Prediction Accuracy of a Sample-size Estimation Method for ROC Studies

Dev P. Chakraborty, PhD

Rationale and Objectives: Sample-size estimation is an important consideration when planning a receiver operating characteristic (ROC) study. The aim of this work was to assess the prediction accuracy of a sample-size estimation method using the Monte Carlo simulation method.

Materials and Methods: Two ROC ratings simulators characterized by low reader and high case variabilities (LH) and high reader and low case variabilities (HL) were used to generate pilot data sets in two modalities. Dorfman-Berbaum-Metz multiple-reader multiple-case (DBM-MRMC) analysis of the ratings yielded estimates of the modality-reader, modality-case, and error variances. These were input to the Hillis-Berbaum (HB) sample-size estimation method, which predicted the number of cases needed to achieve 80% power for 10 readers and an effect size of 0.06 in the pivotal study. Predictions that generalized to readers and cases (random-all), to cases only (random-cases), and to readers only (random-readers) were generated. A prediction-accuracy index defined as the probability that any single prediction yields true power in the 75%–90% range was used to assess the HB method.

Results: For random-case generalization, the HB-method prediction-accuracy was reasonable, $\sim$ 50% for five readers and 100 cases in the pilot study. Prediction-accuracy was generally higher under LH conditions than under HL conditions. Under ideal conditions (many readers in the pilot study) the DBM-MRMC–based HB method overestimated the number of cases. The overestimates could be explained by the larger modality-reader variance estimates when reader variability was large (HL). The largest benefit of increasing the number of readers in the pilot study was realized for LH, where 15 readers were enough to yield prediction accuracy >50% under all generalization conditions, but the benefit was lesser for HL where prediction accuracy was $\sim$36% for 15 readers under random-all and random-reader conditions.

Conclusion: The HB method tends to overestimate the number of cases. Random-case generalization had reasonable prediction accuracy. Provided about 15 readers were used in the pilot study the method performed reasonably under all conditions for LH. When reader variability was large, the prediction-accuracy for random-all and random-reader generalizations was compromised. Study designers may wish to compare the HB predictions to those of other methods and to sample-sizes used in previous similar studies.

Key Words: ROC; sample-size; methodology assessment; statistical power; DBM; MRMC; simulation; Monte Carlo.

The purpose of most imaging system assessment studies is to determine for a given diagnostic task whether radiologists perform better on one imaging system than another and whether the difference is statistically significant. In the receiver-operating characteristic (ROC) observer performance paradigm in which the radiologist assigns a rating to each patient image (ie, confidence level that the patient has disease), the performance index is usually chosen to be the area under the ROC curve (AUC $\equiv$ A). The statistical analysis determines the significance level of the study (ie, the P value for rejecting the null hypothesis [NH] that the difference between the two AUCs is zero [$\Delta A = 0$]). If the P value is smaller than a prespecified value α, typically set at 5%, one rejects the NH and declares the modalities different at the α significance level. Statistical power is the probability of rejecting the NH when the alternative hypothesis (AH) $\Delta A \neq 0$ is true. The difference ΔA under the AH is referred to as the effect size.

Statistical power depends on the numbers of readers and cases, the variability of reader skill levels, the variability of difficulty levels of the cases, the statistical analysis used to estimate the P value, the effect size, and α. The aim of sample-size estimation methodology is to estimate the numbers of readers and cases needed to achieve the desired power for a specified analysis method, ΔA, and α. Sample-size estimation is an important consideration at the planning stage of a study. An underpowered study (too few readers or cases) raise ethical issues because study patients are subjected to unnecessary imaging procedures for a study of questionable statistical strength. Conversely, an excessively overpowered study subjects unnecessarily large numbers of patients to imaging procedures and raises the cost of the study. It is generally considered preferable to err on the conservative side (ie, overpowered studies are preferred to underpowered studies), provided excessive overpowering is avoided. Studies are typically designed for 80% desired power.

Acad Radiol 2010; 17:628–638

Department of Radiology, University of Pittsburgh, Presbyterian South Tower, Room 4771, 200 Lothrop Street, Pittsburgh, PA 15213. Received July 27, 2009; accepted January 15, 2010. Supported in part by grants from the Department of Health and Human Services, National Institutes of Health, R01-EB005243 and R01-EB008688. **Address correspondence to:** D.P.C. e-mail: dpc10@pitt.edu

doi:10.1016/j.acra.2010.01.007

The true effect size is unknown; indeed, if one knew it, there would be no need to conduct an ROC study. Sample-size estimation involves making a critical decision regarding the anticipated effect size ΔA. To quote an earlier work, "any calculation of power amounts to specification of the anticipated effect size" (1). Increasing $|\Delta A|$ will increase statistical power, but may represent an unrealistic expectation of the true difference between the modalities. On the other hand, an unduly small $|\Delta A|$ may be clinically irrelevant besides requiring a very large sample-size to achieve 80% power. These considerations are described in more detail elsewhere (1–4) and are not the subject of this study. In this study, it is assumed that the true effect size is known, a condition always satisfied in the context of a simulation study.

The topic of sample-size estimation may evoke some trepidation in nonstatisticians involved in ROC studies. Statisticians who understand the specialized techniques that have been developed for ROC studies may not be readily available. Lacking this resource, the investigator looks in the literature for "similar studies" and follows precedents. It is not surprising that some published studies, excluding, of course, clinical trials designed by expert statisticians, tend to cluster around similar numbers (eg, 3–5 readers and 50–100 cases). Sample-size methodologies developed by statisticians for ROC studies are valuable tools because they allow nonexperts to plan reasonably powered ROC studies to answer questions such as is one image processing method better than another. However, proper usage of these tools requires a basic understanding of how they work.

Statistical power depends on the value of $|\Delta A|$ divided by the square root of the variance $\sigma^2_{\Delta A}$ of ΔA (power depends on the magnitude of the difference). When this signal-to-noise-ratio-like quantity is large statistical power is large. Reader and case variability contribute to $\sigma^2_{\Delta A}$. By using sufficient numbers of readers and cases $\sigma^2_{\Delta A}$ can be made sufficiently small to achieve the desired statistical power. Sample-size methodology estimates the magnitudes of different sources of variability contributing to $\sigma^2_{\Delta A}$ from a pilot study with a relatively few number of readers and cases. Once the variabilities are known, the sample-size estimation method can calculate the numbers of readers and cases that will reduce $\sigma^2_{\Delta A}$ sufficiently to achieve the desired power for the pivotal study.

There are several sample-size estimation methods for ROC studies representing different approaches to the statistical analysis of the ratings data and estimation of the magnitudes of the different sources of variability. Methods exist for single-reader studies (5–8) and for multiple-reader studies (9–16). This study is concerned with multiple-reader studies that follow the fully crossed factorial design in which all reader interpret all cases in all modalities. This is referred to as the multiple-reader multiple-case (MRMC) study design. Because the matching tends to decrease $\sigma^2_{\Delta A}$, it yields more statistical power and consequently this design is frequently used in conducting ROC studies. Two well-known sample-size estimation procedures for MRMC are the Obuchowski-Rockette (9,12,13) and the Hillis-Berbaum (HB) methods (10). To keep the scope of the work to a reasonable level, this study was limited to assessment of the HB method. The HB method works in conjunction with the Dorfman-Berbaum-Metz (DBM) method of analyzing MRMC data. It uses the variability components estimated by the DBM-MRMC method to predict the sample-size. DBM-MRMC analysis software is available from http://www-radiology.uchicago.edu/cgi-bin/roc_software.cgi and from http://perception.radiology.uiowa.edu. The HB method has been implemented in SAS software available on http://perception.radiology.uiowa.edu.

Hillis and Berbaum illustrated the usage of their method with two clinical datasets. With clinical datasets the true values of reader and case characteristics (eg, variability) are unknown. Therefore, it is not possible to determine whether the true power corresponding to the predicted numbers of readers and cases is close to 80%. True power is defined as the fraction of NH rejections over many independent MRMC-ROC studies conducted using the predicted numbers of readers and cases. Because this requires practically unlimited resources, simulations (ie, Monte-Carlo methods) are widely used to assess statistical methodologies (17–21). The aim of this study was to assess the prediction accuracy of the HB sample-size estimation method. In the following sections, the DBM-MRMC and HB methods are briefly reviewed. The validation procedure is described and results of validation testing of the HB method are reported.

METHODS

Overview of the Validation Methodology

Unless noted otherwise, the simulated pilot data sets consisted of 5 readers interpreting 50 normal and 50 abnormal cases in 2 modalities under the NH condition. The Roe and Metz ratings simulator (17) was used to generate pilot data sets. The baseline area under the ROC curve was AUC = 0.855. AUC was calculated by the trapezoidal rule. The number of readers in the pivotal study was 10, the effect size ΔA was 0.06, α = 5%, and two-tailed NH testing was used. For each pilot data set, DBM-MRMC analysis estimated the magnitudes of the different sources of variabilities. These were used by the HB method to predict the number of cases K, assumed to be equally split between normal and abnormal cases, needed to achieve 80% power, and the true power P corresponding to K cases was determined. True power was defined as the fraction of NH rejections over 2000 independent MRMC studies conducted using the predicted numbers of readers and cases (2000 simulations are often used to ensure a reasonable degree of accuracy of the power estimate). If the true power was close to 80%, the method had made an accurate prediction. In this study, a simulation quality random number generator, based on the one described in (22) was used. It is available in a collection of mathematical functions that can be downloaded from http://www.gnu.org/software/libc/ (23). The period of the