



Best Practices for PRO Instrument Development and Validation

ISOQOL Patient-Reported Outcomes
and Regulatory Guidance Meeting

June 29, 2006
Washington, DC




Contributors

- ❖ William R. Lenderking
- ❖ Neil K. Aaronson
- ❖ Jakob Bue Bjerne
- ❖ Peter Fayers
- ❖ Ron D. Hays



Session Agenda

- ❖ Presentation
 - ❖ 5 key points from the Guidance for debate and discussion
- ❖ Response from Discussant
 - ❖ Nancy Santanello
- ❖ General Discussion

- 
- ❖ FDA to be commended for tackling a difficult set of issues and for doing a laudable job
 - ❖ Our process: we identified 5 key issues and compared the responses to FDA on those issues from different organizations
 - ❖ ISOQOL, EORTC, Pfizer, QualityMetric, PhRMA, ERIQA



Key Issues

- ❖ Role of the patient in instrument development
- ❖ Appropriate recall period
- ❖ Weighting of items in developing a summary score
- ❖ Appropriate validation
 - ❖ When items are used to measure a different concept
 - ❖ When the scale is to be used in a different population



Role of the Patient

- ❖ FDA: "...item generation is incomplete without patient involvement...FDA plans to review instrument development...to determine whether adequate numbers of patients...have supported the opinion that the specific items are adequate and appropriate to measure the concept." (295-300)

Role of the Patient

- ❖ Incorporation of patient input does not guarantee
 - ❖ Complete coverage
 - ❖ Construct validity
 - ❖ Predictive validity
- ❖ What is an adequate number of patients?
- ❖ Item generation is not clearly defined
 - ❖ Patients' words often require considerable editing
- ❖ Some scales developed without patient input still perform well

Role of the Patient

- ❖ Recommendations
 - ❖ Better not to be overly prescriptive regarding what constitutes an adequate number of patients
 - ❖ Define item generation more specifically
 - ❖ Patient input is most important for evaluating completeness of content coverage and item clarity and readability

Appropriate Recall Period

- ❖ FDA: "PRO instruments that require patients to rely on memory...or to average their response over a period of time may threaten the accuracy of the PRO data. It is usually better to...ask patients to describe their current state." (339-343)

Appropriate Recall Period

- ❖ Many tools require patients to recall experiences over time, which avoids the problem of being unduly influenced by a good or bad day
 - ❖ Symptoms vs. functional assessments
- ❖ Longer recall periods have the potential to increase measurement error through recall bias, although this would likely make it more difficult to detect treatment differences
- ❖ There is insufficient scientific evidence to support FDA's position: recall period should be informed by the disease and the question

Appropriate Recall Period

- ❖ Patients' evaluations of the experience during the reference period is what matters most, not the accuracy of recall of every single experience leading to that judgment
- ❖ Any response process relies on cognitive processing and to some extent on memory. Even "current state" is open to interpretation!
- ❖ The FDA position seems to imply that diaries or daily methods are preferable (e.g., Kahneman et al., Science 2004 Dec 3;306(5702):1776-80)

Appropriate Recall Period

- ❖ Recommendations:
 - ❖ The appropriate recall period should be determined by the disease, the question, the domain, and perhaps the trial design
 - ❖ Different forms of assessments could be developed with different recall periods
 - ❖ Symptoms, functional status, and general health perceptions might be validly measured using different recall periods, although using different recall periods in the same study could complicate interpretation

Weighting of items

- ❖ FDA: “Equally weighted scores for each item are appropriate only when the responses to the items are relatively uncorrelated...assigning equal weights to each item may overweight certain items if the number of response options of the values associated with response options varies by item.” (416-422)

Weighting of items

- ❖ Does this statement imply that uncorrelated items should be grouped together (hence undermining internal consistency), or simply that redundant domains should not be included in a domain?
- ❖ No recommendations are provided regarding how to choose an appropriate weighting strategy
 - ❖ In some scales, some content is intentionally overweighted, which approximates a weighting scheme when equal weights are applied

Weighting of items

- ❖ Weighting items separately can be unwieldy, contribute to a false precision, and increases the risk of calculation errors
 - ❖ Literature suggests that weighting does not add significantly to measurement precision (Wainer, Psychol Bull 1976; Arnold in Hand (ed), Measurement theory and practice: 2004; Dawes, Am Psych 1979)
 - ❖ Worse to calculate a total based on the wrong weights than to use no weights at all
- ❖ Within IRT, equal weighting is justified when items conform to the Rasch model
 - ❖ Similar item discrimination
 - ❖ Unidimensionality and local independence
 - ❖ Equal weighting is robust to deviations from assumption of equal item discrimination

Weighting of items

- ❖ Recommendations:
 - ❖ Delete the passage on weighting
 - ❖ Weighting is unnecessary when
 - ❖ Instrument has a fixed number of response options for all items
 - ❖ Items are of broadly similar relevance and importance
 - ❖ Items are reasonably highly correlated
 - ❖ Unequal weighting should be employed on a case-by-case basis, and should be reserved especially for scales with different numbers of response options

Appropriate validation when items or scales are modified

- ❖ FDA: “...recommends additional validation...of a modified PRO instrument when...an instrument that is developed and validated to measure one concept is used to measure a different concept, e.g.:
 - ❖ Single domain is extracted
 - ❖ Response options are changed to assess a different quality
 - ❖ Index is used when validation only applies to individual domains
 - ❖ Items from an existing PRO are used to create a new instrument
 - ❖ One or more items from an existing instrument are used to support a claim for a concept the items were not developed to measure” (581-607)

Appropriate validation when items or scales are modified

- ❖ Extraction of a single domain should not be an issue
- ❖ Additional validation required has not been specified
- ❖ Cognitive testing should be sufficient for minor modifications
- ❖ Recommendations are overly conservative

Appropriate validation when items or scales are modified

- ❖ Recommendations:
 - ❖ Define specifically the difference between major and minor measurement violations
 - ❖ Changing VAS from vertical to horizontal?
 - ❖ Extracting one domain?
 - ❖ Re-ordering items within a scale?
 - ❖ Combining validated domains into a single index?
 - ❖ Requiring field testing for all changes is too onerous. Pilot testing with cognitive debriefing should be sufficient for minor changes.

Appropriate validation when items or scales are modified

- ❖ Recommendations:
 - ❖ Minor changes not requiring field testing
 - ❖ Changing VAS from vertical to horizontal
 - ❖ Extracting one domain
 - ❖ Re-ordering items within a scale
 - ❖ Changing wording of instructions
 - ❖ It would help to identify a benchmark change beyond which quantitative evidence is required, such as combining all domains into a general index

Appropriate validation when scale is used in a different population or item content is changed

- ❖ FDA: "...recommends additional validation...of a modified PRO instrument when...an instrument developed for...one population or condition is used in a different...population or condition, ...[or] an instrument is altered in item content or format...e.g.:"
 - ❖ Patients in the proposed trial have a disease, condition or severity level that is different from...the population used for instrument and validation
 - ❖ Patients in the proposed trial differ in age, gender, race, or developmental or life stage
 - ❖ Number of items (more or fewer) used to assess a concept or domain..." (610-629)

Appropriate validation when scale is used in a different population or item content is changed

- ❖ Invariance literature has shown that severity of a condition does not have an impact on measurement properties between the severity levels, provided each item has variance
- ❖ To require modification or re-validation if a population is "different" is broad and burdensome
 - ❖ EORTC QLQC30 exists in 65 languages
- ❖ Computerized adaptive assessment, based on item banks, becomes impossible
 - ❖ NIH PROMIS is developing such for pain, fatigue, physical function, social function, and emotional well-being

Appropriate validation when scale is used in a different population or item content is changed

- ❖ Recommendations:
 - ❖ No re-validation is required for changes in severity level within limits usually observed in study populations (IQ of 40 vs 130)
 - ❖ Allow sponsors to make the case for re-validation plans on a case-by-case basis, and do not make the requirements overly heavy
 - ❖ Identify with greater specificity those changes in application which are minor
 - ❖ A different language does not automatically constitute a different population requiring additional validation above and beyond what is normally done in translation

Summary

- ❖ FDA has done an excellent job in drafting the guidelines
 - ❖ It is always easier to be critic than creator
- ❖ Recommendations are based on comments and feedback from a broad variety of groups including ISOQOL, PhRMA, Pfizer, EORTC, QualityMetric, ERIQA
- ❖ Please don't be overly prescriptive in applying the guidelines!
- ❖ Once guidelines are revised, there should be a period of review to determine if they are being applied as intended



Discussion

❖ Nancy Santanello