individuals are independent of the rest n individuals. Following the same reasoning, we have

$$\begin{aligned}
& Cov(\bar{X}_1, \bar{X}_3) \\
&= \frac{1}{n^2} Cov\left(\sum_{i=1}^{n_e} I(p_i < c, q_i \geq c), \sum_{i=1}^n I(p_i < c, q_i \geq c) + \sum_{i=1}^n I(p_i \geq c, q_i < c)\right) \\
&= \frac{1}{n^2} Cov\left(\sum_{i=1}^{n_e} I(p_i < c, q_i \geq c), \sum_{i=1}^{n_e} I(p_i < c, q_i \geq c) + \sum_{i=1}^{n_e} I(p_i \geq c, q_i < c)\right) \\
&= \frac{1}{n^2} \left(\sum_{i=1}^{n_e} Var(I(p_i < c, q_i \geq c)) - n_e P(p < c, q \geq c) P(p \geq c, q < c) \right) \\
&= \frac{n_e}{n^2} \theta_1 (1 - \theta_1 - \theta_2)
\end{aligned}$$

Similarly,

$$Cov(\bar{X}_2, \bar{X}_3) = \frac{n_e}{n^2} \theta_2 (1 - \theta_2 - \theta_1).$$

Then

$$\mathbf{\Sigma} = Var \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \end{pmatrix} = \begin{pmatrix} \frac{n_e}{n^2} \theta_1 (1 - \theta_1) & 0 & \frac{n_e}{n^2} \theta_1 (1 - \theta_1 - \theta_2) \\ 0 & \frac{n_e}{n^2} \theta_2 (1 - \theta_1 - \theta_2) & \frac{n_e}{n^2} \theta_2 (1 - \theta_2 - \theta_1) \\ \frac{n_e}{n^2} \theta_1 (1 - \theta_1 - \theta_2) & \frac{n_e}{n^2} \theta_2 (1 - \theta_2 - \theta_1) & \frac{1}{n} (\theta_1 + \theta_2 - (\theta_1 + \theta_2)^2) \end{pmatrix}$$

From the large sample theory, it follows that for large n ,

$$\bar{\mathbf{X}} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \end{pmatrix} \sim N_3(\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \mathbf{\Sigma}) \text{ approximately.}$$

So, by an application of the delta method, for large n ,

$$f(\bar{\mathbf{X}}) = \frac{\bar{X}_1 + \bar{X}_2}{\bar{X}_3} = PCR^c \sim N(f(\boldsymbol{\mu}) = \frac{\mu_1 + \mu_2}{\mu_3}, G' \mathbf{\Sigma} G) \text{ approximately, where}$$

$$\mathbf{G}' = \left(\frac{\partial f}{\partial \mu_1}, \frac{\partial f}{\partial \mu_2}, \frac{\partial f}{\partial \mu_3} \right) = \left(\frac{1}{\mu_3}, \frac{1}{\mu_3}, -\frac{(\mu_1 + \mu_2)}{\mu_3^2} \right)$$

It can be shown that

$$\begin{aligned}
G'\Sigma G &= \frac{1}{\mu_3^2} [Var(\bar{X}_1) + Var(\bar{X}_2) + \frac{(\mu_1 + \mu_2)^2 Var(\bar{X}_3)}{\mu_3^2} + \frac{2(Cov(\bar{X}_1, \bar{X}_3) + Cov(\bar{X}_2, \bar{X}_3))(\mu_1 + \mu_2)}{\mu_3}] \\
&= \frac{1}{n(\theta_1 + \theta_2)^2} \left[\frac{n_e \theta_1 (1 - \theta_1)}{n} + \frac{n_{\bar{e}} \theta_2 (1 - \theta_2)}{n} + \frac{(n_e \theta_1 / n + n_{\bar{e}} \theta_2 / n)^2 (\theta_1 + \theta_2 - (\theta_1 + \theta_2)^2)}{(\theta_1 + \theta_2)^2} \right. \\
&\quad \left. + 2 \left(\frac{n_e \theta_1 (1 - \theta_1 - \theta_2)}{n} + \frac{n_{\bar{e}} \theta_2 (1 - \theta_2 - \theta_1)}{n} \right) \frac{(n_e \theta_1 / n + n_{\bar{e}} \theta_2 / n)}{\theta_1 + \theta_2} \right] = Var(\text{PCR}^c)
\end{aligned}$$

An estimator of this variance, $Var(\hat{\text{PCR}}^c)$ can be obtained by estimating the probabilities θ_1 and θ_2 by the corresponding sample proportions. Then an approximate large sample 95% confidence interval for the true proportion of correct reclassification at cutoff c is given by

$$\text{PCR}^c \pm 1.96 \sqrt{Var(\hat{\text{PCR}}^c)}.$$

References

1. Parmigiani G, Berry D, Aguilar O. Determining carrier probabilities for breast cancer- susceptibility genes BRCA1 and BRCA2. *American Journal of Human Genetics*. 1998; 62:145–158. [PubMed: 9443863]
2. Chen S, Wang W, Lee S, Nafa K, Lee J, Romans K, Watson P, Gruber SB, Euhus D, Kinzler KW, et al. Prediction of germline mutations and cancer risk in the Lynch syndrome. *Journal of the American Medical Association*. 2006; 296:1479–1487. [PubMed: 17003396]
3. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*. 2008; 117:743–753. [PubMed: 18212285]
4. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*. 1989; 81:1879–1886. [PubMed: 2593165]
5. Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PW, D'Agostino RB, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. *New England Journal of Medicine*. 2008; 359:2208–2219. [PubMed: 19020323]
6. Katki HA. Incorporating medical interventions into Mendelian mutation prediction models. *BMC Medical Genetics*. 2007; 8:13. [PubMed: 17378937]
7. Katki HA, Blackford A, Chen S, Parmigiani G. Multiple diseases in carrier probability estimation: accounting for surviving all cancers other than breast and ovary in BRCAPRO. *Statistics in Medicine* Sep. 2008; 27:4532–4548.
8. Tai YC, Chen S, Parmigiani G, Klein AP. Incorporating tumor immunohistochemical markers in BRCA1 and BRCA2 carrier prediction. *Breast Cancer Research*. 2008; 10:401. [PubMed: 18466632]
9. Chen S, Blackford AL, Parmigiani G. Tailoring BRCAPRO to Asian-Americans. *Journal of Clinical Oncology*. 2009 Feb. 27:642–643. [PubMed: 19075251]
10. Biswas S, Tankhiwale N, Blackford A, Barrera AM, Ready K, Lu K, Amos CI, Parmigiani G, Arun B. Assessing the added value of breast tumor markers in genetic risk prediction model BRCAPRO. *Breast Cancer Research and Treatment*. 2012; 133:347–355. [PubMed: 22270937]
11. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007; 115:928–935. [PubMed: 17309939]
12. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*. 2008; 27:157–172. [PubMed: 17569110]

13. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine*. 2011; 30:11–21. [PubMed: 21204120]
14. Berry DA, Iversen ES, Gudbjartsson DF, Hiller EH, Garber JE, Peshkin BN, Lerman C, Watson P, Lynch HT, Hilsenbeck SG, et al. BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. *Journal of Clinical Oncology*. 2002; 20:2701–2712. [PubMed: 12039933]
15. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Annals of Internal Medicine*. 2006; 145:21–29. [PubMed: 16818925]
16. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Annals of Internal Medicine*. 2009; 150:795–802. [PubMed: 19487714]
17. Cook NR, Paynter NP. Performance of reclassification statistics in comparing risk prediction models. *Biometrical Journal*. 2011; 53:237–258. [PubMed: 21294152]
18. Iversen ES, Parmigiani G, Chen S. Multiple model evaluation absent the gold standard through model combination. *Journal of the American Statistical Association*. 2008; 103:897–909.
19. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. *American Journal of Epidemiology*. 2011 Jun; 173(11):1327–1335.
20. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Statistics in Medicine*. 2013 Apr.

Figure 1.

Proportion of correct reclassification (PCR^c) comparing the “ER/PR, Her-2” model to the “ER/PR” model. The figures below each point are the total number of probands reclassified (n_R^c) at that cutoff. The gray vertical segments are 95% pointwise binomial CIs.

Figure 2.

Proportion of correct reclassification (PCR^c) comparing the “ER/PR, Her-2” model to the “ER/PR” model plotted for subgroups of probands with specific combinations of ER/PR and Her-2/neu status. The figures below each point are the total number of probands reclassified (n_R^c) at that cutoff. When no reclassification occurs at a given cutoff, no point is plotted.

Figure 3.

Power vs. Type I error in simulations: The parameters for various scenarios are listed in Table 4. The title of each plot corresponds to the scenario for power, while the type I error is for Scenario 9 in all plots. The light gray line is at type I error of 0.05. IDI: Integrated Discrimination Improvement; PCR: Proportion of Correct Reclassification.

Figure 4.

Power vs. Type I error in simulations: The parameters for various scenarios are listed in Table 4. The title of each plot corresponds to the scenario for power, while the type I error is for Scenario 10 in all plots. The light gray line is at type I error of 0.05. IDI: Integrated Discrimination Improvement; PCR: Proportion of Correct Reclassification.

Figure 5.

A sample of p 's and q 's generated under scenario 5. The orange and green pairs of lines (vertical and horizontal) are to facilitate seeing the total number of observations reclassified (n_R^c) at cutoffs of 0.2 and 0.7. Each pair of lines divides all points in four quadrants. The points in the second and fourth quadrant are the ones reclassified differently by the two models at that cutoff.

Table 1

Tabular representation of calculation of PCR^c . Only the cells marked by “x” are used in this calculation. The total number of subjects in these four cells is n_R^c (denominator of PCR^c). The two “x” in bold correspond to correct reclassification by the new model. The numbers of subjects in these two cells are the two terms of the numerator of PCR^c .

	<u>Events</u>			<u>Non-Events</u>		
	$q_i < c$	q_i	c	$q_i < c$	q_i	c
$p_i < c$			x			x
q_i	c	x		x		

Table 2

IDIr for comparing the two BRCAPRO models — “ER/PR, Her-2” (new) and “ER/PR” (old). The response variable is y , the transformed difference of carrier probabilities between the two models conditional on the carrier status, as described in the text. The baseline for both ER/PR and Her-2 is the positive status.

	Estimate	Standard Error	t-value	p-value
intercept	0.031	0.009	3.509	0.0005
ER/PR(−)	0.078	0.014	5.722	< 0.0001
Her2(−)	−0.033	0.010	−3.397	0.0007
ER/PR(−)*Her2(−)	−0.074	0.015	−4.888	< 0.0001

Table 3

Simulation scenario 0: E : event status, X_1, X_2 : binary variables, (p, q) : probabilities from old and new models being compared. (p, q) are generated jointly on the logit scale from bivariate normal (BVN) distributions. The five parameters of a BVN distribution are the two means, two standard deviations, and correlation coefficient. The choice of parameters for Bernoulli and BVN distributions are discussed in the text.

Group	X_1	X_2	E	$(\text{logit}(p), \text{logit}(q))$
1	Negative	Negative	Bernoulli(0.58)	BVN(−0.16, 0.32, 0.5, 0.5, 0.95)
2	Negative	Positive	Bernoulli(0.08)	BVN(−0.16, −2.44, 0.5, 0.5, 0.95)
3	Positive	Negative	Bernoulli(0.26)	BVN(−1.10, −1.05, 0.5, 0.5, 0.95)
4	Positive	Positive	Bernoulli(0.21)	BVN(−1.10, −1.32, 0.5, 0.5, 0.95)

Table 4

Simulation results: Scenario 0 corresponds to the scenario in Table 3. The other scenarios differ in the means and/or correlation coefficient r for the BVN distribution. The pair of numbers listed under each Group column are the means of the BVN distribution for generating (p, q) on the logit scale. The last three columns list the proportions of times a 95% CI for each statistic do not contain the corresponding null value. The pair of means in the last six scenarios are the same in each group, and hence those are the “null” scenarios.

Scenario	Means of $(\text{logit}(p), \text{logit}(q))$				r	IDI*	cNRI*	PCR*
	Group 1	Group 2	Group 3	Group 4				
0 (Ideal)	-0.16, 0.32	-0.16, -2.44	-1.10, -1.05	-1.10, -1.32	0.95	1.00	0.982	0.994
1	-1.00, -1.00	-1.00, -2.44,	-1.10, -1.05	-1.10, -1.32	0.95	0.570	0.156	0.828
2	-1.00, -1.00	-0.50, -2.44	-1.05, -1.05	-1.32, -1.32	0.95	0.748	0.124	0.966
3	-1.00, -1.00	-0.50, -2.44	-1.05, -1.05	-1.32, -1.32	0.70	0.334	0.116	0.584
4	0.32, 0.32	-0.50, -2.44	-1.05, -1.05	-1.32, -1.32	0.95	0.724	0.122	0.912
5	-0.16, 0.32	-2.44, -2.44	-1.10, -1.05	-1.10, -1.32	0.95	0.992	0.916	0.282
6	0.32, 0.32	-0.16, -2.44	-1.05, -1.05	-1.32, -1.32	0.95	0.838	0.150	0.974
7	0.32, 0.32	-0.50, -2.44	-1.05, -1.05	-1.32, -1.32	0.70	0.284	0.122	0.574
8	-1.00, -1.00	-2.44, -2.44	-1.05, -1.05	-1.32, -1.32	0.95	0.046	0.046	0.062
9	-1.00, -1.00	-0.50, -0.50	-1.05, -1.05	-1.32, -1.32	0.95	0.046	0.048	0.048
10	-1.00, -1.00	-0.50, -0.50	-1.05, -1.05	-1.32, -1.32	0.70	0.034	0.054	0.048
11	0.32, 0.32	-2.44, -2.44	-1.05, -1.05	-1.32, -1.32	0.95	0.044	0.046	0.046
12	0.32, 0.32	-0.50, -0.50	-1.10, -1.10	-1.10, -1.10	0.95	0.036	0.042	0.042
13	0.32, 0.32	-2.44, -2.44	-1.05, -1.05	-1.32, -1.32	0.70	0.048	0.036	0.054

* IDI: Integrated Discrimination Improvement; cNRI: continuous Net Reclassification Improvement; PCR: Proportion of Correct Reclassification