Reader Variance in ROC Studies—Generalizability to Reader Population at High and Low Performance Levels[1]

David Gur, Andriy I. Bandos, Amy H. Klym, Howard E. Rockette

Rationale and Objectives. To investigate the variability between discriminative performances of readers as a function of average performance levels during receiver operating characteristic (ROC) studies.

Materials and Methods. Four subsets of cases from previously ascertained ROC rating data by 12 observers when detecting interstitial disease and pneumothorax on posteroanterior chest films were selected for each abnormality and reanalyzed to assess changes in "reader" variance component. The subsets were selected based on a prestudy subjective assessment of the subtleness of depicted abnormality (positive cases) and the difficulty in determining its absence (negative cases). Reader variance component was estimated using a bootstrap approach for each subset and the results were used to assess a general relationship between variability and average performance level.

Results. The reader variance component decreased substantially (from 0.007704 to 0.000426), as expected, when the areas under the ROC curves (AUC) for detecting pneumothoraces increased from 84% to 97%. On the other hand, reader variance component increased substantially (from 0.000890 to 0.005181) when AUC for detecting interstitial disease increased from 59% to 87%. The large magnitude of and changes in the reader variance component resulted in a consistent nonmonotone relationship as a function of AUC when other related variance components were included in additon to the reader component.

Conclusion. Among several factors affecting generalizability of ROC results to the population of readers, the reader variance component depended nonmonotonically on the average diagnostic performance and is lowest at both very high and very low levels of performance.

Key Words. AUC; bootstrap; reader variability; ROC; variance components.

Observer performance studies in general and receiver operating characteristic (ROC) type studies in particular are becoming increasingly important in technology and practice assessments. Such studies are now routinely used in a variety of investigations including but not limited to the regulatory approval process (1). Recent developments of analytical tools to assess variance components during ROC studies enable one to gain better understanding of the underlying parameters that affect evaluation of diagnostic systems. One important aspect of these studies is an assessment of the effect of cases and readers on the variability of the measure of interest. Better understanding of the variance components and, perhaps as important, changes in variance components as a function of performance levels (eg, area under the ROC curve [AUC]), may improve our

Acad Radiol 2006; 13:1004–1010

[1] From the Departments of Radiology, School of Medicine (D.G., A.H.K.), and Biostatistics, Graduate School of Public Health (A.I.B., H.E.R.), University of Pittsburgh, Pittsburgh, PA 15261. Received April 20, 2006; accepted May 26, 2006. This work is supported in part by Public Health Service grants EB002106 and EB001694 (to the University of Pittsburgh) from the National Institute for Biomedical Imaging and Bioengineering (NIBIB), National Health Institutes, Department of Health and Human Services. **Address correspondence to:** D.G. e-mail: gurd@upmc.edu

doi:10.1016/j.acra.2006.05.015

ability to generate more generalizable inferences as a result of these studies (2–4). Such investigations may also improve our understanding of some of the limitations associated with observer performance studies (5).

For planning an efficient ROC study, it is frequently important to take into consideration the average diagnostic performance. For example, it has been shown that in general the differences in performance between diagnostic systems are magnified in absolute terms by selecting more difficult cases (6,7). Consequently, phase II (challenge) assessments of diagnostic systems were recommended to be performed with a case mix that would result in the average diagnostic performance as measured by the AUC of approximately 0.75. However, because increase in difficulty often results in an increase in variance "under a fixed reader or reader-free setting" (eg, case and mode related), statistical power is not necessarily gained by the use of exceedingly difficult cases. Thus, in planning a study and interpreting the results, it may be important to know the relationship, if any, between variance components and the case mix with the structure of the case population. The issue of generalizability to the population of readers might look similar in nature to the generalizability to the population of cases; however, in most studies, the number of readers is relatively small making inferences regarding generalizability of the result even more dependent upon the magnitude of the between reader variability. In this article, we highlight a potentially important difference regarding generalizability to the population of readers.

We present here an analysis of the estimated "reader" component of the AUC variance for two abnormalities as a function of the absolute performance level measured by the average AUC in a large observer performance study when performance levels ranged from extremely high to extremely low.

MATERIALS AND METHODS

Original Studies

We focus our analysis in this article on a subset of the data ascertained during two prior multimode, multiabnormality observer performance studies that involved the detection of several abnormalities on the same set of posteroanterior (PA) chest images presented at different resolutions (8,9). The original studies were performed several years previously as part of our general effort to understand many aspects of the transition to a digital environment in radiology. In one study, we assessed the effect of data compression on observer performance (8), whereas in the other study we assessed the combined effects of luminance of the display and image resolution on observer performance (9).

Because the details of the studies have been described elsewhere, we present here only a brief overview. In the first study, 529 PA chest images were read by six readers under each of eight different reading modes. These modes included JPEG compressed images at five different levels and presented on laser printed films at one resolution of 100 μm per pixel and in three modes of noncompressed images presented on laser printed films at low (400 μm), medium (200 μm), and high (100 μm) resolutions. In the second study, six different readers read the same 529 PA chest images laser printed and displayed at three different levels of luminance (low, medium, and high, representing different luminance levels expected to be found in radiology reading rooms, at the time) and at three different levels of resolution as in the first study (100, 200, and 400 μm per pixel), resulting in a total of nine different modes. In both studies, each reader scored the image by using a "continuous" rating scale (0–100) to estimate the likelihood of the presence or absence of each of five specific abnormalities. The abnormalities investigated were interstitial disease, nodule, pneumothorax, alveolar infiltrate, and rib fracture. Estimates of the AUCs were calculated for each abnormality, reader, and mode in both studies (8,9).

Three of the modes in the two studies described were identical (low-, medium-, and high-resolution laser-printed images displayed at the same "high" luminance) involving the same set of 529 chest images and using the same rating scale, making it possible to combine the ROC ratings from both studies resulting in a single data set of ROC ratings with 12 readers. All 12 readers were Board-certified radiologists with substantial, albeit varying, experience in interpreting PA chest images. For the purposes of this study, we used the rating data ascertained during the interpretation of 529 chest images evaluated for the presence of two abnormalities (interstitial disease and pneumothorax) at high and low resolution and displayed at high luminance.

During the case selection process of the 529 chest images each positive finding was rated by experienced observers who were provided with the verified "truth" and all other support documentation as "typical" or "subtle" for the detection of the abnormality of interest. Correspondingly, examinations negative for specific abnormalities were rated as "easy" (also termed here typical) or