Published in final edited form as:

Stat Med. 2017 December 10; 36(28): 4468–4481. doi:10.1002/sim.6862.

Simpson's Paradox in the Integrated Discrimination Improvement

J. Chipman^{a,*} and D. Braun^{b,c}

^a Department of Biostatistics, Vanderbilt School of Medicine, Nashville, TN 37203, U.S.A.

^b Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.

^c Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215, U.S.A.

Abstract

The Integrated Discrimination Improvement (IDI) is commonly used to compare two risk prediction models; it summarizes the extent a new model increases risk in events and decreases risk in non-events. The IDI averages risks across events and non-events and is therefore susceptible to Simpson's Paradox. In some settings, adding a predictive covariate to a well calibrated model results in an overall negative (positive) IDI. However, if stratified by that same covariate, the strata-specific IDIs are positive (negative). Meanwhile, the calibration (observed to expected ratio, O/E, and Hosmer-Lemeshow Goodness of Fit Test, GOF), area under the receiver-operating-characteristic curve (AUC), and Brier score improve overall and by stratum. We ran extensive simulations to investigate the impact of an imbalanced covariate upon metrics (IDI, AUC, Brier score, and R^2), provide an analytic explanation for the paradox in the IDI, and use an investigative metric, a Weighted IDI, to better understand the paradox. In simulations, all instances of the paradox occurred under stratum-specific mis-calibration, yet there were mis-calibrated settings in which the paradox did not occur. The paradox is illustrated on Cancer Genomics Network data by calculating predictions based on two versions of BRCAPro, a Mendelian risk prediction model for breast and ovarian cancer. In both simulations and the CGN data, overall model calibration did not guarantee stratum-level calibration. We conclude that the IDI should only assess model performance among a clinically relevant subset when stratum-level calibration is strictly met and recommend calculating additional metrics to confirm the direction and conclusions of the IDI.

Keywords

Integrated Discrimination Improvement; Simpson's Paradox; Risk Prediction; Reclassification; BRCAPro

* Correspondence to: Vanderbilt School of Medicine, 2525 West End Avenue, Suite 11000, Nashville, TN 37203. jonathan.chipman@vanderbilt.edu.

1. Introduction

Risk prediction models have tremendous potential to translate bench research to clinical practice. To an individual, they convey one's risk of diagnosis or prognosis of a disease, and to the public health community, they may be informative for enacting important public health policy. With such potential for immediate impact, the metrics to assess and compare model performance must be clinically meaningful. Traditional metrics used to compare models include metrics which assess the ability to discriminate between events and non-events, metrics which assess model calibration, and metrics which assess the overall accuracy of the models. To these measures of performance, another novel set of metrics, known as decision analytic methods, allow for weighing the benefits of identifying true versus false positives of a given model [1, 2].

Much of the literature on these metrics in recent years is in genetic epidemiology, where there is interest in evaluating new biomarkers that are added to existing risk prediction models. The traditional discrimination metric is the area under the receiver operating curve (AUC aka c-statistic) [3]. It has been shown that this metric is insensitive to change despite adding a marker with substantive effect sizes [4, 5, 6]. To address this limitation, novel reclassification methods including the Net Reclassification Index (NRI), Continuous NRI (cNRI), and Integrated Discrimination Index (IDI) have been proposed, of which the NRI and IDI are now commonly used [7, 8, 9, 10]. For diseases with clinically established risk categories (defined by risk thresholds), the NRI summarizes the proportion of events (non-events) that based on the new model predictions reclassify with a greater (lower) risk category. In the lack of predefined risk categories, the cNRI considers any increase (decrease) in event (non-event) risk as an improved reclassification. The IDI, which is the focus of this paper, summarizes the extent to which a new model increases the average risk in events and decreases the average risk in non-events.

This work is motivated by an observation of Simpson's Paradox in the IDI metric when comparing two versions of the BRCAPRO risk prediction model. BRCAPRO is a Mendelian risk prediction model, which uses Mendelian laws of inheritance to calculate the probability that an individual is a BRCA1 and/or BRCA2 (BRCA1/2) carrier conditioned on family history and known mutation penetrance (the probability of having a disease at a certain age given the person's genotype) and prevalence [11]. The model identifies individuals who are at high risk for breast and ovarian cancer. Predictive markers, interventions, and other new covariates may be added to the model such that the only risks modified are of those having the new covariate. The latest version, 2.0-8, updates the previous carrier and non-carrier penetrance estimates for Contralateral Breast Cancer (CBC) [12]. As a result, the probability of being a BRCA1/2 carrier based on the updated CBC penetrance decreases in families with CBC. In these families, the BRCAPRO predecessor, 2.0-7, overestimated risk. We observed Simpson's Paradox in comparing the performance of BRCAPRO 2.0-8 with 2.0-7 on a set of families from the Cancer Genomics Network (CGN).

Decision analytic methods incorporate calibration, whereas reclassification methods (such as the IDI) depend on calibration. Yate's Slopes, which comprise the IDI, are functions of calibration-in-the-large (average estimated risk equals average true risk) and -small

(estimated risk equals true risk for all individuals) [13]. Conclusions for both types of methods, may even be harmful (in the wrong direction) when the model is poorly calibrated [14, 15]. Although there are approaches to correct mis-calibration, a model may remain poorly calibrated for example when it is used in a new population. This is a known limitation of these metrics [7, 9].

However, even in the setting where the model is well calibrated, such is the case in the motivating example, it has been shown that one of the metrics, cNRI, can lead to conclusions contradictory to the IDI, NRI, and AUC. This paradoxical cNRI result may be explained by the observation that the cNRI is more a measure of covariate effect than overall model discrimination, and an imbalance of the covariate amongst events and non-events may lead to the contradictory results [16]. In addition, Van Calster observed that reporting the NRI may contradict results compared to reporting the two NRI components (sensitivity and specificity) separately [17]. In the spectrum of a metric's ability to principally measure overall model discrimination (AUC) versus marginal covariate effect (cNRI), the IDI has been described to lie somewhere between [18].

In this paper we investigate instances where Simpson's Paradox – where the conclusion of the overall metric contradicts the conclusions when stratified by a key covariate [19] – occurs in the IDI and whether it simultaneously occurs in other metrics, including: AUC, Brier score, and R^2 . The simulations are designed to ensure the paradox is not a result of poorly calibrated models. We first derive how introducing a discrete covariate to a new model may cause the paradox and introduce a Weighted IDI as a means to investigate Simpson's Paradox. The simulations consider two modeling approaches. First, "Mis-Calibrated" fixes coefficients for the new and old model and therefore by design allows the old model to be mis-calibrated. Second, "Calibrated" fits the new and old model coefficients on the data and therefore by design allows both models to be well calibrated. The first approach is similar to comparing two established risk prediction models, and evaluating these models on a separate dataset. The second approach considers using a dataset to fit both the old and new models, and model evaluation is performed on the same dataset (this is common in the biomarker literature [20, 21]). We end with the data application and final remarks.

2. Simpson's Paradox and the Weighted Integrated Discrimination Improvement

2.1. Notations and Definition

We assume a setting in which one compares an old and new risk prediction model for $i = 1, \dots, n$ individuals. In addition, we assume that the only difference between the old and new risk prediction models is due to a covariate of interest (for example, an additional marker, or in the setting of BRCA PRO CBC status) that was added to the risk prediction model. In this work we assume this covariate, cov , is binary.

Let p_{newi} denote the risk of the event of interest for an individual i under the new model, and p_{oldi} denote the risk of the event of interest for an individual i under the old model, where

$\bar{p}_{new}$ is the average new model risk prediction for the n individuals, and $\bar{p}_{old}$ is the average old model risk predictions for the n individuals. In addition, we introduce a subscript indicating whether an individual i has the event of interest (1 if they have the event of interest, and 0 otherwise). The IDI is the extent to which the new model increases the average risk in events and decreases the average risk in non-events, which can be written as

$$IDI = \left(\bar{p}_{new.1} - \bar{p}_{old.1} \right) - \left(\bar{p}_{new.0} - \bar{p}_{old.0} \right) \quad (1)$$

and estimated as

$$\widehat{IDI} = \left(\hat{\bar{p}}_{new.1} - \hat{\bar{p}}_{old.1} \right) - \left(\hat{\bar{p}}_{new.0} - \hat{\bar{p}}_{old.0} \right). \quad (2)$$

An $IDI = 0$ indicates the new model provides no added discrimination, and an $IDI > 0$ indicates improvement in model discrimination [7].

2.2. The Paradox

To this, a new subscript is introduced where 1 indicates presence of the new covariate and 0 indicates lack of the covariate. For example, $\bar{p}_{new.10}$ is the average new model risk predictions for those who have an event but do not have the covariate. The IDI in Equation (1) can be rewritten in terms of individuals with and without the covariate

$$IDI = \frac{n_{10} \cdot \left(\bar{p}_{new.10} - \bar{p}_{old.10} \right) + n_{11} \cdot \left(\bar{p}_{new.11} - \bar{p}_{old.11} \right)}{n_{10} + n_{11}} - \frac{n_{00} \cdot \left(\bar{p}_{new.00} - \bar{p}_{old.00} \right) + n_{01} \cdot \left(\bar{p}_{new.01} - \bar{p}_{old.01} \right)}{n_{00} + n_{01}} \quad (3)$$

where n_{10} and n_{11} are the number of individuals who have an event without and with the covariate respectively, and n_{00} and n_{01} are the number of individuals who do not have an event without and with the covariate respectively.

The stratum-specific IDI for those with the covariate is

$$IDI_{cov=1} = \left(\bar{p}_{new.11} - \bar{p}_{old.11} \right) - \left(\bar{p}_{new.01} - \bar{p}_{old.01} \right). \quad (4)$$

The stratum-specific IDI for those without the covariate is

$$IDI_{cov=0} = \left(\bar{p}_{new.10} - \bar{p}_{old.10} \right) - \left(\bar{p}_{new.00} - \bar{p}_{old.00} \right). \quad (5)$$

Simpson's Paradox occurs in the IDI when the overall IDI in Equation (3) is negative (positive), yet the stratum-specific IDI's in Equations (4-5) are positive (negative). For simplicity, if we assume risks only change amongst those with the covariate, then we can see the paradox occurs based on the relative weighting of the covariate in events and non-events (as would be expected for Simpson's Paradox). Under this assumption, the $IDI_{cov=0}$ is zero, and the overall IDI reduces to a re-weighted estimate of $IDI_{cov=1}$ as

$$IDI = \frac{n_{11} \cdot (\bar{p}_{new.11} - \bar{p}_{old.11})}{n_{10} + n_{11}} - \frac{n_{01} \cdot (\bar{p}_{new.01} - \bar{p}_{old.01})}{n_{00} + n_{01}}. \quad (6)$$

For a given distribution of risks in events and non-events, certain values of $\pi_{11} = n_{11}/(n_{10} + n_{11})$ and $\pi_{01} = n_{01}/(n_{00} + n_{01})$ will yield a negative (positive) IDI (Equation (6)) despite a positive (negative) $IDI_{cov=1}$ (Equation (4)). Therefore, for a given distribution of risks, the proportion of individuals who have the covariate in events and non-events drives the paradox. The impact of the imbalanced covariates will vary depending upon the distribution of risks in the old and new model.

2.3. Weighted IDI

To clarify the problem of Simpson's Paradox in the IDI and better understand the source of the problem, we introduce a Weighted IDI as a weighted sum across each IDI_{cov} rather than calculating the overall IDI directly. To preserve the IDI bounds, the k stratified IDI s are averaged by strata size, $w_k = n_k/n$, where n_k is the number of individuals in stratum k . For example, when the covariate has only $k = 2$ levels, the Weighted IDI is

$$IDI_{weighted} = \sum_{k=0}^1 w_k \cdot IDI_{cov=k} \quad (7)$$

which is estimated as

$$\widehat{IDI}_{weighted} = \sum_{k=0}^1 w_k \cdot \widehat{IDI}_{cov=k} \quad (8)$$

where $\widehat{IDI}_{cov=k}$ is the estimated stratum-specific IDI.

3. Simulation Study

3.1. Data Generation

Our simulation study investigates whether Simpson's Paradox persists in the IDI under various settings and assesses the susceptibility of the AUC, Brier score, R^2 , and the introduced Weighted IDI to Simpson's Paradox.

Let Y indicate a binary event of interest, where $Y = 1$ indicates an event and $Y = 0$ indicates a non-event. Assume one is interested in evaluating model performance of a new model compared to an old model, where the difference between the new and old model is the addition of a new binary covariate X_2 . Let X_1 indicate a covariate that is present in both the old and new models. We parametrize the old and new models as:

$$\begin{aligned} \text{Old Model: } \text{logit}(P(Y=1|X_1)) &= \beta_0 + \beta_1 X_1 \\ \text{New Model: } \text{logit}(P(Y=1|X_1, X_2)) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 \end{aligned} \quad (9)$$

We begin by generating X_2 as a binary variable from a Bernoulli distribution with probability π_c . Based on X_2 we generate X_1 ; $X_1 | X_2 = 0 \sim N(\mu_1, \sigma_1)$ and $X_1 | X_2 = 1 \sim N(\mu_2, \sigma_2)$. We consider various parameters of $\beta = (\beta_0, \beta_1, \beta_2)$, μ_1 , σ_1 , μ_2 , σ_2 , and π_c (provided in Table 1), and generate Y from a Bernoulli distribution under the new model for a given set of parameters.

3.2. Model Comparisons

With the data generated, we consider two approaches for estimating risk predictions based on the old model, $\hat{p}_{oldi}$, and new model, $\hat{p}_{newi}$. The first approach, "Mis-Calibrated", assumes both the old and new model coefficients (Equation (9)) are determined by the fixed β parameters used to generate the data. By design, the new model is well calibrated, whereas the old model is not. In approach "Calibrated", coefficients for both the old and new models are estimated by fitting a *logistic* model to the data generated. By design, both the new and old models are well calibrated. Each approach is replicated 1000 times with 2000 observations in each replicate so that our results would not be confounded by small-sample bias.

The data generation mechanism for approach "Mis-Calibrated" is meant to mimic the setting of the head-to-head comparison of the BRCAPRO models which were a motivation for this work. In such settings, risk may be updated for only a subset of the population with the covariate. Therefore, the new model would have the same coefficients as the old plus a new fixed beta coefficient (β_2) to modify risks for those with the covariate. This means that for simulation approach "Mis-Calibrated", the same β_0 and β_1 are used in the old and new models. Risk predictions were calculated for the generated data based on these two models.

Approach "Calibrated" is meant to mimic other settings in the risk prediction literature, in which one is looking to add a new marker or covariate to an existing model. In these settings model coefficients for both the old and new models are estimated based on the data, and model evaluation is performed on the same data used to estimate the coefficients [20, 21]. Therefore, the data generated was used to fit the two risk prediction models and then risk predictions were calculated for this data based on the two model fits.

Models were compared using the following metrics; IDI, AUC, Brier score, R^2 , and the introduced Weighted IDI (defined in Equation (7)). IDI was calculated based on Equation (1). AUC was calculated using the *ROCR* R package [22]. Brier score was calculated using the following definition $(1/n) \sum_{i=1}^n \{\hat{p}_{oldi} - 1(Y_i=1)\}^2$ (and similarly for the new model).

R^2 was calculated as Nagelkerke R^2 , $(1 - \exp(-LR/n)) / (1 - \exp(-2L_0/n))$, where LR is the likelihood ratio of the new to old model and L_0 is the likelihood of the old model. In addition, we evaluate model calibration by using the following two metrics; observed to expected ratio (O/E) for calibration-in-the-large and the Hosmer-Lemeshow Goodness of Fit Test by decile (GOF, α set to 0.05) for a test of calibration-in-the-small. O/E was calculated

using the following definition $\frac{\sum_{i=1}^n 1(Y_i=1)}{\sum_{i=1}^n \hat{p}_{oldi}}$ (and similarly for the new model). The GOF Test by decile was obtained using the *ResourceSelection* R package.

3.3. Simulation Results

Figure 1 shows results for the model comparisons under approach "Mis-Calibrated" for the 32 simulation settings considered by the various parameter combinations presented in Table 1. Each panel in the figure shows the median, along with 95% confidence intervals using 2.5 and 97.5 percentiles, for a different model comparison metric. For each metric and simulation setting, we calculate both the overall metric for the entire population, as well as the metric for each of the two X_2 stratum. Results for strata-specific estimates are shown in green ($X_2=0$ stratum) and yellow ($X_2=1$ stratum). For the overall population, black represents scenarios in which the overall metric estimate is within the two stratum-specific estimates; the paradox does not occur in these scenarios. Red represents scenarios in which the overall metric estimate is negative (positive) yet the strata-specific estimates are positive (negative), these are situations in which the paradox occurs. In other words, we considered Simpson's Paradox to occur whenever the median of the overall metric (such as the IDI) would draw a conclusion contradictory to conclusions made by the median of each stratum's metric. Blue represents scenarios in which both the overall metric estimate and stratum-specific estimates are either positive (negative), yet the overall metric estimate is more extreme than the stratum-specific metrics. The paradox does not occur in these scenarios because the overall and strata-specific conclusions are consistent. Results under approach "Calibrated" are shown in Figure 2.

By design, the number of events were generated to be the same as expected (calibration-in-the-large) under the new model for both approaches "Mis-Calibrated" and "Calibrated". The median O/E is one for the new model in all simulation settings under both approaches. The refitting of coefficients in approach "Calibrated", given the number of simulated events Y , results in perfectly calibrated models as verified in Figure 3 (calibration-in-the-large) and Figure A.1 (calibration-in-the-small, see appendix). Since approach "Mis-Calibrated" does not refit coefficients, the O/E of the old model deviates from one (Figure 3) and calibration-in-the-small is not guaranteed (Figure A.1).

Overall calibration does not imply stratum-level calibration (Figures 3 and A.1). For those with the covariate, $X_2=1$ stratum, both approaches are well calibrated-in-the-large and -small for the new model. In contrast, both approaches are mis-calibrated-in-the-large and -small for the old model. For those without the covariate, $X_2=0$ stratum, both approaches are well-calibrated-in-the-large and -small for the new model. Also, since approach "Mis-Calibrated" does not change predictions for these individuals, the old model is well-calibrated-in-the-large and -small. However, in about half the simulation settings, the old model is mis-calibrated-in-the-small under approach "Calibrated".

First, we examine the consistency for each metric of the overall and stratum-specific estimates. In both approaches "Mis-Calibrated" and "Calibrated", the IDI is susceptible to Simpson's Paradox (Figures 1 and 2, top panel in red). Under approach "Calibrated" model coefficients are re-estimated when a new covariate is introduced, making it is less susceptible to the paradox than approach "Mis-Calibrated". In all cases where Simpson's Paradox occurred, the stratum-IDI was estimated under the presence of either mis-calibration-in-the-large and/or mis-calibration-in-the-small (Figures 3 and A.1). However, mis-calibration does not imply Simpson's paradox occurs; there are scenarios under both approaches where mis-calibration is present yet Simpson's paradox does not occur (Figures 1 and 2). The AUC, Brier score, and R^2 did not result in Simpson's Paradox under either approach "Mis-Calibrated" or "Calibrated".

In addition to Simpson's Paradox, certain other scenarios yielded an overall IDI consistent in conclusion with the stratum-specific IDIs yet more extreme in magnitude (Figures 1 and 2, top panel in blue). Again, this occurred for both model comparison approaches. The only other metric where this similarly occurred was the AUC. However, this result is not surprising since, given the way the data was generated, the ranking of event risks does not change within the covariate strata. The Weighted IDI provides estimates that are bounded by the stratum-specific IDIs across all simulation scenarios for both approaches (Figures 1 and 2, second panel from the top).

Next, we examine the consistency of the IDI compared to the other metrics considered. Although the Weighted IDI under both approach "Mis-Calibrated" and approach "Calibrated" provides results that are consistent with the strata-specific IDI results, it is important to also examine the consistency of these results with the other metrics. Under approach "Mis-Calibrated", the overall IDI was negative for half of the simulation settings and therefore inconsistent with the model improvement conclusion suggested by the AUC, Brier score, and R^2 . For some settings, the Weighted IDI corrected for the inconsistency between the IDI conclusions and the other metrics considered, while in other settings it did not correct for the inconsistency even though it did provide results consistent with the strata-specific IDIs (Figure 1).

On the other hand, the overall IDI was positive under all settings in approach "Calibrated" (even in settings where the overall IDI was inconsistent with the strata-specific IDIs) and hence consistent with AUC, Brier score, and R^2 conclusions. The Weighted IDI provided results consistent with the strata-specific IDIs, implying it was negative for half of the simulation settings and therefore inconsistent in conclusion with other metrics (Figure 2).

Thus, in approach "Mis-Calibrated", the overall IDI (even in settings where it was consistent with the strata-specific IDIs) and Weighted IDI reflect the same limitation of the cNRI – that the overall conclusion may be inconsistent with conclusions of other metrics. This inconsistency may be explained by imperfectly calibrated old model under approach "Mis-Calibrated". Yet, in one of these simulation scenarios (Betas(3), Norms(1), Proportion Covariate(1)), the old model has an O/E of 0.94 (95% CI: 0.90, 0.98, Figure 3) and passed GOF test in 100% of simulations (Figure A.1). In approach "Calibrated", while the Weighted IDI faces a similar limitation, it is consistent with the strata-specific IDIs.

From these simulation settings, it may be tempting to draw conclusions regarding which model component (beta coefficients, distribution of observed values, or proportion with the covariate) has the biggest impact in driving Simpson's Paradox; however, there is likely a large interplay between each of these components. Holding all else constant, we perform additional simulations to assess the impact of covariate balance upon overall conclusion. Doing so allows us to change the model calibration.

Using the same data generation strategy, we generate a super-population of 100K observations under one set of parameters from Table 1 (Betas(5), Norms(2), and Proportion Covariate(1)). Assuming a constant number of events (roughly 13%), we sample a total of 2000 individuals according to various π_{11} (of those who have an event, the proportion of individuals who have the covariate) and π_{01} (of those who do not have an event, the proportion of individuals who have the covariate), ranging from 0.1 to 0.9 in increments of 0.1. We calculate risk predictions for these various populations based on both approach "Mis-Calibrated" and approach "Calibrated", with 1000 replicates for each π_{11} and π_{01} combination.

Under approach "Mis-Calibrated", π_{11} and π_{01} dramatically affect calibration (Figures 4 and A.2) and the conclusions of the overall IDI (Figure 4). Since risk models are fixed, the impact of π_{11} and π_{01} is linear and there are settings in which Simpson's Paradox occurs (red), as well as settings where the overall estimates are more extreme than stratum-specific estimates (blue). The Weighted IDI provides conclusions consistent with stratum-specific IDI conclusions. However, by design the new model has better discrimination, which was contradicted by the negative Weighted IDI.

The impact of π_{11} and π_{01} under approach "Calibrated" is not linear. Overall, the new and old models remain calibrated, whereas π_{11} and π_{01} affect stratum-specific calibration (Figures 5 and A.2). Figure 5 highlights that for this approach, the sampling of the covariate in events and non-events can yield an overall IDI which reflects Simpson's Paradox (red) or overall estimates which are more extreme than stratum-specific estimates (blue). Once again, the Weighted IDI provides conclusions consistent with stratum-specific IDI conclusions, yet these conclusions were sometimes inconsistent with the improved discrimination under the new model.

4. CGN Data

The BayesMendel Lab created and continuously improves BRCAPRO, a Mendelian risk prediction model for calculating the probability of being a BRCA1/2 carrier conditioned on family history of cancer and known mutation penetrance and prevalence [11, 23, 24, 25, 26]. Risks are estimated as an application of Bayes' Theorem, and are expressed generally as:

$$P[\text{BRCA1/2 carrier} | \text{family history of cancer}] = \frac{P[\text{BRCA1/2 carrier}] P[\text{family history of cancer} | \text{BRCA1/2 carrier}]}{P[\text{family history of cancer}]} \quad (10)$$

The latest version, 2.0-8, updates the estimated Contralateral Breast Cancer (CBC) penetrances of carriers and non-carriers [12]. As a result, risk of being a BRCA1/2 carrier decreases in families with CBC. In these families, the BRCA1/2 predecessor, 2.0-7, overestimated risk [12, 27].

We calculate the risk using the two versions of the model (2.0-7 and 2.0-8) for 2,038 families with personal and family history of cancer obtained from CGN. Among the 2,038 families, 322 families had CBC incidence. There were 428 BRCA1/2 carriers, 95 of which were from families with CBC incidence. Family CBC incidence is highly associated with being a BRCA1/2 carrier (p -value < 0.001). Calibration (O/E), AUC, and Brier scores improved in 2.0-8 among these families which are described elsewhere [12, 27] (Table 2). Both 2.0-7 and 2.0-8 pass the Hosmer-Lemeshow Goodness of Fit Test by decile (p -values: 0.42 for 2.0-7, and > 0.99 for 2.0-8), though when stratified both models only pass GOF in the strata where CBC was not present in families (p -values: 0.97 for non CBC families for both versions 2.0-7 and 2.0-8; and < 0.001 for CBC families version 2.0-7 and 0.018 for CBC families version 2.0-8).

The $\widehat{IDI}$ of -0.007 (95% BCa CI: $-0.016, 0.001$) suggests that 2.0-8 performs worse than 2.0-7, while the AUC, O/E, and Brier scores all suggest 2.0-8 outperforms 2.0-7 (Table 2). When stratifying by CBC, the $\widehat{IDI}$ among families with CBC incidence is 0.021 (95% BCa CI: $-0.015, 0.052$). Though the lower-bound falls below zero, the estimate suggests a trend of model improvement in these families. In the remaining families with no CBC incidence the estimate is zero, which reflects that the risk only changes among families with CBC. The suggested overall and stratum-specific conclusions reflect Simpson's Paradox. However, the $IDI_{weighted}$ (defined in Equation (7)) is 0.003 (95% BCa CI: $-0.002, 0.008$) (Table 2), which is consistent with the conclusions of $\widehat{IDI}$ when stratified by family CBC incidence. In this case, the suggested Weighted IDI conclusion of model improvement is also consistent with conclusions suggested by the AUC and Brier scores.

5. Remarks

The IDI provides a novel alternative to the AUC in measuring model discrimination. It has greater sensitivity to reflect model improvement than the AUC in instances where a model already has a moderate to strong AUC. The work presented here provides the observation, that with an added imbalanced covariate the IDI is susceptible to Simpson's Paradox; whereas, other discrimination metrics do not appear to be susceptible.

This note of awareness comes from simulations and a data application where the additional covariate is imbalanced between events and non-events. Whereas the AUC, Brier score, and R^2 reflected summary measures and conclusions consistent with stratum-specific measures, the IDI was in some simulated scenarios inconsistent with stratum-specific measures and conclusions (Simpson's Paradox). This was observed both when comparing two models with fixed coefficients (similar to a head-to-head comparison of two separately established models) and when comparing two models with coefficients estimated based on the data. In some simulations and the CGN data, the overall IDI conclusion was not only inconsistent with the stratum-specific IDIs, but also inconsistent with AUC, Brier score, and R^2 .

conclusions regarding model improvement. At the same time, there were also simulation settings in which the overall IDI conclusions were inconsistent with the stratum-specific IDIs, yet consistent with the AUC, Brier score, and R^2 .

Model calibration is an important assumption for the IDI. The simulations where the IDI was inconsistent with other metrics occurred when comparing two models with fixed coefficients (approach "Mis-Calibrated"). Under this approach, the new models are calibrated-in-the-large and -small, yet the old models are mis-calibrated-in-the-large and -small. This inconsistency between the IDI (under mis-calibration) with other discrimination metrics is consistent with other recent literature[15]. Mis-calibration was an artifact of our data generation mechanism, where the goal was to generate models similar to the motivating BRCAPRO example. Under these simulations, model calibration may be an explanation for the inconsistent IDI conclusions. We note, though, that this inconsistency occurred even with a reasonably well calibrated old model such as BRCAPRO version 2.0-7 in the CGN data and in a simulation setting (O/E = 0.94, 95% CI: 0.90, 0.98; passed GOF test in 100% of simulations).

Likely multiple model components impact whether Simpson's Paradox occurs in the IDI or whether the overall estimate is beyond the stratum-specific estimates. As shown in simulations, calibration has a clear role as all instances of Simpson's Paradox occur under stratum-level mis-calibration. However, strata-specific mis-calibration alone does not necessarily lead to the paradox. Other influences may include the magnitude of effect for the new covariate and/or the distribution and effect of other covariates in the model. Still, holding all else constant, the amount of covariate imbalance may yield results that reflect Simpson's Paradox or that reflect an IDI of greater magnitude than stratum-specific IDIs. The Weighted IDI used in this paper helps illustrate that a proper re-weighting of the IDI across covariates may bring the IDI to be consistent with stratum-specific conclusions. Before considering imbalanced covariates, the IDI implicitly implies an equal weighting of mean risk change in events as in non-events. This is true even when there are relatively few events compared to non-events. When an imbalanced covariate is introduced to a model, the equal weighting of change in events and non-events may further result in an overall IDI outside the bounds of the stratum-specific bounds.

In introducing the Weighted IDI in a purely investigative fashion, we do not propose it as a new method for adjusting the IDI to covariates nor to explore covariates that may improve the IDI [28]. A limitation of the Weighted IDI, seen in the simulations, is that despite removing Simpson's Paradox between overall and stratum-specific conclusions, it may provide an overall conclusion inconsistent with other metrics. This is also a limitation of the IDI, where in some scenarios, even when the paradox does not occur, results are inconsistent with other metrics. We defined the Weighted IDI to extend beyond two strata; however, our simulation is limited in scope in that we only considered adding a dichotomous covariate. And, though there is in literature a general approach for handling discrimination metrics given censoring [29], our study only considers binary outcomes.

Highly imbalanced covariates are not uncommon, for example, in Prostate Cancer the TMPRSS2:ERG gene fusion is found to only occur in men with Prostate Cancer [30]. In

general, an observational study with a sample bias may artificially increase or decrease the covariate balance across events and non-events. In the CGN data, we saw imbalance of family CBC incidence. Given the results presented in this work, one must be careful when drawing conclusions based only on the IDI. We recommend using the IDI only when calibration-in-the-large and -small are strictly met. Overall model calibration does not guarantee stratum-level calibration (as seen in simulations and the CGN data), and the IDI should only be used to assess model performance among a clinically relevant subset when stratum-level calibration is strictly met. In addition when using the IDI, we recommend calculating additional metrics to confirm the direction and conclusions of the IDI.

Acknowledgements

We gratefully acknowledge support from the National Cancer Institute at the National Institutes of Health [1R21CA177233-01 to Parmigiani] for work by Jonathan Chipman and [5T32CA009337-32 to Parmigiani] [P01 CA134294 to Lin] for work by Danielle Braun. Also, Giovanni Parmigiani, Su-Chun Cheng, Malka Gorfine, the rest of the BayesMendel team, and the anonymous submission reviewers provided astute feedback which allowed us to further explore and present our work with greater clarity. We thank them.

A. Calibration-in-the-small under simulation settings and in simulated super-population

We assess calibration-in-the small for the simulations described in Section 3.

Figure A.1.

Calibration-in-the-small for simulation settings described in Table 1, presented as the proportion of times the new and old model fail the Hosmer-Lemeshow Goodness of Fit (GOF) Test for the two approaches; "Mis-Calibrated" and "Calibrated". Overall calibration-in-the-small does not imply stratum-level calibration-in-the-small. By design, predictions do not change when the covariate is absent under approach "Mis-Calibrated", so calibration-in-the-small is the same under the new and old model.

Figure A.2.

Calibration-in-the-small for super-population with the simulation setting of Betas (5), Norms (2), and Prop Cov (1) (Table 1) with varying proportions of individuals with $X_2 = 1$ who have events (π_{11}) and those with $X_2 = 1$ who do not have events (π_{01}). Results shown for the two simulation approaches; "Mis-Calibrated" and "Calibrated". Overall calibration-in-the-small does not imply stratum-level calibration-in-the-small. By design, predictions do not change when the covariate is absent under approach "Mis-Calibrated", so calibration-in-the-small is the same under the new and old model.

References

1. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*. Nov; 2006 26(6):565–574. [PubMed: 17099194]
2. Pencina MJ, D'Agostino Sr RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine*. Nov; 2010 30(1):11–21. [PubMed: 21204120]

3. Harrell FEJ, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA*. May; 1982 247(18):2543–2546. [PubMed: 7069920]
4. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American journal of epidemiology*. May; 2004 159(9):882–890. [PubMed: 15105181]
5. Greenland P, O'Malley PG. When is a new prediction marker useful? A consideration of lipoprotein-associated phospholipase A2 and C-reactive protein for stroke risk. *Archives of internal medicine*. Nov; 2005 165(21):2454–2456. [PubMed: 16314539]
6. Ware JH. The limitations of risk factors as prognostic tools. *The New England journal of medicine*. Dec; 2006 355(25):2615–2617. [PubMed: 17182986]
7. 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*. 2007; 27(2):157–172.
8. Pencina KM, Pencina MJ, D'Agostino Sr RB. What to expect from net reclassification improvement with three categories. *Statistics in Medicine*. Aug; 2014 33(28):4975–4987. [PubMed: 25176621]
9. Leening MJG, Vedder MM, Witteman JCM, Pencina MJ, Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Annals of internal medicine*. Jan; 2014 160(2):122–131. [PubMed: 24592497]
10. Tzoulaki I, Liberopoulos G, Ioannidis JPA. Use of reclassification for assessment of improved prediction: an empirical evaluation. *International journal of epidemiology*. Aug; 2011 40(4):1094–1105. [PubMed: 21325392]
11. Parmigiani G, Berry D, Aguilar O. Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *American journal of human genetics*. Jan; 1998 62(1):145–158. [PubMed: 9443863]
12. Mazzola E, Chipman J, Cheng SC, Parmigiani G. Recent BRCAPRO upgrades significantly improve calibration. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*. Aug; 2014 23(8):1689–1695.
13. Yates JF. External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*. 1982; 30(1):132–156.
14. Van Calster B, Vickers AJ. Calibration of Risk Prediction Models: Impact on Decision-Analytic Performance. *Medical Decision Making*. Aug.2014 :0272 989X14547 233.
15. 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*. Apr; 2013 33(19):3405–3414. [PubMed: 23553436]
16. Cook NR. Clinically relevant measures of fit? A note of caution. *American journal of epidemiology*. Sep; 2012 176(6):488–491. [PubMed: 22875759]
17. Van Calster B, Steyerberg EW, D'Agostino RB, Pencina MJ. Sensitivity and Specificity Can Change in Opposite Directions When New Predictive Markers Are Added to Risk Models. *Medical Decision Making*. Apr; 2014 34(4):513–522. [PubMed: 24378915]
18. Pencina MJ, D'Agostino RB, Demler OV, Janssens ACJW, Greenland P, Pencina, et al. Respond to “The Incremental Value of New Markers” and “Clinically Relevant Measures? A Note of Caution”. *American journal of epidemiology*. Sep; 2012 176(6):492–494. [PubMed: 22875757]
19. Simpson EH. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1951; 13:238–241.
20. Xanthakis V, Sullivan LM, Vasan RS, Benjamin EJ, Massaro JM, D'Agostino Sr RB, Pencina MJ. Assessing the incremental predictive performance of novel biomarkers over standard predictors. *Statistics in Medicine*. Apr; 2014 33(15):2577–2584. [PubMed: 24719270]
21. Leening MJG, Steyerberg EW, Van Calster B, D'Agostino RB, Pencina MJ. Net reclassification improvement and integrated discrimination improvement require calibrated models: relevance from a marker and model perspective. *Statistics in Medicine*. Aug; 2014 33(19):3415–3418. [PubMed: 25042215]
22. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCr: visualizing classifier performance in R. *Bioinformatics (Oxford, England)*. Oct; 2005 21(20):3940–3941.

23. Katki HA. Incorporating medical interventions into carrier probability estimation for genetic counseling. *BMC medical genetics*. 2007; 8(1):13. [PubMed: 17378937]
24. Tai YC, Chen S, Parmigiani G, Klein AP. Incorporating tumor immunohistochemical markers in BRCA1 and BRCA2 carrier prediction. *Breast cancer research : BCR*. 2008; 10(2):401. [PubMed: 18466632]
25. 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(22):4532–4548. [PubMed: 18407567]
26. Chen S, Blackford AL, Parmigiani G. Tailoring BRCAPRO to Asian-Americans. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. Feb; 2009 27(4): 642–3. author reply 643–4. [PubMed: 19075251]
27. Parmigiani G, Chen S, Iversen ESJ, Friebe TM, Finkelstein DM, Anton-Culver H, Ziogas A, Weber BL, Eisen A, Malone KE, et al. Validity of models for predicting BRCA1 and BRCA2 mutations. *Annals of internal medicine*. Oct; 2007 147(7):441–450. [PubMed: 17909205]
28. Biswas S, Arun B, Parmigiani G. Reclassification of predictions for uncovering subgroup specific improvement. *Statistics in Medicine*. May; 2014 33(11):1914–1927. [PubMed: 24353130]
29. Uno H, Tian L, Cai T, Kohane IS, Wei LJ. A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. *Statistics in Medicine*. Oct; 2012 32(14):2430–2442. [PubMed: 23037800]
30. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science (New York, N.Y.)*. Oct; 2005 310(5748):644–648.

Figure 1.

The median IDI, Weighted IDI, Change in AUC, Change in Brier score, and R^2 along with 95% Confidence Intervals under various simulation settings. Under approach "Mis-Calibrated", the new and old model coefficients are fixed. The stratum-specific metrics (circles) are green (for the $X_2=0$ stratum) and yellow (for the $X_2=1$ stratum). The overall metric (squares) are red when Simpson's Paradox occurs relative to the stratum-specific metrics, blue when the overall conclusion is consistent with the stratum-specific conclusions

yet more extreme in magnitude, and black when the overall conclusion is consistent with the stratum-specific conclusions.

yet more extreme in magnitude, and black when the overall conclusion is consistent with the stratum-specific conclusions.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 3.

Calibration-in-the-large presented as 95% Confidence Intervals of the observed-over-expected ratio (O/E): Under approach "Mis-Calibrated", O/E are 1 by design under the new model, but vary under the old model. Under approach "Calibrated", O/E are 1 by design under both the old and new models –the estimates for the new model appear hidden behind the old model estimates.

Figure 4.

2000 observations are sampled from a super-population with the simulation setting of Betas (5), Norms (2), and Prop Cov (1) (Table 1) with varying proportions of individuals with $X_2 = 1$ who have events (π_{11}) and those with $X_2 = 1$ who do not have events (π_{01}). Risk was calculated under approach "Mis-Calibrated". These proportions drive whether Simpson's Paradox occurs in the IDI, the paradox does not occur in the Weighted IDI.

Figure 5.

2000 observations are sampled from a super-population with the simulation setting of Betas (5), Norms (2), and Prop Cov (1) (Table 1) with varying proportions of individuals with $X_2 = 1$ who have events (π_{11}) and those with $X_2 = 1$ who do not have events (π_{01}). Risk was calculated under approach "Calibrated". These proportions drive whether Simpson's Paradox occurs in the IDI, the paradox does not occur in the Weighted IDI.

Table 1

Simulation parameters.

Betas	β_0	β_1	β_2	Norms	μ_1	σ_1	μ_2	σ_2	Proportion Covariate	π_c
1	-3.50	0.75	-2.00	1	0.00	3.00	3.00	3.00	1	0.10
2	-3.50	0.75	-3.00	2	0.00	2.00	4.00	2.00	2	0.25
3	-3.50	1.50	-2.00							
4	-3.50	1.50	-3.00							
5	-3.50	0.75	2.00							
6	-3.50	0.75	3.00							
7	-3.50	1.50	2.00							
8	-3.50	1.50	3.00							

32 sets of simulations were performed based on these parameters.

Table 2

Comparison of performance (AUC, O/E, Brier score, *IDI*, *IDI_{weighted}*) of BRCAPRO version 2.0-7 and version of 2.0-8 with 95% BCa marginal confidence intervals for 2,038 families in CGN.

	Overall	CBC in Family	No CBC in Family
AUC 2.0-7	0.788 (0.761, 0.812)	0.779 (0.718, 0.831)	0.790 (0.760, 0.817)
AUC 2.0-8	0.793 (0.766, 0.817)	0.779 (0.718, 0.831)	0.790 (0.760, 0.817)
O/E 2.0-7	0.891 (0.825, 0.963)	0.552 (0.469, 0.639)	1.081 (0.986, 1.181)
O/E 2.0-8	0.977 (0.905, 1.054)	0.730 (0.623, 0.845)	1.081 (0.986, 1.181)
Brier score 2.0-7	0.149 (0.136, 0.161)	0.250 (0.217, 0.286)	0.130 (0.117, 0.143)
Brier score 2.0-8	0.140 (0.128, 0.153)	0.198 (0.168, 0.232)	0.130 (0.117, 0.143)
<i>IDI</i>	-0.007 (-0.016, 0.001)	0.021 (-0.015, 0.052)	0
<i>IDI_{weighted}</i>	0.003 (-0.002, 0.008)	-	-

322 of the families had CBC incidence, 1,716 families had no CBC incidence. Simpson's Paradox in the IDI is highlighted in bold. The overall IDI conclusion (new model is worse) is inconsistent with conclusions of AUC and Brier score, yet the Weighted IDI is consistent with the other metrics. Risks do not change amongst families without CBC incidence, and hence neither do the metrics. R^2 is not calculated for this data.