Using Item Response Theory to Summarize Health Care Quality Measures:
An Application to HEDIS® Diabetes Measures in Medicare Advantage

John L. Adams
UCLA Center for Maximizing Outcomes and Research on Effectiveness, 16 September 2014

Collaborators

• Marc N. Elliott, Amelia M. Haviland, Carol A. Edwards
  RAND
  Sarah Gaillot
  CMS

  Funded by CMS: HHSM-500-2005-000281 to RAND
Overview of today’s talk

- Building composite quality of care measures
- The potential of item response theory (IRT)
- IRT analysis of Medicare Advantage diabetes HEDIS© measures
- Conclusions and future directions

Composite quality of care measures (mostly at a more aggregate level)

- Opportunities scoring
  - \#passed/\#triggered
- Observed difficulty of delivery (ODD) adjustment
  - \text{Sum(observed)/sum(expected)} \text{ or } \text{Sum(observed)-sum(expected)}
  - Not quite case mix adjustment, only population pass rates
- Standardization
  - \frac{(\text{rate1-mean(rate1)})/\text{SDrate1}+(\text{rate2-mean(rate2)})/\text{SDrate2}+.....}{......}
- Typically not as much case mix adjustment or fancy measurement methods as health status or outcomes
The item response theory (IRT) model

• A latent trait model
  – Assumes a person has a latent quality score
  – That quality score drives the passing of measures
  – Similar to factor analysis with a single factor
    ▪ Correctly handles binary data
    ▪ Can handle complicated missing data patterns
• The 2 parameter model for binary data looks like a logistic regression except the theta isn’t known:
  \[ P_{jk}(+ | \theta_k) = \frac{e^{\alpha_j(\theta_k - \beta_j)}}{1 + e^{\alpha_j(\theta_k - \beta_j)}} \]
  • Where k is the person and j is the item

Some useful references

HEDIS Diabetes Measures in Medicare Advantage

• We have the HEDIS diabetes measures for reporting years 2008-2010
  – Today I will focus on the 2010 measurement year data
  – Approximately 450k unique individuals that triggered at least one diabetes measure

• Nine measures:
  – HbA1c Testing
  – HbA1c Poor Control >9% (reversed)
  – HbA1c Good Control
  – Retinal eye exam
  – LDL-C Screening
  – LDL-C Control <100mg/dl
  – Kidney disease / Nephropathy
  – Blood pressure control <130/80
  – Blood pressure control <140/90

But some pairs of these measures have a natural ordinal structure

• Glycemic control
  – HbA1c Poor Control >9% (reversed)
  – HbA1c Good Control

• Blood pressure control
  – Blood pressure control <130/80
  – Blood pressure control <140/90

• We will treat these as ordered categorical data and combine them
  – Just like ordinal logistic regression
The quality scores

- Each combination of triggering and pass/not has a quality estimate
  - Like a best linear unbiased predictor (BLUP)
  - On a z-score scale, roughly -2 to +2
Computation

• In the beginning
  – Used Proc Nlmixed in SAS 9.2
  – There are only $3^6 \times 4^2 \approx 4k$ possible patterns of triggering and scoring (but only about 500 occur in nature)
  – Weighted analysis makes the problem tractable

• And then a miracle occurred
  – Experimental Proc IRT in SAS 9.4
  – MIRT and many good things are now possible


High scoring trigger/pass/fail combinations

<table>
<thead>
<tr>
<th>Quality Score</th>
<th>Standard error</th>
<th>HbA1c Testing</th>
<th>HbA1c Poor Control &gt;9% (reversed)</th>
<th>HbA1c Good Control</th>
<th>Retinal eye exam</th>
<th>LDL-C Screening</th>
<th>LDL-C Control &lt;100mg/dl</th>
<th>Kidney disease / Nephropathy</th>
<th>Blood pressure control &lt;130/80</th>
<th>Blood pressure control &lt;140/90</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.47</td>
<td>0.63</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.47</td>
<td>0.63</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.47</td>
<td>0.64</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.42</td>
<td>0.63</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.29</td>
<td>0.61</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.29</td>
<td>0.61</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.29</td>
<td>0.61</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.28</td>
<td>0.69</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.27</td>
<td>0.61</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.27</td>
<td>0.61</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.27</td>
<td>0.61</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Low scoring trigger/pass/fail combinations

<table>
<thead>
<tr>
<th>Quality score</th>
<th>Standard error</th>
<th>HbA1c Testing</th>
<th>HbA1c Poor Control 10% (reversed)</th>
<th>HbA1c Good Control</th>
<th>Retinal eye exam</th>
<th>LDL-C</th>
<th>LDL-C Control &lt;100mg/dl</th>
<th>Kidney disease / Nephropathy</th>
<th>Blood pressure control &lt;130/80</th>
<th>Blood pressure control &lt;140/90</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.85</td>
<td>0.58</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-1.85</td>
<td>0.58</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-1.85</td>
<td>0.58</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-1.84</td>
<td>0.59</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>-1.83</td>
<td>0.59</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>-1.79</td>
<td>0.58</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-1.72</td>
<td>0.58</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>-1.71</td>
<td>0.56</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-1.77</td>
<td>0.56</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-1.77</td>
<td>0.56</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

What about the possibility of more than one latent dimension of quality?
Exploring the second dimension

- Fitting the model parameters was very sensitive
- There is a conditional structure in the data not captured in the model
  - You can’t be scored on a lab value if you didn’t get the test
- Tried turning the LDL screening and control variables into an ordered categorical variable
  - Some conceptual and data issues with this

What I still need to work on

- I am not sure how important the local dependence assumption is for this problem
Conclusions and future directions

• The IRT model for HEDIS diabetes measures has good face validity
• IRT has the potential to sensibly:
  – Address “topped out” measures without dropping them
  – Fold in emergent measures while maintaining the meaning of the scale
• Future directions
  – Continue to explore multidimensionality (MIRT)
  – Explore differential item functioning (race, gender, etc.)
  – Expand models to include system level structure
  – Explore a wider range of measures

Questions?