Patient-Reported Outcomes:
Measurement, Implementation and Interpretation

Joseph C. Cappelleri and Andrew Bushmakin
Pfizer Inc

joseph.c.cappelleri@pfizer.com  andrew.g.bushmakin@pfizer.com

4-hour short course at the New York City Metropolitan Area Chapter,
American Statistical Association, New York, NY,
April 1, 2014
Acknowledgment: Book Reference

Cappelleri JC, Zou KH, Bushmakin AG, Alvir JMJ, Alemayehu D, Symonds T.

*Patient-Reported Outcomes: Measurement, Implementation and Interpretation*

Boca Raton, Florida: Chapman & Hall/CRC; 2014
Outline

• Chapter 1: Introduction
• Chapter 2: Development of a Patient-Reported Outcome
• Chapter 3: Validity
• Chapter 4: Reliability
• Chapter 5: Exploratory and Confirmatory Factor Analyses
• Chapter 6: Item Response Theory
• Chapter 7: Cross-sectional Analysis
• Chapter 8: Longitudinal Analysis
• Chapter 9: Mediation Models
• Chapter 10: Missing Data
• Chapter 11: Enriching Interpretation
Chapter 1: Introduction
Patient-Reported Outcomes in Perspective

• Patient-reported outcomes (PROs): Any report on the status of a patient’s health condition that comes directly from the patient
  – Without interpretation of the patient’s response by a clinician or anyone else

• PRO is an umbrella term that includes a whole host of subjective outcomes
  – Pain, fatigue, depression
  – Treatment satisfaction, aspects of well-being
  – Physical symptoms (e.g., nausea, vomiting)
  – Health-related quality of life
PROs in Clinical Research

• In general, the same clinical design principles that apply to other clinical endpoints also apply to PRO endpoints

• Five characteristics that tend to be associated with PRO measures
  – 1. Missing data
  – 2. Psychometrics
  – 3. Interpretation
  – 4. Multiple comparisons
  – 5. Longitudinal data and analysis
Label Claims

• 1997 to 2002: PRO evidence cited in Clinical Studies section for 30% of the new medical product approvals
  – 11% of the new products were approved on the basis of PROs alone

• 2006 to 2010: Of 116 medical products, 28 (24%) were granted PRO claims
  – 24 of 28 (86%) were symptoms and 20 (71%) were primary endpoints
  – Majority were simple one-item scales and traditionally accepted measures

• Several publications are available such as the FDA guidance and the EMA guidance
  – Qualified drug development tools to improve the process (FDA)
Chapter 2: Development of a Patient-Reported Outcome
Process for Developing a New PRO

Population selection
- Age, severity, culture
- Purposive sampling

Item generation
- Patient interviews, focus groups
- Analysis
- Saturation

Item wording generation
- Reading age
- Clarity of wording

Cognitive interviews
- Confirmation of PRO content relevance, clarity of wording
Item Generation:
Patient Interviews and Focus Groups

- To generate the content of a PRO measure, either individual interviews or group discussion with patients will be required
  - Choice of the two depends on sensitivity of the topic and whether there is a need to react and feed off each other
  - Focus groups require a skilled moderator
  - Individual interviews allow more in-depth discussion and are easier to analyze but take longer to collect data
  - Generally a mixture of the two approaches is beneficial

- Qualitative research is a scientific technique which requires a protocol outlining the study details (e.g., inclusion/exclusion criteria, number of subjects, pre-specification if particular subgroups are to be recruited)
  - Development of interview guide for the right and consistent questions
  - Questions open-ended and broad: “What kinds of sleep difficulties do you experience?”
  - Followed by more detailed or specific questions: “You said that you had problems with staying asleep, can you describe in more detail what specifically these problems are?”
Analysis of Qualitative Data

• Once data are collected, analysis of the verbatim transcripts is then conducted
  – Develop coding whereby similar concepts are given a code name (for fatigue, concepts might be tiredness and unrested)
  – List of patient statements per code and their frequency
  – Develop codes, group them into concepts, and from these concepts a theory about data is developed

• From this process a conceptual framework will emerge
  – Based on clinical input and literature review along with qualitative interviews
Example of a Conceptual Framework

Sleep Disturbance (Concept)

Falling asleep (Domain)
- Item 1: How difficult was it to fall asleep?
- Item 2: How difficult was it to get comfortable?

Staying asleep (Domain)
- Item 3: How difficult was it to stay asleep?
- Item 4: How restless was your sleep?

Impact (Domain)
- Item 5: How rested were you when you woke up?
- Item 6: How difficult was it to start your day?
Other Considerations

• Recall period should be considered

• Consider most appropriate response options

• Saturation

• Item wording

• Cognitive interviews
Chapter 3: Validity
Content Validity

• Content validity is the degree to which the content of a measurement instrument is an adequate reflection of the concept (construct) to be measured
  – Qualitative work is central

• Reflects the instruments ability to measure the stated concepts in the relevant population

• How well does the PRO instrument capture all of the important aspects of the concept from the patient’s perspective?

• Face validity, a component of content validity, is the degree to which a measurement looks as though it is an adequate reflection of the construct to be measured
Construct Validity

• Construct validity is the degree to which the scores of a measurement instruments are consistent with hypotheses
  – Internal relationships
  – Relationship with scores on other instruments
  – Differences between relevant groups

• Assessments made through correlations, changes over time, and differences between groups of patients

• Three types of construct validity
  – Convergent and divergent validity
  – Known-groups validity
  – Criterion validity
## Item-Level Discriminant Test

<table>
<thead>
<tr>
<th>SEAR Item</th>
<th>Item-to-Total Correlations</th>
<th>Domain: Sexual Relationship Satisfaction</th>
<th>Domain: Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I felt relaxed about initiating sex with my partner</td>
<td></td>
<td>0.68</td>
<td>0.50</td>
</tr>
<tr>
<td>2. I felt confident that during sex my erection would last long enough</td>
<td></td>
<td>0.74</td>
<td>0.56</td>
</tr>
<tr>
<td>3. I was satisfied with my sexual performance</td>
<td></td>
<td>0.85</td>
<td>0.54</td>
</tr>
<tr>
<td>4. <em>I felt that sex could be spontaneous</em></td>
<td></td>
<td>0.62</td>
<td>0.49</td>
</tr>
<tr>
<td>5. I was likely to initiate sex</td>
<td></td>
<td>0.63</td>
<td>0.44</td>
</tr>
<tr>
<td>6. I felt confident about performing sexually</td>
<td></td>
<td>0.82</td>
<td>0.57</td>
</tr>
<tr>
<td>7. I was satisfied with our sex life</td>
<td></td>
<td>0.82</td>
<td>0.60</td>
</tr>
<tr>
<td>8. My partner was unhappy with the quality of our sexual relations</td>
<td></td>
<td>0.57</td>
<td>0.34</td>
</tr>
<tr>
<td>9. I had good self-esteem</td>
<td></td>
<td>0.48</td>
<td>0.68</td>
</tr>
<tr>
<td>10. I felt like a whole man</td>
<td></td>
<td>0.56</td>
<td>0.73</td>
</tr>
<tr>
<td>11. <em>I was inclined to feel that I am a failure</em></td>
<td></td>
<td>0.37</td>
<td>0.50</td>
</tr>
<tr>
<td>12. I felt confident</td>
<td></td>
<td>0.51</td>
<td>0.71</td>
</tr>
<tr>
<td>13. <em>My partner was satisfied with our relationship in general</em></td>
<td></td>
<td>0.55</td>
<td>0.63</td>
</tr>
<tr>
<td>14. I was satisfied with our relationship in general</td>
<td></td>
<td>0.52</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Source: Cappelleri et al. 2004
Known-Groups Validity:
Mean Scores and 95% Confidence Intervals

Source: Cappelleri et al. 2004
## Criterion Validity: Concurrent Validity

<table>
<thead>
<tr>
<th>Erectile Function Domain</th>
<th>Clinical Diagnosis of ED</th>
<th>Clinical Diagnosis of no ED</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED (≤ 25)</td>
<td>1000 (true positive)</td>
<td>14 (false positive)</td>
<td>1014</td>
</tr>
<tr>
<td>No ED (26-30)</td>
<td>35 (false negative)</td>
<td>102 (true negative)</td>
<td>137</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1035</strong></td>
<td><strong>116</strong></td>
<td><strong>1151</strong></td>
</tr>
</tbody>
</table>

ED = Erectile Dysfunction.

Estimated odds ratio of 0.54 (95% confidence interval, 0.48 to 0.60):  
For every one-point increase in the erectile function score, the odds of having ED (relative to not having ED) decreased by about half

Sensitivity = true positive rate = \(\frac{1000}{1035} = 0.97\)

Specificity = true negative rate = \(\frac{102}{116} = 0.88\)

Source: Cappelleri et al. 1999
Concurrent Validity:
Receiver Operating Characteristic Curve

Area Under Curve = 0.97

97% chance that a randomly selected subject with ED had a lower erectile function score (and hence more likely to be diagnose with ED) than a randomly chosen subject without ED.

Source: Cappelleri et al. 1999
Chapter 4: Reliability
Reliability

• Validity assesses the extent to which an instrument measures what it is meant to measure

• Reliability assesses how precise or stable the instrument measures what it measures and it typically discussed in terms of reproducibility
  – Internal reliability vs. repeatability reliability

• Repeatability reliability
  – Test-retest reliability
  – Inter-rater reliability
  – Equivalent-forms reliability
Intraclass Correlation Coefficient (ICC): Continuous Variables

- Reliability expresses how well patients with true systematic differences can be distinguished from each other (after accounting for measurement error)

- ICC is a reliability parameter that measures the strength of agreement between repeated measurements on the same set of patients

\[
\text{ICC for a single score} = \frac{\text{Between Patient Variability}}{\text{Between Patient Variability} + \text{Within Patient Variability}}
\]

\[
\text{ICC for a single score} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_e^2}
\]

\[
\text{ICC for an average score} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_e^2}{m}}
\]
Reliability of the Erection Hardness Score

ICC based on an average number of responses ($m$)

Source: Mulhall et al. 2007
Illustration of a Bland-Altman Plot

Difference of screening and baseline scores vs. Mean of screening and baseline scores.
### Simple Kappa for Binary Data

**Gold Standard**

<table>
<thead>
<tr>
<th>Erectile Function Domain</th>
<th>Clinical Diagnosis of ED</th>
<th>Clinical Diagnosis of no ED</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED (≤ 25)</td>
<td>1000 (true positive)</td>
<td>14 (false positive)</td>
<td>1014</td>
</tr>
<tr>
<td>No ED (26-30)</td>
<td>35 (false negative)</td>
<td>102 (true negative)</td>
<td>137</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1035</td>
<td>116</td>
<td>1151</td>
</tr>
</tbody>
</table>

ED = Erectile Dysfunction.

The Kappa statistic is calculated as:

\[
k = \frac{p_o - p_e}{1 - p_e}
\]

Substituting the values:

\[
k = \frac{\left(\frac{1000 + 102}{1151}\right) - \left(\frac{911.8 + 13.8}{1151}\right)}{1 - \left(\frac{911.8 + 13.8}{1151}\right)} = 0.78
\]

Source: Cappelleri et al. 1999
Internal Consistency Reliability: Cronbach’s Alpha Coefficient

- Applies to consistency of responses on the same multi-item scale
  - Items are intended to tap the same unidimensional construct

\[
\text{Cronbach’s alpha} = \frac{n \bar{r}}{1 + (n - 1) \bar{r}}
\]

- \( n \) the number of items
- \( \bar{r} \) the average inter-item correlation

- Five components of SEAR questionnaire gave a range from 0.76 to 0.91 (Cappelleri et al. 2004)
Chapter 5:
Exploratory and Confirmatory Factor Analyses
Chapter 6:
Item Response Theory
What is Item Response Theory?

- A statistical theory consisting of mathematical models expressing the probability of a particular response to a scale item as a function of the (latent or unobserved) attribute of the person and of certain parameters or characteristics of the item

- Assumptions
  - Unidimensionality, local independence, and model fit
# Dichotomous Item Response Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Mathematical Form</th>
<th>Item Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-parameter logistic</td>
<td>[ P_{ij} (\theta_j) = \frac{1}{1 + e^{-(\theta_j - b_i)}} ]</td>
<td>Difficulty ((b))</td>
</tr>
<tr>
<td></td>
<td>[ \ln \left[ \frac{P_{ij}(\theta_j)}{1 - P_{ij}(\theta_j)} \right] = (\theta_j - b_i) ]</td>
<td></td>
</tr>
<tr>
<td>Two-parameter logistic</td>
<td>[ P_{ij}(\theta_j) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}} ]</td>
<td>Difficulty ((b)), Discrimination ((a))</td>
</tr>
<tr>
<td>Three-parameter logistic</td>
<td>[ P_{ij} (\theta_j) = c + (1 - c) \left[ \frac{1}{1 + e^{-a_i(\theta_j - b_i)}} \right] ]</td>
<td>Difficulty ((b)), Discrimination ((a)), Guessing ((c))</td>
</tr>
</tbody>
</table>

*Note: \(\theta_j\) is the latent attribute of the person*
Item Characteristic Curves
Two Items of Differing Difficulty: Rasch Model

Item A: Walk
Item B: Run

Probability of Positive Response vs. Physical Functioning ($\theta$)

POOR EXCELLENT
Item Characteristic Curves
Two Items of Differing Discrimination and Difficulty

Probability of Positive Response

Physical Functioning (\(\theta\))

POOR

EXCELLENT

Item A: Walk
\((\theta=0.70, p=0.80)\)
\((\theta=-0.69, p=0.20)\)

Item B: Run
\((\theta=1.93, p=0.80)\)
\((\theta=0.08, p=0.20)\)
Polytomous Response Model

• Partial credit model – generalization of the one-parameter (Rasch) dichotomous IRT model
  – Category threshold parameters are akin to difficulty thresholds in binary IRT models
  – These parameters reflect the attribute level at which a particular response category of an item becomes as likely (50% chance) to be responded to as previous category
  – Rating scale model is a special case

• Graded response model – extension of the two-parameter dichotomous IRT model

\[
P_{ik} (\theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ik})}} - \frac{1}{1 + e^{-a_i(\theta - b_{i(k+1)})}}
\]
During the last seven days, how much of the time have you accomplished your daily activities as a result of your physical health?
Item Information Function:
A Good Item (Item 1) and a Poorer Item (Item 2)

\[ I(\theta)_i = a_i^2 P_i (1 - P_i) \]
## Common IRT Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Item Response Format</th>
<th>Model Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasch / 1-Parameter Logistic</td>
<td>Dichotomous</td>
<td>Discrimination power equal across all items. Threshold varies across items.</td>
</tr>
<tr>
<td>2-Parameter Logistic</td>
<td>Dichotomous</td>
<td>Discrimination and threshold parameters vary across items.</td>
</tr>
<tr>
<td>Graded Response</td>
<td>Polytomous</td>
<td>Ordered responses. Discrimination varies across items.</td>
</tr>
<tr>
<td>Nominal</td>
<td>Polytomous</td>
<td>No pre-specified item order. Discrimination varies across items.</td>
</tr>
<tr>
<td>Partial Credit (Rasch Model)</td>
<td>Polytomous</td>
<td>Discrimination power constrained to be equal across items.</td>
</tr>
<tr>
<td>Rating Scale (Rasch Model)</td>
<td>Polytomous</td>
<td>Discrimination equal across items. Item threshold steps equal across items.</td>
</tr>
<tr>
<td>Generalized Partial Credit</td>
<td>Polytomous</td>
<td>Variation of Partial Credit Model with discrimination varying among items.</td>
</tr>
</tbody>
</table>
Illustration of Person-Item Map on Physical Functioning (Rasch Rating Scale Model)

Notes:
1. More easy items than difficult ones; items on moderate activity needed
2. Some items have the same difficulty
3. Patients tend to cluster at the higher end of the scale, indicating that they will endorse most items
Fit Indices

- More common to assess item fit than model fit
- Key concept is residual
- Rasch models have input and output mean square statistics for item fit (and also person fit)
- Monotonicity of average attribute estimates as category level of item increases
- Rasch models offer person separation (reliability) index
Example

- National Eye Institute-Visual Function Questionnaire (NEI-VFQ)

- Consider its six-item near-vision subscale

- Each item has the same set of response options
  - 0 = stop doing because of eyesight
  - 25 = extreme difficulty
  - 50 = moderate difficulty
  - 75 = a little difficulty
  - 100 = no difficulty at all

- Domain score ranges from 0 (worst) to 100 (best)
Probability Curves of Three Items: Near-vision Subscale of NEI-VFQ
Chapter 7: Cross-Sectional Analysis
Types of PRO Data

• Binary, ordinal (Likert, adjectival), continuous scales may influence the statistical method used

• Visual analogue scale

How severe is your pain right now? Place a vertical mark on the line below to indicate the severity of your pain.

No Pain At All |

Worst Pain Imaginable
Comparing Two or More Samples

• Nonparametric methods
  – Does not require data be normally distributed
  – Test for differences in distributions between groups

• Parametric methods
  – More powerful than nonparametric tests when data are approximately normally distributed
  – Often more interpretable than nonparametric test
  – Central limit theorem

• Under most circumstance data from ordinal rating scales can be analyzed as if they were interval-level measurement without introducing severe bias
Chapter 8: Longitudinal Analysis
Analytic Considerations

• Longitudinal data are quite common in and central to PRO studies, especially clinical trials

• Interest centers on how disease or intervention affects an individual’s functioning and well-being over time

• Longitudinal analysis also considers how groups change over time and how between-group factors – like treatment – affect groups over time

• Generally two types for PRO measures taken as a continuous outcome
  – Repeated measures model
  – Random coefficients model
Consider a hypothetical example of a clinical two-arm study for a new treatment:

- Six hundred subjects are selected from the population of interest.
- Each subject is given an active drug or placebo during the 4-week study.
- The outcome variable is a PRO measure \((Y)\) and the covariates are time and treatment.

One simple form of a regression model for subject \(i\) at measurement occasion \(j\) on treatment \(k\) can be denoted as a sum of four terms:

\[
Y_{ijk} = a + b_j + r_k + e_{ijk},
\]

where

- \(Y_{ijk}\) is the PRO response for subject \(i\) at the measurement occasion \(j\) on treatment \(k\) \((i = 1, 2, \ldots, 600; j = 1, 2, 3, 4; k = 1, 2)\)
- \(a\) is the overall mean
- \(b_j\) is the fixed time effect at week \(j\)
- \(r_k\) is the fixed effect of the treatment \(k\)
- \(e_{ijk}\) is the error term associated with outcome measurement \(Y_{ijk}\)
Repeated Measures Model with Time as Categorical Covariate

\[ Y_{ijk} = a + b_j + r_k + e_{ijk} \]

**Model Y = Visit Treatment**

**Repeated Visit / Subject=ID Type=UN**

```
Proc Mixed data=mixed_ds_1;
    Class Visit Treatment ID ;
    Model Y = Visit Treatment / Solution ddfm=kr;
    Repeated Visit / Subject=ID Type=UN ;
Run;
```
Unstructured Variance-Covariance Matrix

\[
\begin{bmatrix}
\sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\
\sigma_{21} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\
\sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} \\
\sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \\
\end{bmatrix}
\]
SAS Output from Simulated Data:
600 subjects, 2 treatments, 4 time points (weeks)

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN(1,1)</td>
<td>ID</td>
<td>3.4192</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>ID</td>
<td>0.03207</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>ID</td>
<td>2.4887</td>
</tr>
<tr>
<td>UN(3,1)</td>
<td>ID</td>
<td>0.08599</td>
</tr>
<tr>
<td>UN(3,2)</td>
<td>ID</td>
<td>-0.01222</td>
</tr>
<tr>
<td>UN(3,3)</td>
<td>ID</td>
<td>3.9962</td>
</tr>
<tr>
<td>UN(4,1)</td>
<td>ID</td>
<td>0.1501</td>
</tr>
<tr>
<td>UN(4,2)</td>
<td>ID</td>
<td>0.02173</td>
</tr>
<tr>
<td>UN(4,3)</td>
<td>ID</td>
<td>0.06495</td>
</tr>
<tr>
<td>UN(4,4)</td>
<td>ID</td>
<td>4.5675</td>
</tr>
</tbody>
</table>

Solution for Fixed Effects

| Effect      | Visit | Treatment | Standard Estimate | Error | Pr > |t| |
|-------------|-------|-----------|-------------------|-------|------|---|
| Intercept   |       |           | 14.0102           | 0.09549 | .0001 |
| Visit       | 1     |           | -1.4979           | 0.1132  | .0001 |
| Visit       | 2     |           | -0.5233           | 0.1081  | .0001 |
| Visit       | 3     |           | 1.0694            | 0.1186  | .0001 |
| Visit       | 4     |           | 0                  |         |      |
| Treatment   | 1     |           | 1.9975            | 0.07760 | .0001 |
| Treatment   | 2     |           | 0                  |         |      |
Published Example
Urge to Smoke -- Repeated Measures Analysis for Week 1 through Week 7

<table>
<thead>
<tr>
<th>Treatment</th>
<th>n</th>
<th>Least-Squares Mean (SE)</th>
<th>Comparison vs. Placebo</th>
<th>95% CI</th>
<th>P Value</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Varenicline</td>
<td>341</td>
<td>1.11 (0.04)</td>
<td>−0.54 (0.06)</td>
<td>−0.66 to −0.42</td>
<td>.001</td>
<td>−0.67</td>
</tr>
<tr>
<td>Bupropion SR</td>
<td>318</td>
<td>1.41 (0.05)</td>
<td>−0.24 (0.06)</td>
<td>−0.36 to −0.12</td>
<td>.001</td>
<td>−0.30</td>
</tr>
<tr>
<td>Placebo</td>
<td>337</td>
<td>1.65 (0.05)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Source: Gonzales et al. 2006
Random Coefficients Models: Random Intercept-Slope Model

\[ Y_{ij} = a + b \times t_{ij} + \alpha_i' + \beta_i' \times t_{ij} + e_{ij} \]

Model: \( Y = \text{Week} \)

Random: \( \text{INTERCEPT Week/Subject=ID Type=UN} \)

**Proc Mixed**
\[
\begin{align*}
\text{data} &= \_\text{tmp}_2; \\
\text{Class} &= \text{ID}; \\
\text{Model} &= Y = \text{Week} / \text{Solution ddfm=kr}; \\
\text{Random} &= \text{INTERCEPT Week / Subject=ID Type=UN Solution}; \\
\text{Run};
\end{align*}
\]
Random Intercept-Slope Model

\[ Y = \alpha + \beta t \]

\( \alpha_{150} + \beta_{150} t \)
\( \alpha_{200} + \beta_{200} t \)
\( \alpha_{100} + \beta_{100} t \)
\( \alpha_{90} + \beta_{90} t \)
\( \alpha_{60} + \beta_{60} t \)
\( \alpha_{1} + \beta_{1} t \)
Estimated Mean FKSI-15 Scores: Random Intercept-Slope Model

Source: Cella et al. 2008
# Mean Treatment Differences for PRO Instruments: Random Intercept-Slope Model

<table>
<thead>
<tr>
<th>Instruments</th>
<th>Overall Estimated Means</th>
<th>Difference* (95% Confidence Interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sunitinib</td>
<td>IFN-alfa</td>
</tr>
<tr>
<td>FKSI-DRS</td>
<td>29.4</td>
<td>27.4</td>
</tr>
<tr>
<td>FKSI-15</td>
<td>45.3</td>
<td>42.1</td>
</tr>
<tr>
<td>FACT-G</td>
<td>82.3</td>
<td>76.8</td>
</tr>
<tr>
<td>EQ-5D Index</td>
<td>0.76</td>
<td>0.73</td>
</tr>
<tr>
<td>EQ-VAS</td>
<td>73.4</td>
<td>68.7</td>
</tr>
</tbody>
</table>

*Difference between means may not equate exactly because of rounding error

Source: Cella et al. 2008
Chapter 9: Mediation Models
Chapter 10: Missing Data
Introduction

• Missing data poses challenges in the analysis and interpretation of data
  – Potential loss of statistical power or sensitivity to detect clinically meaningful treatment differences
  – Potential bias for estimates of treatment effect

• Two types of missing data
  – Item non-response
  – Questionnaire non-response
Study Design to Minimize Data

• **Primary prevention**
  – Treat PRO measures like other endpoints
  – Identify and train key personnel to oversee the process
  – Adopt standard administration of PROs across sites
  – Minimize patient burden

• **Secondary prevention**
  – Prospectively documenting specific reasons for missing data
    • Example: “Patient refusal due to poor health”
    • Example: “Patient refusal unrelated to health”
  – Collect auxiliary factors that contribute or explain “missingness”
    • Example: toxicity, evaluation of patient health status by others
Missing Data Patterns and Mechanism

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)
Missing Items

• Missing data on at least one specific item

• Treat missing item as missing
  – Loss of power and threat of serious bias

• Pro-rate if at least half of items are answered
  – Example: Physical function scale on EORTC QLQ-C30
  – Impute mean of completed items to missing items
  – Well-suited for multi-item scales where there is no clear ordering or hierarchy of item difficulty
Missing Domains or Questionnaires

• Complete Case Analysis
  – Can have value in sensitivity analysis
  – Breaks down randomization and reduces sample size
  – Assumes MCAR

• Imputation
  – Single (MCAR) – last (baseline) observation carried forward
  – Multiple (MAR)

• Maximum Likelihood Methods
  – Longitudinal mixed effect models (MAR)
  – Can be robust
Missing Domains or Questionnaires: MNAR Models

• Pattern Mixture Models
  – Allows parameters to vary according to missing data patterns, with model conditional upon each pattern (e.g., early, late, completers)
  – Pattern-specific estimates are weighted and combined

• Selection Models
  – Links measurement and missingness processes by having the PRO response from the measurement model serve as a predictor in the dropout (missingness) model

• Shared Parameter Models
  – Measurement of PRO values (observed and unobserved) and time to dropout are assumed independent given the random effects (which are the same in both models), which are assumed to drive the measurement of outcome and missing data
Chapter 11:
Enriching Interpretation
Journal References:
Illustrations Cited


