Published in final edited form as:

Stat Med. 2017 December 10; 36(28): 4503–4508. doi:10.1002/sim.7341.

First things first: Risk model performance metrics should reflect the clinical application

Kathleen F. Kerr and

Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98115, voice: 206-543-2501, fax: 206-543-3286

Holly Janes

Fred Hutchinson Cancer Research Center, Vaccine and Infectious Disease and Public Health Sciences Divisions, 1100 Fairview Ave N M2 C200, Seattle WA 98109, voice: 206-667-6353, fax: 206-667-4378

Abstract

Developing new measures of risk model performance is an active line of research, often motivated by the conventional wisdom that AUC is an “insensitive” measure of the additional predictive capacity offered by new biomarkers. Without endorsing AUC, we argue that this charge is not substantiated. Three articles in this issue discuss alternative metrics of risk model performance: $\text{NRI}(p)$ (two-category net reclassification index at the event rate), IDI (integrated discrimination index), and R-squared statistics. Guided by the principle that performance metrics should match the intended use of a risk prediction model, we argue that routine use of these indices is not justified. Instead, we recommend decision-theoretic measures to evaluate risk prediction models for applications in which clinically-relevant risk thresholds have been established for classifying individuals. In the absence of established risk thresholds, additional research is needed to develop suitable metrics.

Keywords

risk prediction; biomarkers; incremental value; net benefit; relative utility

Three articles in this issue address metrics for evaluating risk models [1–3]. We offer our perspective on key aspects of the articles.

1. Routine Use of $\text{NRI}(p)$ is Not Justified—Event Rate is Generally Not An Appropriate Risk Threshold

NRI statistics, which come in various flavors, are used to evaluate the incremental value of new biomarkers for predicting individuals’ risks of a clinical event. All NRI statistics are based on contrasting predicted risks for risk models with and without the new biomarker(s). NRI statistics are widespread in applications despite multiple authors highlighting issues:

they are prone to misinterpretation, they do not summarize the clinical impact of either risk model, and their statistical properties are such that they can give misleading results [4–8].

Pencina, Steyerberg, and D’Agostino [1] examine one particular member of the NRI family, $\text{NRI}(p)$, which is the 2-category NRI statistic that uses the rate of the clinical event, p , as the threshold for declaring an individual to be “high risk.” The authors show that $\text{NRI}(p)$ does not suffer from some of the problems that befall other NRI statistics, and therefore recommend $\text{NRI}(p)$ as a generally useful statistic for quantifying biomarker incremental value. We disagree. If the event rate p is not an appropriate threshold to identify individuals who should be considered as high risk, then any metric that uses p as a risk threshold is also inappropriate, including $\text{NRI}(p)$. Although Pencina et al mention that the “threshold at the event rate p may not be clinically relevant,” we think this fundamentally important point warrants more attention.

Suppose that the purpose of a risk model is to identify “high risk” individuals – with risks above some risk threshold R – for an intervention. We refer to an individual who will go on to have the clinical event as a “case.” Intuitively, if the expected benefit of the intervention for a case is very large relative to the burden of the intervention for a non-case, then a relatively lower risk threshold is appropriate to recommend intervention. Similarly, if the burden of the intervention for a non-case is substantial relative to the benefit for a case, a relatively higher risk threshold should be used because it becomes more important to avoid intervention for non-cases. A result from classic decision theory formalizes this intuition: if B is the expected benefit of intervention for a case and C is the expected cost of unnecessary

intervention for a non-case, then the optimal risk threshold is $R = \frac{C}{C+B}$ [9–11].

If investigators choose the event rate p as the high risk threshold, they implicitly assume that

the cost-benefit ratio is $\frac{C}{B} = \frac{p}{1-p}$ [10]. For example, if the event rate is 5%, then setting $R = p$ implies that the expected benefit of intervention for a case is 19 times greater than the cost of intervention for a non-case. Again, whether or not this is reasonable depends on the clinical context and the intervention’s efficacy and side-effects, which are quite distinct from

the event rate. The relation $\frac{C}{B} = \frac{p}{1-p}$ might be true or at least reasonable in some applications, but only coincidentally. We see no reason to suppose this relation will be true generally, or even frequently. We give an example (Box 1) that shows that adopting $\text{NRI}(p)$ as the performance metric might lead to incorrect conclusions.

When there is a clinically appropriate high risk threshold R there are two fundamental quantities of interest in assessing a risk model: the sensitivity and specificity of the risk model with respect to R ($\text{sens}(R)$ and $\text{spec}(R)$). We think investigators want to know both of these numbers. For a single number summary of a risk model’s performance, we favor Relative Utility (RU), also known as standardized Net Benefit (sNB) [10, 12, 13]:

$$RU = sNB = sens(R) - \frac{1-p}{p} \frac{R}{1-R} (1 - spec(R)) \quad (*)$$

RU varies between 0 and 1 and can be interpreted as the proportion of population Net Benefit attained by the risk model relative to the highest possible Net Benefit, corresponding to a model with perfect sensitivity and specificity. The risk threshold R appears in this expression because it is assumed that it has been chosen rationally, i.e., R satisfies

$\frac{R}{1-R} = \frac{C}{B}$. The expression (*) simplifies in the special case where the costs and benefits of intervention are such that the appropriate risk threshold is $R = p$,

$$RU = sNB = sens(p) + spec(p) - 1,$$

and the difference in relative utility between two risk models is then

$$\Delta RU = \Delta sNB = \Delta sens(p) + \Delta spec(p) = NRI(p).$$

While $NRI(p) = RU$ when $R = p$, we emphasize that this decision-theoretic interpretation of $NRI(p)$ is only valid when the relative magnitudes of costs and benefits justify using p as the high risk threshold. The relation does not justify use of $NRI(p)$ in general.

2. $NRI(p)$ Is Not a Generally Useful Summary of a Relative Utility Curve

Pencina et al [1] note that p corresponds to the highest point on the plot of Relative Utility (RU) over a range of risk thresholds. This observation gives $NRI(p)$ an interpretation as the difference in maximum relative utilities for two risk models [1]. This would be a useful result if $NRI(p)$ gave an upper bound for the difference in relative utilities for two models. If that were true, then when $NRI(p)$ is small we could conclude that the difference in relative utility for two risk models is small regardless of the exact appropriate risk threshold. Unfortunately, $NRI(p)$ does not provide such a bound, as Pencina et al point out, since $NRI(p) = \max(RU) - \max(RU)$ – for an example, see Figure 4 in [1]. Therefore, $NRI(p)$ does not tell us about the difference in relative utility between two risk models except in the special situation where p is the appropriate risk threshold. We therefore disagree with the general recommendation in [1] to use $NRI(p)$ as a summary of a plot of relative utility across risk thresholds. Instead, we posit that the best summary of a relative utility plot is its value at the clinically appropriate risk threshold, which must be determined by considering the costs and benefits of the intervention and not from the plot itself [13].

3. Adding Uninformative Biomarkers to Risk Models Diminishes Performance; Performance Measures Should Reflect This

We re-visit the interpretation of the simulation study, which Pencina et al [1] suggest supports use of $NRI(p)$. They perform simulations similar to [14] where risk models with

and without four biomarkers are fit to a development sample of size 1000. Under the simulation model, the four biomarkers are unrelated to the outcome and have nothing to contribute to prediction. Independent data are used to evaluate the difference in performance of fitted models with and without the four biomarkers. Three performance metrics are considered: the change in area under the ROC curve or AUC (Δ AUC), category-free NRI, and NRI(p). The authors highlight the fact that Δ AUC is negative on average. This is fully appropriate: on average, a model containing the useless biomarkers is *worse* than a model that does not, because the useless biomarkers simply add noise. The estimated regression coefficient for a useless biomarker is 0 on average but never exactly 0 in any particular fitted model. Therefore, the Δ AUC measure correctly captures that models with the useless biomarkers are, on average, worse than the models without them. Pencina et al incorrectly call this “negative bias.” In fact, what is concerning is that the category-free NRI has positive expected value [7]. That is, on average the category-free NRI suggests that the model containing four useless biomarkers has superior performance, which is known to be false by construction.

4. The IDI, Simpson’s Paradox, and Risk Model Calibration

In a related article, Chipman and Braun [2] show that another measure for contrasting risk models – the integrated discrimination index or IDI – can exhibit Simpson’s Paradox: the IDI can have consistent sign (positive or negative) in all strata of a population defined by some covariate but have the opposite sign in the population considered as a whole. We find this neither surprising nor alarming. What we find most useful is the authors’ reminder of the importance of evaluating calibration for risk models in general and, in particular, the IDI’s sensitivity to calibration and its potential to mislead when models are poorly calibrated.[5]

5. Is the AUC Really Insensitive to Improvements in Model Performance?

It is often stated that the Δ AUC is an “insensitive” metric of the incremental value of a biomarker, a point that is reiterated in [2] to motivate interest in the IDI. We think this claim warrants further scrutiny. In our experience, the proclaimed insensitivity of Δ AUC is most often cited in the following circumstances.

1. Investigators observe that a biomarker has a strong marginal association with the outcome. Yet adding the biomarker to an existing risk model produces a small increase in AUC.
2. Investigators observe that a biomarker is highly statistically significant when added to an existing risk model. That is, the p-value is very small for the biomarker’s regression coefficient in a model containing established predictors. Yet adding the biomarker to the existing risk model produces a small increase in AUC.
3. For nested risk models with and without some promising new biomarker(s), a test based on contrasting the AUCs for the two models is not statistically significant.

None of these situations actually justify the claim that AUC is an insensitive measure of a biomarker's incremental value [8]. The first situation is based on the intuitive – but incorrect – expectation that a biomarker with large marginal predictive capacity should also have large incremental value. In fact, a biomarker can have positive marginal predictive capacity and have 0 incremental value, or, at the other extreme, no marginal predictive capacity and positive incremental value (as measured by *any* metric)[15, 16]. In general, incremental value is not a monotone function of a biomarker's marginal predictive capacity. There can be two biomarkers M1 and M2 where M1 is superior to M2 in individual predictive capacity yet M2 has higher incremental value than M1. This last fact is perhaps the more interesting and important paradox to highlight. Observing small increments in AUC for a biomarker with impressive marginal predictive capacity does not indict the measure.

The second situation is an instance of conflating statistical significance and clinical or practical significance. A p-value summarizes the evidence against a null hypothesis; it does not measure effect size. Therefore, strong evidence that a biomarker has positive incremental value is not the same as evidence that a biomarker has large incremental value.

Unfortunately, it seems to be a common experience that biomarkers that show significant evidence of an association with an outcome accounting for existing predictors turn out *not* to contribute meaningfully to prediction.[17]

The third situation reflects the common practice of testing the null hypothesis that $AUC=0$ after fitting nested risk models to a dataset (where the larger model includes the new biomarker(s)). The test of $AUC=0$ that contrasts the estimated AUCs for the two risk models has been shown to be highly conservative when the same data are used to fit the models and perform the test [18–20], and the null hypothesis is typically not rejected. The test suffers from poor power even when independent data are used to perform the test [19]. Fortunately, it has been shown that the test based on AUC is unnecessary because it tests the same null hypothesis as the test that the coefficient(s) of the new biomarker(s) in the regression model are equal to zero [19]. The likelihood ratio test for the regression coefficients is the uniformly most powerful test; alternative tests are not needed. The poor operating characteristics of the test based on contrasting AUCs does not indict the AUC measure itself, only the test.

6. AUC and R-squared Metrics Do Not Assess the Clinical Impact of Using a Risk Model

While we disagree with characterizing AUC as an “insensitive” measure, there are other important reasons to seek alternatives to AUC. We wholeheartedly endorse the principle stated by Pencina, Fine, and D’Agostino [3] that that “it is sensible to choose a metric that best corresponds to the way in which a given model will be applied.” AUC is a measure of a risk model's capacity to discriminate between cases and non-cases but AUC but does not summarize the clinical impact of using the risk model. Therefore, AUC does not honor this principle. Unfortunately, to the best of our knowledge neither do any of the R-squared metrics explored in [3].

7. Discussion

There seems to be consensus on the need to move away from AUC to other metrics, but the field is struggling to identify those metrics. In applications where there are established risk categories, we think it is both appropriate and desirable to incorporate those categories into assessments of risk model performance. A good starting place is to examine how risk models place cases and non-cases into the established risk categories.[21–23] We further recommend evaluating the Net Benefit to the population of using the risk model, or the risk model's Relative Utility, along with the constituents of these metrics (i.e., sensitivities and specificities). When subpopulations have different costs and benefits, and therefore different appropriate risk thresholds, Net Benefit should be examined stratified by those subpopulations [13].

For settings without established risk categories, more research is needed on methods for evaluating risk model performance, including compelling metrics. When the purpose of a risk model is to guide intervention, statisticians should engage their collaborators in discussions about reasonable risk thresholds. If there is no consensus on the best risk threshold for intervention, the performance of a risk model can be assessed over a range of plausible thresholds. Decision curves [10] and Relative Utility curves [12] are graphical instruments that achieve this. An added benefit of these discussions is that they encourage thought and deliberation about the clinical application of risk prediction and the goals for incorporating a risk model into clinical or public health practice.

Acknowledgments

This work was supported by funding from the National Institutes of Health R01HL085757 to KFK and R01CA152089 to HJ.

We thank the authors of the articles we have discussed for their attention to these important issues. We look forward to continued advances in this area in coming years.

References

1. Pencina MJ, Steyerberg EW, D'Agostino RB. Net reclassification index at event rate: properties and relationships. *Stat Med*. 2016
2. Chipman J, Braun D. Simpson's paradox in the integrated discrimination improvement. *Statistics in Medicine*. 2016
3. Pencina MJ, Fine JP, D'Agostino RB. Discrimination slope and integrated discrimination improvement – properties, relationships and impact of calibration. *Stat Med*. 2016
4. Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net Reclassification Indices for Evaluating Risk Prediction Instruments A Critical Review. *Epidemiology*. 2014; 25:114–121. [PubMed: 24240655]
5. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Stat Med*. 2014; 33:3405–3414. [PubMed: 23553436]
6. Hilden J. Commentary: On NRI, IDI, and “good-looking” statistics with nothing underneath. *Epidemiology*. 2014; 25:265–267. [PubMed: 24487208]
7. Pepe MS, Fan J, Feng Z, Gerds T, Hilden J. The Net Reclassification Index (NRI): A Misleading Measure of Prediction Improvement Even with Independent Test Data Sets. *Statistics in Bioscience*. 2015; 7:282–295.

8. Vickers AJ, Pepe M. Does the net reclassification improvement help us evaluate models and markers? *Ann Intern Med.* 2014; 160:136–137. [PubMed: 24592500]
9. Peirce CS. The numerical measure of the success of predictions. *Science.* 1884; 4:453–454.
10. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006; 26:565–574. [PubMed: 17099194]
11. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med.* 1980; 302:1109–1117. [PubMed: 7366635]
12. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. *J R Stat Soc Ser A Stat Soc.* 2009; 172:729–748.
13. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. *J Clin Oncol.* 2016; 34:2534–2540. [PubMed: 27247223]
14. Pepe MS, Janes H, Li CI. Net risk reclassification p values: valid or misleading? *J Natl Cancer Inst.* 2014; 106:dju041. [PubMed: 24681599]
15. Bansal A, Pepe MS. When does combining markers improve classification performance and what are implications for practice? *Stat Med.* 2013; 32:1877–1892. [PubMed: 23348801]
16. Kerr KF, Bansal A, Pepe MS. Further insight into the incremental value of new markers: the interpretation of performance measures and the importance of clinical context. *Am J Epidemiol.* 2012; 176:482–487. [PubMed: 22875756]
17. 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. *Am J Epidemiol.* 2004; 159:882–890. [PubMed: 15105181]
18. Demler OV, Pencina MJ, D'Agostino RB. Misuse of DeLong test to compare AUCs for nested models. *Stat Med.* 2012; 31:2577–2587. [PubMed: 22415937]
19. Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. *Statistics in Medicine.* 2013; 32:1467–1482. [PubMed: 23296397]
20. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. *BMC Med Res Methodol.* 2011; 11:13. [PubMed: 21276237]
21. Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology.* 2014; 25:114–121. [PubMed: 24240655]
22. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. *Am J Epidemiol.* 2011; 173:1327–1335. [PubMed: 21555714]
23. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. *Ann Intern Med.* 2008; 149:751–760. [PubMed: 19017593]

Box 1

Suppose we are working in a field where there is a well-established risk model. Two separate groups have published extensions of the model using new biomarkers, so there are two new models A and B. All three models are assessed on a large, independent dataset. For simplicity, we assume that the validation dataset is large so that the performance of the risk models can be assessed very precisely. This assumption focuses attention on the interpretation of metrics without the distraction of assessing uncertainty.

The investigators calculate $\text{NRI}(p)$ for both new models and find it is 0.07 for both models. They thus conclude that both models A and B offer the same degree of improvement over the old risk model.

	Old Model	New Model A	New Model B
NRI(p)	–	0.07	0.07
sensitivity(p)	0.70	0.75	0.79
1-specificity(p)	0.10	0.08	0.12

The table shows that Model A and B arrive at an NRI(p) value of 0.07 very differently. Compared to the old model, Model A improves both sensitivity and specificity: the sensitivity is 5% higher and specificity is 2% higher. Model B improves the sensitivity by 9%, but specificity is reduced by 2%. Choosing NRI(p) as an appropriate measure of model performance implies that the reduction in specificity with Model B is more than offset by its improvement in sensitivity. This might be true, but more information about the costs and benefits of intervention is needed to make this judgement.

Suppose the event rate p is 1%. Then in 10,000 people from the population, on average 100 are cases. The old model, Model A, and Model B are expected to identify 70, 75, and 79 of these cases, respectively. Among the 9900 non-cases, the old model would identify 990 false positives, Model A would identify 792 false positives, and Model B would identify 1188 false positives. Using NRI(p) to compare Model B and the old model implies that the 9 extra cases Model B identifies over the old model are more than worth the additional 198 false positives. Furthermore, equal NRI(p) for Models A and B implies that the additional 396 false positives arising from Model B compared to Model A are offset by the additional 4 cases Model B detects. Again, whether these tradeoffs are acceptable depends entirely on the costs and benefits of the intervention, but NRI(p) does not incorporate this information. If the intervention is, say, additional imaging that detects curable disease in cases, we can easily imagine these tradeoffs would be deemed acceptable. On the other hand, these tradeoffs would likely not be deemed acceptable for more a more serious intervention such as surgery, particularly if the benefit to cases is limited.