Module 6d: Patient-Reported Outcomes: Reliability, Validity, and Responsiveness

Category 2: Patient-Centeredness

Prepared by Albert Wu, MD, MPH
Clifton Bingham, MD

Presented by Albert Wu, MD, MPH
Module 6d

Patient-Reported Outcomes (PC-3): Reliability, Validity, and Responsiveness

This lecture is part of a series related to the PCORI Methodology Standards Academic Curriculum.
Medical Outcomes Trust Scientific Review Criteria

- Conceptual and measurement model
- Reliability
- Validity
- Responsiveness
- Interpretability
- Respondent and administrative burden
- Alternative forms / cultural and language adaptations
COnsensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) Criteria for Quality of PRO

COSMIN checklist manual

Lidwine B. Mokkink
Caroline B Terwee
Donald L Patrick
Jordi Alonso
Paul W Stratford
Dirk L Knol
Lex M Bouter
Henrica CW de Vet
Estimating Reliability
Measurement of any phenomenon always contains a certain amount of chance error.

Every observed score on any measuring instrument is made up of two quantities:
- True score (T)
- Some random error (E)

\[ X = T + E \]
Systematic error = bias

This differentiates reliability from validity

If systematic error is known, it can be adjusted for
Reliability

- Proportion of true variance in an observation

- Signal (versus noise)
  - Example
    - Music versus noise while listening to radio in your car
The extent to which a measuring procedure yields the same results on repeated trials

Assessed by:
- Reproducibility
- Internal consistency
Reproducibility

- The extent to which a given score will be observed:
  - Test-retest
    - At two different points in time when no change has occurred
  - Alternative forms
    - By two equivalent but nonidentical measuring instruments
Estimating Reliability

- Parallel measurements
  - Parallel items are functions of the same true score
  - Differences between them are the result of purely random error

- You can estimate the reliability of empirical measurements if you have:
  - Two or more items for a single concept
  - A single item measured at two points in time

- In general, the greater the number of separate measurements of a given phenomenon, the more accurate (and greater) the estimate of its reliability will be
For a multi-item scale, the extent that items “hang together” as a measure of the intended concept

- Cronbach’s alpha = \( N\rho / [1 + \rho(N - 1)] \)
  - Where \( \rho \) is the mean inter-item correlation, and \( N \) is the number of items

- Range is 0 to 1
Calculate Internal Consistency for MHI-5

- Items a, b, c, d, and e
- Mean of correlations between a-b, a-c, a-d, a-e, b-c, b-d, b-e, c-d, c-e, d-e = \( \rho \)
- If \( \rho = 0.60 \), then
  - Alpha = \( 5\rho / [1 + \rho(4)] \) = 0.88
Any measurement contains some random error

Reliability is a function of the amount of true variance in an observed score

Reliability is commonly estimated using:
- Test-retest
- Internal consistency

Cronbach’s alpha is a commonly used measure of internal consistency
- As a rule of thumb, 0.70 is adequate for group comparisons
Validity and Responsiveness
Validity

- The extent to which an instrument measures what it is supposed to measure
- A PRO measure is valid for an intended purpose
- Validating a health measure is the process of accumulating different kinds of evidence to determine the most appropriate interpretations of a health score
Kinds of Validity

- Content validity
- Construct validity
- Criterion validity
- Responsiveness
Content Validity

- The extent to which a measure represents a specified domain or universe of content
Construct Validity

- Performance relative to a conceptual model of hypothesized relationships

- Example:
  - Physical function should decline with age
    - Convergent/discriminant validity
    - Agreement/difference among measures assessing same/different concept
Convergent Validity

- Extent to which measures of the same concept correlate with each other

- Example:
  - Five items to measure mental health correlated 0.7 - 0.9 with each other
Relative lack of correlation of a measure with other measures that are intended to be different

Example:
- Items to measure physical function more correlated with each other than with mental health items
A method to test construct validity by examining multitrait analysis relationship among measures assessing similar and different concepts

Blends convergent and discriminant validity when more than one measurement method has been used

A multitrait-multimethod matrix:
- Intercorrelations of two or more health concepts measured by two or more methods
Construct Validity

- Known groups
  - Difference in scores for groups known to differ

- Example:
  - People with cataracts bilaterally have worse patient-reported visual functioning than people with cataract in one eye only
Criterion Validity

- The extent to which a measure relates to an accepted gold standard
Responsiveness

- Ability to discern (a small but clinically important) change
Responsiveness

- Sensitivity to a small but clinically important change

- Effect size = \(\frac{\text{mean at time 1} - \text{mean at time 2}}{\text{SD at time 1}}\)

- Standardized response mean = \(\frac{\text{mean at time 1} - \text{mean at time 2}}{\text{SD of difference}}\)
Practical Considerations

- Mode of administration
- Time to administer
- Language
- Respondent burden
- Availability of supporting materials
Evidence for the validity of PROMS is generated when they are related in expected ways to other sources of information.

PROMs can ...
- Predict important future events (clinical outcomes, hospitalization, job loss, death)
- Detect difference between treatments
- Detect changes over time (responsive)