Visualizations are increasingly important in helping users manage large data streams. As a result, researchers often need to compare the performance of several visualizations. We present two statistical techniques, multiple-reader multiple-case receiver operating characteristic curve analysis, and generalized linear mixed models, to compare the accuracy and speed of decisions using data visualizations. These techniques have several advantages over simpler strategies for assessing decision quality, and should be made part of the quantitative evaluation of visualizations.

Contact email: pieczkiewicz.david@mcrf.mfldclin.edu
Improved Techniques for Quantitatively Comparing Data Visualizations

DS Pieczkiewicz PhD, LV Rasmussen, JB Starren MD PhD
Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI

Introduction
As the volume and complexity of medical data increase, so does the need for efficacious ways of visualizing and presenting this data to clinicians. While many evaluations of visualizations in medical informatics measure user satisfaction and preference, they do not always assess the accuracy of decision-making using visualizations, nor do they assess the latency, the time needed to make a decision. We present an experimental framework and statistical techniques which can be used to assess accuracy and latency, adjusted from similar work in radiology evaluating the performance of imaging modalities.

Experimental Framework
The overall experimental design is called Multiple Reader Multiple Case (MRMC). In individual reading sessions, all readers review all cases in all available modalities. The MRMC framework makes the following assumptions:
- Two or more visualizations (modalities) are being compared.
  - A “signal” is being detected in the data, such as disease presence.
  - Human readers decide whether the signal is present or not.
- The decisions can be expressed along a scale.
- A binary gold standard exists for the signals/decisions.

Plan
Define question
- “For pulmonary function test and symptom data from post-transplant patients in a long-term transplant home monitoring program, do the data suggest that the patient is experiencing an infection or rejection event?”

Prepare
Define standard
- Independent testing, case review, expert consensus, etc.

Execute
Develop modalities
- Define all modalities under consideration
- Assemble cases
  - Half positive for signal, half negative
- Assemble readers
  - Readers competent in interpreting underlying data

Generate displays
- Generate all cases in all modalities

Randomize presentation order
- Minimize ordering effects

Collect decisions
- Decisions on continuous or quasi-continuous scale
- Collect long label
  - Can obtain usable preference statistic, during or after reading session

Mixed Models
- Decision accuracy and latency are analyzed as linear mixed models, which separate independent variables (case, reader, modality) inside and random factors. By specifying the analytic models this way, correlations between reader and cases can be accounted for, and the results can be generalized to all cases and all readers.

Advantages of MRMC Analysis
- More statistically powerful than standard AUCs and regressions
- Can better tolerate missing data
- Methods available in free software or standard statistical packages
- Provides a framework for analyzing many decision tasks with visualization

Disadvantages of MRMC Analysis
- Requires somewhat advanced understanding of statistics and ROC curves
- MRMC models more complicated to report than aggregate analyses
- Readers and reader time are precision but scarce resources
- Not appropriate for all decision tasks

Conclusions
The MRMC framework and mixed models provide a statistically powerful means of analyzing decision accuracy and latency data, and provide many theoretical and practical advantages over simpler forms of analysis and reporting. As health care systems increasingly employ visualizations of dense clinical data, effective means of analyzing and comparing display efficacy are necessary and important additions to informatics research methodology.

References

Acknowledgments
DSP was supported in part by a NIBIR/NIAMN training grant to the Computing and Informatics in Medicine and Biomedical Informatics in the Biological and Medicine Training Program at the University of Wisconsin, Madison (NIH 1T15LM07351).