Title

Comparative Statistical Properties of Expected Utility and Area Under the ROC Curve for Laboratory Studies of Observer Performance in Screening Mammography

Journal

Academic Radiology, 21(4)

ISSN

1076-6332

Authors

Abbey, Craig K
Gallas, Brandon D
Boone, John M
et al.

Publication Date

2014-04-01

DOI

10.1016/j.acra.2013.12.011

Peer reviewed

Published in final edited form as:
Acad Radiol. 2014 April ; 21(4): 481–490. doi:10.1016/j.acra.2013.12.011.

Comparative statistical properties of expected utility and area under the ROC curve for laboratory studies of observer performance in screening mammography

Craig K Abbey, PhD[1,*], **Brandon D Gallas, PhD**[2], **John M Boone, PhD**[3], **Loren T Niklason, PhD**[4], **Lubomir M Hadjiiski, PhD**[5], **Berkman Sahiner, PhD**[2], and **Frank W Samuelson, PhD**[2]

[1]Dept. of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106

[2]US FDA, Center for Devices and Radiological Health, Office of Science and Engineering Laboratories, 10903 New Hampshire Avenue WO62-3617, Silver Spring, MD 20993

[3]Dept. of Radiology, UC Davis Medical Center, 4860 Y Street, Suite 3100, Sacramento, CA 95817

[4]Hologic Inc., 35 Crosby Drive, Bedford, MA 01730

[5]Dept. of Radiology, University of Michigan Comprehensive Cancer Center, 1500 East Medical Center Drive, MIB C476, Ann Arbor, MI 48109

Abstract

Rationale and Objectives—Our objective is to determine whether expected utility (EU) and the area under the ROC (AUC) are consistent with one another as endpoints of observer performance studies in mammography. These two measures characterize ROC performance somewhat differently. We compare these two study endpoints at the level of individual reader effects, statistical inference, and components of variance across readers and cases.

Materials and Methods—We reanalyze three previously published laboratory observer performance studies that investigate various x-ray breast imaging modalities using EU and AUC. The EU measure is based on recent estimates of relative utility for screening mammography.

Results—The AUC and EU measures are correlated across readers for individual modalities (r = 0.93) and differences in modalities (r = 0.94 to 0.98). Statistical inference for modality effects based on multi-reader multi-case analysis is very similar, with significant results ($p < 0.05$) in exactly the same conditions. Power analyses show mixed results across studies, with a small increase in power on average for EU that corresponds to approximately a 7% reduction in the number of readers. Despite a large number of crossing ROC curves (59% of readers), modality effects only rarely have opposite signs for EU and AUC (6%).

Conclusions—We do not find any evidence of systematic differences between EU and AUC in screening mammography observer studies. Thus, when utility approaches are viable (i.e. an appropriate value of relative utility exists), practical effects such as statistical efficiency may be used to choose study endpoints.

*Corresponding Author. Email: abbey@psych.ucsb.edu. Phone 805 893 3853..

Keywords

Expected utility; area under the ROC curve; observer performance studies

1. Introduction

Laboratory observer performance studies utilizing receiver operating characteristic (ROC) methodology have become a mainstay for demonstrating improvements in technology or methodology for radiological imaging (1-5). These studies are now found widely in the radiological literature and are often used as evidence in submissions to regulatory agencies such as the US FDA (6). Most commonly, ROC studies use a rating of suspicion, e.g. probability of malignancy, to determine the tradeoff between sensitivity and false-positive fraction in a diagnostic task. In order to generalize results to the population of patients and readers, a substantial effort is required to collect a sample of relevant cases and to evaluate a sample of readers from which inferences regarding the imaging modality are obtained (7-10).

These inferences are based on an index, or figure of merit, extracted from the ROC curve. The figure of merit summarizes an ROC curve with a single number representing overall performance that can be compared across readers, cases, and modalities. The predominant figure of merit for ROC studies of observer performance has been the area under the curve (AUC) which is the diagnostic sensitivity averaged over all possible false-positive fractions (11, 12). The AUC is independent of disease prevalence and can also be interpreted as a measure of class separability, since it represents the probability that a case from the abnormal population will considered more suspicious than a case from the normal population.

However, AUC has also been subjected to criticism since it is not directly related to diagnostic utility and misclassification costs (13, 14). In averaging sensitivity across the full range of possible false-positive fractions, AUC incorporates information at levels that may be well beyond what is considered reasonable for a given task (15, 16). This has led to the use of partial-area indices (pAUC) as alternative figures of merit along with sensitivity at a fixed false-positive fraction (15, 17). These have been less widely adopted, perhaps because of the ambiguity in defining the appropriate portion of the curve to be evaluated. Furthermore, AUC has more recently been criticized by Hand (18) who contends that it is fundamentally incoherent with respect to misclassification costs, requiring cost functions that depend on the classifier. These critiques suggest that findings may change if a utility-based endpoint is used instead of AUC.

Despite some attempts (19-21), utility-based approaches have not been widely adopted within the medical imaging observer-performance-evaluation community, at least in part for the reason suggested by Metz (3) and others (22), because they require that the values for the possible outcomes (true positive, false positive, true negative, false negative) are known or at least agreed upon. However, recent investigations in screening mammography (23) have used a technique suggested by Lusted (24) to estimate relative utility from large clinical trials (25) and observational studies (26). One of these studies gives an estimate of relative utility with a reported relative error of 14%, which is sufficiently precise to consider using in performance assessments of ROC data in terms of expected utility (EU).

In this work, we reanalyze previously published observer performance studies relevant to x-ray mammography to compare AUC and EU as study endpoints A previous study investigating statistical power in a simulation environment finds generally good statistical power for EU relative to AUC (27). The purpose of our analysis is to see if there are any

systematic differences in the results that depend on the figure of merit with real data. In all cases the studies use a fully-crossed multi-reader multi-case design, in which all readers score all cases in all modalities. We compare AUC and EU across readers, evaluate the inferences (i.e. p-values) for modality comparisons, and compare components of variance. We also investigate the frequency of crossing ROC curves and the frequency of opposite modality effects for AUC and EU; these statistics characterize how often the study results depend on the performance measure.

2. Materials and Methods

2.A. Figures of merit for observer performance in ROC studies

An ROC curve is a well-known characterization of decision making performance in a classification task (e.g. “recall for diagnostic workup” or “do not recall”). However, for the purpose of comparing different imaging modalities it is desirable to summarize the ROC curve with a single number, the figure of merit. This work considers two such figures of merit, the traditional area under the ROC curve (AUC), and an expected utility (EU) measure described below. Both are extracted from an estimated ROC curve fit to observer rating data.

Figure 1 shows how the two figures of merit are defined graphically. As suggested by its name, the AUC is determined from an ROC curve by calculating the area under it. When a parametric model is used for the ROC curve the area can generally be determined directly from the model parameters. Larger values of AUC indicate better performance.

The EU measure (27) is based on the utility of various task outcomes (true-positive, false-positive, etc.) and disease prevalence. These parameters are combined to define iso-utility lines in the ROC domain of true-positive fraction (TPF) and false positive fraction (FPF). The slope (β) of the iso-utility lines is defined by the relative utility of the task and the prevalence of disease (12, 22, 24). Let R be the set of all TPF and FPF points of an ROC curve, the EU measure is defined as

$$\mathrm{EU} = \max_{(\mathrm{TPF},\mathrm{FPF})\in R} (\mathrm{TPF} - \beta \mathrm{FPF}). \quad (1)$$

The point at which this maximum is achieved is referred to as the optimal operating point. As seen in Figure 1, EU may be interpreted graphically as the y-intercept of a line with slope β that passes through the optimal operating point. Larger values of EU indicate better performance.

It is clear from Equation 1 that the iso-utility slope must be known if a numerical value of EU is to be determined from a given ROC curve. We use a value of $\beta = 1.03$ as suggested in previous studies (28) estimating relative utility in screening mammography from large clinical trials in the United States, which is the country of origin for all the data we analyze. However, caution must be used in generalizing EU results to other medical tasks or other countries because of different disease prevalence or different weightings of decision outcomes.

2.B. Reader data

Three investigations of reader performance related to screening mammography were analyzed here for the purpose of comparing EU to AUC. All of the data we use were collected by the original authors under IRB approved protocols. Each investigation consisted of 1 to 3 studies comparing various imaging modalities. Each study was analyzed as a fully crossed factorial design in which all readers scored all cases in all modalities. Note that in a

few instances there were missing data in the form of a missing case score for a given reader and modality (0.4% of responses across all data). In these situations, the missing scores were filled in by taking the average score for the case across the remaining readers in that modality. Table 1 summarizes the general characteristics of the data.

2.B.1. DMIST Reader Studies—The Digital Mammographic Imaging Screening Trial (DMIST) reader performance studies (29, 30) were acquired as part of the DMIST project (25, 31) conducted by the American College of Radiology Imaging Network (ACRIN) and funded by the NIH. The larger purpose of this effort was a comparison of screen-film mammography – the standard of care at the time – to digital mammography systems that were emerging on the market. DMIST included both a prospective clinical trial as well as the retrospective reader studies used here. We analyze three reader performance studies that investigate devices from different manufacturers including GE (Senographe 2000D, GE Healthcare, Waukesha, WI), Fuji (Computed Radiography System for Mammography, FujiFilm Medical, Stamford, CT), and Fisher (SenoScan, Fischer Medical Technologies, Denver CO). For identification purposes, we identify these studies as D1, D2, and D3 respectively. Some data from a fourth manufacturer (Hologic) was available, but the limited number of available cases, 28 in total, made this study inadequate for our purposes. The data we use are identical that used in the primary publications (29, 30) for the three studies we analyzed.

Each study used independent patient images as well as readers that were largely independent (4 radiologists read in more than one study), and thus we analyzed them separately. Ground truth was established by biopsy or a negative follow-up mammogram. Each of the three DMIST studies compared screen-film mammography (which we identify as modality 1: M1), soft-copy digital mammography (M2), and hard-copy digital mammography (M3). Each modality was read independently of the others, with time between readings to negate any memory effects (29, 30), and scored on a 7-point malignancy scale. Readers generally used the entire categorical scale; the average number of categories used across readers was 6.9. The rate of missing responses was 0.5% (50 out of 9,918). Previous publications using this data have found no significant effect between screen-film mammography and soft-copy digital mammography (29), or between soft-copy digital mammography and hard-copy digital mammography (30).

2.B.2. University of Michigan CAD Study—The University of Michigan (UM) data consists of one study evaluating computer-aided diagnosis (CAD) using digitized screen-film mammograms (32). The study investigated discrimination of malignant and benign abnormalities. This is generally considered a diagnostic task, not a screening task. The distinction is important because the expected utility measure we use here is based on the relative utility of screening mammography. The appropriate relative utility for diagnostic mammography is not known to our knowledge. Nonetheless, the mammographic views and dose levels used in the study were consistent with screening, and therefore we will think of these images as a subpopulation containing abnormalities that would be encountered in screening. Furthermore, the purpose of this work is to compare EU and AUC results. For these reasons, we use the EU measure as defined above for the UM data.

All cases were biopsy proven. The three modalities of the study consist of mammography alone (M1), mammography before receiving a CAD malignancy score (M2), and mammography after receiving a CAD score (M3). M1 was read independently of M2 and M3, with time between readings to minimize any memory effects. The latter two modalities were read in a sequential paradigm in which the reader scored a case on the basis of the mammogram alone, and was then given the CAD input, and asked to re-score the case given this new information. Under the assumption that the observer is not affected by knowing that

a CAD score is coming, M2 may be considered a replication of M1, and we would not expect any difference in performance. Cases were scored on a 100-point malignancy scale (1% to 100%). On average readers used 48.9 of the 100 available categorical scores. There was no missing data in this study. Previous publication of this data (32) found no significant difference between M1 and M2 (i.e. mammography-alone and pre-CAD score). A significant improvement was found comparing M3 to either M1 or M2.

2.B.3. Hologic Tomosynthesis Studies—The Hologic data consists of two studies investigating the addition of digital breast tomosynthesis (DBT) images to DM for breast cancer screening (33). The two-view DM images (CC and MLO) were acquired at five participating sites using a commercial system (Selenia; Hologic, Bedford, MA). The additional DBT images were acquired on an investigational tomosynthesis system from the same manufacturer and used 15° tube rotation, 0.7mm aluminum filtration, 11 image taken over a 10 second acquisition time, and at a dose equivalent to DM . The first study (H1) compared digital mammography alone (M1) to digital mammography with 2-view DBT (M2). The second study (H2) compared digital mammography alone (M1), digital mammography with 1-view (MLO only) DBT (M2), and digital mammography with 2-view DBT (M3).

The two studies used independent readers, but there was some overlap in the cases. The 48 positive (i.e. cancer) cases used in H1 were also used in H2. The negative cases were selected at random from groups of different case types (negative at screening, negative at recall, and negative at biopsy), and may have repeated cases as well. Thus these two studies should not be considered independent. Positive cases were biopsy proven, and negative cases in women that did not undergo biopsy were verified by 1 year of observation.

All studies were read in a sequential paradigm in which the DM images (M1) for the case were scored first. The reader was shown the additional DBT views (in H1) or view (in H2) for M2 and asked to rescore the case based on the combined image data. In H2, the reader was then shown the final DBT view and asked again to rescore the case on the basis of all available image data for M3. The readers provided a 101-point probability of malignancy score (0%-100%) in addition to scores related to the Breast Imaging Reporting and Data System (BI-RADS). Our analysis focuses on the probability of malignancy scores. On average, readers used 17.2 of the 101 possible categorical scores. The rate of missing responses was 0.4% (31 of 8,394). Some preliminary results using EU from the Hologic studies have been previously presented (34).

2.C. ROC Analysis and Inference

In each of the studies described above, ROC curves were fitted by maximum likelihood to the categorical data of each reader in each modality. The contaminated binormal model (CBM) was used as the probability model for the study (35-37). The CBM posits a latent decision variable that is monotonically related to a standard normal distribution for normal cases, and a mixture of a standard normal and a shifted normal for abnormal cases. In our implementation (27), the two parameters of the model affecting the abnormal distribution are the contamination fraction, a, and the shift parameter u, which control the mixture of the two normal distributions and the degree of shift.

Pseudo-values for statistical inference were obtained by jackknifing and normalized as described by Hillis and Berbaum (10, 38). Statistical modeling and inference was conducted using the approach of Dorfman, Berbaum, and Metz (DBM), which consists of a three-way mixed effects analysis of variance (8). The DBM method was used to test for modality differences in the data using both AUC and EU and with significance defined at $p < 0.05$.

DBM was also used to obtain the components of variance estimates for comparison of the two figures of merit.

2.D. Modality effects and crossing ROC curves

Utility based measures have been argued for on the basis that ROC curves will often cross in evaluations of competitive systems (18). In such cases it would be possible for a comparison based on a utility measure to reach a different conclusion than one based on AUC. When one ROC curve is above another at every point between 0 and 1, modality effects for EU and AUC will generally be concordant (i.e. have the same sign) as long as the optimal operating points are not at the extreme points of the ROC curve (the points (0,0) or (1,1) in Figure 1). It is therefore of interest to investigate how often ROC curves cross in examples of actual published data, and how often AUC and EU result in effects with opposite signs.

Thus, we have evaluated the fraction of cases in which ROC curves cross, and the fraction in which the modality comparisons have opposite signs on a reader-by-reader basis. In the CBM model, two ROC curves, with parameters (a_1, u_1) and (a_2, u_2) are guaranteed to cross if and only if $(a_1 - a_2)(u_1 - u_2) > 0$. As a result, from the estimated parameters for two ROC curves it can be readily determined whether the two curves cross, and by comparing the EU and AUC figures of merit, we can see whether they have opposite signs.

3. Results

Table 2 gives results of the DBM analysis for the various studies. It reports the mean figure of merit for each modality in each study, as well as the DBM *p*-values for various modality comparisons. Note that we do not make any attempt to correct for multiple comparisons here because our purpose is not to evaluate the modalities, but rather to compare the inferences obtained under the two figures of merit. The values we report are consistent with values determined in the various publications associated with these studies, even though the published studies used different probability models for the fitted ROC curves and different approaches to handling missing data. On average, the absolute deviation between the AUCs in Table 2 and the published values for the corresponding study are 0.005. The single largest difference between them (D3: Modality 3) is 0.026. The *p*-values in Table 2 also lead to the same findings of significance that are in the published reports. This congruence suggests that the results are reasonably stable to the different procedures to analyze the data.

Figure 2 consists of various scatterplots comparing AUC and EU. Figure 2A shows the AUC (*x*-axis) and EU (*y*-axis) for each reader in each condition. The two performance measures appear to be highly correlated with an overall Pearson correlation coefficient of 0.93. Figure 2B-2D shows scatterplots of pairwise modality differences in AUC and EU for each reader. In each case the scatterplot is well fit by a line with slopes ranging from 1.37 to 1.73 and small offsets. Pearson correlation coefficients for these plots range from 0.94 to 0.98. Thus effect sizes appear to be somewhat amplified with the EU metric, although this fact needs to be put in the context of how variability scales between the two measures.

Table 3 gives the six DBM components of variance for AUC and EU in each study, consisting of the reader variance (R), case variance (C), the treatment by reader interaction (TR), the treatment by case interaction (TC), the reader by case interaction (RC) and the residual error (TRC). The components are scaled to the individual item level (modality, reader, and case). The values are determined from linear combinations of mean-square estimates used in the ANOVA model, and are truncated to zero when these estimates lead to negative values (10, 38). In all studies, the largest source of variance is the residual TRC variance, followed by the case variance, the reader by case interaction, and the treatment by

case interaction. Additional statistical properties, including standard errors of modality effects averaged across readers and cases are given in the appendix.

Figure 3 plots the rate of crossing ROC curves in each study along with the rate of opposite signs in the observed effects for the EU and AUC measures. The rate of crossing ROC curves is generally quite high, ranging from 33% to 94% with an average of 59%. Not surprisingly, rates of crossing ROC curves were higher in the DMIST studies where effect sizes were smaller, suggesting that the resulting ROC curves were usually closer together. However, it is notable that even in the UM and Hologic studies where significant effects were found, ROC curves still crossed at a fairly high rate.

In contrast, the rate of opposite effect signs between EU and AUC is fairly low, ranging from 0% to 14% across studies with an average of 6%. This finding is not surprising in light of Figure 2, where the vast majority of observed effect sizes (See Fig. 2B-2D) fall in the 1st and 3rd quadrants (upper right and lower left) indicating concordant effects. Furthermore, Figure 2 shows that the relatively small number of discordant effects (in the 2nd and 4th quadrant of the plots) occur for readers with quite small observed modality effects. Thus, for the EU measure used here, a large number of crossing ROC curves translates to relatively few discordant observations that occur when these noisy effect sizes are small.

4. Discussion

There has been debate about the use of AUC because of its tenuous connection to misclassification costs (13, 14, 18), and there is relatively little practical experience with utility-based methods in the kinds of ROC studies used to make claims about different imaging modalities. A notable exception is the work of by Halpern et al. (19), which analyzed previously published data over a range of possible ROC slopes and motivated our investigation. These authors treated the ROC slope in Equation 1 as a free parameter that ranged from 0.1 to 3. The ambiguity of the ROC slope, and by inference the diagnostic utility underlying the clinical task, was subsequently pointed out as a limiting factor for the approach by Metz (3). Our work focuses on screening mammography where we have additional knowledge of a clinically justified ROC slope (23, 28) along with additional tools for the evaluation of statistical power in ROC studies (39) that have been developed since the Halpern et al. publication. The purpose of our study was to see if there was any evidence for different study results or different statistical efficiency in the datasets available to us.

4.A. Effect sizes and components of variance

The comparative results in Figure 2 and Tables 2 and 3 establish basic properties of EU relative to AUC in these studies. The two measures are clearly correlated (Fig. 2), and lead to identical inference regarding the imaging modality (Table 2). EU generally appears to have larger effect sizes (Fig. 2), but also has larger components of variance as well (Table 3). For a direct comparison of EU and AUC components of variance, Figure 4 plots the ratio of the EU and AUC components across studies. The components of variance that are used in the DBM procedure to test for modality differences are indicated with an asterisk (*). Ratios are not plotted if either component was small (> 0.0001) or truncated to zero. This plot gives a sense of the inflation of variance for the EU measure relative to AUC.

If we use slopes of the linear relationships in Figure 2 as a guide, we would expect ratios in Figure 4 that are in the range of 1.9 to 3.0 to balance the increase in effect size. A ratio above this range suggests that the increased effect size for EU is not large enough to balance the increase in variability of that component. Conversely a ratio below this range suggests less relative variance for EU compared to AUC. The reader and case variance components are generally above this range. The interaction terms that are used for determining the

significance of modality differences (indicated by *) are variable, with ratios covering this range as well as above and below it.

The nominal slope (β) used to compute the EU measure has a small effect on these results. In addition to the slope of 1.03 used, we have evaluated the effect of a 20% increase or decrease in the ROC slope used to compute EU (data not shown). The different slopes have no effect on the pattern of significance in Table 2. The average slope of the line relating observed effects from AUC to those from EU (analogous to Figure 2 B-D) changes by less than 5%, and the intercept changes by less than 0.003. The average of the components of variance (analogous to Table 3) changes by less than 14%.

4.B. Power calculation

As a way of putting the relative inflation of effect size and variance in context, we have used them in a power calculation to hypothetically size an observer performance study. We used the approach of Hillis and Berbaum (39), which is based on hypothesized effect sizes and components of variance. The power calculation considered a comparison of two modalities with components of variance indicated by each of the studies considered. The effect size for the UM and Hologic studies was set to the largest observed difference between modalities in the study, for each endpoint. For example, in the UM study the AUC effect size was set to $\Delta AUC = 0.0496$ (M3 – M1 in Table 2) while the EU effect size was set to $\Delta EU = 0.0790$. In the DMIST studies, the observed effect sizes were relatively small relative to components of variance, and non-significant. In these studies we posited a default AUC modality effect of $\Delta AUC = 0.1$, and then use the linear relationships in Figure 2 to determine the EU modality effect. For example, in study D1 we obtain an EU modality effect of $\Delta EU = 0.135$ ($= 0.137 \times \Delta AUC - 0.002$). The number of cases in each power analysis was set to the number actually used in the study. The number of readers was varied over a range from 3 to 20.

Figure 5 gives the results of this power analysis. The plots show power as a function of the number of readers for AUC (Fig. 5A) and for EU (Fig. 5B). As expected, statistical power for both measures increased considerably as the number of readers was increased. To facilitate comparisons between AUC and EU, we also show a plot of the number of readers needed for 80% power in each study (Fig. 5C). This plot shows that the estimated number of readers needed varies considerably over the different studies, from 4 in study H1 using EU to 19 in study D3 using AUC. There is also variability in which figure of merit results in the lowest number of readers for a given study. In 4 of the 6 studies, EU results in fewer readers needed for 80% power, with the other 2 studies favoring AUC. Averaging across all studies, we find a 7% reduction of number of readers needed for 80% power with the EU measure.

4.C. Crossing ROC curves

As seen in Figure 3, the majority of modality comparisons (59%) ROC curves were crossing. However, this translated to relatively few observations in which modality effects changed sign (6%). Furthermore, inspection of Figure 2 suggests that the few observed sign changes occurred for very small observed effects, and Table 3 shows that these few differences did not alter the pattern of significance.

It is important to recognize the limitation of our analysis that focused on one particular clinical task. Nonetheless, within this domain the different endpoints did not appear to change the findings in any of the studies. Even though it is possible to construct examples in which the choice of AUC or EU can change which modality is considered best (13, 20), these results provides some evidence that the two are interchangeable in practice.

4.D. Summary and conclusions

In these three published observer performance studies related to breast cancer screening, acquired for diverse purposes, and with different experimental designs, we find remarkable concordance between the traditional AUC and an EU figure of merit that uses a relative utility derived from large scale clinical trials. The figures of merit themselves as well as differential effects between imaging modalities agree well across readers. There is little evidence of systematic differences in performance between AUC and EU, despite the fact that the majority of reader ROC curves are crossing in these studies, leaving open the possibility of discordant effects. Furthermore EU and AUC give identical inferences with regard to modality effects. Thus EU and AUC would appear to be interchangeable as endpoints of an observer performance study in these data.

In comparing the statistical properties of the two endpoints, we find that EU increases the effect size between modalities, but it also generally increases components of variance. Averaging over all studies considered suggests that there may be a small benefit in statistical power to the EU measure. This is consistent with previous simulation studies (27).

The investigation of EU described here has been limited to studies related to x-ray mammography for breast cancer screening. Furthermore, EU has been set to represent utility in screening mammography as derived from large clinical trials in the United States. Thus it is not guaranteed that the results will extend to different medical tasks, or substantially different misclassification costs that may occur in other countries. Nonetheless, the results provide some measure of confidence in using EU within the domain it has been tested in, and motivates the investigation of these extensions.

Acknowledgments

CKA: NSF (NSF-FDA SIR: CBET-1238502). JMB: NIH R01-EB002138

5. Appendix

The purpose of this appendix is to report some further statistical properties of the reader studies used in this work. It is hoped that these will give the interested reader additional information for comparing EU and AUC.

Table A1 shows the standard error of the reader and case averaged estimate of performance for a single modality, and the standard error of the difference between modalities. These can be computed from the components of Table 3, with appropriate normalization for the number of readers and cases in the study. We also give the correlation coefficient between reader and case averaged modality effects, which gives a sense of how much dependence arises as a function of the common readers and cases.

Table A1

Additional study statistics. The table shows standard error estimates for performance in a given modality (SE-Mod.), standard error estimates for modality differences (SE Dif.), and the Pearson correlation coefficient (CC) between modalities.

Study	Endpoint	SE-Mod.	SE-Dif.	CC
D1	AUC	0.036	0.029	0.69
	EU	0.058	0.041	0.75
D2	AUC	0.045	0.029	0.80

Study	Endpoint	SE-Mod.	SE-Dif.	CC
	EU	0.080	0.041	0.87
D3	AUC	0.049	0.049	0.50
	EU	0.072	0.059	0.66
UM	AUC	0.022	0.013	0.83
	EU	0.043	0.030	0.76
H1	AUC	0.031	0.027	0.63
	EU	0.054	0.039	0.74
H2	AUC	0.028	0.016	0.84
	EU	0.051	0.029	0.83

References

1. Goodenough DJ, Rossmann K, Lusted LB. Radiographic applications of receiver operating characteristic (ROC) curves. Radiology. 1974; 110(1):89–95. [PubMed: 4808546]
2. Metz CE. ROC methodology in radiologic imaging. Invest Radiol. 1986; 21(9):720–33. [PubMed: 3095258]
3. Metz CE. ROC analysis in medical imaging: a tutorial review of the literature. Radiol Phys Technol. 2008; 1(1):2–12. [PubMed: 20821157]
4. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology. 2003; 229(1):3–8. [PubMed: 14519861]
5. Shiraishi J, Pesce LL, Metz CE, Doi K. Experimental design and data analysis in receiver operating characteristic studies: lessons learned from reports in radiology from 1997 to 2006. Radiology. 2009; 253(3):822–30. [PubMed: 19864510]
6. Gallas BD, Chan HP, D'Orsi CJ, et al. Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. Acad Radiol. 2012; 19(4):463–77. [PubMed: 22306064]
7. Beiden SV, Wagner RF, Campbell G, Metz CE, Jiang Y. Components-of-variance models for random-effects ROC analysis: the case of unequal variance structures across modalities. Acad Radiol. 2001; 8(7):605–15. [PubMed: 11450961]
8. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. Invest Radiol. 1992; 27(9):723–31. [PubMed: 1399456]
9. Gallas BD. One-shot estimate of MRMC variance: AUC. Acad Radiol. 2006; 13(3):353–62. [PubMed: 16488848]
10. Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. Acad Radiol. 2008; 15(5):647–61. [PubMed: 18423323]
11. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143(1):29–36. [PubMed: 7063747]
12. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978; 8(4):283–98. [PubMed: 112681]
13. Hilden J. The area under the ROC curve and its competitors. Med Decis Making. 1991; 11(2):95–101. [PubMed: 1865785]
14. Hilden J. Evaluation of diagnostic tests - the schism. Society for Medical Decision Making Newsletter. 2004; (4):5–6.
15. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology. 1996; 201(3):745–50. [PubMed: 8939225]
16. Obuchowski NA, McClish DK. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. Stat Med. 1997; 16(13):1529–42. [PubMed: 9249923]

17. McClish DK. Analyzing a portion of the ROC curve. Med Decis Making. 1989; 9(3):190–5. [PubMed: 2668680]
18. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine learning. 2009; 77(1):103–23.
19. Halpern EJ, Albert M, Krieger AM, Metz CE, Maidment AD. Comparison of receiver operating characteristic curves on the basis of optimal operating points. Acad Radiol. 1996; 3(3):245–53. [PubMed: 8796672]
20. Sunshine J. Contributed Comment. Adacemic Radiology. 1995; 2(Suppl):S72–S4.
21. Wagner RF, Beam CA, Beiden SV. Reader variability in mammography and its implications for expected utility over the population of readers and cases. Med Decis Making. 2004; 24(6):561–72. [PubMed: 15534338]
22. Swets, JA.; Pickett, RM. Evaluation of diagnostic systems : methods from signal detection theory. New York: Academic Press; 1982.
23. Abbey CK, Eckstein MP, Boone JM. An equivalent relative utility metric for evaluating screening mammography. Med Decis Making. 2010; 30(1):113–22. [PubMed: 19706880]
24. Lusted, LB. Introduction to medical decision making. Springfield, Ill.; Thomas: 1968.
25. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med. 2005; 353(17):1773–83. [PubMed: 16169887]
26. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst. 2004; 96(24):1840–50. [PubMed: 15601640]
27. Abbey CK, Samuelson FW, Gallas BD. Statistical Power Considerations for a Utility Endpoint in Observer Performance Studies. Acad Radiol. 2013
28. Abbey CK, Eckstein MP, Boone JM. Estimating the relative utility of screening mammography. Med Decis Making. 2013; 33(4):510–20. [PubMed: 23295543]
29. Hendrick RE, Cole EB, Pisano ED, et al. Accuracy of soft-copy digital mammography versus that of screen-film mammography according to digital manufacturer: ACRIN DMIST retrospective multireader study. Radiology. 2008; 247(1):38–48. [PubMed: 18372463]
30. Nishikawa RM, Acharyya S, Gatsonis C, et al. Comparison of soft-copy and hard-copy reading for full-field digital mammography. Radiology. 2009; 251(1):41–9. [PubMed: 19332845]
31. Pisano ED, Gatsonis CA, Yaffe MJ, et al. American College of Radiology Imaging Network digital mammographic imaging screening trial: objectives and methodology. Radiology. 2005; 236(2): 404–12. [PubMed: 15961755]
32. Hadjiiski L, Chan HP, Sahiner B, et al. Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: an ROC study. Radiology. 2004; 233(1):255–65. [PubMed: 15317954]
33. Rafferty EA, Park JM, Philpotts LE, et al. Assessing Radiologist Performance Using Combined Digital Mammography and Breast Tomosynthesis Compared with Digital Mammography Alone: Results of a Multicenter, Multireader Trial. Radiology. 2012
34. Abbey CK, Samuelson FW, Gallas BD, Boone JM, Niklason LT. Statistical properties of a utility measure of observer performance compared to area under the ROC curve. Book Statistical properties of a utility measure of observer performance compared to area under the ROC curve. City: International Society for Optics and Photonics. 2013; 86730D-D-9
35. Dorfman DD, Berbaum KS. A contaminated binormal model for ROC data: Part III. Initial evaluation with detection ROC data. Acad Radiol. 2000; 7(6):438–47. [PubMed: 10845403]
36. Dorfman DD, Berbaum KS. A contaminated binormal model for ROC data: Part II. A formal model. Acad Radiol. 2000; 7(6):427–37. [PubMed: 10845402]
37. Dorfman DD, Berbaum KS, Brandser EA. A contaminated binormal model for ROC data: Part I. Some interesting examples of binormal degeneracy. Acad Radiol. 2000; 7(6):420–6. [PubMed: 10845401]
38. Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. Stat Med. 2005; 24(10):1579–607. [PubMed: 15685718]

39. Hillis SL, Obuchowski NA, Berbaum KS. Power estimation for multireader ROC methods an updated and unified approach. Acad Radiol. 2011; 18(2):129–42. [PubMed: 21232681]

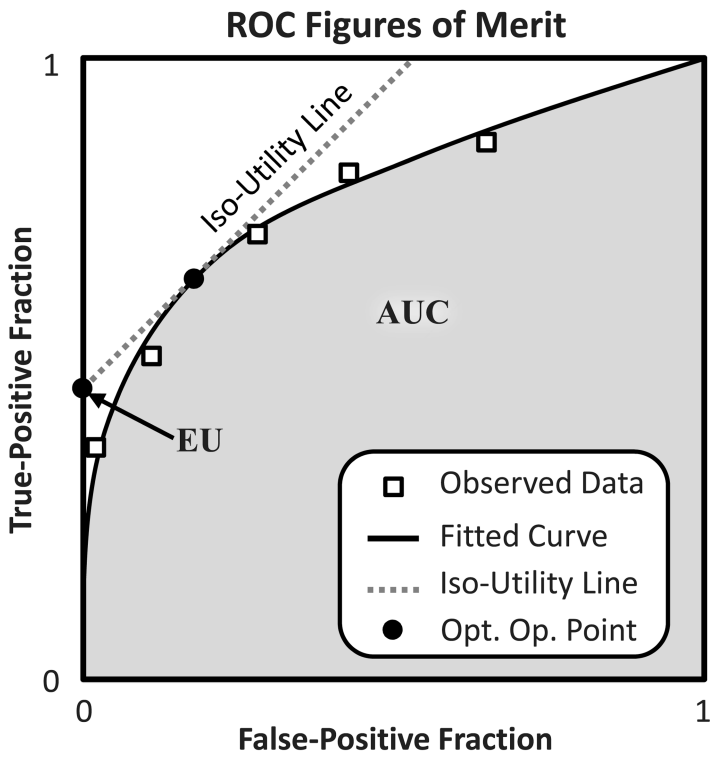

Figure 1.
This diagram shows how AUC and EU are determined for a given ROC curve. A smooth ROC curve is fitted to observed (hypothetical) data using the contaminated binormal model and maximum likelihood fitting. The area under the ROC curve (AUC) is depicted in gray. Under the assumption that task utilities result in iso-utility lines with a given slope, the y-intercept of the highest iso-utility line that intersects the ROC curve defines the expected utility (EU) measure. Note that the iso-utility line is tangent to the ROC curve at the optimal operating point.

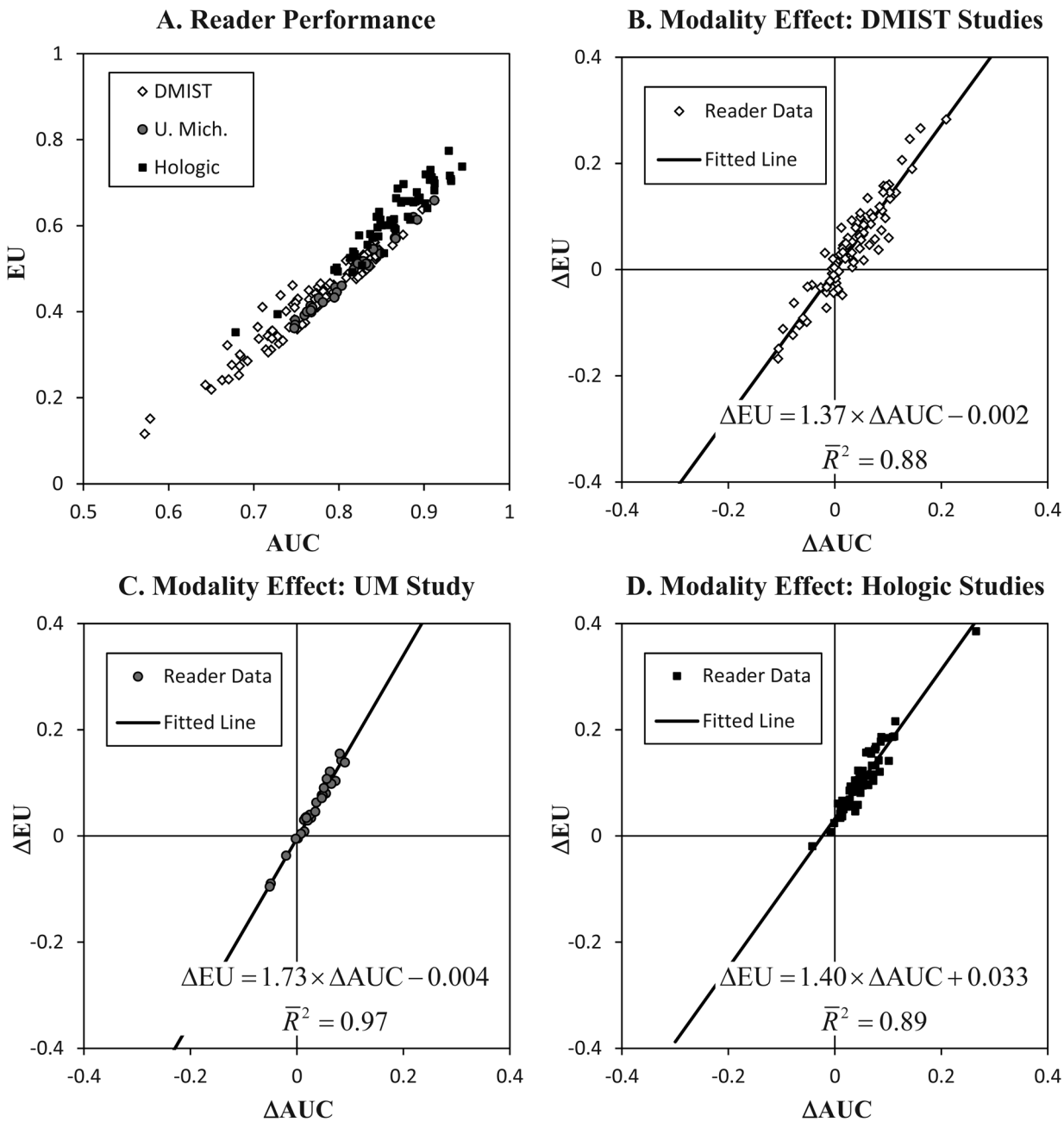

Figure 2.

Modality Effects. The AUC and EU figures of merit are shown for each reader and modality in the scatterplot of performance measures (A). Pairwise differences between modalities for each reader are shown for each of the three studies considered (B-D) with differences arranged so that the average difference across readers for any comparison is positive. In each study, the equation of the least-squares fitted line relating effect sizes is given.

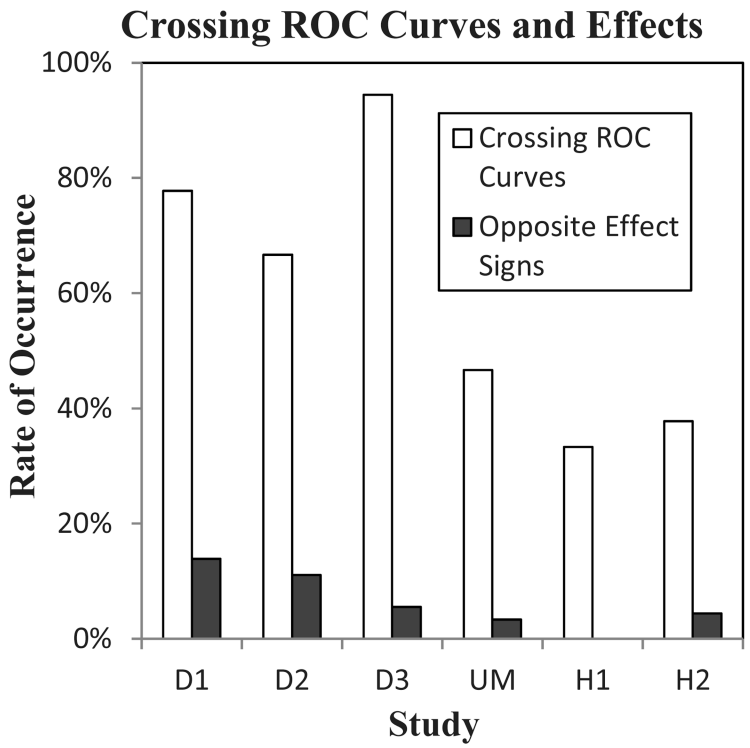

Figure 3.
Crossing ROC curves. The plot shows the fraction readers with crossing ROC curves in each study as well as the fraction of readers with modality differences in AUC and EU that have different (opposite) signs. Note that both of these are elevated in the DMIST studies where there is less of a modality effect.

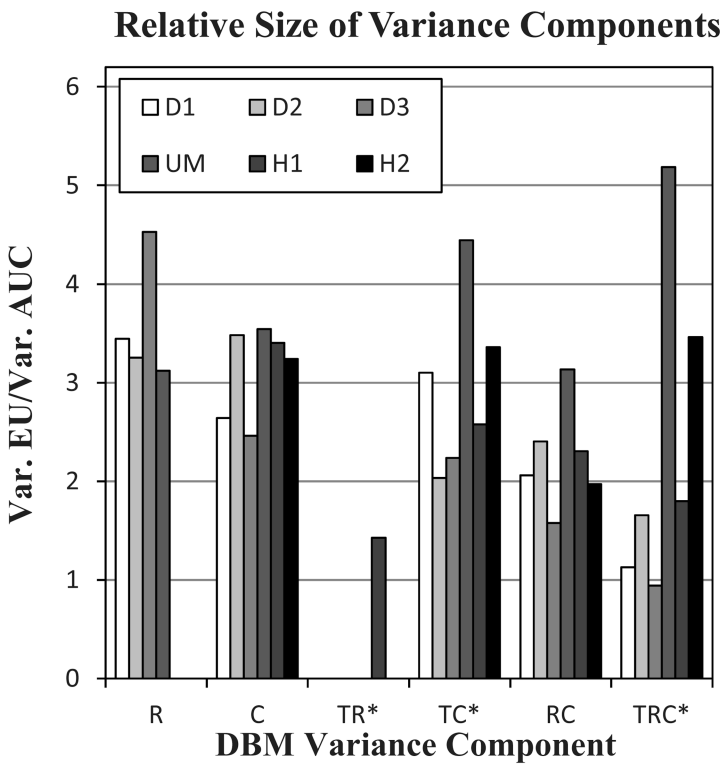

Figure 4.
The relative size of components of variance. Ratios of the elements of Table 2 are shown for each component of variance. The ratio is only shown for variance components greater than 0.0001. The three components directly related to modality comparisons in the MRMC design are indicated (*).

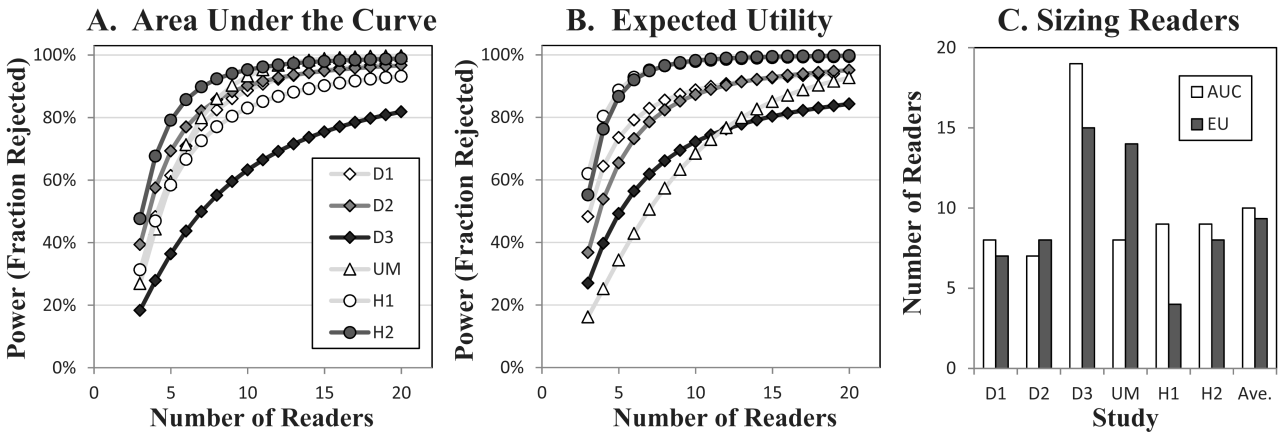

Figure 5.

Power Analysis. Plots of statistical power based on the Hillis and Berbaum method [35] are plotted as a function of the number of readers for area under the ROC curve (A) and expected utility (B) figures of merit in each study (Legend in A applies to both plots). The number of cases used in the power calculation is the same as the number in the actual study. The effect size was set to the largest difference in reader averaged performance across modalities, except in the DMIST studies where a default of 0.1 was used for AUC and 0.136 was used for EU based on the regression line in Figure 2B. The number of readers needed to get 80% power (C) varies considerably from study to study. On average, EU results in a 7% reduction in the number of readers needed to achieve 80% power.

Table 1

Reader Studies. For each source of data, the table gives the study identifier, a brief description of each modality, the number of readers, and the number of cases with positive (P) and negative (N) case numbers indicated in parentheses. Note that SFM = screen-film mammography; DM = digital mammography; CAD = computer-aided diagnosis; DBT = digital breast tomosynthesis.

Data	Study ID	Modality 1 (M1)	Modality 2 (M2)	Modality 3 (M3)	Readers	Cases(P, N)
DMIST	D1	SFM: GE	Soft-Copy DM: GE	Hard Copy DM: GE	12	120 (48, 72)
	D2	SFM: Fuji	Soft-Copy DM: Fuji	Hard Copy DM: Fuji	12	98(27, 71)
	D3	SFM: Fisher	Soft-Copy DM: Fisher	Hard Copy DM: Fisher	6	115 (42, 73)
U. Mich.	UM	No CAD	pre-CAD	w/CAD	10	253 (138, 115)
Hologic	H1	DM alone	DM with 2-View DBT	None	12	312 (48, 264)
	H2	DM alone	DM with 1-View DBT	DM with 2-View DBT	15	310 (51, 259)

Acad Radiol. Author manuscript; available in PMC 2015 April 01.

Table 2

DBM Inference. For each study considered, the table gives the AUC or EU figures of merit averaged across readers in each modality as well as the p-values from tests for modality differences in the whole study (All) and each paired modality comparison (M1-M2, etc.). Shaded cells indicate p-values less than the nominal 0.05 level.

Study ID	FOM	M1-Ave.	M2-Ave.	M3-Ave.	*p*: All	*p*: M1-M2	*p*: M1-M3	*p*: M2-M3
D1	AUC	0.82	0.78	0.79	0.1476	0.1109	0.2241	0.4914
	EU	0.51	0.45	0.45	0.2598	0.2408	0.2062	0.8624
D2	AUC	0.78	0.74	0.76	0.2573	0.1262	0.5739	0.1862
	EU	0.43	0.37	0.41	0.2995	0.1405	0.5820	0.2664
D3	AUC	0.75	0.72	0.69	0.4132	0.5615	0.1448	0.5046
	EU	0.41	0.35	0.33	0.4012	0.4095	0.1612	0.7125
UM	AUC	0.79	0.81	0.84	0.0015	0.1184	0.0043	0.0022
	EU	0.44	0.48	0.52	0.0056	0.1672	0.0105	0.0094
H1	AUC	0.81	0.89	NA	0.0029	NA	NA	NA
	EU	0.51	0.68	NA	< 0.0001	NA	NA	NA
H2	AUC	0.83	0.86	0.89	< 0.0001	0.0057	0.0004	0.0262
	EU	0.54	0.60	0.67	< 0.0001	0.0031	< 0.0001	0.0043

Table 3

DBM Components of Variance for AUC and EU. Table entries of 0 occur when the variance component estimate is less than 0.0001.

Study	Endpoint	R	C	TR	TC	RC	TRC
D1	AUC	0.0012	0.0954	0	0.0221	0.0370	0.3222
	EU	0.0040	0.2521	0	0.0685	0.0762	0.3634
D2	AUC	0.0009	0.1470	0	0.0229	0.0351	0.2036
	EU	0.0029	0.5120	0.0008	0.0465	0.0844	0.3371
D3	AUC	0.0016	0.0962	0	0.0364	0.0701	0.6160
	EU	0.0071	0.2369	0.0012	0.0813	0.1108	0.5799
UM	AUC	0.0010	0.0741	0	0.0036	0.0538	0.1836
	EU	0.0032	0.2625	0	0.0159	0.1687	0.9523
H1	AUC	0	0.1841	0.0011	0.0591	0.0841	0.2770
	EU	0.0012	0.6267	0.0015	0.1524	0.1940	0.4981
H2	AUC	0	0.1932	0	0.0251	0.1215	0.2157
	EU	0.0010	0.6263	0	0.0843	0.2397	0.7469