Evaluating Displays of Clinical Information

David S. Pieczkiewicz, PhD
NIBIB / CIBM Postdoctoral Fellow
Biomedical Informatics Research Center
Marshfield Clinic Research Foundation
&
University of Wisconsin, Madison
What is Evaluation?

- The systematic determination of the merit, worth, or significance of an entity
- Quantitative and qualitative approaches
- Experimental and non-experimental (e.g., controlled and non-controlled)
- Focus groups, RCTs, and everything in between
# Levels of Diagnostic Efficacy

<table>
<thead>
<tr>
<th>Efficacy Type</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical efficacy</td>
<td>physical validity?</td>
</tr>
<tr>
<td>Diagnostic accuracy</td>
<td>statistical performance?</td>
</tr>
<tr>
<td>Diagnostic-thinking accuracy</td>
<td>affects physicians’ estimates?</td>
</tr>
<tr>
<td>Therapeutic efficacy</td>
<td>affects patient management?</td>
</tr>
<tr>
<td>Patient-outcome efficacy</td>
<td>affects patient health?</td>
</tr>
<tr>
<td>Societal efficacy</td>
<td>wider social cost/benefit?</td>
</tr>
</tbody>
</table>

---

From Fryback and Thornbury (1991)
Evaluation for EHRs

- EHRs usually assessed in terms of efficacy
- How well do they “work”?  
  - Clinical utility  
  - Clinical Outcomes  
  - Usability  
  - User acceptance
- Many EHR evaluations stop at user acceptance
  
  *This is good, but incomplete!*
Elting et al. (1999)

Table

<table>
<thead>
<tr>
<th></th>
<th>Conventional treatment</th>
<th>Investigational treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total no</td>
<td>50</td>
<td>60</td>
</tr>
<tr>
<td>% Fail</td>
<td>38</td>
<td>12</td>
</tr>
<tr>
<td>Good prognosis</td>
<td>30 (30)</td>
<td>35 (11)</td>
</tr>
<tr>
<td>Poor prognosis</td>
<td>20 (45)</td>
<td>25 (12)</td>
</tr>
</tbody>
</table>

(Negatively framed tables displayed failure rates in red. Positively framed tables displayed response rates in green)

Bar graph

- Good prognosis: Conventional treatment: 30, Investigational treatment: 35
- Poor prognosis: Conventional treatment: 20, Investigational treatment: 25

Pie chart

- Investigational treatment (n = 50)
  - Good prognosis
  - Poor prognosis
- Conventional treatment (n = 60)
  - Good prognosis
  - Poor prognosis

Icon

- Conventional treatment
- Investigational treatment

Legend:
- Response
- Fail
Measuring Efficacy

• **Accuracy**: How often or well the target task is completed (action, decision, etc.)

• **Latency**: How long it takes to perform the task, independent of accuracy

• **Preference**: What users feel comfortable with

from Starren and Johnson (2000)
Decision Accuracy

- Percent correct
- Easy to measure and report
- Misses many decision distinctions (true and false positives and negatives, etc.)
- Sensitivity, specificity, positive predictive value, negative predictive value
- Provides more information
- Provides measures for particular cutoffs and prevalences
ROC Analysis

- Receiver-operating characteristic (ROC) curves describe accuracy over all cutoffs.
- Area under curve describes overall accuracy of decisions.
- Multiple curves can compare the performance of two or more visualizations.
MRMC ROC Analysis

- Multiple-reader multiple-case (MRMC) ROC analysis developed for radiology
- Multiple readers assess multiple cases in each modality (visualization) of interest
- Decisions given on probability scale
- Decisions collated to generate ROC curve areas and variance information
- Determines if different modalities have statistically different accuracies
The MRMC Design

A case $c$ contains the medical information needed to assess a patients’ condition at a particular time.
The MRMC Design

For multiple cases $c_i$, some cases are positive for the feature of interest and some are negative.
The MRMC Design

Each case $c_i$ is viewed under each modality $m_j$. 
The MRMC Design

Decisions $d_{ij}$ and other data are collected in random order to wash out viewing-order influences.
The MRMC Design

Process is repeated for each reader $r_k$, with a different random case ordering for each
MRMC ROC Software

- DBM MRMC—University of Iowa
  - Windows application, ready-to-run
  - SAS program for sample size estimation
- OBUMRM—Cleveland Clinic Foundation
  - FORTRAN program
  - Must be compiled to use
- Both packages freely available
Decision Latency

- $t$-tests and ANOVAs most accessible
- Repeated measures ANOVA takes correlation patterns into account
- Also provides better accounting for sources of variance
- Does not handle missing data very well
Mixed Models

- Type of generalized linear model which can encompass repeated measures ANOVAs
- Also takes correlations into account
- Factors can be “fixed” or “random”
- More efficient use of experimental data
- Much more robust to missing data
Mixed Models

- MRMC design translates into fully-crossed mixed model
- Latency modeled by fixed modality factor and random reader and case factors
- $P$-values of modality slopes are tests of whether modalities differ by latency
- Can more easily investigate other factors
- MRMC ROC analysis actually a form of mixed modeling
# Mixed Model Commands

<table>
<thead>
<tr>
<th>Software</th>
<th>Command</th>
</tr>
</thead>
<tbody>
<tr>
<td>R and S-Plus</td>
<td><code>lme()</code></td>
</tr>
<tr>
<td>SAS</td>
<td><code>proc mixed</code></td>
</tr>
<tr>
<td>SPSS</td>
<td><code>mixed</code></td>
</tr>
<tr>
<td>Stata</td>
<td><code>xtmixed</code></td>
</tr>
</tbody>
</table>
Lung Transplant Home Monitoring Program

- Created by the University of Minnesota and Fairview-University Transplant Center
- Patients use a portable electronic spirometer to record pulmonary and symptom information
- Data uploaded and triaged weekly
Tabular Modality

from Pieczkiewicz et al. (2007)
Graphical Modalities

from Pieczkiewicz et al. (2007)
The following ANOVA routines are available. You may select multiple options. You must select at least one option.

- Analysis treating both readers and cases as random samples
- Analysis treating only cases as a random sample
- Analysis treating only readers as a random sample

Other ANOVA options are available
- Display the Obuchowski-Rockette components of variance
Analysis 1: Random Readers and Random Cases

(Results apply to the population of readers and cases)

a) Test for $H_0$: Treatments have the same AUC

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Mean Square</th>
<th>F value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>1</td>
<td>0.47140141</td>
<td>6.39</td>
<td>0.0526</td>
</tr>
<tr>
<td>Error</td>
<td>5.00</td>
<td>0.07372649</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Error term: $\text{MS(TR)} + \max[\text{MS(TC)}-\text{MS(TRC)},0]$  

Conclusion: The treatment AUCs are not significantly different, $F(1,5) = 6.39, p = .0526$.

b) 95% confidence intervals for treatment differences

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Estimate</th>
<th>StdErr</th>
<th>DF</th>
<th>t</th>
<th>Pr &gt; t</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 2</td>
<td>-0.06268</td>
<td>0.02479</td>
<td>5.00</td>
<td>-2.53</td>
<td>0.0526</td>
<td>-0.12639 , 0.00104</td>
</tr>
</tbody>
</table>

$H_0$: the two treatments are equal.  
Error term: $\text{MS(TR)} + \max[\text{MS(TC)}-\text{MS(TRC)},0]$  

c) 95% treatment confidence intervals based on reader x case ANOVAs  
for each treatment (each analysis is based only on data for the  
specified treatment

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Area</th>
<th>Std Error</th>
<th>DF</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.78356094</td>
<td>0.02755194</td>
<td>16.12</td>
<td>(0.72518772 , 0.84193415)</td>
</tr>
<tr>
<td>2</td>
<td>0.84623745</td>
<td>0.03697621</td>
<td>12.60</td>
<td>(0.76609538 , 0.92637952)</td>
</tr>
</tbody>
</table>

Error term: $\text{MS(R)} + \max[\text{MS(C)}-\text{MS(RC)},0]$
Accuracy Results

\[ C = 20 \ (10^+/10^-), \ M = 3, \ R = 12 \]

\[ F_{2,22} = 0.147 \]

\[ P = 0.86 \]
. xi: xtmixed lntime i.modality || _all:R.case || _all:R.reader
i.modality       _Imodality_1-7      (naturally coded; _Imodality_1 omitted)

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log restricted-likelihood = -526.85469
Iteration 1:   log restricted-likelihood = -526.85469

Computing standard errors:

Mixed-effects REML regression
Group variable: _all

Number of obs      =       720
Number of groups   =         1

Obs per group: min =       720
     avg =     720.0
     max =       720

Wald chi2(2)       =     48.91
Log restricted-likelihood = -526.85469          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  lntime |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------------------------------------------------------------------
     _Imodality_6 |  -.1332807   .0433225    -3.08  0.002      -0.2181913    -.0483702
     _Imodality_7 |   .1689817   .0433225     3.90  0.000       0.0840711    .2538923
     _cons |   3.813324   .1536722    24.81  0.000       3.512132    4.114516
------------------------------------------------------------------------------

Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
   _all: Identity               |
     sd(R.case) |   .1280731   .0287307      .0825102    .1987962
     sd(R.reader) |   .5121313   .1107496      .3352023    .7824484
     sd(Residual) |   .4745745   .012803      .450133     .5003431

LR test vs. linear regression:     chi2(2) =    474.66   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
Latency Results

$C = 20 (10^+/10^-)$, $M = 3$, $R = 12$

- Interactive Graph: $45.30$
- Static Graph: $39.65$
- Table: $53.64$

$\beta_{\text{static}} = -0.133$
$P = 0.002$

$\beta_{\text{table}} = 0.168$
$P < 0.001$
# Preference Results

<table>
<thead>
<tr>
<th>Modality</th>
<th>Average Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive Graph</td>
<td>1.1</td>
</tr>
<tr>
<td>Static Graph</td>
<td>2.2</td>
</tr>
<tr>
<td>Table</td>
<td>2.8</td>
</tr>
</tbody>
</table>

(R = 12 readers)
Glucose Data Viewer
Disadvantages

- Methods not as “easy” as traditional ones
- Sample size requirements can be unclear
- MRMC ROC software takes skill to use
- Mixed models more computationally-intensive, and possibly nonconvergent
- May not apply to some aspects of EHR evaluation and research
Conclusions

• Efficacy studies usually stop at user satisfaction and/or user preference
• Accuracy and latency can be useful, objective measures of EHR efficacy
• ROC methodologies can be applied to measure decision accuracy in EHRs
• Mixed models can be used to assess latency
• Software now readily available for these purposes
Acknowledgments

- Stan Finkelstein, PhD
- Marshall Hertz, MD
- Justin Starren, MD, PhD
- Luke Rasmussen
- Kevin Berbaum, PhD
- Kevin Schartz, PhD
- Nancy Obuchowski, PhD

- Computation and Informatics in Biology and Medicine Program, University of Wisconsin

- National Institute of Biomedical Imaging and Bioengineering, NIH

- National Library of Medicine