Using the Mean-to-Sigma Ratio as a Measure of the Improperness of Binormal ROC Curves

Stephen L. Hillis, PhD, Kevin S. Berbaum, PhD

Rationale and Objectives: A basic assumption for a meaningful diagnostic decision variable is that there is a monotone relationship between the decision variable and the likelihood of disease. This relationship, however, generally does not hold for the binormal model. As a result, receiver operating characteristic (ROC)-curve estimation based on the binormal model produces improper ROC curves that are not concave over the entire domain and cross the chance line. Although in practice the "improperness" is typically not noticeable, there are situations where the improperness is evident. Presently, standard statistical software does not provide diagnostics for assessing the magnitude of the improperness.

Materials and Methods: We show how the mean-to-sigma ratio can be a useful, easy-to-understand and easy-to-use measure for assessing the magnitude of the improperness of a binormal ROC curve by showing how it is related to the chance-line crossing. We suggest an improperness criterion based on the mean-to-sigma ratio.

Results: Using a real-data example, we illustrate how the mean-to-sigma ratio can be used to assess the improperness of binormal ROC curves, compare the binormal method with an alternative proper method, and describe uncertainty in a fitted ROC curve with respect to improperness.

Conclusions: By providing a quantitative and easily computable improperness measure, the mean-to-sigma ratio provides an easy way to identify improper binormal ROC curves and facilitates comparison of analysis strategies according to improperness categories in simulation and real-data studies.

Key Words: Receiver operating characteristic curve; diagnostic radiology; mean-to-sigma ratio; binormal model; proper ROC model.

For diagnostic studies that evaluate and compare medical imaging modalities (eg, mammography) that require a human reader (typically a radiologist) to interpret generated images with respect to disease likelihood or severity, a commonly used method for estimating a receiver operating characteristic (ROC) curve is to use maximum likelihood estimation based on the assumption of a latent binormal model (1–4); we refer to this method as the *binormal method*. The latent binormal model assumption states that there exists a monotone transformation that, when applied to the decision variable of interest, results in a latent decision variable that is normally distributed for nondiseased cases as well as for diseased cases, with the means and variances allowed to differ for the two distributions. For example, consider a study where a radiologist is asked to assign likelihood-of-disease confidence levels to images using a discrete five-level ordinal integer scale (eg, 1 = "definitely not diseased", ..., 5 = "definitely diseased"); for this situation, it is typical to assume that these ratings represent the binning of values of a latent (ie, unobserved) continuous decision variable representing the reader's likelihood-of-disease perception. Often the ROC-curve summary measure of interest is the area under the curve (AUC).

For large samples, the binormal method has been shown to perform well for decision variable distributions that can vary greatly from the binormal distribution (5–8). We refer to the ROC curve corresponding to a latent binormal model decision variable as the *binormal ROC curve*. Throughout, we assume that the decision variable of interest is continuous and that larger values of it are more indicative of disease.

In most practical situations, a meaningful decision variable should be an increasing function of the likelihood ratio (likelihood of being diseased divided by likelihood of not being diseased) (9). A decision variable model having this property and its corresponding ROC curve are said to be *proper* (10). A function whose first derivative is decreasing throughout an open interval is called *concave* or *concave downward* in that interval,

Acad Radiol 2011; 18:143–154

From the Center for Research in the Implementation of Innovative Strategies in Practice (CRIISP), Department of Veterans Affairs (VA) Iowa City Medical Center (152), 601 Highway 6 West, Iowa City, IA 52246-2208 (S.L.H.); Departments of Biostatistics (S.L.H.) and Radiology (K.S.B.). Received December 4, 2009; accepted September 1, 2010. Supported by the National Institutes of Health, grant R01EB000863. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs. **Address correspondence to:** S.L.H. e-mail: steve-hillis@uiowa.edu

doi:10.1016/j.acra.2010.09.002

and a function whose first derivative is increasing throughout an open interval is called *convex* or *concave upward* in that interval (11). Because the slope of a ROC curve for a continuous decision variable is equal to the likelihood ratio at the corresponding threshold, it follows that the slope of a proper ROC curve decreases as the false-positive fraction (fpf) increases, that is, a proper ROC curve will be concave everywhere ($0 \leq$ fpf ≤ 1) (9). If the decision variable is not an increasing function of the likelihood function, then its model and corresponding ROC curve are said to be *improper* (10).

The latent binormal model is improper if the nondiseased and diseased distribution variances differ; furthermore, there is a single fpf value such that the ROC curve is concave on one side and convex on the other side (9). In addition, as we show later, there is a single fpf value where the ROC curve crosses the chance line, implying that for a range of fpf values the decision variable performs worse than guessing.

Although binormal model ROC curves are improper unless the diseased and nondiseased variances are equal, in practice the "improperness" is so small that it is not apparent when looking at the ROC curve. However, there are situations when the improperness is apparent, with the ROC curve visibly crossing below the chance line and having an obvious "hook." For these situations, we deem the ROC curve and its corresponding binormal model to be *noticeably* or *slightly improper*, depending on how easily the improperness can be seen. Pan and Metz (12) note that "because ROC curves do not show shapes of this kind when they are estimated from reliable data sets, hooks and degeneracy can be considered artifacts of the conventional binormal ROC model."

Presently, researchers often ignore or do not check for improperness in fitted binormal ROC curves, even though there can be situations where the magnitude of the improperness is large enough to make the validity of conclusions based on the improper ROC curve questionable. Furthermore, standard statistical software packages do not provide any diagnostics for assessing the magnitude of the improperness; thus, the researcher can only know the extent of the improperness from visually examining ROC-curve plots, which often is not done when the researcher is primarily interested in an ROC-curve summary index, such as the AUC.

There is not general agreement on an appropriate analysis strategy for ROC data that will satisfactorily account for the inherent improperness of binormal ROC curves. At one end of the spectrum is the strategy of using the binormal method and ignoring any improperness in resulting ROC curves, and at the other end is the strategy of always using a proper method that never results in improper ROC curves. In between are other strategies, such as using a proper method only when the binormal method produces a clearly visible improper ROC curve. Although improperness can be visually assessed from graphs, a discussion of the different analysis strategies requires a quantitative improperness measure that is easy to compute and interpret. Our purpose is to investigate the properties of the mean-to-sigma ratio as a quantitative measure of improperness. However, we do not attempt to discuss which analysis strategy should be used, because that would require separate treatment.

TABLE 1. Rating Data for a Radiologist from Van Dyke et al (13)

	Rating					
	1	2	3	4	5	Total
Normal	39	19	9	1	1	69
Diseased	7	7	3	5	23	45

In summary, our main purpose is to show how the mean-to-sigma ratio can be a useful, easy-to-understand, and easy-to-use measure for assessing the magnitude of the improperness of binormal ROC curves. The outline of the article is as follows. We illustrate the inherent improperness of the binormal model with an example, show how the mean-to-sigma ratio can be used as a measure for assessing the degree of improperness, and discuss alternative proper models. Using data from a multireader multimodality study, we illustrate the usefulness of the mean-to-sigma ratio for assessing improperness and for comparing the binormal method with an alternative method based on a proper model.

MATERIALS AND METHODS

Example of a Noticeably Improper Binormal ROC Curve

To illustrate the inherent improperness in binormal ROC curve estimation, consider Table 1, which shows the rating data for one reader from a study (13) that will be described in more detail in the Results section. Figure 1 shows the corresponding fitted binormal ROC curve; note that there is a visible hook and chance-line crossing near the upper right-hand corner of the unit square. In Figure 1 the ROC curve crosses the chance line at the point (fpf, tpf) = (0.976, 0.976), where tpf stands for true-positive fraction, shown by the intersection of the "crossing" reference line with the ROC curve. Furthermore, this ROC curve is concave for fpf < 0.735, but is convex for fpf > 0.735. Letting ROC(t) denote the tpf corresponding to fpf = t, the point (.735,ROC(.735)) on the ROC curve separates the concave and convex portions of the curve and thus is an *inflection point*; this point is shown by the intersection of the "inflection" reference line and the ROC curve in Figure 1.

In general we will denote the fpf = tpf value where the ROC curve crosses the chance line by t_0 and refer to t_0 as the *chance-line crossing fpf*; similarly, we will denote the fpf value where the ROC curve changes from convex to concave (for either increasing or decreasing fpf) by t_1 and refer to t_1 as the *inflection-point fpf*. (More precisely, t_1 is the fpf coordinate of the inflection point.) For example, in Figure 1 the chance-line crossing fpf is t_0 = .976 and the inflection-point fpf is t_1 = .735.

Figure 2 shows the latent nondiseased and diseased distribution densities that yield the ROC curve in Figure 1, with y