Does my biomarker have clinical value? With the advent of -omics techniques, this important question now arises far more frequently in oncology research. In 1988, two statisticians and a gynecologic oncologist (1) collaborated to develop an omnibus answer: calculate the receiver operator curve with and without the biomarker and compare the areas under the curve (AUCs). Assessing the increase in the AUC's statistical significance built on pioneering research by Hoeffding in the late 1940s (2). His statistical breakthrough showed that for a wide class of nonparametric “statistics” the standard Normal distribution was applicable for assessing statistical significance (P values) when the sample size was sufficiently large. DeLong et al. (1) showed that the increase in the AUC belonged to this general class of “statistics,” enabling the powerful theory of U statistics (U for unbiased) to gauge accurately its statistical significance with a simple Normal distribution z test. The appeal of the AUC was its general applicability because it was nonparametric; the power of U statistics was that the derived standard error accounted correctly for the implicit correlation between the two curves (with and without the biomarker) evaluated on the same patients. The oncologist made the vital contribution of defining a clear clinical indication for the biomarker—namely, “when to perform surgical correction of intestinal obstruction in patients known to have ovarian carcinoma.”(1)

However, the generality of the AUC results in a less powerful test than one constructed for a specific intended use. The AUC is the average sensitivity across all specificities from 0% to 100%, which dilutes the signal for a particular clinical question when the range of relevant specificities is narrow. For example, early detection of ovarian cancer in the general postmenopausal population with a blood test requires a biomarker with very high specificity near 98% (3,4). Specificities much lower are not relevant for this clinical application, and AUCs average specificities across all specificities, diluting any signal in sensitivity at 98% specificity. This conservative nature of the increase in AUC due to dilution across irrelevant specificities, I believe, has led to development of alternative, and hopefully more sensitive, methods for assessing biomarker value. One such method is net reclassification (NR) (5), the proportion of case patients where the estimated risk with the biomarker is greater than without it plus the proportion of control subjects where the estimated risk with the biomarker is less than without it. Risk means the probability of having the target disease estimated by a statistical model such as logistic regression. In many circumstances, NR has been appealing because it is statistically significant when the increase in AUC is not.

In this issue of the Journal, Pepe et al. (6) point out NR’s appeal has grown rapidly—understandably if it corrects the conservativeness of the AUC test. An important function of statistical testing is, however, to provide an objective measure of empirical evidence in the presence of uncertainty. If a statistical test of a biomarker is statistically significant more frequently than its stated level (P value) when the biomarker has no connection to disease, it fails one of the fundamental requirements of an objective statistical test. Through a rigorous simulation of biomarkers with no connection to the risk of having disease, Pepe et al. have convincingly demonstrated NR fails this essential requirement and, in this model, yields statistically significant results four times more frequently than expected by chance. Their conclusion that NR has misled biomarker research seems entirely warranted, and their advice to halt its use justified. One may argue that further simulations with greater variety of endpoint frequency, a broader range of the number of null biomarkers, and alternative statistical models besides logistic regression for estimating risk should be conducted before such a conclusion is reached.

However, given the ubiquity of the biomarker question, the size of the failure in this example, and NR’s potential to influence the course of biomarkers in clinical medicine and cancer research, it would seem prudent for the research community to follow the authors’ advice and set aside this misleading “statistic.” Although the authors do not address the cause of the failure, it is interesting to speculate whether the standard error is too small, resulting in more statistically significant results than expected, due to not accounting correctly for the implicit correlation in the comparison of risks with and without the biomarker on the same patients.

Part of the appeal of NR is that it seems to address one of AUC’s deficiencies—namely, being more clinically relevant. Often the approach to assess biomarker utility is to test for statistical significance with a general-purpose test such as AUC increase and then state change in a clinically relevant but different metric without a direct assessment of its statistical significance. With Pepe et al.’s conclusions about NR, bridging this gap by developing models to assess the statistical significance of clinically relevant metrics remains an important challenge for statistical biomarker research. Each clinical question comes with its specific trade-offs for balancing false positives and false negatives with true positives and true negatives (7), incidence of endpoint, existing clinical tests, and where in the disease course a new biomarker test may best fit. These considerations result in a specific range of clinically relevant
specificities (and sensitivities). A blood test for early detection of ovarian cancer where there is no existing clinical test will have different requirements from a blood test for early detection of breast cancer where mammography is an existing screen. The application may be to women with dense breasts where mammography is known to have lower sensitivity. Trade-offs between true and false biomarker test results, and therefore desired specificities, will be different from the trade-offs in early detection of ovarian cancer.

With the explosion in biomarker research due to the -omic revolutions, directly addressing the statistical significance of the clinically relevant metric where specificity requirements are tailored to each clinical context results in greater power and thus fewer patients tested and greater sensitivity to detect a clinically relevant biomarker signal. This efficient and clinically directed approach is a vital direction in which biomarker statistical research needs to head. Until recently the alternative to a nonparametric approach was a small selection of rigid parametric models, which often do not fit real data well and are therefore not preferred. But recent advances in statistical computing [R (8), WinBUGS (9), STAN (10)] have given biomarker statistical researchers the capability of developing statistical models that are flexible while structured to incorporate biological knowledge, fit data well, and be more powerful than general-purpose tests.

For example, a panel of biomarkers is likely required to cover the spectrum of disease in most cancers. Often a biomarker distribution in cancer patients has a proportion of case patients in whom it is overexpressed and other case patients in whom it is expressed at the same level as in control subjects. A mixture of two $t$ distributions to represent the biomarker distribution in case patients captures this modicum of biology and is simultaneously robust to outliers ($t$ distribution) and flexible (mixture). Extending this flexible robust model to a panel of biomarkers requires multivariate mixtures of multivariate $t$ distributions, a statistical model now implementable in STAN. Although STAN’s perspective is Bayesian, frequency properties of Bayesian-derived statistical methods are important (11), and their evaluation through methods such as in the study by Pepe et al. will be an important step in justifying application of such approaches.

What is now needed to complement general-purpose AUCs from the 1980s, are statistical assessments of biomarker utility that begin with a specific clinical issue as the driving force. Clearly defining the clinical issue and quantifying the trade-offs between its sequelae requires, in the spirit of DeLong et al. (1), close collaboration between clinician scientists, biomarker discoverers, and biomarker statistical researchers. The well-defined intended use then leads to clinically relevant metrics, e.g., sensitivity at small range of high specificities. The statistical computing systems developed this decade can produce inference for increases in clinically relevant metrics even with statistical models incorporating flexible and therefore realistic biomarker distributions. Then they can be checked for frequentist properties as in Pepe et al. This is one way biomarker statistical research can be updated from the 1980s to 2014 to address with clinical precision the now ubiquitous question, “Does this cancer biomarker have clinical utility?,” without, as Pepe et al. clearly show, being misled.

References

Funding
This work was supported by grant CA152990 from the Early Detection Research Network of the National Cancer Institute.

Notes
The ideas in this editorial are solely the responsibility of the author. The funders had no role in the writing of the editorial or the decision to submit it for publication. Massachusetts General Hospital has licensed software developed by the author to Abcodia Inc, and the author is a consultant to Abcodia Inc.

Affiliation of author: Biostatistics Center, Massachusetts General Hospital, Boston, MA.