Challenges and Opportunities for Using Common PRO Measures in Comparative Effectiveness Research

Greg Simon, MD, MPH
Group Health Research Institute

David Cella, PhD
Northwestern University

Reesa Laws, BS
Kaiser Permanente Center for Health Research
History of measurement-based care for depression

- Nationally:
  - 1990-2000: Effectiveness trials of collaborative care programs
  - 2000-2005: Large-scale implementation trials

- In our health systems:
  - 2000-2005: Guidelines recommend use of standard outcome measures (PHQ9)
  - 2005-2010: Implementation of PHQ9 in EHRs
  - 2010-2015: Implementation of standard care processes and monitoring/reporting performance
PHQ9 Depression Questionnaire

- 9-item self-report questionnaire
- Maps to DSM criteria for major depression
- Has become default standard in most large healthcare systems
Health system motivations for collecting PHQ data:

- Internal quality initiatives
- External quality metrics
- Purchaser & health plan wellness initiatives
MHRN Role in promoting measurement-based care:

- Producers (and promoters) of effectiveness evidence
- Content experts for guideline development
- Technical consultation regarding measure selection
- Technical assistance with reporting and analytics
- Highlighting health system performance in research presentations and publications

PHQ definitely had momentum – so we threw all of our weight behind it.
Health system data streams for PHQ data:

- Visit-based questionnaires
- Online portal questionnaires
- Health Risk Appraisal questionnaires
Sources of PHQ9 data in MHRN health systems
Data quality problems

- Variable (and unknown) conditions of administration
- Missing items
- Duplicate records
- Complementary records
Why bother with this messiness?

In 6 MHRN health systems:

Approximately 1.9 million observations for approximately 600,000 unique patients

For free!
Use Cases

- Pragmatic trial of outreach to prevent suicide attempt
  - Weekly extraction of PHQ9 data to identify outpatients at risk for suicide attempt

- Population-based suicide risk calculator
  - Link PHQ9 data and other predictors to develop point-of-care risk prediction tool

- Racial and ethnic variation in depression care
  - Examine racial and ethnic variation in treatment adherence and clinical effectiveness

- Personalized care for treatment-resistant depression
  - Identify patterns of prior treatment response predicting response to next-step treatment
New measurement domains:

- Alcohol use disorders
- Externalizing disorders in children
- Mania/mixed symptoms in bipolar disorder
- Attention deficit disorder in adults

These are health system priorities, not ours.
Health system motivations:

- Internal quality initiatives
- External quality metrics
- Purchaser & health plan wellness initiatives

Note: Research is not on this list!
(We are a little tail on a very big dog!)
Common measures and common metrics to enable CER in everyday healthcare settings

David Cella PhD
Northwestern University
Goal:
Conduct comparative effectiveness research using data collected by the health care system

• Depression as the use case
Current state: PHQ-9 is dominant measure
We can call that a common measure, but...

- Several large and small providers resist PHQ-9
  - Kaiser Northern Ca
  - Cleveland Clinic
- Suicide question → PHQ-8
- Length → (PHQ-2; PHQ-4)
- Long-term relevance (DSM 4→5→?)
A solution

• PRO Rosetta Stone (PROsetta Stone®) links Patient-Reported Outcomes Measurement Information System (PROMIS) measures with other related “legacy” instruments

• PHQ-9 score linked to the PROMIS Depression measures using procedures based on item response theory (and equipercentile) methods (Choi et al)
  – Cross-walk tables
  – Allows PHQ-9 scores to be expressed as standardized T-score linked to the PROMIS metric.
A Rosetta Stone for Linking Patient-Reported Outcome Measures

www.prosettastone.org

USPHS Grant No. RC4 CA157236-01
Depression is One of the 83 Calibrated PROMIS Banks or Scales

- T Score
  - Mean = 50
  - SD = 10

- Referenced to the US general population
- Can administer as 4-10 item short forms or Computer Adaptive Testing (CAT)

www.nihpromis.org
Interpreting PROMIS T-Scores

- **Within Normal Limits**
- **Mild**
- **Moderate**
- **Severe**

About 80% of the general population fall within normal limits, while about 20% fall into the severe category.

*These are general guidelines to aid in interpreting PROMIS T-scores. Within a given condition or PROMIS domain, thresholds may differ.*
## Raw Score to T-Score Conversion Table for PHQ-9 to PROMIS (IRT Fixed Parameter Calibration Linking)

<table>
<thead>
<tr>
<th>PHQ-9 Score</th>
<th>PROMIS T-score</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>37.4</td>
<td>6.4</td>
</tr>
<tr>
<td>1</td>
<td>42.7</td>
<td>5.3</td>
</tr>
<tr>
<td>2</td>
<td>45.9</td>
<td>4.8</td>
</tr>
<tr>
<td>3</td>
<td>48.3</td>
<td>4.7</td>
</tr>
<tr>
<td>4</td>
<td>50.5</td>
<td>4.3</td>
</tr>
<tr>
<td>5</td>
<td>52.5</td>
<td>4.0</td>
</tr>
<tr>
<td>6</td>
<td>54.2</td>
<td>3.8</td>
</tr>
<tr>
<td>7</td>
<td>55.8</td>
<td>3.7</td>
</tr>
<tr>
<td>8</td>
<td>57.2</td>
<td>3.6</td>
</tr>
<tr>
<td>9</td>
<td>58.6</td>
<td>3.5</td>
</tr>
<tr>
<td>10</td>
<td><strong>59.9</strong></td>
<td><strong>3.4</strong></td>
</tr>
<tr>
<td>11</td>
<td>61.1</td>
<td>3.3</td>
</tr>
<tr>
<td>12</td>
<td>62.3</td>
<td>3.3</td>
</tr>
<tr>
<td>13</td>
<td>63.5</td>
<td>3.2</td>
</tr>
<tr>
<td>14</td>
<td>64.7</td>
<td>3.2</td>
</tr>
<tr>
<td>15</td>
<td>65.8</td>
<td>3.2</td>
</tr>
<tr>
<td>16</td>
<td>66.9</td>
<td>3.2</td>
</tr>
<tr>
<td>17</td>
<td>68.0</td>
<td>3.1</td>
</tr>
<tr>
<td>18</td>
<td>69.2</td>
<td>3.2</td>
</tr>
<tr>
<td>19</td>
<td>70.3</td>
<td>3.2</td>
</tr>
<tr>
<td>20</td>
<td>71.5</td>
<td>3.2</td>
</tr>
<tr>
<td>21</td>
<td>72.7</td>
<td>3.3</td>
</tr>
<tr>
<td>22</td>
<td>74.0</td>
<td>3.4</td>
</tr>
<tr>
<td>23</td>
<td>75.3</td>
<td>3.5</td>
</tr>
<tr>
<td>24</td>
<td>76.7</td>
<td>3.6</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>PROM (Patient-Reported Outcome Measure)</td>
<td>PHQ-9, a standardized tool to assess depression</td>
<td></td>
</tr>
<tr>
<td>PRO-PM (Patient-reported Outcome Performance Measure)</td>
<td>Example: Percentage of patients with diagnosis of major depression or dysthymia and initial PHQ-9 score &gt; 9 with a follow-up PHQ-9 score &lt; 5 at 6 months (NQF #0711)</td>
<td></td>
</tr>
</tbody>
</table>

NQF prefers that PRO-PMs NOT be tied exclusively to a single PROM

Adapted from National Quality Forum
Test Drive: Cleveland Clinic Study (Katzan et al)

To determine the group-level and patient-level concordance in performance of 2 depression-related PRO-PMs assessed using different depression PROMs:

1. Patient Health Questionnaire-9 (PHQ-9)
2. PROMIS Depression Short-Form (PROMIS SF)
3. PHQ-9 co-calibrated on the PROMIS metric (PHQ-9_{PROMIS}).
Results

PROM scores across levels of PHQ-9

PROMIS Depression ShortForm

PHQ-9 cocalibrated on PROMIS Metric

(\text{PHQ-9}_{\text{PROMIS}})
## Analytic Methods

### 1. Concordance calculations continued

Depression thresholds were defined using crosswalk tables*

<table>
<thead>
<tr>
<th>Category</th>
<th>PHQ-9</th>
<th>T-scores</th>
<th>T-scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PROMIS SF and</td>
<td>PHQ-9\textsubscript{PROMIS}</td>
</tr>
<tr>
<td>Positive depression screen</td>
<td>&gt; 9</td>
<td>≥ 59.9</td>
<td></td>
</tr>
<tr>
<td>Remission</td>
<td>&lt; 5</td>
<td>&lt; 52.5</td>
<td></td>
</tr>
<tr>
<td>Progress towards remission</td>
<td>↓ 50%</td>
<td>↓ 50% of PHQ-9 equivalent</td>
<td></td>
</tr>
</tbody>
</table>

* Per Choi et al, Psychol Assess 2014:26(2);513–527
## Results

**Depression Diagnosis:** Percentage of patients with depression at the time of initial assessment during the study period

<table>
<thead>
<tr>
<th>N=5,376</th>
<th>(\text{Group Level})</th>
<th>(\text{Patient Level})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{PROM})</td>
<td>Depression Score</td>
<td>Depression Threshold</td>
</tr>
<tr>
<td>PHQ-9, median [IQR]</td>
<td>7 [3,12]</td>
<td>(PHQ9&gt;9)</td>
</tr>
<tr>
<td>PROMIS ShortForm Mean (SD)</td>
<td>52.7 (11.2)</td>
<td>(PROMIS \geq 59.9^*)</td>
</tr>
<tr>
<td>PHQ-9\text{PROMIS} Mean (SD)</td>
<td>55.3 (10.5)</td>
<td>(PROMIS \geq 59.9^*)</td>
</tr>
</tbody>
</table>

*based upon cut-offs used by Choi*
## Results

**Depression Diagnosis:** Percentage of patients with depression at the time of initial assessment during the study period

<table>
<thead>
<tr>
<th>N=5,376</th>
<th><strong>Group Level</strong></th>
<th><strong>Patient Level</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>PROM</td>
<td>Depression Score</td>
<td>% Meeting Threshold</td>
</tr>
<tr>
<td>PHQ-9, median [IQR]</td>
<td>7 [3,12]</td>
<td>PHQ9&gt;9</td>
</tr>
<tr>
<td>PROMIS ShortForm Mean (SD)</td>
<td>52.7 (11.2)</td>
<td>PROMIS ≥59.9*</td>
</tr>
<tr>
<td>PHQ-9 PROMIS Mean (SD)</td>
<td>55.3 (10.5)</td>
<td>PROMIS ≥59.9*</td>
</tr>
</tbody>
</table>

*based upon cut-offs used by Choi*
## Results

### Depression Diagnosis: Percentage of patients with depression at the time of initial assessment during the study period

<table>
<thead>
<tr>
<th>N=5,376</th>
<th>Group Level</th>
<th>Patient Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROM</td>
<td>Depression Score</td>
<td>Depression Threshold</td>
</tr>
<tr>
<td>PHQ-9, median [IQR]</td>
<td>7 [3,12]</td>
<td>PHQ9&gt;9</td>
</tr>
<tr>
<td>PROMIS ShortForm Mean (SD)</td>
<td>52.7 (11.2)</td>
<td>PROMIS (\geq 59.9^*)</td>
</tr>
<tr>
<td>PHQ-9_PROMIS Mean (SD)</td>
<td>55.3 (10.5)</td>
<td>PROMIS (\geq 59.9^*)</td>
</tr>
</tbody>
</table>

*based upon cut-offs used by Choi
## Results

Depression Remission: The percentage of patients with initial PHQ-9 score >9 who have a follow-up PHQ-9 score < 5 *(based on NQF #0711)*

<table>
<thead>
<tr>
<th>N = 701</th>
<th>Group Level</th>
<th>Patient Level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PROM</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHQ-9</td>
<td>% Depression at Baseline (n)</td>
<td>% Depression Follow-up (n)</td>
</tr>
<tr>
<td>PHQ-9</td>
<td>41.5% (291)</td>
<td>36.7% (257)</td>
</tr>
<tr>
<td>PROMIS ShortForm</td>
<td>35.2% (247)</td>
<td>33.7% (236)</td>
</tr>
<tr>
<td>PHQ9_PROMIS</td>
<td>38.1% (267)</td>
<td>35.2% (247)</td>
</tr>
</tbody>
</table>

*includes only patients who had diagnosis of depression by both PHQ9 and relevant PROM
# Results

Depression Remission: The percentage of patients with initial PHQ-9 score >9 who have a follow-up PHQ-9 score < 5 (based on NQF #0711)

<table>
<thead>
<tr>
<th>PROM</th>
<th>Group Level</th>
<th>Patient Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N = 701</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% Depression at Baseline (n)</td>
<td>% Depression Follow-up (n)</td>
</tr>
<tr>
<td>PHQ-9</td>
<td>41.5% (291)</td>
<td>36.7% (257)</td>
</tr>
<tr>
<td>PROMIS ShortForm</td>
<td>35.2% (247)</td>
<td>33.7% (236)</td>
</tr>
<tr>
<td>PHQ9PROMIS</td>
<td>38.1% (267)</td>
<td>35.2% (247)</td>
</tr>
</tbody>
</table>

*includes only patients who had diagnosis of depression by both PHQ9 and relevant PROM
## Results

Depression Remission: The percentage of patients with initial PHQ-9 score >9 who have a follow-up PHQ-9 score < 5 (based on NQF #0711)

<table>
<thead>
<tr>
<th>N = 701</th>
<th>Group Level</th>
<th>Patient Level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PROM</strong></td>
<td>% Depression at Baseline (n)</td>
<td>% Depression Follow-up (n)</td>
</tr>
<tr>
<td>PHQ-9</td>
<td>41.5% (291)</td>
<td>36.7% (257)</td>
</tr>
<tr>
<td>PROMIS ShortForm</td>
<td>35.2% (247)</td>
<td>33.7% (236)</td>
</tr>
<tr>
<td>PHQ9PROMIS</td>
<td>38.1% (267)</td>
<td>35.2% (247)</td>
</tr>
</tbody>
</table>

*includes only patients who had diagnosis of depression by both PHQ9 and relevant PROM
Results

Progress towards Remission: The percentage of patients with initial PHQ-9 score >9 who have a follow-up PHQ-9 score that is reduced by ≥50% (based on NQF #1885)

<table>
<thead>
<tr>
<th></th>
<th>N=701</th>
<th>Group Level</th>
<th>Patient Level</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>% with Depression at Baseline, (n)</td>
<td>% Progress towards Remission, (n)</td>
<td>Difference (PHQ-9 - other PROM)*</td>
</tr>
<tr>
<td>PROM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHQ-9</td>
<td>41.5% (291)</td>
<td>14.8% (43/291)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PROMIS ShortForm</td>
<td>35.2% (247)</td>
<td>15.8% (39/247)</td>
<td>-1.0%</td>
<td>88.9% (176/198)</td>
</tr>
<tr>
<td>PHQ9PROMIS</td>
<td>38.1% (267)</td>
<td>15.7% (42/267)</td>
<td>-0.9%</td>
<td>94.6% (246/260)</td>
</tr>
</tbody>
</table>

*includes only patients who had diagnosis of depression by both PHQ9 and relevant PROM
## Results

Progress towards Remission: The percentage of patients with initial PHQ-9 score >9 who have a follow-up PHQ-9 score that is reduced by ≥ 50% (based on NQF #1885)

<table>
<thead>
<tr>
<th></th>
<th>Group Level</th>
<th>Patient Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% with Depression at Baseline, (n)</td>
<td>% Progress towards Remission, (n)</td>
</tr>
<tr>
<td>PHQ-9</td>
<td>41.5% (291)</td>
<td>14.8% (43/291)</td>
</tr>
<tr>
<td>PROMIS ShortForm</td>
<td>35.2% (247)</td>
<td>15.8% (39/247)</td>
</tr>
<tr>
<td>PHQ9PROMIS</td>
<td>38.1% (267)</td>
<td>15.7% (42/267)</td>
</tr>
</tbody>
</table>

*includes only patients who had diagnosis of depression by both PHQ9 and relevant PROM
## Results

Progress towards Remission: The percentage of patients with initial PHQ-9 score >9 who have a follow-up PHQ-9 score that is reduced by $\geq 50\%$ \textit{(based on NQF #1885)}

<table>
<thead>
<tr>
<th>N=701</th>
<th>Group Level</th>
<th>Patient Level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PROM</strong></td>
<td>% with Depression at Baseline, (n)</td>
<td>% Progress towards Remission, (n)</td>
</tr>
<tr>
<td>PHQ-9</td>
<td>41.5% (291)</td>
<td>14.8% (43/291)</td>
</tr>
<tr>
<td>PROMIS ShortForm</td>
<td>35.2% (247)</td>
<td>15.8% (39/247)</td>
</tr>
<tr>
<td>PHQ9PROMIS</td>
<td>38.1% (267)</td>
<td>15.7% (42/267)</td>
</tr>
</tbody>
</table>

*includes only patients who had diagnosis of depression by both PHQ9 and relevant PROM
Conclusion

• High concordance in performance at the group level for depression PRO-PMs measured using PHQ-9, PROMIS SF and PHQ-9/PROMIS.

• Findings support the ability to use linkage of scale scores to assess performance of PRO-PMs using different PROMs.

• This can enable depression measures choice flexibility for the HCS, and enable CER.