

## Determining Decision Limits for New Biomarkers: Clinical and Statistical Considerations

David A. Morrow<sup>1\*</sup> and Nancy R. Cook<sup>2</sup>

The pathway from discovery to clinical adoption of new biomarkers is marked by a series of significant benchmarks that relatively few cardiovascular biomarkers have successfully navigated (1). For those biomarkers that do emerge as viable candidates for clinical use, determining appropriate decision limits takes on substantial importance. In our experience, this task is demanding and is regularly executed poorly. In this issue of *Clinical Chemistry*, Tang and colleagues assess the prognostic performance of myeloperoxidase (MPO)<sup>3</sup> as a biomarker for stable ischemic heart disease and perform internal validation of a cutpoint (2). This example provides an opportunity to address the clinical and statistical issues that affect the determination of clinical decision limits for new biomarkers.

### Clinical Considerations

Ambiguity concerning decision limits frustrates clinicians and negatively affects the clinical adoption of a biomarker. At the time of clinical introduction, a new biomarker ideally should have well-characterized decision limits that (a) are pragmatic to apply, (b) have undergone validation in multiple studies, (c) have been evaluated in the relevant population(s) and application(s), and (d) have achieved synergy between available scientific data and regulatory labeling.

#### RELEVANCE OF PRAGMATISM

Although it is not always possible in clinical medicine, practitioners crave certainty in interpreting test results and much prefer to think in dichotomous terms. Therefore, a tension sometimes exists between the desire to keep cutpoints simple (so that they are easily remembered and convenient to apply) and the limitations imposed by the complexity of best capturing the

diagnostic, prognostic, or therapeutic information offered by the biomarker. This tension can create controversy in the modeling of a new biomarker as reasonable, often opposing, arguments arise from the clinical and statistical perspectives. We believe that both points of view should be taken into consideration in the evaluation of a new biomarker; however, we assert that a strong motivation exists in clinical medicine for the development of dichotomous or categorical cutpoints to facilitate decision-making.

In some cases, the biology of a biomarker is such that a clear threshold exists in the relationship between its concentration and a diagnosis or outcome. In this situation, a dichotomous cutpoint is preferable from all perspectives, although the absolute value may confer some additional information. For example, because cardiac troponin was not detectable in the blood of healthy individuals until recently, any detectable increase in concentration was indicative of myocardial injury, diagnostic of myocardial infarction in the appropriate setting, and associated with an adverse prognosis (3). More often, however, when a predictor variable is continuous, risk increases throughout the range of the predictor concentrations (e.g., cholesterol). Nevertheless, despite the near-linear relationship, clinical guidelines that use cholesterol are based on specific thresholds rather than on a continuous model (4). Discrete decision limits are needed for clinicians to remember them, to develop guidelines that are possible to transmit and implement, and to create benchmarks for quality of care. Therefore, as for cholesterol, clinicians have embraced the convenience of defined cutpoints for most cardiovascular biomarkers to support decision-making (5) and have sometimes used categorical approaches (e.g., low risk or “rule-out,” intermediate or gray zone, high risk or “rule-in”) to partially account for the graded nature of the prognostic and diagnostic relationships (6).

#### CLINICAL REQUIREMENTS FOR VALIDATION

The aim of clinical pragmatism does not supersede the need for decision limits that adequately reflect the information offered by the new biomarker and that have undergone sufficient exploration and validation. Initial reports typically estimate the risk relationship between a biomarker and outcomes with cutpoints that have been optimized for the data set in which they were de-

<sup>1</sup> TIMI Study Group, Cardiovascular Division, and <sup>2</sup> Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA.

\* Address correspondence to this author at: TIMI Study Group/Cardiovascular Division, Brigham and Women's Hospital, 75 Francis St., Boston, MA 02115. Fax 617-734-7139; e-mail dmmorrow@partners.org.

Received November 1, 2010; accepted November 2, 2010.

Previously published online at DOI: 10.1373/clinchem.2010.155879

<sup>3</sup> Nonstandard abbreviations: MPO, myeloperoxidase; AUC, area under the ROC curve; NRI, net reclassification improvement; IDI, integrated discrimination improvement.

rived. This approach is highly likely to overestimate the strength of the risk relationship, but its use is understandable when no prior data are available to prespecify cutpoints. When subsequent studies perpetuate this limitation by developing new, internally derived cutpoints rather than providing data for the previously proposed cutoffs, the practice becomes problematic from the perspective of developing well-validated decision limits for clinical application.

Robust validation of decision limits for new biomarkers typically requires (a) examination in separate studies, (b) a sample size sufficient to reliably discriminate differences in performance between proposed cutpoints, (c) relevance of the studied population to the clinical application, and (d) evaluation for the specific intended indications. These requirements are intensive with respect to resources. Consequently, this work frequently is completed after the initial clinical introduction. It is not unusual that pertinent influences of patient characteristics on the performance of accepted cutpoints are recognized only after commercial availability of a biomarker test, as exemplified by studies of natriuretic peptides in women and individuals with renal dysfunction (5) and studies of C-reactive protein across race and ethnicity (7). In addition, one cannot assume that a single cutpoint will be useful across a variety of applications (e.g., for both diagnosis and risk stratification) and conditions (e.g., stable and unstable ischemic heart disease). Moreover, the appropriateness of these cutpoints for selecting specific therapies is a separate question. Because the risk relationships may vary, it is necessary to study the biomarker for each of the intended applications and populations to completely assess the proposed decision limits.

## Statistical Considerations

In their study (2), Tang et al. sought to select the best single cutoff value for MPO. The authors chose a cutoff that maximized the area under the ROC curve (AUC). When there is a single dichotomous predictor, this exercise is the same as maximizing the Youden index (sensitivity + specificity - 1), which is considered "optimal" in the sense that it minimizes the overall misclassification rate (8). What is optimal in any situation, however, depends on other factors, such as the relative costs of false-negative and false-positive classifications. The best cutoff to use for treatment decisions should take into account both the prevalence of the disease and the relative importance of sensitivity vs specificity, which can be assessed through alternative-weighting studies (8).

Other possibilities for choosing cutoffs can be based on the shape of the relationship of a predictor with the disease. If the shape exhibits a threshold, then

that point is a natural choice for a cutoff. If the relationship is monotonic, however, choosing a single cutpoint can be arbitrary, and from a statistical perspective it is preferable in such cases to use a continuous measure, either alone or in a comprehensive model, to estimate risk. Risk strata for clinical use can then be established from the predicted risk. Alternatively, thresholds can be determined at the level of the individual patient via net-benefit curves, which display the benefit of a particular treatment strategy across a range of thresholds (9).

## ADJUSTING FOR OPTIMISM

Tang et al. use 5-fold cross-validation to select the best cutpoint for MPO by estimating the AUC in test samples from the study population. Cross-validation is a well-known device for internal validation that can lead to unbiased estimates of performance (10). Instead of evaluating the fit in the test samples, however, the authors used the test data to select the best cutoff based on the AUC. Whenever the best measure is selected, the achieved value is optimistic, or too high in the case of the AUC, because of regression to the mean. If selection is used to find the final model, even if the selection occurs in test sets, a separate data set is needed to obtain an unbiased estimate of fit or of effect. An unbiased estimate of the AUC could alternatively be obtained by calculating the mean over all of the test samples.

## RECLASSIFICATION ANALYSIS

The authors use reclassification to evaluate model improvement with the selected MPO cutoff. Reclassification (11) was initially described for situations with established clinical cutpoints, and the net reclassification improvement (NRI) (12) quantifies the probability that patients are appropriately assigned to categories of higher or lower risk. When no prespecified cutpoints exist, care must be taken in their selection, because the NRI can vary with the cutpoints chosen. The authors use a ratio of 6:3:1 to form categories without clearly justifying their choices of these levels. The alternative measure of calibrating the predicted-to-observed risks within categories (13) varies less with category definition and could indicate how accurate the predicted risks are.

The integrated discrimination improvement (IDI) is a category-free measure that uses the continuous estimates of predicted risk. It is equivalent to a difference in  $R^2$  measures, which are familiar from linear regression (14). Such measures are rarely used for binary or survival outcomes, however, because their values typically tend to be very low. The authors' estimated IDI (10%) is rather large, but it is unclear whether this measure is also overestimated with these data. The random survival forests methodology used in the report to estimate the 3-year risk has, like most data-mining meth-

ods, more potential for overfitting than the usual Cox modeling because it incorporates multiway interactions that fit the observed data much more closely. Therefore, the use of this single cutoff, determined from a single data set without adequate internal validation, must be regarded as needing additional investigation.

### Summary and Recommendations

The availability of well-validated decision limits is vital to optimal integration of a new biomarker into clinical practice. Approaches to internal validation and data-mining methods, such as those used by Tang et al. (2), lead to overfitting and overestimation of risk relationships and are generally not sufficient for selecting final clinical cutpoints. Such methods, when applied correctly, can be reasonable for suggesting cutpoints for external validation. Biomarkers that have monotonic linear relationships with risk are best handled as continuous variables when incorporated into comprehensive risk models. As consistently demonstrated in clinical practice and professional society guidelines, however, practitioners will almost always seek thresholds to provide structure for clinical decision-making, such as those existing for cholesterol. Therefore, such cutpoints warrant development and validation. Although the approach is demanding, we recommend assessment of clinical decision limits by external validation in 2 or more data sets that are appropriate to each of the proposed clinical application(s), with attention paid to the possibility of differences in risk relationships in clinically relevant subpopulations.

**Author Contributions:** *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

**Authors' Disclosures or Potential Conflicts of Interest:** *Upon manuscript submission, all authors completed the Disclosures of Potential Conflict of Interest form. Potential conflicts of interest:*

**Employment or Leadership:** None declared.

**Consultant or Advisory Role:** D.A. Morrow, Beckman Coulter, Instrumentation Laboratories, Ortho Clinical Diagnostics, Roche Diagnostics, and Siemens.

**Stock Ownership:** None declared.

**Honoraria:** None declared.

**Research Funding:** D.A. Morrow, Beckman Coulter, BG Medicine, Ortho Clinical Diagnostics, Roche Diagnostics, Siemens, and Singulex.

**Expert Testimony:** None declared.

**Role of Sponsor:** The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

### References

- Morrow DA, de Lemos JA. Benchmarks for the assessment of novel cardiovascular biomarkers. *Circulation* 2007;115:949–52.
- Tang WHW, Wu Y, Nicholls SJ, Hazen SL. Plasma myeloperoxidase predicts incident cardiovascular risks in stable patients undergoing medical management for coronary artery disease. *Clin Chem* 2010;57:33–9.
- Morrow DA, Antman EM. Evaluation of high-sensitivity assays for cardiac troponin. *Clin Chem* 2009;55:5–8.
- Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive summary of the third report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA* 2001;285:2486–97.
- Morrow DA, Cannon CP, Jesse RL, Newby LK, Ravkilde J, Storrow AB, et al. National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines: clinical characteristics and utilization of biochemical markers in acute coronary syndromes. *Clin Chem* 2007;53:552–74.
- Januzzi JL, van Kimmenade R, Lainchbury J, Bayes-Genis A, Ordonez-Llanos J, Santalo-Bel M, et al. NT-proBNP testing for diagnosis and short-term prognosis in acute destabilized heart failure: an international pooled analysis of 1256 patients: the International Collaborative of NT-proBNP Study. *Eur Heart J* 2005;27:330–7.
- Albert MA, Glynn RJ, Buring J, Ridker PM. C-reactive protein levels among women of various ethnic groups living in the United States (from the Women's Health Study). *Am J Cardiol* 2004;93:1238–42.
- Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am J Epidemiol* 2006;163:670–5.
- Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26:565–74.
- Harrell FE Jr. *Regression modeling strategies*. New York: Springer; 2001.
- Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;115:928–35.
- Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new biomarker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157–72.
- Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Ann Intern Med* 2009;150:795–802.
- Pepe MS, Feng Z, Gu JW. Comments on 'Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., *Statistics in Medicine* (DOI: 10.1002/sim.2929). *Stat Med* 2008;27:173–81.