Validation of Monte Carlo Estimates of Three-Class Ideal Observer Operating Points for Normal Data

Darrin C. Edwards

Rationale and Objectives: Traditional two-class receiver operating characteristic (ROC) analysis is inadequate for the complete evaluation of observer performance in tasks with more than two classes.

Materials and Methods: Here, a Monte Carlo estimation method for operating point coordinates on a three-class ROC surface is developed and compared with analytically calculated coordinates in two special cases: (1) univariate and (2) restricted bivariate trinormal underlying data.

Results: In both cases, the statistical estimates were found to be good in the sense that the analytical values lay within the 95% confidence interval of the estimated values about 95% of the time.

Conclusions: The statistical estimation method should be key in the development of a pragmatic performance metric for evaluation of observers in classification tasks with three or more classes.

Key Words: ROC analysis; three-class classification; ideal observer decision rules.

Receiver operating characteristic (ROC) analysis has, for many years, been the standard for evaluating observer performance in a medical decision task with two classes to which observations belong (1). A particularly familiar example is the canonical radiologic task of identifying whether an abnormality, such as a fracture or lesion, is present in an image.

Not all medical, or even radiological, tasks are so readily restricted to two outcomes, however. A particular task might require distinguishing among multiple types of abnormality (2) or distinguishing normal tissue from abnormalities of different types (3), or, in the computer-aided diagnosis (CAD) task that originally motivated much of the work in this area, one might need to distinguish malignant and benign actual lesions from the false-positive detections produced by an automated scheme (4–6). Traditional two-class ROC analysis is inadequate for the complete evaluation of observer performance in such tasks. Unfortunately, although the broader theoretical characteristics of observer behavior in a three-class classification task were outlined many decades ago (7), the extension of this knowledge to a complete understanding and the implementation of such knowledge into practical tools for addressing real clinical problems have remained elusive.

Why should this be the case? After all, two-class ROC analysis is successful not just because of its practical usefulness but also because of the theoretical simplicity and elegance it possesses. One measures some quality of the objects being classified and compares this decision variable with a critical threshold. If the decision variable is above the critical threshold, the object is classified as positive; the fraction of actually positive objects called positive is the true-positive fraction (TPF), while the fraction of actually negative objects (incorrectly) called positive is the false-positive fraction (FPF). The FPF and TPF values give the observer's ROC operating point at that critical threshold; if a different critical threshold is chosen, a different operating point is obtained, and the collection of all (FPF, TPF) pairs for all values of the critical threshold is the ROC curve (1,7,8). The observer that achieves the best possible ROC performance, given the variability of the decision variables across the actually positive and negative populations, is called the ideal observer (7,8). Objectively comparing two different observers based on their ROC curves is simple if one curve is always above the other, but it is problematic if the two curves cross; however, the area under the ROC curve (AUC) is a readily calculated quantity that allows such comparison and, furthermore, is theoretically justifiable since the AUC can be shown to be the chance of the observer making a correct decision in a two-alternative forced-choice task (i.e., when presented with a pair of observations, one actually positive and one

Acad Radiol 2013; 20:908–914

From the Department of Physiology, 303 E Superior St., Northwestern University, Chicago, IL 60611. Received January 31, 2013; accepted April 16, 2013. **Address correspondence to:** D.C.E. e-mail: darrin.c.edwards@gmail.com

http://dx.doi.org/10.1016/j.acra.2013.04.002

actually negative, the chance of correctly deciding which is which) (8). Given a simple model for the underlying data, the binormal model, ROC curves can be computed under both assumptions in which the observer's decision variable is directly related to the underlying data [the conventional binormal model (9)] and in which the decision variable is an ideal observer decision function of the underlying data (the proper binormal model (10)).

Turning to the three-class task, difficulties appear at the outset and accumulate at a discouraging rate. The performance of an observer is characterized by six of the nine possible conditional classification probabilities given a particular decision rule (the probabilities of deciding an observation belongs to one class when it is actually drawn from that or another class, analogous to TPF and FPF), not three as one might naively expect (11–15). [The nine classification probabilities are related by three equations, one for each class, because they are conditional probabilities (16), just as TPF + FNF = 1 where FNF is the false-negative fraction; thus, three of the classification probabilities may be eliminated, leaving six.] A number of researchers have proposed applying restrictions to the form of the observer, to reduce the dimensionality of the performance description from six to three in a more principled fashion (17–21). However, despite the resulting observer behavior being consistent with that of the ideal observer given the restrictions imposed (22,23), these descriptions of observers and their performance are nevertheless not fully general. In the general case, six conditional classification probabilities (the most principled choice of which is the set of six misclassification probabilities) form the coordinates of a six-dimensional ROC space, and the observer's ROC "surface" has as many as 5 degrees of freedom. For the ideal observer, these coordinates are the probabilities of a random decision variable vector lying in one of a set of three wedge-shaped regions (7,24); this is analogous to the two-class case, but the integrals involved in their calculation are far less tractable. (As for the behavior of nonideal observers, it is not even completely clear what the general form of the decision function should be.) To evaluate the performance of such an observer, one might naively hope to calculate the volume under this (five-dimensional) ROC surface and use that for comparison with other observers in a manner analogous to the two-class AUC. Unfortunately, this quantity was shown not to correlate in any useful way with intuitive concepts of "performance" due to degeneracy issues in ideal observer ROC surfaces (25).

This result implies that a performance metric is required that is not just a simple generalization of AUC. Nevertheless, since the six misclassification probabilities that form the coordinates of the observer's ROC surface remain the fundamental descriptors of the observer's performance, it is to be expected that a valid performance metric will still be derived in some more sophisticated fashion from the ROC surface. Conceptual characterization of such a performance metric is not impossible (26), but regardless of the theoretical details, it is clear that any such performance metric will require a detailed description of the observer's ROC surface. Lacking fully general analytic methods for calculating three-class ideal observer operating points, we can safely assume that numerical or statistical methods will be required—essentially Monte Carlo methods, since the operating point coordinates are conditional probabilities [i. e., integrals of particular probability density functions (PDFs) over particular regions]. Of course, such a statistical operating point estimation method would need to be validated in practice; this requires, in turn, an analytic method for calculating the "true" value of the operating point.

Our most recently published work finally gave Charles Metz and me hope of breaking this vicious circle; in it, we fully and analytically characterized the ROC operating point behavior of a three-class ideal observer acting on *univariate* normal underlying data (27). However, unlike the two-class task for which the sufficient statistic used by the ideal observer is always univariate, the three-class ideal observer makes use of a pair of decision variables. To consider the fully general case, therefore, it is necessary to be able to take into account *bivariate* data. That is, while it is perhaps imaginable that a particular medical imaging task could be found such that the highly multivariate image data could be classified by an ideal observer acting on univariate decision variable data, this is highly unlikely to be true in general. In any case, a fully general approach would still be needed in to establish that the univariate method was adequate and valid for that particular task.

A brief review of the univariate work is next, along with a more detailed description of a second special case in which three-class ideal observer operating points may be calculated analytically, in the case of bivariate normal data under certain restrictions on both the data distributions and the form of the decision rule. The basis of the statistical method for estimating ideal observer operating points is also described. Then, a set of simulation studies is described in which analytically calculated and statistically estimated operating points are compared for a variety of underlying data distributions and particular choices of ideal observer decision criteria. The results of these simulations studies are presented later, followed by a discussion of their implications. The conclusions that can be drawn from them are finally given. The goals of the present work are to show that the general validity of the statistical estimation method can be adopted as a working hypothesis and that this validity can continue to be evaluated as new analytic methods (beyond the two considered here) become available.

THEORY

In a two-class classification task, with underlying data $\mathbf{x}$, the ideal observer makes decisions by comparing a function of the data, called the likelihood ratio (LR), with a critical threshold γ. The LR is the ratio of the PDFs of the underlying data, conditional on the classes from which the data are drawn; that is, $\mathrm{LR}(\vec{\mathbf{x}}) \equiv p(\vec{\mathbf{x}}|\text{class 1}) / p(\vec{\mathbf{x}}|\text{class 2})$. If $\mathrm{LR}(\mathbf{x}) > \gamma$, the observation is called positive, otherwise it is called negative. (A bold typeface is used to denote random variables.)