OXFORD

SOT | Society of Toxicology
www.toxsci.oxfordjournals.org

TOXICOLOGICAL SCIENCES, 156(1), 2017, 11–13

doi: 10.1093/toxsci/kfw225
Advance Access Publication Date: November 4, 2016
Forum

FORUM

Net Reclassification Index and Integrated Discrimination Index Are Not Appropriate for Testing Whether a Biomarker Improves Predictive Performance

Peter M. Burch,* Warren E. Glaab,† Daniel J. Holder,† Jonathan A. Phillips,‡ John-Michael Sauer,§ and Elizabeth G. Walker§,[1]

*Pfizer Inc. Worldwide Research & Discovery, Groton, CT 06340; †Merck & Co, Inc, West, Point PA 19486; ‡Boehringer Ingelheim Pharmaceuticals Inc, Ridgefield, CT 06877; and §Predictive Safety Testing Consortium, Critical Path Institute, Tucson, AZ 85718

[1]To whom correspondence should be addressed at Predictive Safety Testing Consortium, Critical Path Institute, 1730 East River Rd., Tucson, AZ 85718. Fax 520.547.3456; E-mail: ewalker@c-path.org.

ABSTRACT

One of the goals of the Critical Path Institute's Predictive Safety Testing Consortium (PSTC) is to promote best practices for evaluating novel markers of drug induced injury. This includes the use of sound statistical methods. For rat studies, these practices have centered around comparing the area under the receiver-operator characteristic curve for each novel injury biomarker to those for the standard markers. In addition, the PSTC has previously used the net reclassification index (NRI) and integrated discrimination index (IDI) to assess the increased certainty provided by each novel injury biomarker when added to the information already provided by the standard markers. Due to their relatively simple interpretations, NRI and IDI have generally been popular measures of predictive performance. However recent literature suggests that significance tests for NRI and IDI can have inflated false positive rates and thus, tests based on these metrics should not be relied upon. Instead, when parametric models are employed to assess the added predictive value of a new marker, following (Pepe, M. S., Kerr, K. F., Longton, G., and Wang, Z. (2013). Testing for improvement in prediction model performance. *Stat. Med.* 32, 1467–1482), the PSTC recommends that likelihood based methods be used for significance testing.

Key words: biomarkers < Safety Evaluation; NRI; IDI; statistics.

MOTIVATION

Two recent article in *Toxicological Sciences*, Burch *et al.* (2015) and Phillips *et al.* (2016), evaluated novel biomarkers to detect drug induced injury. The former article evaluated four blood biomarkers of drug induced skeletal muscle (SKM) injury and the latter evaluated two urinary biomarkers of drug-induced kidney injury (DIKI). Both articles were based on data compiled from multiple rat studies contributed by the Critical Path Institute's Predictive Safety Testing Consortium and used similar statistical methods. These methods included comparing the area under the receiver-operator characteristic curve for each novel injury biomarker to those for the standard markers and assessing the added certainty provided by each biomarker using the net reclassification index (NRI) and integrated discrimination index (IDI) introduced by Pencina *et al.* (2008). Use of the NRI and IDI was motivated by their relatively simple interpretations and wide spread use. However, we have recently been made aware that although estimates of NRI and IDI can be used to describe predictive performance, significance tests for NRI and IDI may have inflated false positive rates (Pepe *et al.*, 2014; Pencina *et al.*, 2015) and thus, should not be relied upon. The purpose of this communication is to promote best practices, by

	NRI and IDI results published			Added likelihood results		
Marker	Fraction improved positive findings	Fraction improved negative findings	Total IDI	Marker coefficient (estimate ± SE)	Likelihood ratio test statistic	Likelihood ratio test *P*-value
CKM	0.828	0.730	0.2063	0.75 ± 0.06	242.62	<1.0E-17
FABP3	0.725	0.775	0.2217	0.91 ± 0.08	213.59	<1.0E-17
MYL3	0.688	0.818	0.2701	0.70 ± 0.06	258.43	<1.0E-17
sTnI	0.706	0.787	0.2030	0.51 ± 0.05	185.40	<1.0E-17

	NRI and IDI results published			Added likelihood results		
Marker	Fraction improved positive findings	Fraction improved negative findings	Total IDI	Marker coefficient (estimate ± SE)	Likelihood ratio test statistic	Likelihood ratio test *P*-value
OPN	0.659	0.756	0.158	0.73 ± 0.10	88.83	<1.0E-17
NGAL	0.735	0.646	0.066	0.61 ± 0.12	33.32	7.8E-09

recommending that significance tests based on NRI and IDI be replaced by a valid test, such as the likelihood ratio test, and provide such test results for the SKM and DIKI injury biomarkers supporting the conclusions in the articles mentioned earlier.

THE LIKELIHOOD RATIO TEST IS THE BEST ALTERNATIVE TO TESTING NRI AND IDI

Pepe *et al.* (2014) notes that a Google Scholar search conducted in 2013 yielded 1810 citations to the original NRI article, demonstrating its tremendous popularity. Since NRI is directly related to the proportion of individuals for which a novel biomarker improves prediction and IDI to the mean improvement, they provide measures of predictive performance that can be easily interpreted. However, in addition to having inflated false positive rates, Pepe *et al.* (2013) has argued that the predictive performance hypothesis these statistics attempt to test is redundant for biomarkers that have already been shown to be a risk factor conditional on the standard biomarkers. In other words, showing that a biomarker is a significant risk factor in an appropriate model that includes the standard biomarkers is sufficient to conclude that the biomarker increases predictive performance. Further, to show that a biomarker is a significant risk factor, Pepe *et al.* (2013) recommends that when parametric models are employed, likelihood-based methods should be used since they have high power and well-performing methods are readily available. Thus, for situations like the SKM and DIKI injury biomarkers we advocate the use of the likelihood ratio to test whether a novel biomarker is a risk factor after controlling for the standard biomarkers. For both the SKM and DIKI analyses, the likelihood ratio test consists of calculating the ratio of the likelihood of the logistic model which contains the standard markers plus the novel marker to the likelihood of the logistic model which contains the standard markers without the novel marker and comparing the ratio to the appropriate reference distribution. The null hypothesis for this test is that the logistic regression coefficient for the novel marker is zero in the model containing it and the standard markers. A significant likelihood ratio test effectively implies that the novel biomarker improves prediction. Use of NRI or IDI, in addition, to describe the improvement is optional, but at least in their present form, significance tests based on these statistics should not be used.

LIKELIHOOD RATIO TESTS FOR SKM AND DIKI MARKERS

Burch *et al.* (2015) showed that the four blood injury biomarkers skeletal troponin I (sTnI), myosin light chain 3 (Myl3), creatine kinase M Isoform (Ckm), and fatty acid binding protein 3 (Fabp3) can help detect SKM injury. NRI and IDI were used to characterize the improvement in predictive performance each biomarker provides when added to a model that contains the standard biomarkers, the enzymatic assays for CK and aspartate transaminase. The table below shows results from the nested linear logistic regression modeling framework described in the article which yielded the estimates of fraction of improved predicted values (the two components of NRI) and total IDI given in the article as well as the likelihood based test results which were not given in the article.

These results are consistent and show that each of these novel biomarkers improves detection of SKM injury when added to the standard biomarkers. Replacement of the NRI and IDI testing results with likelihood ratio test results is an improvement, since as discussed earlier, the likelihood ratio test is a valid test, while the NRI and IDI tests may have inflated false positive rates.

Phillips et al. (2016) assessed the added information to detect renal tubular epithelial degeneration or necrosis when osteopontin (OPN) or neutrophil gelatinase-associated lipocalin (NGAL) is added to a linear logistic regression model that contains the standard biomarkers of blood urea nitrogen and serum creatinine. The table below gives the fraction improved and IDI statistics from the article and adds the results of likelihood based tests for these biomarkers.

These results indicate that OPN and NGAL improve detection of DIKI injury when added to the standard markers. Addition of the likelihood ratio test statistics to this table shows that the improvement added by each marker is statistically significant using a valid test. The NRI and IDI testing results were left out of this table since their inflated false positive rates make them hard to interpret.

SUMMARY

In summary, tests of NRI and IDI have been shown to have inflated false positive rates and thus should not be trusted to test for the improved predictive performance of a novel biomarker. When parametric models are employed to assess the added predictive value of a novel biomarker, such as in Burch *et al.* (2015) and Phillips *et al.* (2016), following Pepe *et al.* (2013) we recommend likelihood based methods for significance testing. We show that all of the novel biomarkers proposed in these two articles have highly statistically significant likelihood ratio tests, supporting the articles' conclusions that each novel injury biomarker improves predictive performance when added to a model which contains the standard biomarkers.

REFERENCES

Burch, P. M., Hall, D. G., Walker, E. G., Bracken, W., Giovanelli, R., Goldstein, R., Higgs, R. E., King, N. M. P., Lane, P., Sauer, J.-M., *et al.* (2015). Evaluation of the Relative performance of drug-induced skeletal muscle injury biomarkers in rats. *Toxicol. Sci.* **150**(1), 247–256.

Pencina, M., D'Agostino, R. B., Sr., D'Agostino, R. B., Jr., and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat. Med.* **27**, 157–172.

Pencina, M. J., Neely, B., and Steyerberg, E. W. (2015). Correspondence RE: Net risk reclassification P values: valid or misleading? *J. Natl. Cancer Inst.* **107**. doi:10.1093/jnci/dju355.

Pepe, M. S., Janes, H., and Li, C. (2014). Net risk reclassification P values: Valid or misleading?. *J. Natl. Cancer Inst.* **106**, dju041.

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

Phillips, J. A., Holder, D. J., Ennulat, D., Gautier, J.-C., Sauer, J.-M., Yang, Y., McDuffie, E., a Sonee, M., Gu, Y.-Z., Troth, S. P., *et al.* (2016). Rat urinary Osteopontin and neutrophil gelatinase-associated lipocalin improve certainty of detecting drug-induced kidney injury. *Toxicol. Sci.* **151**(2), 214–223.