Overview

1. Why measure patient-reported health status?
2. Different types of PROs
3. Development and evaluation of PROs
4. “New” methods using Item Response Theory
5. Interpreting scores on PRO measures

Why measure patient-reported health status?
How do we assess benefit?

Copies of virus in blood
Tumor size
Blood pressure
PeakVO2

For endpoints to inform decisions, they must matter to patients, clinicians, and payers.
**PeakV02**

$r = .53$

Flynn et al, 2015

---

**Overactive Bladder Syndrome**

Pad weight to measure leakage volume

What is a meaningful reduction in volume?

---

**Treatment Benefit**

Feel  Function  Survive

FDA
**Patient-Reported Outcome (PRO)**

“A measurement based on a report that comes directly from the patient (i.e., study subject) about the status of a patient’s health condition without amendment or interpretation of the patient’s response by a clinician or anyone else.”

**PROMIS® Fatigue Measure**

<table>
<thead>
<tr>
<th>In the past 5 days...</th>
<th>Never</th>
<th>Rarely</th>
<th>Sometimes</th>
<th>Often</th>
<th>Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>How often did you feel tired?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>How often did you experience extreme exhaustion?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>How often did you run out of energy?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>How often did your fatigue limit you at work (include work at home)?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>How often were you too tired to think clearly?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Different Types of PROs**
1. Dilution of effects of biological interventions

2. Correlation between successive boxes decreases

Adapted from Wilson & Cleary (JAMA 1995)

3. Development and Evaluation of PRO Measures

1. Determine what PRO concept we want to measure and why
2. Collect qualitative data to understand meaning of the PRO concept

Fatigue

3. Write items you think will measure the concept

Fatigue

Item 1
Item 2
Item 3
Item 4
Item 5
Item 6
Item 7
Item 8

4. Test items for understanding (cognitive interviews)

Item 1
Item 2
Item 3
Item 4
Item 5
Item 6
Item 7
Item 8
Item 9
Item 10
5. Administer items to a large sample of people

Item 1
Item 2
Item 3
Item 4
Item 5
Item 6
Item 7
Item 8

6. Use psychometric (statistical) analyses to see how well items are working and develop scoring method

How well do items fit this model?

Fatigue Score

7. Evaluate the reliability and validity of the measure

Validity
Measures what it's supposed to measure

Reliability
Measures with little error (a.k.a. precision)
**Types of Validity**

Content

Face

Construct

Convergent/discriminant

Known groups

Predictive

Responsiveness

**Convergent Validity: PROMIS Depression Domain**

If I have not changed, I should get the same score...

using different sets of items from the same measure

Internal Consistency (Cronbach’s alpha)

Test-Retest

regardless of who scores it

Interrater
Everyone must complete the same items.

All items are necessary to obtain a score.

Score might not be on the same metric as other measures of the same thing.
An item bank is a large collection of items measuring a single domain.

Any and all items can be used to provide a score for that domain.

Dynamic, not fixed.
Using a PRO Item Bank

Item Bank

Fixed-Length Measures

Computerized Adaptive Tests (CATs)

Ready-made Make-your-own

Next item administered depends on answer to previous item
Fatigue Item Bank

- **Chemotherapy trial** Items 1-10
- **Osteoarthritis trial** CAT
- **Diabetes trial** Item 7
- **Heart failure trial (NYHA Class III)** Items 6-12

**Same metric, same meaning**

Different Sites, Different Measures

- **Measure 1 (in HCS 1)**
  - Item 1
  - Item 2
  - Item 3
  - Item 4
  - Item 5
  - Metric 1

- **Measure 2 (in HCS 2)**
  - Item 1
  - Item 2
  - Item 3
  - Item 4
  - Metric 2

- **Measure 3 (in HCS 3)**
  - Item 1
  - Item 2
  - Item 3
  - Item 4
  - Item 5
  - Item 6
  - Metric 3

Item Bank

- **Measure 1 (in HCS 1)**
  - Item 1
  - Item 2
  - Item 3
  - Item 4
  - Item 5

- **Measure 2 (in HCS 2)**
  - Item 1
  - Item 2
  - Item 3
  - Item 4

- **Measure 3 (in HCS 3)**
  - Item 1
  - Item 2
  - Item 3
  - Item 4
  - Item 5
  - Item 6
  - Other items

**One Metric**

- Item 1
- Item 2
- Item 3
- Item 4
### Traditional PRO Measure vs. PRO Item Bank

<table>
<thead>
<tr>
<th>Traditional PRO Measure</th>
<th>PRO Item Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>All items are required to compute a score</td>
<td>Any and all subsets of items can generate a score</td>
</tr>
<tr>
<td>Everyone must take same items</td>
<td>Different people can get different items</td>
</tr>
<tr>
<td>Use it “off the shelf”</td>
<td>Use items in bank to create measure for specific use</td>
</tr>
<tr>
<td>Scores not easily comparable to scores from another measure of the same domain</td>
<td>Cross-walk between scores from different measures in the same item bank</td>
</tr>
</tbody>
</table>

### Examples of PRO Resources Based on Item Banks

- Adult and pediatric item banks measuring domains relevant across chronic diseases
- Freely available
- [www.nihpromis.org](http://www.nihpromis.org)
Multidimensional set of brief measures assessing cognitive, emotional, motor and sensory function from ages 3-85

Freely available

www.nihtoolbox.org

Core set of instruments for use in chronic neurological conditions (supplemental set for specific diseases, patient subgroups)

Freely available

www.neuroqol.org

Differential Item Functioning
In the past 7 days, did you cry?

Yes  No

(Depression item)

Differential Item Functioning

Item behaves differently for 2 or more groups.

Probability of “YES”
In the past 7 days, did you cry?

Probability of "YES"

Depression

Differential Item Functioning

Item behaves differently for 2 or more groups.

The "map" between depression and item is different for 2 or more groups.
In the past 7 days, did you cry?

Probability of "YES"

Depression

Females

Males

Interpreting Scores on PRO Measures

Effects of Exercise Training on Health Status in Patients With Chronic Heart Failure

HF-ACTION Randomized Controlled Trial
Exercise vs Usual care

Changes in Kansas City Cardiomyopathy Questionnaire
(Score Range: 0 - 100)

1.93
(95% C.I., 0.84, 3.01)

Exercise arm has statistically greater rate of change between baseline and 3 months.

A. KCCQ Overall Summary Score

Spertus et al., 2005
Mean Diff = 1.93
Exercise = 54%
Control = 29%

Review

1. Why measure patient-reported health status?
2. Different types of PROs
3. Development and evaluation of PROs
4. “New” methods using IRT
5. Interpreting scores on PRO measures

Questions

kevin.weinfurt@duke.edu