HHS Public Access

Author manuscript

Proc SPIE Int Soc Opt Eng. Author manuscript; available in PMC 2020 May 09.

Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 2019 February ; 10952: . doi:10.1117/12.2513106.

Estimating latent reader-performance variability using the Obuchowski-Rockette method

Stephen L. Hillis[a], **Badera Al Mohammad**[b], **Patrick C. Brennan**[b]

[a]Departments of Radiology and Biostatistics, University of Iowa, Iowa City, IA, USA

[b]Department of Medical Imaging and Radiation Sciences, Faculty of Health Sciences, The University of Sydney, Lidcombe, NSW, Australia

Abstract

We describe how the Obuchowski-Rockette (OR) method of analysis for multi-reader diagnostic studies can be used to estimate the variability of latent reader-performance outcomes, such as the area under the ROC curve (AUC). For a specific reader the latent or true reader performance outcome can conceptually be thought of as the average of the estimates that would result if the reader were to read a very large number of case samples. We note that for the sample sizes used in typical diagnostic studies, the latent reader-performance outcome is equal to the observed outcome minus measurement error. An often-cited study that assesses the variability of various reader-performance outcomes, including the AUC, is the study by Craig Beam et. al., "Variability in the Interpretation of Screening Mammograms by US Radiologists," published in 1996. However, a problem with this type of study is that the variability estimates include measurement error. Thus this approach overestimates latent reader variability and gives variability estimates that are dependent on case sample size. The proposed method overcomes this problem. We illustrate the proposed method for 29 radiologists in Jordan, with each reading 60 chest computed tomography (CT) scans. Using the OR method we estimate the middle 95% range for latent AUC values to be 0.07; i.e., we estimate that 95% of radiologists differ by less than 0.07 in their ability to successfully discriminate between a pair of diseased and nondiseased cases. In contrast, the estimate for the 95% range for the observed AUCs is 0.18. Thus we see how the conventional method of describing variability of reader performance estimates can greatly overstate the variability of the true abilities of the readers.

Keywords

Variability; diagnostic radiology; Obuchowski-Rockette; reader performance; AUC

Further author information: (Send correspondence to S.L.H.) S.L.H.: steve-hillis@uiowa.edu, Telephone: 1 641-226-0968.

1. INTRODUCTION

Variability among radiologists with respect to their performance abilities has typically been described in terms of variability of corresponding observed performance estimates, such as sensitivity, specificity, and area under the receiver operating characteristic curve (AUC). An often cited example of this approach is the paper "Variability in the Interpretation of Screening Mammograms by US Radiologists" by Beam et al (Beam)[1] which has been cited 427 times according to Google Scholar. In the Beam study, 108 radiologists evaluated the mammograms of 79 women.

For example, Fig. 2 in the Beam paper displays the observed sensitivities and specificities for the readers in the Beam study. The sample ranges of the observed sensitivity and specificity (for the normal women, excluding women with benign biopsy results) percentages were 53.3 and 63.2, respectively, and the sample range of the AUCs was 0.205 (expressing the AUCs as the actual area under the ROC curve, not the percentage of area.) These results led the authors to make the following conclusion: "Our findings indicate that there is wide variability in the accuracy of mammogram interpretation in the population of US radiologists."

We believe, however, that the approach used by Beam et. al. is misleading because it describes variability of *estimated* (or *observed*) performance abilities that include measurement error. Such estimates will exhibit less variability as the number of cases from which they are computed increases, making it difficult to compare results from studies with different numbers of cases. More importantly, these estimates are more variable that the *true* (or *latent*) performance outcomes, which can conceptually be thought of as the average of the estimates that would be obtained if the readers were to read many case samples.

In this paper we show how variability of the latent performance outcomes can be easily estimated using the well established Obuchowski-Rockette (OR)[2] method for analyzing multi-reader diagnostic studies. In Section 2 we discuss the statistical basis for the proposed method, in Section 3 we describe a real data set that we will use for illustrating the method, in Section 4 we present results, and we make concluding remarks in Section 5.

2. STATISTICAL METHODS

In this section we describe how the OR method can be used to estimate the variability of the latent reader-performance outcomes, and compare this measure of variability with the often reported *observed variability*, i.e., variability of the individual reader performance estimates. The study design of interest in this paper involves multiple readers reading images from the same cases and assigning a confidence-of-disease rating to each case. We assume a reader-performance outcome, such as AUC, is computed for each reader based on the reader's confidence-of-disease ratings. The OR method is a general analysis method for analyzing data from this and other study designs that allows conclusions to generalize to both the case and reader populations. Although the OR method is typically used to compare reader performance between different imaging modalities, in this paper we focus on the version of the OR model that is applicable when we are considering outcomes for a single modality.

The OR method was first proposed by Obuchowski and Rockette[2] in 1995. Hillis and colleagues[3,4] showed that the method developed by Dorfman, Berbaum and Metz[5] in 1992 was just a particular application of the OR method and that the OR model had the advantage of having much more interpretable parameters. In addition, Hillis[6] proposed a new formula for computing degrees of freedom for the OR model that improved its performance. It is this improved version of the original OR method which is presently in use and which we assume throughout. Hillis[7] derived the statistical properties for OR models corresponding to several different study designs, including the one-modality design considered in this paper.

2.1 Estimating latent reader-performance variability using the Obuchowski-Rockette method

2.1.1 The one-modality OR model—

Let $\hat{\theta}_j$ denote the reader performance estimate (e.g., AUC estimate) for reader *j*. The Obuchowski-Rockette (OR) model for the study design of interest in this paper, where each reader reads all of the cases using the same imaging method, is given by

$$\hat{\theta}_j = \mu + R_j + \varepsilon_j, \tag{1}$$

where μ is the population mean, R_j denotes the random effect of reader *j*, and ε_j is the error term (see Ref. 7 for a derivation of the properties of this model). Here we are assuming that both cases and readers are treated as random samples. We assume that the R_j are mutually independent and normally distributed with mean zero and variance σ_R^2. The ε_j are assumed to be normally distributed with mean zero and variance σ_ε^2, and are assumed to be independent of the reader effects.

The quantity $\mu + R_j$ is the latent performance value for reader *j*, free of any measurement error. Letting $\sigma_R = \sqrt{\sigma_R^2}$ it follows that the variance and standard deviation of the latent reader performance values, the $\mu + R_j$ values, across the population of readers are given by σ_R^2 and σ_R, respectively. The error term ε_j can be interpreted as measurement error attributable to the random selection of cases, to the interactions between cases and readers, and to within-reader variability that describes how a given reader interprets the same image in different ways on different occasions.

Because each reader reads the same cases, we expect there to be a nonnegative correlation between the error terms when treating cases as random sampling units. This correlation between the reader performance values is incorporated into Model 1 by allowing the error terms to be have a nonnegative covariance, denoted by Cov_2, between each pair of error terms.

2.1.2 Estimates of the OR parameters—

The Obuchowski-Rockette analysis method provides estimates of σ_R^2 and Cov_2. The covariance is typically estimated by a resampling method, such as the jackknife, bootstrap, or the method of Delong et al.[8] Letting $\widehat{\text{Cov}}_2$ denote this estimate, the estimate for σ_R^2 is given by

$$\hat{\sigma}_R^2 = \mathrm{MS}(R) - \hat{\sigma}_\varepsilon^2 + \widehat{\mathrm{Cov}_2} \tag{2}$$

where MS (R) denotes the mean square due to readers; i.e.,

$$\mathrm{MS}(R) = \frac{1}{r-1}\sum_{j=1}^{r}\left(\hat{\theta}_j - \hat{\theta}_\bullet\right)^2, \tag{3}$$

where r is the number of readers and $\hat{\theta}_\bullet$ denotes the mean of the reader outcome estimates: $\hat{\theta}_\bullet = \frac{1}{r}\sum_{j=1}^{r}\hat{\theta}_j$. Letting $\hat{\sigma}_R$ denote the square root of $\hat{\sigma}_R^2$, we can estimate the interval which contains the middle 95% of the population latent read-performance values by

$$(\hat{\theta}_\bullet - 1.96\hat{\sigma}_R, \hat{\theta}_\bullet + 1.96\hat{\sigma}_R).$$

We will refer to the length of this interval as the 95% range; i.e.,

$$95\,\%\ \text{range} = 2(1.96)\hat{\sigma}_R. \tag{4}$$

Similarly, for comparison we will estimate the 95% range of the observed reader-performance outcomes, based on the sample standard deviation. To be approximately correct, these confidence-interval results assume that the averages of both the latent and observed reader performance measures are approximately normally distributed, that they have the same mean, and that their corresponding variability estimates are close to the true values.

Model (1) can be fit using available software for various reader performance outcomes. We used the freely available OR-DBM MRMC 2.51 software[9] to fit model (1) to our data.

2.2 Comparison with observed reader-performance variability

Often researchers report the sample variance of the reader-performance estimates. The sample variance is given by MS(R), defined by Eq. 3. This sample variance estimates the variance of the outcome when a randomly chosen reader reads the particular study sample of cases. As such, it is an estimate of the observed reader-performance variability for readers reading the study cases, and includes variability of the latent reader performance values plus measurement error attributable to within-reader variability and to reader-by-case interaction. (It does not include variability due to the random selection of cases because it treats the cases as fixed.) This estimate will always be at least as great as the estimate given in Eq. 2 because it will always be the case that $\hat{\sigma}_\varepsilon^2 \geq \widehat{\mathrm{Cov}_2}$, which implies $\mathrm{MS}(R) \geq \hat{\sigma}_R^2$. Moreover, we see from Eq. 2 that the OR estimate of the variance of the latent reader performance values, $\hat{\sigma}_R^2$, can be interpreted as the estimated variance, MS(R), of the reader performance estimates (treating cases as fixed), minus the estimate of measurement error variance given by $\left(\hat{\sigma}_\varepsilon^2 - \widehat{\mathrm{Cov}_2}\right)$. The reason for the subtraction of Cov_2 from $\hat{\sigma}_\varepsilon^2$ is because Cov_2 represents that

part of the error term for the OR model (which treats cases as random) attributable to the random selection of cases.

Because observed variability (estimated by MS(R)) includes measurement error, observed variability is not useful for describing variability in latent reader abilities, other than providing an estimated upperbound on the latent reader-performance variability.

3. EXAMPLE ILLUSTRATING THE PROPOSED APPROACH

Our example data set comes from the study described by Al Mohammad et al.[10] In this study 30 radiologists with varying experience levels (median = 7 years, minimum = 2 years and maximum = 30 years) were recruited from different hospitals in Jordan. Eleven radiologists were from a specialized cancer center and 19 were from non-specialized cancer centers (general hospitals). All readers read the same test set of 60 chest tomography (CT) cases, of which 30 were cancer cases and 30 were cancer-free cases. One radiologist was deemed an outlier and omitted from analyses, resulting in n = 29 radiologists.

Readers were instructed to identify and locate all perceived malignant nodules, and to provide a confidence-of-disease for each perceived nodule. For cases that appeared to be normal, readers were told to assign a score of 1. A perceived benign nodule was to be given a score of 2, and a perceived malignant nodule was to be given a score ranging from 3 to 5, with a higher score indicating higher confidence of malignancy. Readers were told to ignore abnormalities other than lung nodules or any other findings suspicious for lung cancer. Further details about the study are available in Al Mohammad et al.[10]

For our purposes we consider only the AUC and sensitivity (for specificity = 0.794) reader-performance outcomes. Both measures were computed from each reader's ROC curve, which was estimated using the PROPROC[11-13] procedure. The outcomes and the reader variance component were computed using the OR-DBM software.[9]

4. RESULTS

Table 1 presents the estimates for the observed and latent reader-performance outcomes. For example, the first line gives the mean, median, minimum, maximum, range, standard deviation and 95% range for the observed AUC values; except for the 95% range, these values are just the usual sample statistics. The second line gives the mean, standard deviation and 95% range for the AUC latent values. The estimated mean for the latent values is just the sample mean of the observed values. However, for the latent values the standard deviation is equal to $\sqrt{\hat{\sigma}^2_R}$, which was computed by the OR-DBM software using Eqn. 2. For both lines the standard deviations (and hence corresponding 95% ranges) are much smaller for the latent AUC values (0.046 vs. 0.018), as well as for the sensitivity values (0.097 vs. 0.025), resulting in observed-to-latent standard-deviation ratios of 2.6 and 3.9, respectively.

Figure 1 displays plots of the observed AUC and sensitivity values and representative realizations of the latent-value distribution that show what the latent values might look like. In particular, the "latent" plots have the same sample mean and variance as the estimated mean and variance of the latent-value distributions. The plots of the representative

realizations of the latent-value distribution in Figure 2 were created by plotting the following Y values, defined by the transformation

$$Y_j = \frac{\hat{\theta}_j - \hat{\theta}_\bullet}{\sqrt{\mathrm{MS}(R)} / \hat{\sigma}_R}.$$

Figure 1 clearly conveys how much less variability exists between the readers with respect to true performance ability then is erroneously suggested by the observed value plots.

5. DISCUSSION

We have shown how the latent variability of readers can be easily estimated using the Obuchowski-Rockette method. The OR estimate for the variance of the latent reader outcomes is given by the OR estimate of the reader variance component, given by Eq. 2, which we showed (Sect. 2.2) can be interpreted as an estimate of the variance of the observed reader-performance values minus an estimate of the measurement error.

We showed in our example how describing variability of readers based on the sample variance of the observed reader outcomes can result in variability estimates considerably higher than the latent reader performance variability estimate. Thus such estimates, which do not exclude measurement error, can paint a misleading picture of large variation in performance abilities among radiologists, suggesting that there may be serious weaknesses in their training. The proposed method provides a solution to this problem by providing an estimate of the variability of the abilities of radiologists that is not inflated by measurement error.

ACKNOWLEDGMENTS

This research was supported in part by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under Award Number R01EB025174.

REFERENCES

[1]. Beam CA, Layde PM, and Sullivan DC, “Variability in the interpretation of screening mammograms by us radiologists: Findings from a national sample,” Archives of Internal Medicine 156(2), 209–213 (1996). [PubMed: 8546556]

[2]. Obuchowski NA and Rockette HE, “Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an anova approach with dependent observations,” Communications in Statistics-Simulation and Computation 24(2), 285–308 (1995).

[3]. Hillis SL, Obuchowski NA, Schartz KM, and Berbaum KS, “A comparison of the dorfman-berbaum-metz and obuchowski-rockette methods for receiver operating characteristic (ROC) data.,” Statistics in Medicine 24, 1579–1607 (2005). [PubMed: 15685718]

[4]. Hillis SL, Berbaum KS, and Metz CE, “Recent developments in the dorfman-berbaum-metz procedure for multireader study analysis,” Academic Radiology 15, 647–661 (2008). [PubMed: 18423323]

[5]. Dorfman DD, Berbaum KS, and Metz CE, “Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method,” Investigative Radiology 27(9), 723–731 (1992). [PubMed: 1399456]

[6]. Hillis SL, “A comparison of denominator degrees of freedom methods for multiple observer ROC analysis,” Statistics in Medicine 26(3), 596–619 (2007). [PubMed: 16538699]

[7]. Hillis SL, “A marginal-mean anova approach for analyzing multireader multicase radiological imaging data,” Statistics in Medicine 33(2), 330–360 (2014). [PubMed: 24038071]

[8]. DeLong ER, DeLong DM, and Clarke-Pearson DL, “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.,” Biometrics 44(3), 837–845 (1988). [PubMed: 3203132]

[9]. Schartz KM, Hillis SL, Pesce LL, and Berbaum KS “OR-DBM (Version 2.51),” [Computer Software], http://perception.radiology.uiowa.edu, accessed 10 February 2019.

[10]. Al Mohammad B, Hillis SL, Reed W, Alakhras M, and Brennan PC, “Radiologist performance in the detection of lung cancer using ct,” Clinical Radiology 74(1), 67–75 (2019). [PubMed: 30470412]

[11]. Pan XC and Metz CE, “The ”proper” binormal model: parametric receiver operating characteristic curve estimation with degenerate data,” Academic Radiology 4(5), 380–389 (1997). [PubMed: 9156236]

[12]. Metz CE and Pan XC, “”proper” binormal ROC curves: theory and maximum-likelihood estimation,” Journal of Mathematical Psychology 43(1), 1–33 (1999). [PubMed: 10069933]

[13]. Hillis SL, “Equivalence of binormal likelihood-ratio and bi-chi-squared ROC curve models,” Statistics in Medicine 35(12), 2031–2057 (2016). [PubMed: 26608405]

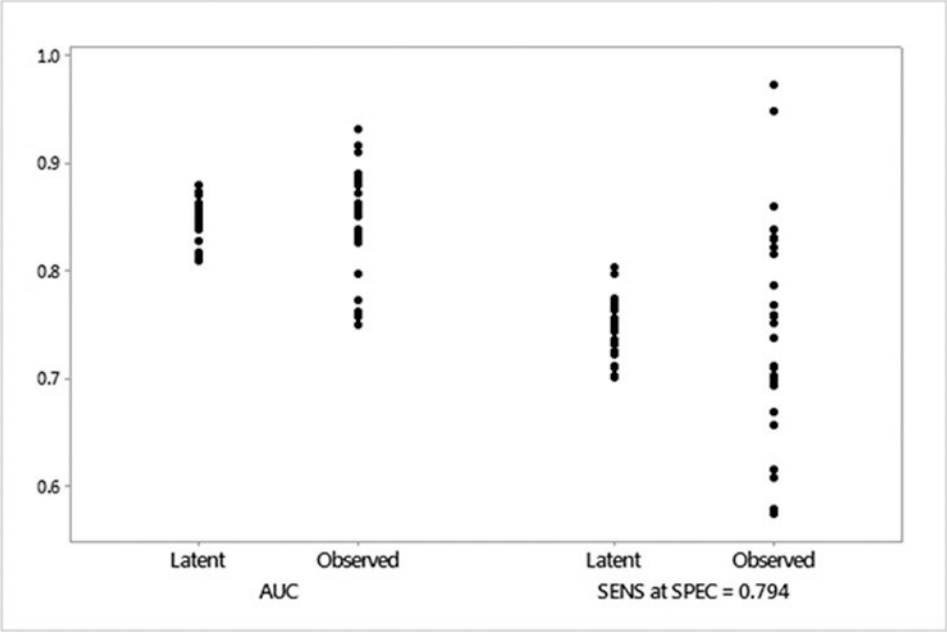

Figure 1.

Observed and representative latent values for AUC and sensitivity (at specificity = 0.794) outcomes. The representative latent values are not the true latent values, but rather are values such that the sample mean and variance are the same as the corresponding estimated mean and variance for the latent value distribution.

Table 1.

Estimates for observed and latent values for AUC and sensitivity (at specifity = 0.794) outcomes for n = 29 radiologists. The 95% range is equal to 2 × (1.96)× stddev.

Outcome	Type of results	Mean	Median	Min	Max	Range	Stddev	95% range
AUC	Observed values	0.846	0.851	0.749	0.931	0.182	0.046	0.180
	Latent values	0.846					0.018	0.070
SENS (spec = 0.794)	Observed values	0.744	0.751	0.572	0.973	0.401	0.097	0.381
	Latent values	0.744					0.025	0.098