Evaluating the decision accuracy and speed of clinical data visualizations

David S Pieczkiewicz,1 Stanley M Finkelstein2

ABSTRACT
Clinicians face an increasing volume of biomedical data. Assessing the efficacy of systems that enable accurate and timely clinical decision making merits corresponding attention. This paper discusses the multiple-reader multiple-case (MRMC) experimental design and linear mixed models as means of assessing and comparing decision accuracy and latency (time) for decision tasks in which clinician readers must interpret visual displays of data. These tools can assess and compare decision accuracy and latency (time). These experimental and statistical techniques, used extensively in radiology imaging studies, offer a number of practical and analytic advantages over more traditional qualitative methods such as percent-correct measurements and ANOVAs, and are recommended for their statistical efficiency and generalizability. An example analysis using readily available, free, and commercial statistical software is provided as an appendix. While these techniques are not appropriate for all evaluation questions, they can provide a valuable addition to the evaluative toolkit of medical informatics research.

MEASURES OF ACCURACY
In imaging studies, it is common to examine clinician decision accuracy under one or more imaging types, called modalities. This is exactly analogous for CDVs, which are themselves modalities. The receiver-operating characteristic (ROC) curve is a common way to assess and summarize the accuracy of imaging modalities, diagnostic tests, and clinical judgments. In the imaging setting, ROC curves are generated from reader responses to decision tasks when examining cases under different modalities. (A discussion of the motivation and mechanics of test performance measures, including ROC curves, is provided in the online appendix.) Modality-specific ROC curves give an indication of the accuracy of the modality being measured. Formally, the area under the curve (AUC) represents the probability that a reader will give a higher rank or score to a randomly chosen positive case than to a randomly chosen negative one. The AUC is a number between 0 and 1, and effectively represents the overall probability of accurate decisions under a given modality. Confidence intervals (CIs) for AUCs can also be generated, and statistical hypothesis tests performed to compare the AUCS of different modalities.

MEASURES OF LATENCY
When comparing latency across different modalities, a straightforward approach is to compare measures via t tests or ANOVAs. Another is to perform repeated measures ANOVAs (RM-ANOVAs) on the latencies. These relax the assumptions of independence made in standard t tests and ANOVAs by allowing an analyst to specify which variables are repeated across values of other variables, thus taking any inherent correlations within readers and cases into account. An increasingly popular solution for such situations is to use linear mixed models to analyze the latency data. Like RM-ANOVAs, mixed models are able to account for the correlations that exist when factors such as cases and readers are repeated across modalities. This increases the statistical power of the analyses and reduces the overall number of readers and cases required. Unlike RM-ANOVAs, they can tolerate missing data, and as long as the data are statistically missing at random, can provide robust estimates without imputation or deletion of incomplete cases. Given their regression-style formulation, other covariates of interest can be inserted into the model and analyzed as well.

THE MULTIPLE-READER MULTIPLE-CASE DESIGN
For both accuracy and latency measures, a common experimental design exists which can help
maximize the statistical efficiency, power, and generalizability of the resulting analyses. In this design, multiple readers each examine all cases under all modalities of interest. Called the multiple-reader multiple-case (MRMC) design, it has been successfully used in radiology, and increasingly, general diagnostic tests and CDSSs to comparatively assess two or more modalities of interest. Though the MRMC design was designed for studies using ROC AUCs to assess accuracy data, it is equally suited to analyze latency. A flowchart of the general MRMC experimental procedure, including the collection and analysis of decision and latency measures, is provided in box 1.

**Mixed models and software for analysis**

In mixed models, factors of interest can be classified as ‘fixed’ or ‘random’ effects. A fixed effect can be thought of as one which enumerates all the possible levels or values of that effect. Modality would be classed as a fixed effect, as the modalities being tested in an MRMC design represent the complete set of possibilities to be investigated. In contrast, reader and case would be classed as random effects, as the individual readers and cases used in an experiment are drawn from larger theoretical pools of each, and thus contain some extra amounts of uncertainty and variation. The mixed model can then account for the correlations observed in MRMC experiments, and make parameter estimates more generalizable.

Both accuracy and latency can be analyzed using mixed models. However, if accuracy is measured through modality-specific AUCs, specialized software is required to generate ROC curves, calculate AUCs, and properly enter them into a mixed model framework. Two programs available for analyzing accuracy data are DBM MRMC, from the Medical Image Perception Laboratory at the University of Iowa, and OBUMRM from the Cleveland Clinic Foundation. Both are freely available at their respective websites. DBM MRMC is a stand-alone application, currently for Windows only, while OBUMRM is supplied as FORTRAN source code. DBM MRMC is in continuous development, and has several features not found in OBUMRM, which has not been updated for a number of years. Recently, the DBM MRMC group has released a module to perform MRMC accuracy analysis from within SAS. Currently, there are no SPSS, Stata, or R packages to perform this analysis. In contrast, latency measures can be used directly in mixed models, and analyzed using standard statistical software packages, including SAS, via proc mixed; SPSS, via the mixed command; Stata, via xtmixed; and R, via the lme_4 package.

In both types of analysis, the end result is a regression or ANOVA table listing the various factors, their parameter estimates, and their associated p values. A detailed analysis of a hypothetical experiment, using simulated data, is provided in the online appendix using the MRMC framework, ROC analysis, and mixed models.

**DISCUSSION**

**Decision task and measurement scale**

Readers should present their decisions along continuous or quasi-continuous scales. The use of finer scales (eg, 0–100, as opposed to 1–5) often yields better AUC estimates, and reduces the incidence of ‘degenerate’ ROC curves, in which too few control points are defined to obtain reasonable AUC measures. However, expressing decisions along such scales may be unfamiliar to clinician readers. In our own studies, one reader commented that expressing a decision along a 0–100 scale was difficult, as his own mental model of the decision process was to be fairly certain that disease was present or not. As a result, many of the reader’s decisions clustered around 0 or 100. Despite this, readers should be encouraged to think in probabilistic terms as they read each case.

**Defining a gold standard**

It is important to clearly define the gold standard with which reader decisions will be compared. In practice, it is often difficult or impossible to obtain ‘the truth’ for a particular case or decision. Still, clinician review of full medical records, or consensus decisions by multiple clinicians, can result in

---

**Box 1 General procedural outline of MRMC accuracy/latency experiments**

**PLAN OVERALL STUDY**

- Define diagnostic question(s)
- Frame decisions on continuous or quasi-continuous scale
- Form gold standard criteria
- Form reader and case inclusion criteria

**PLAN EXPERIMENT**

- Develop display modalities (M = 2 or 3)
- Assemble pool of readers R (if a pilot study, R = 4–6)
- Assemble pool of cases C (if pilot, C = 5–6 each positive/negative cases; 10–12 total)
- Generate MxC displays of all case-modality combinations

**PERFORM EXPERIMENT**

- For each reader:
  - Randomize viewing order of case-modality combinations
  - For each case-modality combination: collect decision measure(s); collect latency measure(s); collect other measures as necessary
  - Collect preference data at end of each reader’s viewing session

**ANALYZE DATA**

- For accuracy:
  - Format and enter data into DBM MRMC or similar software
  - Specify readers and cases as random effects; modality as fixed effect
  - Obtain F score and p value for omnibus test of modality AUC equivalence
  - Obtain AUCs and CIs for each modality
  - If F test is significant, examine CIs of pairwise AUC differences

- For latency:
  - Format and enter data into statistical software package
  - Log-transform latency measures if appropriate
  - Specify readers and cases as random effects; modality as fixed effect
  - Specify fully crossed factors
  - Run mixed model and perform pairwise tests of modality equivalence as appropriate
  - Obtain mean latencies and CIs for each modality
  - Obtain mean latencies and CIs in original units (via inverse transformation)

- Analyze other data, including preference data, as appropriate
- Use accuracy analysis variance measures to perform power/sample size analyses
- Use power estimates to inform sample size (C and/or R) of subsequent experiments
a reasonable truth state. However the gold standard is defined and obtained, it should be through means which are independent of the modalities being evaluated in the MRMC study, to avoid potential bias.

**Defining the pool of readers**

Having a mix of reader specialties and levels of experience can serve to increase the generalizability of findings. However, the increased variation which may accompany wide inclusion criteria may also hide differences which may be present within particular subgroups of readers. Ultimately, the experimenter must determine the most appropriate mix of readers for their studies. Stratifying by reader subgroup, or adding reader attributes as model parameters are two options.

**Sample sizes and statistical power**

For a given level of statistical power, there is an approximate tradeoff between the number of readers and the number of cases required. Increasing the number of readers will generally decrease the number of cases required, and vice versa. The addition of a single reader will generally yield a larger increase in statistical power than a single case: an added reader will result in $M \times C$ additional readings, while an added case will result in only $M \times R$ additional readings—usually a smaller number.

The DBM MRMC software website provides SAS programs to suggest sample sizes for MRMC studies, based on decision data. Given estimates of variance components from readers, cases, modalities, and their interactions—which can be calculated by the DBM MRMC program from pilot data—the programs can list tables of reader numbers and case numbers which will yield a desired level of statistical power. (See the online appendix for a sample size analysis of the simulated data.)

In our experience, the number of readers tends to be the limiting factor, rather than the number of cases. Reader recruitment, incentivizing, and scheduling is frequently the most difficult aspect of conducting an MRMC study. Increasing the number of cases will increase the amount of time required to read a set of case/modality combinations, which may also serve to limit reader participation. The number of modalities being tested will have an impact as well.

Pilot studies are recommended when feasible. A limited number of modalities ($M=2$ or 3), readers ($R=4–6$) and cases ($C=5–6$ each positive and negative; $10–12$ cases total) will often provide reasonable estimates of variance components from which to calculate required sample sizes for subsequent studies, as well as provide estimates of required reading times. From this information, experimenters can more intelligently choose a reader/case sample size that makes sense in the context of their study.

At present, there is no general method for determining power and sample sizes for parameters such as latency. However, the sample size estimates obtained for accuracy measures should serve as a reasonable surrogate for those of latency measures.

**Signal prevalence**

Necessary sample sizes can also be affected by the underlying prevalence of the signal being detected. Sample sizes are minimized when the prevalence of positive cases is 0.5; that is, when cases are half positive and half negative. Although this prevalence is not typical for many clinical signal detection tasks, it has not been found to significantly change ROC estimates.

**Randomizing case presentation order**

The order in which case/modality combinations are presented to readers can have an impact on accuracy and latency analyses, either through a learning effect, or in biasing readers to subsequent readings. Readers will also tend to judge cases with increasing speed as they progress through an experimental session. To control for potential ordering effects, each reader should have a separate randomization of case/modality presentations.

**Collecting other data**

Depending on an experimenter’s interests, other decisions, latencies, preferences, and comments can also be collected under the MRMC framework. Additionally, variables such as reader experience and gender can also be collected, and in the case of latency, entered directly into the mixed models as potential factors. If interactive, computer-based modalities are being investigated, events such as mouse clicks can be recorded and time-stamped to analyze feature utilization. User preferences can also be captured at the end of a reading session in a variety of ways, from simple rankings of modality preference to formal usability surveys. Care should be taken not to overburden scarce readers with an unreasonable number of tasks and questions.

**Transforming latency data**

Human reaction times, including latency, often have right-skewed distributions. Performing a logarithmic transformation on such data often yields a more normal distribution which is better suited to mixed model analysis. After statistics and p values are calculated, means and CIs can be inverse-transformed back into their original scales. Since calculated CIs will be symmetric about the transformed mean, they will be somewhat right-skewed when inverse-transformed.

**THE ROLE OF THE MRMC EXPERIMENTAL DESIGN IN CDV EVALUATION**

Unlike latency analyses, accuracy analyses provided by current software packages do not handle missing data, nor do they allow for the addition of other covariates to the model. Further, accuracy analyses using ROC curves ultimately evaluate binary decision tasks. While many decision tasks can be framed in terms of binary signal detection, not all can be. Likewise, AUC measures derived from ROC curves may be less useful in some cases than considering properties such as positive predictive value, sensitivity, or specificity, especially when trying to minimize false positives and negatives.

A theoretical objection to the MRMC general experimental design is that it is not naturalistic; that is, asking readers to assess many cases in serial fashion and provide their decisions on ordinal scales can potentially take the decision tasks out of their original clinical contexts. This can be true, but for some tasks, such as the regular triage of longitudinally monitored patients, the MRMC design is not unrealistic. In any case, this does not negate the advantage that MRMC methods provide regarding control over potential confounding factors which may arise in natural clinical settings. Provided the appropriate care is taken in selecting representative cases and readers, MRMC methods can provide a measure of quantitative rigor, reproducibility, and generalizability which complements other evaluation methods such as case studies and randomized trials.

**CONCLUSIONS**

There will always be a need for clinicians to view and assess data themselves. As such, the evaluation of CDVs should be a vital component of evaluations of CDSS efficacy. This brief presented the MRMC experimental framework, which can be used to design and conduct experiments to detect systematic differences in decision accuracy and latency between several
data visualization modalities in an arbitrary decision task. The MRMC design, as well as the mixed-model approach used to analyze MRMC data, are flexible and powerful methods which have a number of theoretical and practical advantages over other evaluation methods, although they also have their own drawbacks and learning curves. Overall, they can provide a useful addition to the evaluation toolkits of informatics researchers.

Acknowledgments We thank numerous helpful discussions with MRMC and usability methodology experts, as well as the three anonymous reviewers of an earlier draft of this manuscript for their comments and suggestions.

Funding This study was supported through a graduate fellowship from the Biomedical Engineering Institute at the University of Minnesota. DSP’s present work is supported through a National Library of Medicine training grant to the Computation and Informatics in Biology and Medicine Training Program at the University of Wisconsin, Madison (1L15LM007359), and a Security Health Plan Fellowship in Interactive Clinical Design.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES