Published in final edited form as:

Acad Radiol. 2010 May ; 17(5): 628–638. doi:10.1016/j.acra.2010.01.007.

Prediction accuracy of a sample-size estimation method for ROC studies

Dev P. Chakraborty, Ph.D.

University of Pittsburgh, Department of Radiology University of Pittsburgh Presbyterian South Tower, Room 4771 200 Lothrop Street Pittsburgh, PA 15213 dpc10@pitt.edu 412-605-1553 (phone) 412-605-1554 (fax)

Abstract

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 2 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 range 75% to 90% was used to assess the HB method.

Results—For random-case generalization the HB-method prediction-accuracy was reasonable, ~ 50% for 5 readers in the pilot study. Prediction-accuracy was generally higher under low reader variability conditions (LH) than under high reader variability conditions (HL). 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 observed large variability of the DBM-MRMC modality-reader variance estimates, particularly 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 ~ 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.

Keywords

ROC; sample-size; methodology assessment; statistical power; DBM; MRMC; simulation; Monte Carlo

INTRODUCTION

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, i.e., 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, i.e., 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 pre-specified 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 null hypothesis 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 and / or cases) raise ethical issues since 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, i.e., overpowered studies are preferred to underpowered studies, provided excessive overpowering is avoided. Studies are typically designed for 80% desired power.

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 [1] “any calculation of power amounts to specification of the anticipated effect size”. 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 in [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 non-statisticians 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, e.g., 3-5 readers and 50-100 cases. Sample-size methodologies developed by statisticians for ROC studies are valuable tools since they allow non-experts to plan reasonably powered ROC studies to answer questions like is an 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 $\sigma_{\Delta A}^2$ 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 $\sigma_{\Delta A}^2$. By using sufficient numbers of readers and cases $\sigma_{\Delta A}^2$ can be made sufficiently small to achieve the desired statistical power. Sample-size methodology estimates the magnitudes of different sources of variability contributing to $\sigma_{\Delta A}^2$ 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 $\sigma_{\Delta A}^2$ 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 interprets all cases in all modalities. This is referred to as the multiple-reader multiple-case (MRMC) study design. Since the matching tends to decrease $\sigma_{\Delta A}^2$ 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 (e.g., 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. Since this requires practically unlimited resources, simulations (i.e., 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 two modalities under the NH condition. The Roe and Metz ratings simulator [17] was used to generate pilot data sets. The base-line 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, $\alpha = 5\%$ and two-tailed null hypothesis testing was used. For each pilot data set DBM-MRMC analysis estimated the magnitudes of

the different sources of variabilities. These were utilized 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 Ref. [22] was used. It is available in the GNU scientific library [23] (routine `gsl_rng_mt19937`). The period of the generator is approximately 10^{6000} .

DBM-MRMC analysis

For a given modality the ratings assigned by the reader to the cases yield a single AUC value which, in a sense, averages out the contribution of each case. The jackknife procedure is a method for recovering the individual case contributions, albeit on a different quantity (not AUC) termed a pseudo-value which is defined by

$$\theta_{ijk} = KA_{ij} - (K - 1)A_{ij(k)} \quad \text{Eqn. 1}$$

In Equation 1 θ_{ijk} is the pseudo-value corresponding to modality i ($i = 1, 2, \dots, I$), reader j ($j = 1, 2, \dots, J$) and case k ($k = 1, 2, \dots, K$), where I is the number of modalities (2 in this study), J is the number of readers and K is the number of cases. A_{ij} is AUC for modality i and reader j when all cases are included in the analysis. $A_{ij(k)}$ is AUC for modality i and reader j when case k is excluded from the analysis. The jackknife procedure is repeated for all modalities, readers and cases, yielding a 3-dimensional matrix containing IJK pseudo-values. In recent extensions to the original [24] DBM-MRMC method by Hillis and Berbaum [25] the transformation

$\theta_{ijk}^* = \theta_{ijk} + A_{ij} - \bar{\theta}_{ij}$ is applied to the pseudo-values (the bar denotes the average over the dotted index) and θ_{ijk}^* is referred to as a normalized pseudo-value. In this work normalized pseudo-values were used and the asterisk symbol is suppressed. Other changes from the original algorithm that are described in [25] were also implemented.

The extended DBM-MRMC random effects model for the (normalized) pseudo-values is [24-26]:

$$\theta_{ijk} = \theta + \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + \varepsilon_{ijk} \quad \text{Eqn. 2}$$

The notation $x \sim N(0, \sigma^2)$ denotes that x is a sample from a normal distribution with zero mean and variance σ^2 . According to eqn. 2 each pseudo-value θ_{ijk} is assumed to be the sum of 8 terms: a constant term θ representing the average over all modalities, readers and cases, estimated by $\bar{\theta}_{...}$; a modality effect τ_i estimated by $\bar{\theta}_{i..}$; a random contribution $R_j \sim N(0, \sigma_R^2)$ from the j^{th} reader; a random contribution $C_k \sim N(0, \sigma_C^2)$ from the k^{th} case; and 2nd order interaction contributions, modality-reader $(\tau R)_{ij} \sim N(0, \sigma_{\tau R}^2)$ which accounts for modality dependence of a reader's contribution; modality-case $(\tau C)_{ik} \sim N(0, \sigma_{\tau C}^2)$; reader-case $(RC)_{jk} \sim N(0, \sigma_{RC}^2)$ and an error term $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$. For normalized pseudo-values it can be shown that for two modalities

$\left| \bar{\theta}_{2..} - \bar{\theta}_{1..} \right| = \Delta A$, an estimate of the effect size. The estimate is subject to sampling variability but when averaged over many simulations the estimated effect size approaches the ratings-simulator (see below) specified effect size $\Delta A = 0.06$. The variances $\sigma_R^2, \sigma_C^2, \sigma_{\tau R}^2, \sigma_{\tau C}^2, \sigma_{RC}^2$ and

σ_{ε}^2 are referred to as variance components. DBM-MRMC software analyzes the pseudo-value matrix using a customized mixed model analysis of variance procedure [25]. The procedure yields the p-value for rejecting the $H_0: \tau_1 = \tau_2$ and estimates of the variance components.

In Eqn. 2 it is assumed that both readers and cases are random effects, i.e., both contribute random terms to θ_{ijk} . This is termed random-reader random-case analysis (abbreviated to random-all) which allows generalization of a significant ($p < 0.05$) finding to the population of readers and cases. Analysis that generalizes to the population of cases but is specific to the particular set of readers in the study is termed fixed-reader random-case analysis (abbreviated to random-cases). Analysis that is specific to the case set used in the study but generalizes to the population of readers is termed random-reader fixed-case analysis (abbreviated to random-readers).

HB sample-size estimation method

Since the term R_j in Eqn. 2 contributes random but equal values to θ_{ijk} it does not contribute to the difference $\bar{\theta}_{1..} - \bar{\theta}_{2..}$ between the modalities. Therefore σ_R^2 does not contribute to σ_{AA}^2 and statistical power is unaffected. However, the component $(\tau R)_{ij}$ contributes different random values to the two modalities which increases σ_{AA}^2 and decreases power. Likewise σ_C^2 has no effect on statistical power but $\sigma_{\tau C}^2$ decreases power, as does the pure error term ε_{ijk} . The HB sample-size algorithm [10,27] generates a table of different combinations of numbers of readers and cases that yield the desired power for random-all, random-cases and random-readers analyses.

An in-house implementation of the DBM-MRMC algorithm that calculated the quantities necessary for this study was used. Using the trapezoidal area under the ROC curve as the performance index, the implementation has been verified for several data sets against the DBM-MRMC web site software [28]. Likewise, an in-house implementation of the HB method that has been verified against the HB web site SAS program for several data sets was used. To avoid floating point arithmetic errors a lower bound (0.5) was placed on the denominator degrees of freedom of the F-distribution.

Roe and Metz decision variable simulator

Pilot datasets were simulated using the Roe and Metz simulator [17]. The equation describing the simulator is similar to Eqn. 2 but unlike Eqn. 2 which applies to pseudo-values, the Roe and Metz simulator applies to ratings-like quantities termed z-samples. A z-sample is a continuous variable whereas a rating is usually an integer (e.g., 1 = high confidence normal, ..., 3 = ambivalent, ..., 6 = high confidence abnormal). A continuous variable can always be converted to an integer by binning. The other difference from Eqn. 2 is that the z-sample variance components are assumed to be known (in fact they define the z-sample simulator) while the pseudo-value variance components are unknown and have to be estimated from the ratings. The process can be schematically described by: z-sample variance components + z-sample simulator $\rightarrow$ ratings; ratings + trapezoidal rule $\rightarrow$ AUC; jackknifed AUC $\rightarrow$ pseudo-values; pseudo-values + DBM-MRMC $\rightarrow$ pseudo-value variance components; pseudo-value variance components + HB $\rightarrow$ sample-size estimates.

Since ratings tend to be higher on abnormal cases than on normal cases, the RM simulator needs an additional truth index t : $t = 1$ for a normal case and $t = 2$ for an abnormal case. A z-sample for the i^{th} modality, j^{th} reader, k^{th} case and truth t is denoted z_{ijkt} . Roe and Metz modeled z_{ijkt} by a sum of terms similar to those appearing in Eqn. 2, except every term has a t subscript and represent contributions to the z-sample, not to pseudo-values. There is a constant term μ_t , a modality effect term, $(\Delta\mu)_{it}$ and random terms representing contributions of readers, cases,

modality-reader, modality-case, reader-case and error, assumed to be sampled from zero-mean Gaussian distributions with variances $(\sigma_R^z)^2, (\sigma_C^z)^2, (\sigma_{rR}^z)^2, (\sigma_{rC}^z)^2, (\sigma_{RC}^z)^2$ and $(\sigma_\varepsilon^z)^2$, respectively. The 6 variances are collectively referred to as a variance structure. The z-superscripts emphasize the distinction between the variance components of the Roe and Metz simulator, which generate z-samples or ratings, and the variance components appearing in the DBM random-effects model for the pseudovalues. Since both readers and cases were regarded as random factors, the simulator just described generates random-all data sets. Modifications needed to generate data sets for random-cases and random-readers are described in Appendix A.

Roe and Metz [17] have suggested twelve combinations of AUC values and z-sample variance structures that, based on their experience, are representative of clinical data sets. Two of these variance structures, denoted LH and HL, that exhibited complementary characteristics were chosen for this study. LH has relatively low reader variance and relatively high case variance and the opposite is true for HL. This notation is different from Roe and Metz who use the first term to denote data correlation and the second term to denote reader variance, e.g., in their notation HL would be high data correlation and low reader variance. In the author's opinion using variances to describe both terms and reserving the first term for reader and the second term for cases, makes the presentation easier to follow.

The z-sample variance components listed in Table 1 were obtained from Table 1 in Ref. [17].

For either variance structure $(\sigma_C^z)^2 + (\sigma_{rC}^z)^2 + (\sigma_{RC}^z)^2 + (\sigma_\varepsilon^z)^2 = 1$. This ensures that for a given reader and modality each of the two Gaussian distributions corresponding to normal and abnormal cases, respectively, has unit variance. This allows $\mu_2 - \mu_1$ to be interpreted as the separation of two unit variance normal distributions defining the NH modality and $\mu_2 - \mu_1 + (\Delta\mu)_{22}$ defines the separation for the AH modality. For AUC = 0.855 it can be shown it can be shown [29] that $\mu_2 - \mu_1 = 1.496$, and for an effect size $\Delta A = 0.06$ in AUC units the effect size in z-sample units is $(\Delta\mu)_{22} = 0.444$. The combination of variance structure (LH or HL), AUC (0.855), the desired generalization (random-all, random-cases or random-readers) and effect size (0.06) defines the z-sample simulator.

To avoid possible problems with degenerate datasets the z-samples were not binned and AUC was estimated using the trapezoidal rule. With 100 cases in each pilot dataset there were 100 operating points and the error in estimating AUC is expected to be minimal, and since power involves differences between AUCs, the effect on power and sample-size is expected to be smaller.

Determination of True Power

Two thousand (2000) independent pilot data sets were simulated. Each consisted of z-samples for 5 readers interpreting 50 normal and 50 abnormal images in two modalities (i.e., 5/100 pilots) under the null hypothesis condition $(\Delta\mu)_{22} = 0$, which corresponds to $\Delta A = 0$. DBM-MRMC analysis was applied to each pilot data set and the pseudovalue variance components were recorded. Using these and assuming 10 readers and $\Delta A = 0.06$ in the pivotal study, the HB method estimated the number of cases K_i ($i = 1, 2, \dots, 2000$) for 80% power. To validate the method each K_i one could conduct 2000 alternative hypothesis (AH) simulations with $(\Delta\mu)_{22} = 0.444$ for 10 readers and K_i , and determine true power P_i . A more efficient way is to obtain a few power calibration data points and fit them to an interpolation function. For example, for the LH simulator and random-all generalization, 2000 AH simulations were conducted for n ($n = 15$) values of numbers of cases K_c ($c = 1, 2, \dots, n$) chosen from preliminary trials to bracket the power range extending from ~ 0.3 to > 0.95 . For each c true power P_c was estimated by the number of null hypothesis rejections divided by 2000. The n data points (K_c ,

P_c) were least-squares fitted (SigmaPlot for Windows Version 10.0, Systat Software, Inc., San Jose, CA) to exponential rise to maximum functions of varying order, as detailed below. The appropriate function was used to interpolate the true power P_i for each K_i and to obtain an estimate of $K_{0.80}$, the true number of cases that yielded 80% true power.

Validation of the HB method

For each pilot data set i ($i = 1, 2, \dots, 2000$) the HB method predicted the number of cases K_i and the true power P_i was determined from the appropriate interpolation function. Sometimes the HB method failed to find a value K_i below $K_{MAX} = 2000$, the termination criterion of the HB web site SAS software. We refer to this as “clipping” and denote the condition by $K_i = K_{MAX}$. The final HB prediction, K_{HB} , was defined as the median of K_i over the trials where $K_i < K_{MAX}$ and the corresponding power P_{HB} was determined from the appropriate interpolation function.

The probability that the true power corresponding to an individual sample-size estimate falls inside a specified range encompassing 80% is a measure of the prediction-accuracy of the sample-size estimation method. Since a moderate overestimate is preferable to an underestimate, in this study the range was defined as 0.75 to 0.90. The corresponding numbers of cases $K_{0.75}$ and $K_{0.90}$ were determined from the appropriate interpolation function, and $Q_{0.75,0.90}$ was defined as the fraction of 2000 simulations where K_i was included in the interval $(K_{0.75}, K_{0.90})$, i.e., $Q_{0.75,0.90} = \text{Prob}(K_{0.75} < K_i < K_{0.90})$. Since the choice of the range was arbitrary, an alternate choice $Q_{0.70,0.95}$ was also considered. Likewise, the current termination criterion for the HB method web site implementation ($K_{MAX} = 2000$) is arbitrary, hence additional studies were conducted in which K_{MAX} was varied.

DBM-MRMC analyses of 2000 independent pilot data sets, each with 5 readers and 100 cases, were conducted. For each pilot data set the analysis provided estimates of the pseudo-value variance components. The medians of these estimates over those trials satisfying $K_i < K_{MAX}$ were calculated. The pseudo-value variance components listed in Table 1 represent grand averages of multiple (about 5) independently estimated medians. The coefficients of variance (CVs) of the medians are listed in parentheses. Other statistics defined in Appendix B were recorded, namely, the non-centrality parameter, the critical value of the F-statistic ($F_{1-\alpha;1,ddf}$), the denominator degrees of freedom (ddf), and the clipping fraction f_{max} , i.e., the fraction of the 2000 pilot data sets where $K_i = K_{MAX}$. For two modalities the numerator degrees of freedom of the F-distribution is unity.

RESULTS

The DBM-MRMC estimates of $\sigma_{\tau R}^2$ had the most variability, as measured by the coefficient of variance (CV), especially when $\sigma_{\tau R}^2$ was large, Table 1. For either simulator the CV of $\sigma_{\tau R}^2$ was larger than that of $\sigma_{\tau C}^2$ or $\sigma_{\tau E}^2$, but the values were smaller and the ordering was the same:

$\sigma_{\tau R}^2 < \sigma_{\tau C}^2 < \sigma_{\tau E}^2$. The CV of $\sigma_{\tau R}^2$ for HL was about 2.4 times the CV of $\sigma_{\tau R}^2$ for LH. Note that the listed values are averages and CVs of 5 independent determinations of the medians, each calculated over 2000 independent pilot data sets excluding the clipped predictions (the 2000's).

The CVs of the individual estimates of $\sigma_{\tau R}^2$ were much larger, e.g., the CV of the 1246 modality-reader estimates for HL random-all was ~ 1.6 vs. 0.064 for the median.

Fig. 1 shows power calibration data points (open circles) for LH random-all, effect size = 0.06 and 10 readers in the pivotal study. Each data point was obtained from 2000 AH simulations. The solid line, the power interpolation function, is a 5-parameter exponential rise to maximum function $P = a + b \cdot (1 - \exp(-c \cdot K)) + d \cdot (1 - \exp(-e \cdot K))$. Using the binomial formula, for 2000 AH

trials an approximate 95% confidence interval around 80% power is (0.782, 0.818). Fig. 2 is the corresponding plot for HL random-all. In this case a 4-parameter function $P=a*(1-\exp(-b*K))+c*(1-\exp(-d*K))$ sufficed, and likewise for HL random-readers (not shown). For all other conditions a 3-parameter function $P= a + b*(1-\exp(-c*K))$ sufficed. These functions were used to estimate the number of cases for a specified power, e.g., $K_{0.80}$, $K_{0.75}$ and $K_{0.90}$ corresponding to 80%, 75% and 90% power, respectively.

Fig. 3 and Fig. 4 show interpolation functions for random-all, random-cases and random-readers for the LH and HL variance structures, respectively. Random-all always yielded the lowest curve implying more cases needed for a desired power. When either readers or cases was regarded as fixed, i.e., a source of variability was removed, the corresponding curve was upward-left shifted, implying fewer cases needed for a desired power. However, the relative reduction due to fixing readers or cases depended on the variance structure. Fig. 3 shows that for LH the random-cases curve is slightly upward-left shifted relative to random-all, implying a small reduction, but the random-readers curve is considerably upward-left shifted, implying a large reduction, consistent with the entries in Table 2 showing that for random-cases the decrease in $K_{0.80}$ (162 $\rightarrow$ 149) was smaller than for random-readers (162 $\rightarrow$ 41). Fig. 4 corresponding to the HL variance structure shows the opposite behavior: the random-cases curve is considerably upward-left shifted relative to random-all, while the random-readers curve is less upward-left shifted. Therefore the random-cases condition is expected to result in a larger reduction in the number of cases relative to random-readers, consistent with the values in Table 2 showing that for random-cases the decrease in $K_{0.80}$ (190 $\rightarrow$ 69) was larger than for random-readers (190 $\rightarrow$ 138). LH is characterized by a lower (by a factor of 8.9) ratio of

modality-reader to modality-case pseudovalue variances $(\sigma_{\mu R}^2/\sigma_{\mu C}^2)_{LH}=0.00507$, compared to HL, $(\sigma_{\mu R}^2/\sigma_{\mu C}^2)_{HL}=0.045$, see Table 1. Therefore, for the LH simulator setting the already small modality-reader variance to zero, as done for random-case analysis [25], is expected to have a smaller affect on $K_{0.80}$ than setting the larger modality-case variance to zero, as done for random-reader analysis, and the opposite behavior is expected for HL.

Table 2 summarizes the predictions of the HB method for Monte Carlo simulated pilot data sets with 5 readers, 50 normal and 50 abnormal cases, two modalities and effect size $\Delta A = 0.06$. Listed are the number of cases $K_{0.80}$ for 80% true power, the median HB-predicted number of cases K_{HB} , the corresponding true power P_{HB} , the clipping fraction $f_{\max}$ for $K_{\max} = 2000$, and the prediction-accuracy $Q_{0.75,0.90}$. The trends are difficult to discern and will become clearer when the effect of using more readers in the pilot study are described. For both simulators random-case generalization had the largest prediction-accuracy ($Q_{0.75,0.90} > 48\%$) and the median power was $P_{HB} \sim 89\%$, i.e., the HB method overestimated the number of cases irrespective of whether case variability was low (HL) or high (LH). Qualitatively similar results were obtained for the LH simulator for random-reader generalization: $Q_{0.75,0.90} \sim 45\%$ and $P_{HB} \sim 85\%$. However, since in this instance $K_{0.80}$ was only 41 cases the results may be less reliable. For the LH simulator and random-all generalization the prediction-accuracy was lower $Q_{0.75,0.90} \sim 38\%$ and the median power was $P_{HB} \sim 90\%$. For the HL simulator and random-all generalization the prediction-accuracy was even lower ($Q_{0.75,0.90} \sim 26\%$) and the median power was smaller than desired ($P_{HB} \sim 77\%$), and for random-reader generalization both prediction-accuracy ($Q_{0.75,0.90} \sim 21\%$) and median power ($P_{HB} \sim 70\%$) were low. To summarize prediction-accuracy was higher under low reader variability conditions (LH), and for random-cases, when reader variability becomes irrelevant (since one sets it to zero). Prediction-accuracy was lower under high reader variability conditions (HL). On the whole the HB method was conservative, i.e., it overestimates the number of cases. As shown next, the two underestimates under HL (for random-all and random-readers) change to overestimates when the number of readers was increased. [For LH random-reader the low value of $K_{0.80}$ (41) could be increased by decreasing the effect-size to 0.05, but then it was not possible

to achieve 80% power for HL random-all, as the calibration curve corresponding to Fig. 2 plateaus below 80%. Consequently a comparison at fixed effect-size was not possible.]

Table 3 shows the effect of increasing the numbers of readers in the pilot study R_{PIL} under random-all and random-reader generalization scenarios. [Random-case results are not shown since 5 readers gave relatively good prediction-accuracies $\sim 50\%$, and while increasing the number of readers decreases the variability of $\sigma_{\tau R}^2$, this variance component is irrelevant in random-case analysis.] The prediction-accuracy $Q_{0.75,0.90}$ increased as R_{PIL} increased, implying more pilot study readers are always helpful. The largest benefits were realized for LH, where 15 readers were enough to yield prediction accuracy $> 50\%$ under all generalization conditions, but the benefits were lesser for HL. For LH the maximum prediction-accuracy $Q_{0.75,0.90}$ corresponding to $R_{PIL} = 30$ was 51% for random-all and 62% for random-readers. The corresponding values for HL were 46% and 39%, respectively. For a given numbers of readers in the pilot study and a given generalization scenario, prediction-accuracy was always larger for LH than for HL. For the LH simulator $f_{\max}$ decreased rapidly to almost zero as R_{PIL} increased, but $f_{\max}$ was relatively constant for HL. As the numbers of readers in the pilot study increased the median HB predictions remained conservative ($P_{HB} \sim 89\%$) for LH and approached conservativeness (84%) for HL. Simulations were also conducted for HL with 200 readers and 50 cases for which prediction accuracy improved to 50% for both random-all and random-readers and the median power was $P_{HB} \sim 87\%$. To summarize, under the ideal condition of many readers in the pilot study the HB-method was conservative under all conditions and prediction-accuracy approached 62%. The largest benefits of increasing the number of readers were realized for LH but the benefits were lesser for HL.

The reasons for the overpowering and the clipping at 2000 cases are examined in more detail in Appendix B. In simple terms the reason has to do with the previously noted higher variability of the modality-reader variance component $\sigma_{\tau R}^2$, Table 1. An overestimate of $\sigma_{\tau R}^2$ tends to decrease the power, because it tends to wash out the fixed effect-size, and the HB-method compensates by increasing the number of cases. A sufficiently large overestimate can lead to clipping. To illustrate, consider the following results for the HL simulator and random-all generalization for which substantial clipping occurred ($f_{\max} \sim 38\%$). The medians of $\sigma_{\tau R}^2$, $\sigma_{\tau C}^2$ and σ_E^2 were calculated separately over the clipped and the non-clipped predictions. The median of $\sigma_{\tau R}^2$ was substantially larger (~ 7 times) for the clipped predictions. The median modality-case and median error variance components were the same to within 8%. Over both non-clipped and clipped predictions, $\sigma_{\tau R}^2$ was negative in 413 of 2000 trials, $\sigma_{\tau C}^2$ was negative in 15 trials and σ_E^2 was never negative. The variance component $\sigma_{\tau R}^2$ was much less likely (4% vs. 30%) to be negative for the clipped vs. non-clipped predictions while $\sigma_{\tau C}^2$ was about half as likely to be negative.

Since $\sigma_{\tau R}^2 = 0$ for random-case predictions this also explains why clipping never occurred for random-cases, Table 2. Since variability of $\sigma_{\tau R}^2$ was larger when $\sigma_{\tau R}^2$ was larger, and since $\sigma_{\tau R}^2$ was larger for HL than for LH, clipping is expected to occur with greater frequency for HL than for LH, as was observed, see Table 2. Increasing the number of readers in the pilot study tends to decrease the variability of $\sigma_{\tau R}^2$. This benefits LH more ($f_{\max} \rightarrow 0$) since $\sigma_{\tau R}^2$ is already small, but does not benefit HL as much.

Fig. 5 and Fig. 6 show normalized histograms of K_i ($i = 1, 2, \dots, 2000$) for the LH and HL simulators, respectively, under the random-all condition (the rather similar random-reader histograms are not shown). In Fig. 5 the area under the histogram between the lines labeled $K_{0.75} (= 141)$ and $K_{0.90} (= 226)$ is the prediction-accuracy index $Q_{0.70, 0.95} = 38\%$; $K_{HB} = 225$

and $K_{0.80} = 162$. Note the peak at 2000 cases representing the clipped predictions which contribute 13% to the area. For the HL simulator $Q_{0.70, 0.95} = 26\%$, $K_{HB} = 159$ and $K_{0.80} = 190$. The area contributed by the peak is 39%. Fig. 7 shows the histogram of K_i for the LH simulator under the random-case condition for which $K_{0.75} = 131$, $K_{0.90} = 205$, $Q_{0.70, 0.95} = 53\%$, $K_{HB} = 194$, $K_{0.80} = 149$ and $f_{max} = 0$. For random cases there were no instances of clipping. Fig. 8 is similar to Fig. 7 except it applies to the HL simulator under the random-case condition for which $K_{0.75} = 62$, $K_{0.90} = 95$, $Q_{0.70, 0.95} = 48\%$, $K_{HB} = 92$, $K_{0.80} = 69$ and $f_{max} = 0$.

The prediction-accuracy ordering trends evident in Table 2 were also observed for $Q_{0.70, 0.95}$, except the values were larger, corresponding to the larger power range. For the HL simulator under the random-all condition, as long as $K_{MAX} > 500$, the results of the HB method (K_{HB} , P_{HB} and $Q_{0.75, 0.90}$) were relatively insensitive to the choice of K_{MAX} .

DISCUSSION

The primary goal of this study was validating the Hillis-Berbaum (HB) method for three generalization scenarios: random-all, random-cases and random-readers for two simulators with complementary characteristics: LH had higher case and lower reader variability while HL had lower case and higher reader variability. A distinction was made between the predictions K_{HB} and P_{HB} , representing medians over 2000 independent pilot data sets, excluding those yielding clipped predictions, and the prediction-accuracy $Q_{0.75, 0.90}$, the probability that any single sample-size estimate will yield true power in the range 75% and 90%. It is evident from Table 2 that it was possible for the median prediction to be good, e.g., for HL random-all $P_{HB} \sim 77\%$, while prediction-accuracy can be quite low, $Q_{0.75, 0.90} \sim 26\%$. The prediction-accuracy $Q_{0.75, 0.90}$ may be a more relevant quantifier of the performance of a sample-size estimation method than how close P_{HB} is to 80%. Due to the unusual spread evident in the histograms (the clipping at 2000 cases) for all except random-case generalizations, in the author's opinion the standard deviation of P_{HB} is not particularly meaningful. Since the better values for $Q_{0.75, 0.90}$ were typically $\sim 50\%$, from a practical point of view conditions yielding $Q_{0.75, 0.90} \sim 50\%$ are referred to as making "reasonable" predictions.

Under ideal conditions (many readers in the pilot study) the DBM-MRMC based HB method overestimated the number of cases. The estimates of $\sigma_{\tau R}^2$ had the most variability, particularly when $\sigma_{\tau R}^2$ was large, i.e., for the HL simulator. Large variability and a tendency to overestimate the number of cases can lead to gross overestimates. The large peaks at 2000 in Fig. 5 and Fig. 6 represent gross overestimates. The fact that gross overestimates can occur is not unique to this study. In a non-ROC context an example is given [30] where the true sample-size was 100 cases but for pilot studies with 20 cases about 22% of sample-size estimates exceeded 1000 cases.

For 5 readers in the pilot study the HB method made reasonable predictions when reader variability was low and generalization to either cases or readers (not both) was desired. One can reduce the variability of $\sigma_{\tau R}^2$ by using more readers in the pilot study. The largest benefits were realized for LH, where 15 readers were enough to yield prediction accuracy $> 50\%$ under all generalization conditions, but the benefits were lesser for HL. For the larger reader variability HL simulator even 30 readers and 100 cases (6000 interpretations) were barely enough for a reasonable prediction for random-all and not enough for random-readers. Sample-size estimation under large reader variability conditions may be problematical for the DBM-MRMC based HB method. Study designers may wish to compare the HB method predictions to those of other methods and to sample-sizes used in previous similar studies.

This study assumed that the true effect-size was known. As noted earlier, in practice the effectsize is unknown and one has to make an educated guess. Since the effect-size enters the sample-size equations as the square, the effect of an incorrect choice would be large. The results of this study are specific to the DBM-MRMC based HB method and to two variance structures. Other sample-size estimation methods, e.g., the Obuchowski-Rockette method, could behave differently and need to be examined. Roe and Metz have listed other variance structures, e.g., high reader variability and high case variability, and different AUC values (0.702 and 0.962), the effects of which were not studied. This study assumes that the LH and HL simulators are representative of real clinical data sets with low reader variability and high case variability, and the converse, respectively. The simulator parameters were obtained from the Roe and Metz [17] paper and it is possible that the HL simulator parameters exaggerates reader variability for HL (i.e., the readers are not as variable as the parameters suggest). If that is the case the HB-methods predictions would be more accurate than the numbers reported in this study. Imaging technology has changed considerably since 1997 and in the author's opinion the variance structures need to be updated, especially for the newer modalities (e.g., digital breast tomosynthesis) that did not exist in 1997. Simulators need to be designed to represent different diagnostic tasks, modalities, readers with different expertise levels and different strata of case (e.g., from different institutions). Since even pilot studies are expensive to conduct, it is desirable that observer study ratings data obtained by different investigators be made publicly available.

CONCLUSIONS

The HB method tends to overestimate the number of cases, i.e., it errs on the conservative side. Random-case generalization had the highest prediction accuracy. Provided about 15 readers are used in the pilot study with 100 cases, the HB-method performed reasonably under low reader variability conditions. 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.

Acknowledgments

The author is grateful to Hong-Jun Yoon, MSEE, for programming support on this project. This work was supported in part by grants from the Department of Health and Human Services, National Institutes of Health, R01-EB005243 and R01-EB008688.

APPENDIX A

Dataset simulator for random-readers and cases

The Roe and Metz random-all simulator [17] is described by:

$$z_{ijkl} = \mu_l + (\Delta\mu)_{il} + R_{jl}^z + C_{kl}^z + (\mu R)_{ijl}^z + (\mu C)_{ikl}^z + (RC)_{jkl}^z + (\mu RC)_{ijkl}^z. \quad \text{Eqn. A1}$$

To simulate z-samples for random-cases one sets $(\sigma_R^z)^2 = (\sigma_{\tau R}^z)^2 = 0$ which does not affect the normalization requirement $(\sigma_C^z)^2 + (\sigma_{\tau C}^z)^2 + (\sigma_{RC}^z)^2 + (\sigma_{\epsilon}^z)^2 = 1$. The simulation model was

$$z_{ijkl} = \mu_l + (\Delta\mu)_{il} + C_{kl}^z + (\tau C)_{ikl}^z + (RC)_{jkl}^z + \epsilon_{ijkl}^z. \quad \text{Eqn.A2}$$

To generate data for random-readers and fixed cases one sets $(\sigma_c^z)^2 = (\sigma_{\tau c}^z)^2 = 0$. However, since $(\sigma_{RC}^z)^2 + (\sigma_\varepsilon^z)^2 < 1$, to restore normalization the variances $(\sigma_{RC}^z)^2$ and $(\sigma_\varepsilon^z)^2$ implicit in Eqn. A3 were inflated by dividing by $(\sigma_{RC}^z)^2 + (\sigma_\varepsilon^z)^2$, resulting in new variances $(\sigma_a^z)^2$ and $(\sigma_b^z)^2$ defined by

$$\begin{aligned} (\sigma_a^z)^2 &= \frac{(\sigma_{RC}^z)^2}{(\sigma_{RC}^z)^2 + (\sigma_\varepsilon^z)^2} \\ (\sigma_b^z)^2 &= \frac{(\sigma_\varepsilon^z)^2}{(\sigma_{RC}^z)^2 + (\sigma_\varepsilon^z)^2}. \end{aligned} \quad \text{Eqn. A3}$$

Note that $(\sigma_a^z)^2 + (\sigma_b^z)^2 = 1$ ensuring that the simulator is properly normalized. The simulation model was

$$z_{ijkt} = \mu_t + (\Delta\mu)_{it} + R_{jt}^z + (\tau R)_{ijt}^z + (RC)_{jkt}^z + \varepsilon_{ijkt}^z. \quad \text{Eqn. A4}$$

where

$$\begin{aligned} (RC)_{jkt} &\sim N(0, (\sigma_a^z)^2) \\ \varepsilon_{jkt} &\sim N(0, (\sigma_b^z)^2). \end{aligned} \quad \text{Eqn. A5}$$

APPENDIX B

Hillis Berbaum sample-size estimation formulae for random-all

The non-centrality parameter (nc) is (R = number of readers and K = number of cases in the pivotal study)

$$nc = (0.5RK(\Delta A)^2) / (K\sigma_{\tau R}^2 + R\sigma_{\tau C}^2 + \sigma_\varepsilon^2). \quad \text{Eqn. B1}$$

As noted earlier in connection with Table 1, $\sigma_{\tau R}^2$ was more variable than $\sigma_{\tau C}^2$ and σ_ε^2 . Although $\sigma_{\tau C}^2$ and σ_ε^2 are larger than $\sigma_{\tau R}^2$, the effect of an overestimate of $\sigma_{\tau R}^2$ is amplified by the fact that it is multiplied by K , which could be in the hundreds, while $\sigma_{\tau C}^2$ is multiplied by a much smaller number, the number of readers R in the pivotal study. Therefore when an overestimate of $\sigma_{\tau R}^2$ occurs the non-centrality parameter decreases and the power predicted by the HB method will be smaller and the method will compensate by increasing the number of cases.

If $\sigma_{\tau C}^2 > 0$ the denominator degrees of freedom (ddf) is

$$ddf = (R - 1) (K\sigma_{\tau R}^2 + R\sigma_{\tau C}^2 + \sigma_\varepsilon^2)^2 / (K\sigma_{\tau R}^2 + \sigma_\varepsilon^2)^2. \quad \text{Eqn. B2}$$

If $\sigma_{\tau C}^2 \leq 0$

$$ddf = (R - 1).$$

Eqn. B3

The critical value of the F-distribution is $F_{1-\alpha;1,ddf}$ and a significant result is obtained when the observed value of the F-statistic exceeds the critical value. According to B2, when an overestimate of $\sigma_{\tau R}^2$ occurs ddf approaches $R - 1$. The critical value $F_{1-\alpha;1,ddf}$ increases as ddf decreases: e.g., $F_{0.95;1,2} = 18.6$ and $F_{0.95;1,1} = 161$. Therefore when ddf is small it is less likely that the observed F-statistic will exceed the critical value. The power predicted by the HB method equations will be smaller and the method will compensate by increasing the number of cases. If the overestimate of $\sigma_{\tau R}^2$ is sufficiently large clipping will occur.

The median non-centrality parameter was smaller (67%) and the median ddf was smaller (~48%). The median ddf for the clipped predictions was 9.3 and when $\sigma_{\tau R}^2 < 0$, ddf could be smaller than 9. The critical value $F_{1-\alpha;1,ddf}$ was larger (~1.2 times). The statistical power is

$$power \approx \text{Prob}(F_{1,ddf;nc} > F_{1-\alpha;1,ddf}).$$

Eqn. B4

where $F_{1,ddf;nc}$ is a sample from the non-central F-distribution with non-centrality parameter nc and numerator and denominator degrees of freedom equal to 1 and ddf, respectively. The non-central F-distribution used was from the library of cumulative density functions available at the web site (http://people.sc.fsu.edu/~burkardt/f_src/dcdflib/dcdflib.html). The function $F_{1-\alpha;1,ddf}$ was from IDL software: Version 6.4 Win32 (x86), available from ITT Visual Information Solutions, Boulder, CO.

For random-case and random-reader analyses, for the non-centrality parameter one sets $\sigma_{\tau R}^2 = 0$ and $\sigma_{\tau C}^2 = 0$, respectively, and for the denominator degrees of freedom one sets $ddf = (K_{PIV} - 1)$ and $ddf = (R_{PIV} - 1)$, respectively.

REFERENCES

1. ICRU. Statistical Analysis and Power Estimation. JOURNAL OF THE ICRU 2008:37–40.
2. Lenth RV. Some Practical Guidelines for Effective Sample Size Determination. The American Statistician 2001;55(3):187–193.
3. Obuchowski NA. How Many Observers Are Needed in Clinical Studies of Medical Imaging? Am. J. Roentgenol 2004;182(4):867–869. [PubMed: 15039154]
4. Obuchowski N. Determining sample size for roc studies: what is reasonable for the expected difference in tests in ROC areas? Academic Radiology 2003;10(11):1327–1328. [PubMed: 14626309]
5. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148(3):839–43. [PubMed: 6878708]
6. Obuchowski NA. Computing Sample Size for Receiver Operating Characteristic Studies. Inv Radiol 1994;29(2):238–243.
7. Obuchowski NA, McLish DK. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. Statistics in Medicine 1997;16:1529–1542. [PubMed: 9249923]
8. Metz CE. ROC Methodology in Radiologic Imaging. Investigative Radiology 1986;21(9):720–733. [PubMed: 3095258]
9. Obuchowski NA. Multireader, Multimodality Receiver Operating Characteristic Curve Studies: Hypothesis testing and Sample Size Estimation Using an Analysis of Variance Approach with dependent Observations. Acad Radiol 1995;2:S22–S29. [PubMed: 9419702]
10. Hillis SL, Berbaum KS. Power Estimation for the Dorfman-Berbaum-Metz Method. Acad. Radiol 2004;11(11):1260–1273. [PubMed: 15561573]

11. Beiden SV, Wagner RF, Campbell G. Components-of Variance Models and Multiple-Bootstrap Experiments: An Alternative Method for Random-Effects, Receiver Operating Characteristic Analysis. *Academic Radiology* 2000;7(5):341–349. [PubMed: 10803614]
12. Obuchowski NA. Sample size calculations in studies of test accuracy. *Statistical Methods in Medical Research* 1998;7(4):371–392. [PubMed: 9871953]
13. Obuchowski NA. Sample Size Tables For Receiver Operating Characteristic Studies. *Am. J. Roentgenol* 2000;175(3):603–608. [PubMed: 10954438]
14. Zhou, X-H.; Obuchowski, NA.; McClish, DK. *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons; 2002.
15. Pepe, MS. *Oxford Statistical Science Series*. Oxford University Press; New York: 2003. The statistical evaluation of medical tests for classification and prediction..
16. Obuchowski N. Reducing the Number of Reader Interpretations in MRM Studies. *Acad Radiol* 2009;16:209–217. [PubMed: 19124107]
17. Roe CA, Metz CE. Dorfman-Berbaum-Metz Method for Statistical Analysis of Multireader, Multimodality Receiver Operating Characteristic Data: Validation with Computer Simulation. *Acad. Radiol* 1997;4:298–303. [PubMed: 9110028]
18. Dorfman, DD.; Berbaum, KS.; Lenth, RV.; Chen, YF. *Medical Imaging 1999: Image Perception and Performance*. SPIE; San Diego, CA, USA: 1999. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: split-plot experimental design..
19. Obuchowski NA, Rockette HE. Hypothesis Testing of the Diagnostic Accuracy for Multiple Diagnostic Tests: An ANOVA Approach with Dependent Observations. *Communications in Statistics: Simulation and Computation* 1995;24:285–308.
20. Hillis SL, Berbaum KS. Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification. *Acad Radiol* 2005;12(12):1534–1541. [PubMed: 16321742]
21. Dorfman DD, Berbaum KS, Lenth RV, Chen YF, Donaghy BA. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: factorial experimental design. *Acad Radiol* 1998;5(9):591–602. [PubMed: 9750888]
22. Matsumoto M, Nishimura T. Mersenne Twister - a 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulation* 1998;8(1):3–30.
23. Galassi, M.; Davies, J.; Theiler, J.; Gough, B.; Jungman, G.; Booth, M.; Rossi, F. *GNU Scientific Library Reference Manual*. 1.6 ed.. Network Theory Limited; Bristol, UK: 2005.
24. Dorfman DD, Berbaum KS, Metz CE. ROC characteristic rating analysis: Generalization to the Population of Readers and Patients with the Jackknife method. *Invest. Radiol* 1992;27(9):723–731. [PubMed: 1399456]
25. Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. *Acad Radiol* 2008;15(5):647–661. [PubMed: 18423323]
26. Roe CA, Metz CE. Variance-Component Modeling in the Analysis of Receiver Operating Characteristic Index Estimates. *Acad. Radiol* 1997;4(8):587–600. [PubMed: 9261459]
27. Hillis, SL.; Berbaum, KS. *MRMC Sample Size Program User Guide*. [Dec 28, 2009]. available from <http://perception.radiology.uiowa.edu>, last
28. Berbaum, KS.; Metz, CE.; Pesce, LL.; Schartz, KM. *DBM MRMC User's Guide, DBM-MRMC 2.1 Beta Version 2*. [Dec 28, 2009]. available from http://www-radiology.uchicago.edu/cgi-bin/roc_software.cgi and from <http://perception.radiology.uiowa.edu>, last
29. Burgess AE. Comparison of receiver operating characteristic and forced choice observer performance measurement methods. *Med. Phys* 1995;22(5):643–655. [PubMed: 7643805]
30. Kreaemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Arch Gen Psychiatry* 2006;63:484–489. [PubMed: 16651505]

Fig. 1.

This shows power calibration data points (open circles) for LH random-all, effect size = 0.06 and 10 readers in the pivotal study. Each data point was obtained from 2000 AH simulations. The solid line, the power interpolation function, is a 5-parameter exponential rise to maximum function $P = a + b \cdot (1 - \exp(-c \cdot K)) + d \cdot (1 - \exp(-e \cdot K))$. P is defined as the true power.

Fig. 2.

This is similar to Fig. 1 except it applies to HL random-all. The solid line is a 4-parameter exponential rise to maximum function $P=a*(1-\exp(-b*K))+c*(1-\exp(-d*K))$.

Fig. 3.

This shows interpolation functions for random-all, random-cases and random-readers for the LH and HL variance structures. Random-all always yielded the lowest curve implying more cases are needed for 80% power, $K_{0.80} = 162$. The random-cases curve is slightly upward-left shifted relative to random-all, implying a small reduction in number of cases, $K_{0.80} = 149$, but the random-readers curve is considerably upward-left shifted, implying a large reduction, $K_{0.80} = 41$. See text for explanation.

Fig. 4.

This shows interpolation functions for random-all, random-cases and random-readers for the LH and HL variance structures. Random-all always yielded the lowest curve implying more cases are needed for 80% power, $K_{0.80} = 162$. The random-cases curve is slightly upward-left shifted relative to random-all, implying a small reduction in number of cases, $K_{0.80} = 149$, but the random-readers curve is considerably upward-left shifted, implying a large reduction, $K_{0.80} = 41$. The observed behavior follows from the low reader variability and high case variability of this simulator.

Fig. 5.

This figure shows the normalized histogram of K_i ($i = 1, 2, \dots, 2000$) for the LH simulator under the random-all condition. Each value of i corresponds to an independent pilot data set. The area under the histogram between the lines labeled $K_{0.75}$ ($= 141$) and $K_{0.90}$ ($= 226$) is the prediction-accuracy index $Q_{0.70, 0.95} = 38\%$; $K_{HB} = 225$ and $K_{0.80} = 162$. The peak at 2000 cases representing the clipped predictions which contribute 13% to the area (f_{max}).

Fig. 6.

This is similar to Fig. 5 except it applies to the HL simulator for which $K_{0.75} = 141$, $K_{0.90} = 572$, $Q_{0.70, 0.95} = 26\%$, $K_{HB} = 159$, $K_{0.80} = 190$ and $f_{\max} = 39\%$. The small underestimate changes to an overestimate when more readers are included in the pilot study.

Fig. 7.

This figure is similar to Fig. 5 except it applies to the LH simulator under the random-case condition for which $K_{0.75} = 131$, $K_{0.90} = 205$, $Q_{0.70, 0.95} = 53\%$, $K_{HB} = 194$, $K_{0.80} = 149$ and $f_{\max} = 0$. For random cases there were no instances of clipping (this was also true for HL).

Fig. 8.

This figure is similar to Fig. 7 except it applies to the HL simulator under the random-case condition for which $K_{0.75} = 62$, $K_{0.90} = 95$, $Q_{0.70, 0.95} = 48\%$, $K_{HB} = 92$, $K_{0.80} = 69$ and $f_{\max} = 0$.

Table 1

Listed are the z-sample and the pseudo-value variance components for the two simulators used in this work. The pseudo-value variance components represent grand averages of multiple (~5) independent medians obtained from DBM-MRMC analysis of 2000 pilot data sets, each with 5 readers and 100 cases. The coefficients of variance (CVs) of the medians are listed in parentheses. Only the last three pseudo-value variance components are relevant to sample-size estimation. The CV of $\sigma^2_{\tau R}$ was larger than that of $\sigma^2_{\tau C}$ or σ^2_e . The CV of $\sigma^2_{\tau R}$ was about 2.4 times as large for HL than for LH.

z-sample variance components						
Simulator	$(\sigma_R^z)^2$	$(\sigma_C^z)^2$	$(\sigma_{RC}^z)^2$	$(\sigma_{\tau R}^z)^2$	$(\sigma_{\tau C}^z)^2$	$(\sigma_e^z)^2$
LH	0.0055	0.3	0.2	0.0055	0.3	0.2
HL	0.0300	0.1	0.2	0.0300	0.1	0.6
pseudo-value variance components						
Simulator	σ_R^2	σ_C^2	σ_{RC}^2	$\sigma_{\tau R}^2$	$\sigma_{\tau C}^2$	σ_e^2
LH	2.818E-4 (1.78E-1)	5.346E-2 (1.21E-2)	4.675E-2 (7.16E-3)	1.974E-4 (2.72E-1)	3.896E-2 (1.26E-2)	4.167E-2 (2.54E-3)
HL	1.689E-3 (5.32E-2)	1.544E-2 (1.37E-2)	7.235E-2 (6.86E-3)	4.755E-4 (6.37E-2)	1.057E-2 (2.03E-2)	9.682E-2 (3.22E-3)

Table 2

This table summarizes the predictions of the HB method for Monte Carlo simulated pilot data sets with 5 readers, 50 normal and 50 abnormal cases, two modalities and effect-size $\Delta A = 0.06$. Listed are the number of cases $K_{0.80}$ for 80% true power; the median HB-predicted number of cases K_{HB} ; the corresponding true power P_{HB} ; the clipping fraction $f_{\max}$ for $K_{\text{MAX}} = 2000$ and the prediction-accuracy $Q_{0.75,0.90}$. This table shows that prediction-accuracy was higher under low reader variability conditions (LH) and for random-cases, for which reader variability is irrelevant. Prediction-accuracy was lower under high reader variability conditions (HL). On the whole the HB method overestimates the number of cases. The two underestimates under HL changed to overestimates when the number of readers was increased sufficiently, see Table 3.

Simulator	Generalization	$K_{0.80}$	K_{HB}	P_{HB}	$f_{\max}$	$Q_{0.75,0.90}$
LH						
	ALL	162	225	0.899	0.133	0.377
	CASES	149	194	0.886	0.000	0.528
	READERS	41	47	0.845	0.046	0.447
HL						
	ALL	190	159	0.772	0.389	0.260
	CASES	69	92	0.891	0.000	0.480
	READERS	138	80	0.704	0.383	0.208

Table 3

This table shows the effect of increasing the numbers of readers in the pilot study R_{PIL} . The prediction-accuracy $Q_{0.75,0.90}$ increased as R_{PIL} increased implying more pilot study readers are always helpful. The largest benefits were realized for LH, where 15 readers were enough to yield prediction accuracy > 50% under all generalization conditions, but the benefits were lesser for HL. For the same R_{PIL} and generalization scenario, $Q_{0.75,0.90}$ was larger for LH than for HL. Under the ideal condition of many readers in the pilot study the HB-method was conservative under all conditions and prediction-accuracy approached 62%.

Simulator	Generalization	R_{PIL}	K_{HB}	P_{HB}	f_{max}	$Q_{0.75,0.90}$
LH	ALL	5	225	0.899	0.133	0.377
		10	219	0.892	0.049	0.436
		15	219	0.892	0.019	0.483
		20	217	0.889	0.005	0.520
		30	219	0.892	0.001	0.509
	READERS	5	47	0.845	0.046	0.447
		10	51	0.870	0.011	0.540
		15	51	0.870	0.002	0.573
		20	51	0.870	0.001	0.592
		30	53	0.881	0.000	0.620
HL	ALL	5	159	0.772	0.389	0.260
		10	194	0.803	0.395	0.320
		15	204	0.810	0.403	0.355
		20	219	0.819	0.418	0.379
		30	283	0.847	0.367	0.455
	READERS	5	80	0.704	0.383	0.208
		10	118	0.776	0.392	0.303
		15	147	0.809	0.389	0.361
		20	160	0.821	0.390	0.342
		30	174	0.831	0.410	0.387