Selection of a Rating Scale in Receiver Operating Characteristic Studies: Some Remaining Issues[1]

Howard E. Rockette, PhD, David Gur, ScD

Rationale and Objectives. The aim of this study is to compare the ratings of a group of readers that used two different rating scales in a receiver operating characteristic (ROC) study and to clarify some remaining issues when selecting a rating scale for such studies.

Materials and Methods. We reanalyzed a previously conducted ROC study in which readers used both a 5-point and a 101-point scale to identify abdominal masses in 95 cases. Summary statistics include the distribution of scores by reader for each of the rating scales, the proportion of tied scores when using the 5-point scale that correctly resolved when using the 101-point scale and the proportion of paired normal-abnormal cases where the two rating scales resulted in a different selection of an abnormal case.

Results. As a group, the readers used 84 of the rating categories when using the 101-point scale but the categories used differed for individual readers. All readers tended to resolve the majority of ties on the 5-point scale in favor of correct decisions and to maintain correct decisions when a more refined scale was used.

Conclusions. The reanalysis presented here provides additional evidence that readers in a ROC study can adjust to a 101-point scale and the use of such a refined scale can increase discriminative ability. However, the decision of selecting an appropriate scale should also consider the underlying abnormality in question and relevant clinical considerations.

Key Words. Observer performance; ROC; rating scale.

In some fields, including but not limited to radiology, the application of receiver operating characteristic (ROC) type rating systems often assume an underlying continuous scale that is approximated by a discrete categorization. Historically a 5- (or a 6-) point rating scale had been used for this purpose and this method may have advantages when it is closely related to a set of commonly used diagnostic decisions/recommendations (1). More recently, a 101-point scale has been suggested for this purpose (2–4). Because of the large number of categories, a 101-point scale can be treated as a continuous scale and therefore avoid some of the analytic complexities associated with a discrete ordinal scale. Although several authors have discussed the limitations associated with either of these two approaches (1,5,6), some general issues remain. Furthermore, the possibilities that some decisions in radiology should be viewed more appropriately as an inherently binary decision (7) potentially increases the magnitude of the differences that can occur between discrete and continuous scales because a binary decision may be viewed as using a 2-point (discrete) scale. The purpose of this article is to clarify several issues in regard to the selection of a rating scale in an ROC study by comparing the actual ratings used by a group of readers in a study that employed both

Acad Radiol 2008; 15:245–248

[1] Departments of Biostatistics, Graduate School of Public Health (H.E.R.), and Radiology (D.G.), Imaging Research, F.A.R.P. Building, 3362 Fifth Avenue, University of Pittsburgh, Pittsburgh, PA 15261. Received September 4, 2007; accepted October 2, 2007. Supported in part by grant numbers EB002106 and EB001694 (to the University of Pittsburgh) from the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Institutes of Health. **Address correspondence to:** D.G. e-mail: gurd@upmc.edu

doi:10.1016/j.acra.2007.10.011

a 5-point and a 101-point scale to identify abdominal masses (2). We also present some summary statistics useful in describing the effect of refining a given ordinal scale.

First, it should be recognized that the statistical aspects of contrasting different rating systems depends on the true underlying categorization. If the true underlying scale is continuous, then the use of a discrete scale has by definition less information and will ultimately be inferior when compared using standard statistical measures. Wagner et al (6) demonstrated this in a comparison between a 5-point and a continuous scale when data are generated from an underlying continuous scale. Conversely, if the true underlying scale is discrete, then using a larger number of possible ratings may increase the variance. Gur et al (7) demonstrated this in a simulation study comparing a dichotomous rating with a continuous rating when the true underlying scale is dichotomous. Thus the scale with the most desirable statistical properties often depends on the scale that is conceptually considered as "correct" (or perhaps clinically relevant).

However, simulations usually do not take into consideration possible behavioral changes of raters. For example, a scale with too many categories and beyond the ability of the rater to distinguish among the ratings may result in additional variability because of an increase in the "within" reader variability resulting from lack of consistency. Although one published articles compares a 5-point ordinal scale to a 101-point scale for several abnormalities (5), there is still limited information available to assess raters' behavior and performance when using different rating scales with different underlying assumptions (8,9). Furthermore, different readers may not behave similarly under the same rating conditions (9). One of the objectives in this article will be to contrast the ratings used by individual readers using a 5-point scale with the ratings obtained for the same set of cases with a 101-point scale. The study on which this analysis is based was previously published and showed no statistically significant difference in the estimated areas under the ROC (AUC) curves for the two scales, but the behavior of the readers when using these scales was not described in detail. We also present a useful approach to summarize the potential benefit of scale refinement as well as a possible change in discriminating effect due to increases in variability.

Specifically, we wish to address the following questions:

1. How much of the 101-point scale did the readers actually use in the study in question and did the number of categories actually used by the readers differ?
2. Was there an approximate range of values in the 101-point scale that corresponded to specific discrete rating categories and, if so, did it differ by reader?
3. Did use of the more refined 101-point scale tend to improve the discrimination between disease and nondisease cases or did the large number of rating categories result in an unacceptable number of classifications that were inconsistent with the original 5-point scale?

In answering these questions, we use several simple summary statistics that we believe are useful for contrasting the effect of using a refined scale as compared with a 5-category scale in a ROC setting.

METHODS

Analysis was conducted on ratings by five readers interpreting 95 examinations in which identification of the presence or absence of one or more abdominal masses was the primary diagnostic task. Ratings were done by each reader using both a 101-point scale and a 5-point scale, with higher ratings indicating a greater likelihood of the presence of an abdominal mass. Each reader interpreted approximately 20 cases per session and either the 5-point, or the 101-point, rating scale was used throughout each of the sessions. A minimum of 3 weeks was required between the scoring of the same case session and the sequence in which the two scales were used was randomized. There were 57 cases with and 38 cases without the abnormalities in question. Detailed methodology of the actual original study has been provided elsewhere (2). The original study focused on a comparison of the areas under the ROC curves for the two rating scales, whereas the present study investigates the possible impact of changes in the rating scale on individual cases by different readers and by the group of readers as a whole. The summary statistics that were used in this analysis are based on the pairs of normal-abnormal cases, and therefore can be related directly to the nonparametric estimate of the AUC based on the Wilcoxon statistic.