Beyond the Usual Prediction Accuracy Metrics: Reporting Results for Clinical Decision Making

How well an algorithm or a test predicts or diagnoses disease is not an adequate determination of its clinical usefulness. The relevant question is, are people better or worse off if the test is used as part of clinical care? Stock measures of prediction accuracy, such as sensitivity, specificity, area under the receiver-operating characteristic (ROC) curve, Brier scores, or predicted-versus-observed tests for model calibration, do not measure patient outcomes after a test. Derived measures, such as likelihood ratios, or traditional association measures, such as odds ratios, also do not serve that purpose. Fortunately, a large and growing body of analytic and graphical approaches more closely links test results to outcomes (1). Raji and colleagues, in their report on the Liverpool Lung Project risk model in this issue (2), use several such methods to evaluate strategies based on alternative risk prediction models for the use of computed tomography screening for lung cancer.

Central to their analysis is the "threshold model" for making clinical decisions (3). The risk threshold is the degree of disease risk at which the clinician and patient would decide to proceed with some action; in this case, computed tomography screening. At this risk level, the projected benefits of action exactly equal the harms (that is, zero net benefit). Above that threshold, the benefits of screening (early detection of disease in persons with cancer) exceed the harms (needless exposure to radiation and consequences of false-positive results in persons without cancer). Below the threshold, the harms outweigh the benefits, simply because there are fewer persons with disease who can benefit and more disease-free persons to harm (4).

The methodological innovation embodied in decision curves, which Vickers originally proposed for use in clinical medicine (5) and Raji and colleagues implemented, is the recognition that choosing the risk threshold for a decision inherently includes weighing the benefits and harms. If one uses the same risk threshold in the same population for a given decision, one is implicitly weighing the benefits and the harms in the same way. For example, if we are willing to screen only when the risk for disease exceeds 10%, we are saying that for each person who is appropriately screened (who benefits), one is willing to screen inappropriately (harm) up to 9 people. That 1:9 ratio (the odds corresponding to a probability of 10%) represents the seriousness of the harm of a false-positive result compared with the benefit of a true-positive result, which is implicit in using such a threshold. This relative weight can be used to calculate a measure of net benefit—the proper clinical metric for comparing different rules governing the use of a test.

Another value of using this approach is that it forces us to consider clinically relevant alternative strategies, a focus at the heart of comparative effectiveness. The extreme strategies (screen everyone or screen no one) are always options. A screen-all strategy gives maximum benefit to persons with disease but also does maximum harm to those without, whereas a screen-no one strategy gives no benefit to persons with disease but does no harm to those without. Between these extremes lies the strategy of interest and its competing alternatives. We can compare decision curves among these extreme and intermediate strategies to help us sort out which is optimal under what conditions.

Panel C in Figure 2 of Raji and colleagues' article shows the decision curves for each screening strategy in the Liverpool Lung Project prospective cohort. Why does the screen-all strategy line decline so steeply, and how are we to interpret that finding? An increasing screening risk threshold implicitly indicates increasing harm from false-positive results relative to the benefit of a true-positive result. The screen-all strategy fixes the number of true- and false-positive results and maximizes benefits. But as we judge the false-positive screens to be more harmful, the value of this strategy steadily decreases. Indeed, it enters negative territory (relative to no screening) at just under the 5% threshold. The screen-all strategy would be expected to produce net harm in this population compared with no screening if we judged the harm from false-positive results to be more than 5/95 of the benefit of true-positive results.

However, using the Liverpool Lung Project risk model to determine screening is almost always better than screening no one or everyone. How much better? Reading off the graph, the net benefit at the 5% threshold is about 1.7%, versus −0.2% for screening everyone and zero for screening no one. Thus, the net benefit over screening all is 1.9% [1.7% − (−0.2%)], as reported for the European Early Lung Cancer study in Table 3. This means that, for 100 people assessed with this tool, if we valued a true-positive result worth incurring up to 19 false-positive results, then using this screening tool would generate the equivalent of a net benefit of 1.9 true-positive results over using the screen-all strategy. That benefit comes entirely from reducing false-positive results. The algorithm line converges to the "no screening" line near the 20% mark because no one in the cohort has a predicted probability exceeding 20%. Using such a threshold to trigger screening is therefore equivalent to screening no one.

This calculation does not in itself indicate the proper threshold. That choice requires formal decision analysis (6, 7). But with decision curves, we need not choose a risk threshold that embodies a single tradeoff of benefits and harms. By plotting net benefits across all possible risk

thresholds, the decision curve considers all possible weighting of tradeoffs, allowing us to compare alternative decision strategies—whether governed by risk models, diagnostic tests, or biomarkers—without formal decision-analytic evaluations.

Decision curves can be applied not only to binary outcome studies, such as those on lung cancer risk, but also to those that use time-to-event outcomes or case–control designs. In these other designs, results are commonly reported with only hazard or odds ratios. Neither ratio reflects a patient-centered paradigm that allows patients to make informed decisions (8). Odds ratios that are normally considered quite high (for example, >5) at the population level can translate into poor discriminators at the individual level (9). If such models are being proposed for clinical use, they should be evaluated with more useful metrics at clinically relevant decision thresholds. Fortunately, software for creating decision curves across competing models is readily available (10). In addition, although Raji and colleagues did not use them, plots of net benefits and risk thresholds can include confidence bounds (11). These bounds of uncertainty are especially important when sample sizes are small or outcomes are rare.

Utility curves are an alternative method for assessing the clinical value of different decision tools. As with decision curves, one can compare strategies with relative utilities over a range of risk thresholds without having to quantify harms or benefit. Although examples with detailed calculations appear in the clinical literature (12), no user-friendly programs for standard statistical software are available. However, researchers can produce these curves by combining the software offered by Cronin and Vickers (10) with that by Pepe and colleagues (13) for generating true- and false-positive rates for ROC curves.

When decision tools are proposed for clinical application, analyses should not end with tables of the model regression parameters, ROC curve areas, c-statistics, sensitivity, specificity, and the like. Investigators should translate their findings into more clinically meaningful measures. As Raji and colleagues note, there is no substitute for comprehensive model validation in different settings or for studies of the predictor's use with real clinical outcomes. However, by adding such innovations as decision and relative utility curves to traditional reports of new methods and models for prediction or screening, we can assess the relative net health benefits of various clinical decision strategies with more sophistication.

A. Russell Localio, PhD
University of Pennsylvania
Philadelphia, PA 19104

Steven Goodman, MD, MHS, PhD
Stanford University School of Medicine
Stanford, CA 94305

Potential Conflicts of Interest: Disclosures can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M12-1744.

Requests for Single Reprints: A. Russell Localio, PhD, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, 635 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104; e-mail, rlocalio@mail.med.upenn.edu.

Current author addresses are available at www.annals.org.

Ann Intern Med. 2012;157:294-295.

References

1. **Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al.** Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128-38. [PMID: 20010215]
2. **Raji OY, Duffy SW, Agbaje OF, Baker SG, Christiani DC, Cassidy A, et al.** Predictive accuracy of the Liverpool Lung Project risk model for stratifying patients for computed tomography screening for lung cancer. A case–control and cohort validation study. Ann Intern Med. 2012;157:242-50.
3. **Pauker SG, Kassirer JP.** The threshold approach to clinical decision making. N Engl J Med. 1980;302:1109-17. [PMID: 7366635]
4. **Greenland S.** The need for reorientation toward cost-effective prediction: 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;27:199-206. [PMID: 17729377]
5. **Hunink MG, Glasziou PP, Siegel JE, Weeks JC, Pliskin JS, Elstein AS, et al.** Decision Making in Health and Medicine: Integrating Evidence and Values. Cambridge, UK: Cambridge Univ Pr; 2001.
6. **Sox H, Marton KI, Blatt MA, Higgins MC.** Medical Decision Making. Philadelphia: American Coll Physicians; 2006.
7. **Vickers AJ, Elkin EB.** Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565-74. [PMID: 17099194]
8. **Helfand M, Berg A, Flum D, Gabriel S, Normand SL, eds; PCORI Methodology Committee.** Draft Methodology Report: Our Questions, Our Decisions: Standards for Patient-Centered Outcomes Research. Patient-Centered Outcomes Research Institute; 23 July 2012.
9. **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;159:882-90. [PMID: 15105181]
10. Software for Decision Curve Analysis. Memorial Sloan-Kettering Cancer Center; 2012. Accessed at www.mskcc.org/research/epidemiology-biostatistics/health-outcomes/collaborative/decision-curve-analysis on 16 July 2012.
11. **Vickers AJ, Cronin AM, Elkin EB, Gonen M.** Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak. 2008;8:53. [PMID: 19036144]
12. **Baker SG.** Putting risk prediction in perspective: relative utility curves. J Natl Cancer Inst. 2009;101:1538-42. [PMID: 19843888]
13. **Pepe M, Longton G, Janes H.** Estimation and comparison of receiver operating characteristic curves. Stata J. 2009;9:1. [PMID: 20161343]

Current Author Addresses: Dr. Localio: Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, 635 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104.
Dr. Goodman: Stanford University, 259 Campus Drive, HRP/Redwood Building, Stanford, CA 94305.