Interpreting Change in Scores on COA Endpoint Measures

SIXTH ANNUAL

PATIENT-REPORTED OUTCOME CONSORTIUM WORKSHOP

April 29 - 30, 2015 ■ Silver Spring, MD

CRITICAL PATH INSTITUTE
a decade of excellence

10 YEARS

Disclaimer

The views and opinions expressed in the following slides are those of the individual presenters and should not be attributed to their respective organizations/companies, the U.S. Food and Drug Administration, the Critical Path Institute, the PRO Consortium, or the ePRO Consortium.

These slides are the intellectual property of the individual presenters and are protected under the copyright laws of the United States of America and other countries. Used by permission. All rights reserved. All trademarks are the property of their respective owners.

Session Participants

- Cheryl D. Coon, PhD
  - Director, Healthcare Analytics, Adelphi Values
- Joseph C. Cappelleri, PhD, MPH
  - Senior Director of Biostatistics, Pfizer Inc
- Scott Komo, DrPH
  - Senior Statistical Reviewer, Office of Biostatistics, CDER, FDA
- Laura Lee Johnson, PhD
  - Associate Director, Division of Biometrics III, Office of Biostatistics, CDER, FDA

Session Overview

1. Present a brief history on interpreting change in scores on COA measures used as endpoints

2. Illustrate two anchor-based methods for defining clinically important responders

3. Discuss three novel methods for interpreting change that could be added to our toolbox

4. Provide regulatory reflection on interpreting COA scores in a clinical trial

History

- While the difference between treatment groups can be evaluated using significance tests, thresholds are needed to interpret if change on a COA is meaningful

- Minimal (clinically) important difference (MID or MCID) was introduced in the 1980's to determine the smallest amount of change that patients perceive as a benefit[1]

- Draft FDA PRO guidance in 2006 covered both MID for group-level change and a responder definition for individual-level change

- Final FDA PRO guidance in 2009 focused only on responder and individual-level change

[1]Jaeschke et al. Control Clin Trials. 1989;10(4):407-15.

Semantics

- Confusion arises because the term MID is often used in reference to both individual-level change and group-level differences

- Further, shouldn't we be striving for "important" change rather than "minimal" change?
  - "Minimal" is vague, and there was consensus at a recent gathering of psychometricians in our field that we should refrain from using this term

- Thus, some realignment is needed on the semantics for interpreting change on COAs

Semantics

- **<u>Group-level change:</u>** Clinically important difference (CID)[1] is the difference in change scores between two treatment groups, or the change score within one treatment group, that can be considered clinically important

- **<u>Individual-level change</u>:** Clinically important responder (CIR) threshold[1] is the amount of change a patient would have to report to indicate that a treatment benefit has been experienced

- *The FDA Guidance tends to focus on individual-level change, and so shall this session*

[1]Cappelleri et al. Patient-Reported Outcomes: Measurement, Implementation and Interpretation. 2013.

Mainstream Approaches for Defining a Responder Threshold

- **Anchor-based methods**: Anchor change scores on the COA to an external criterion that identifies study subjects who have experienced an important change in their condition

- **Distribution-based methods**: Use the distribution of COA scores to classify the size of meaningful change rather than the statistical or clinical significance of that change

- **Cumulative distribution functions (CDFs)**: For each possible change score on the COA, plot the percentage of subjects achieving that amount of improvement or greater and examine the separation between groups for each possible threshold

Limitations with Mainstream Approaches

- **Anchor-based methods**: A meaningful and sufficiently-related (i.e., correlated) anchor is not always available in our studies

- **Distribution-based methods**: This approach does not connect back to the patient perspective and is more related to scale precision

- **CDFs**: This approach also does not connect back to the patient perspective, and there's misunderstanding in how to interpret them

Are there novel or interesting methods that we should be considering for interpreting change on COAs?

Two Methods for Responder Analysis of Patient-Reported Outcome Measures

Joseph C. Cappelleri, PhD, MPH
Senior Director of Biostatistics, Pfizer Inc

SIXTH ANNUAL
PATIENT-REPORTED OUTCOME CONSORTIUM WORKSHOP

April 29 - 30, 2015 ■ Silver Spring, MD

How Do You Determine the *Responder Definition* for a PRO Instrument?

Anchor-Based Methods

- Anchor-based methods explore the association between the PRO measure of interest and an anchor measure

- To be useful, the anchors chosen should be easier to interpret than the PRO measure itself and be appreciably correlated with it

- Magnitude of responder definition on a PRO measure depends on its correlation with the anchor, its variability, and the variability of the anchor

Two Anchor-Based Approaches

- One using logistic regression
  - Anchor is binary outcome
  - PRO measure is quantitative predictor
  - Receiver Operating Characteristic (ROC) Curve

- Another using linear regression
  - PRO measure is quantitative outcome
  - Anchor predictor could be quantitative or categorical
  - Cross-sectional model or longitudinal model

ROC Curve Illustration (Logistic Regression): Pain Intensity Numerical Rating Scale (PI-NRS)

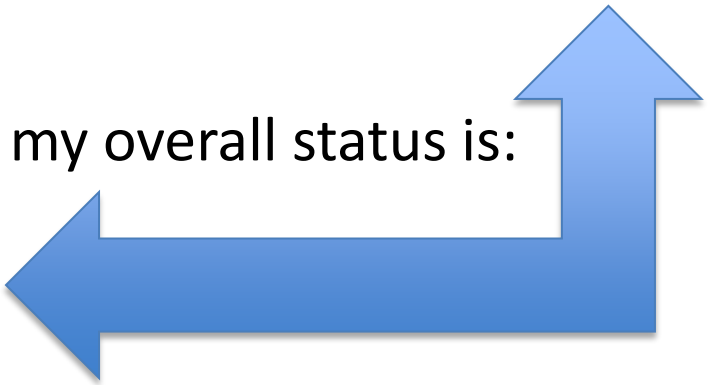
- Farrar JT et al. *Pain* 2001; 94:149-158

- 11-point pain scale: 0 = no pain to 10 = worst pain
  - Baseline score = mean of 7 diary entries prior to drug
  - Endpoint score = mean of last 7 diary entries
  - Interest centers on change score
  - Primary endpoint in pregabalin program

- 10 chronic pain studies with 2724 subjects
  - Placebo-controlled trials of pregabalin
  - Several conditions (e.g., fibromyalgia and osteoarthritis)

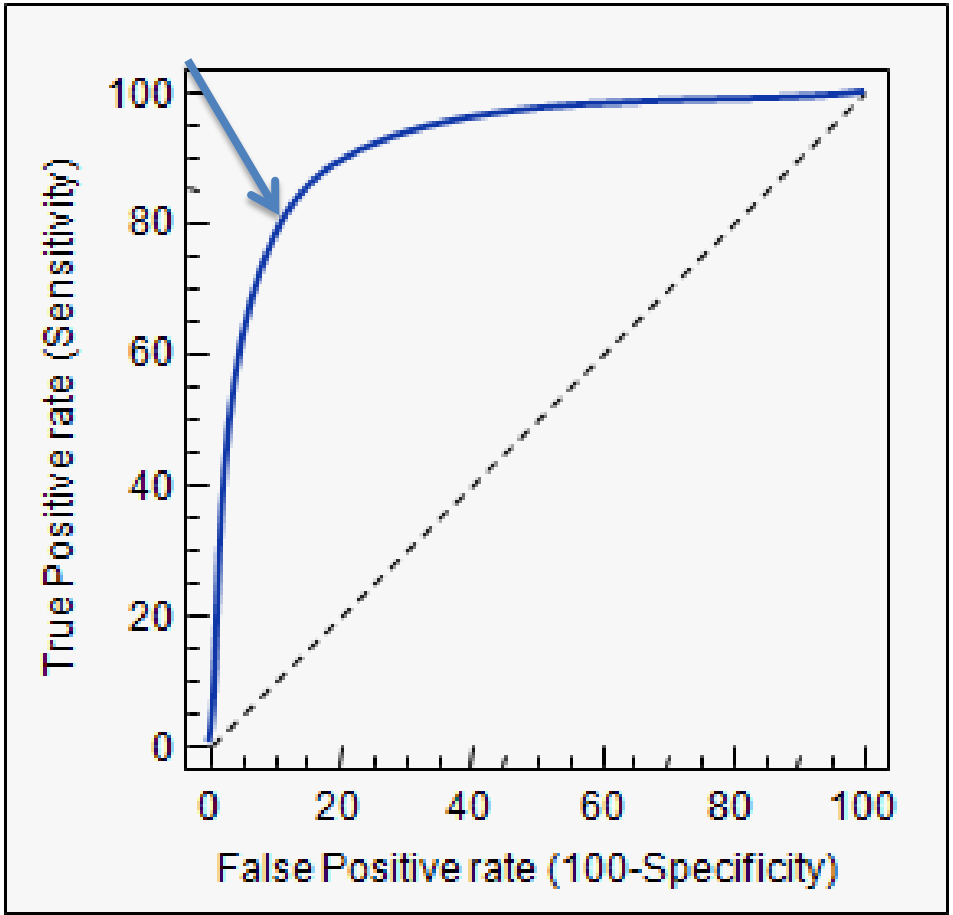
ROC Illustration: PI-NRS

- Patient Global Impression of Change (anchor)
  - Clinical improvement of interest
  - Best change score for distinguishing ‘much improved’ or better on PGIC
- Since the start of the study, my overall status is:
  1. Very much improved
  2. Much improved
  3. Minimally improved
  4. No change
  5. Minimally worse
  6. Much worse
  7. Very much worse



PI-NRS score type	Model	Area under the curve	Change	Sensitivity (%)	Specificity (%)
Raw change	Very much improved	0.873	−2.76	79.2	80.1
Raw change	Much or very much improved	0.853	−1.74	77.0	78.6
Raw change	Minimally, much or very much improved	0.832	−1.0	77.9	75.3
Percent change	Very much improved	0.890	−46.51	81.5	81.5
Percent change	Much or very much improved	0.859	−27.9	78.4	78.4
Percent change	Minimally, much or very much improved	0.832	−14.5	76.8	76.8

Arrow below indicates sensitivity & specificity for two-point improvement or 30% improvement



Longitudinal Illustration (Repeated Measures Linear Regression): Itch Severity Score

- Consistent with FDA guidance with respect (as could be the logistic regression approach)
- Mamolo et al. *Journal of Dermatological Treatment*. In press. doi:10.3109/09546634.2014.906033.
- Clinically important responder on the Itch Severity Scale (ISS)
- Itch Severity Score was the outcome
  - A 11-point numeric rating (0=no itching to 10=worst possible itching)

Longitudinal Illustration: Itch Severity Score

- Anchor involved two steps

- Step One
  - Patient Global Assessment (PtGA)
  - Evaluates the overall extent of cutaneous disease at a given time, with categories of ‘‘clear’’, ‘‘almost clear’’, ‘‘mild’’, ‘‘moderate’’ and ‘‘severe’’

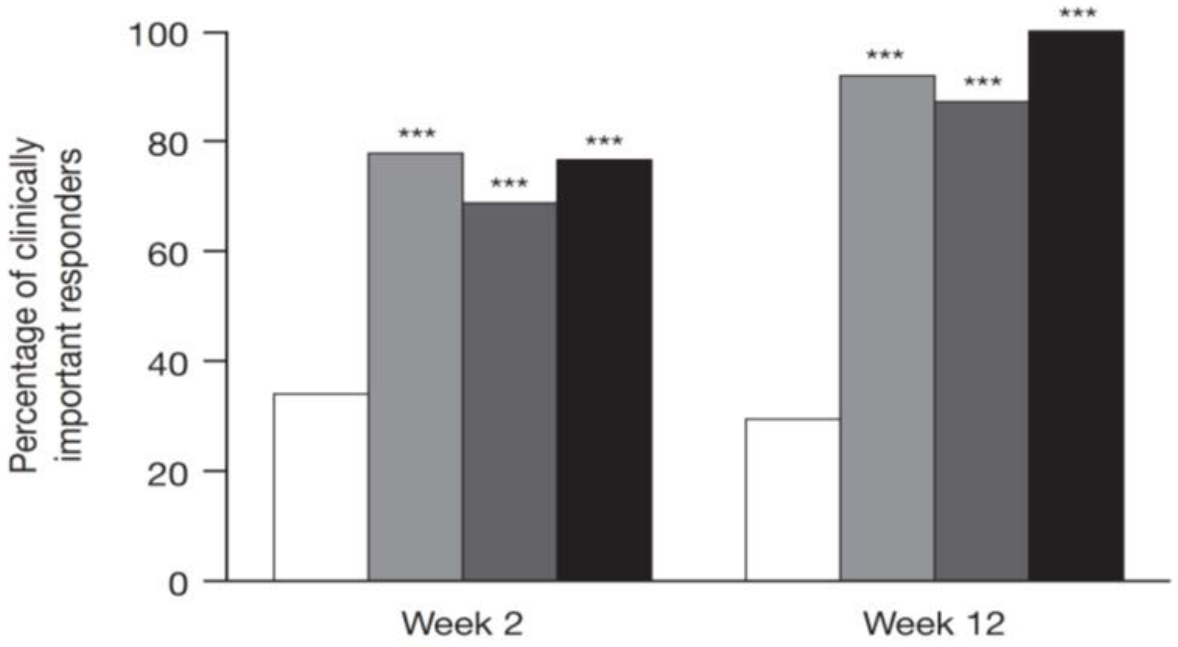
- Step Two
  - Use PtGA to create the Subject Global Impression of Change (SGIC)
  - Compare post-baseline PtGA relative to baseline PtGA
  - If improve, define SGIC as ‘‘better’’ (1); if worse, define SGIC as “worse” (-1); if unchanged, define SGIC as “the same” (0)
  - SGIC motivated by FDA Guidance on PRO measures

Longitudinal Illustration: Itch Severity Score

- A repeated-measures model was used to estimate the relationship between percent change from baseline ISS and the SGIC as an anchor
- Assessed at baseline and weeks 4, 8, 12 and 16
- The percent difference on the ISS corresponding to a one-category change on the SGIC was used to define a clinically important responder
- It was estimated to be 29.85% (95% CI: 23.30, 36.40) – approximately a 30% improvement

Results: Proportion of Individuals with at Least a 30% Improvement

Shaded boxes represent active treatments; white box represents placebo

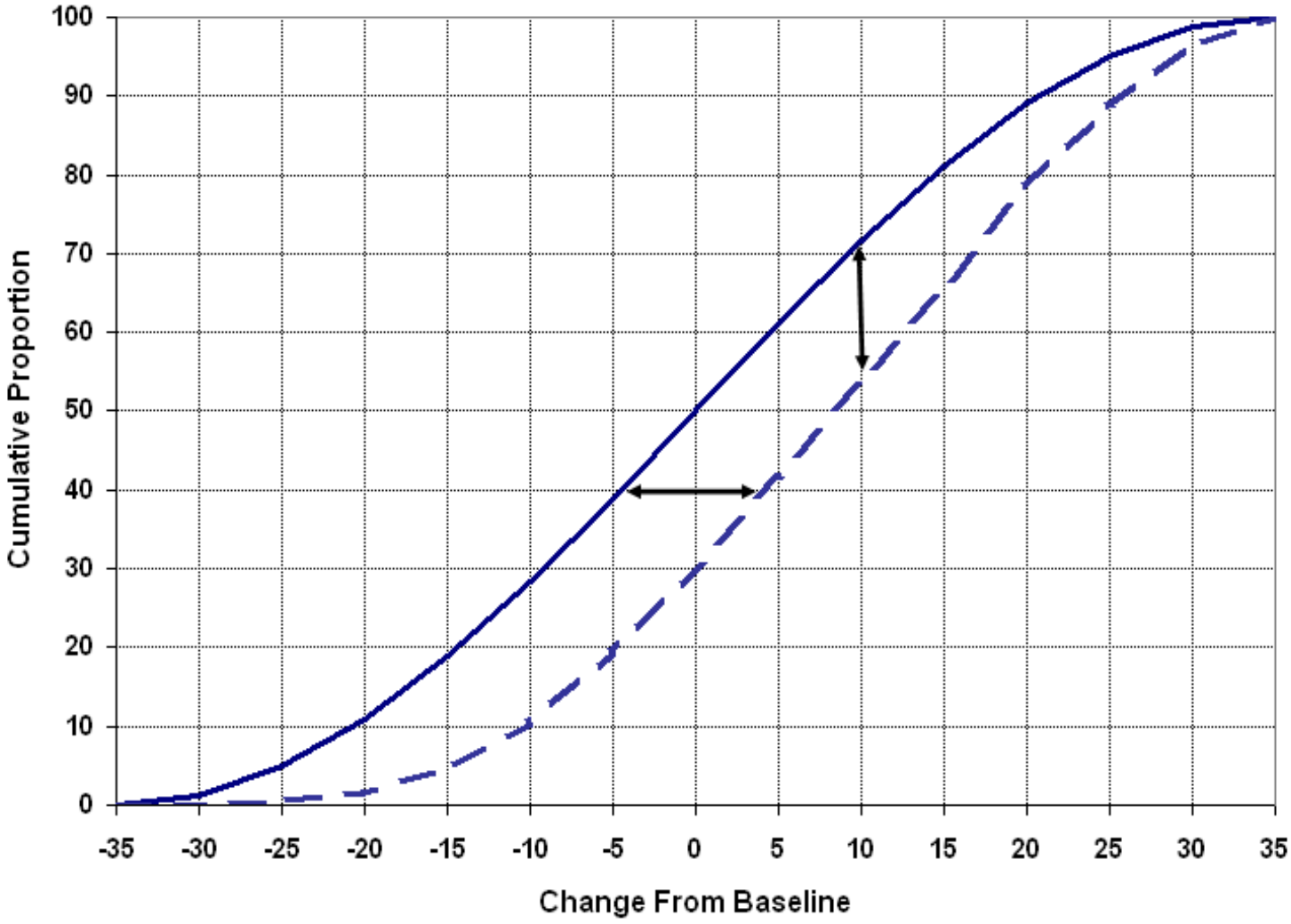


***$P < .0001$

Cumulative Distribution Function: Adjunct to Whichever Anchor Method is Used

- In general, consider cumulative distribution function for each treatment regardless of responder definition
- Can be applied descriptively for observed data
- Provides a descriptive assessment on robustness of responder definition
- Part of FDA Guidance on PRO measures

Illustrative Cumulative Distribution Function: Experimental Treatment (solid line) better than Control Treatment (dash line) -- Negative changes indicate improvement



Vertical Arrow: Difference in response for a change of 10 points or less (better) -- 70% in experimental vs. 55% in control

Horizontal Arrow: 40% of experimental subjects had a change score of -5 or less (better), while 40% of control subjects had a change score of +5 or less (better)

Two General References on Interpretation

- Cappelleri JC, Bushmakin AG. Interpretation of patient-reported outcomes. *Statistical Methods in Medical Research*. 2014; 23:460-483.

- 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 Press. 2013.

Summary

- Highlighted and illustrated two methods for defining a clinically important responder
- Incorporates anchor-based methodology
- Logistic regression (ROC curve)
  - Anchor as outcome, target PRO as predictor
  - Select responder cutoff on PRO that best distinguishes levels of the anchor (based on sensitivity and specificity)
- Linear regression
  - Target PRO as outcome, anchor as predictor
  - Select responder cutoff on PRO that gives the difference in means between adjacent levels of the anchor

Three Novel Methods for Establishing Responder Thresholds on COAs

Cheryl D. Coon, PhD
Director, Healthcare Analytics, Adelphi Values

SIXTH ANNUAL
PATIENT-REPORTED OUTCOME CONSORTIUM WORKSHOP

April 29 - 30, 2015 ■ Silver Spring, MD

Considerations in Selecting Approaches

Beyond anchor-based and distribution-based approaches, as well as CDFs, are there alternative methods that we should be considering for interpreting change on COAs?

- What approach should we use when a meaningful anchor isn't available?
- What approach should we use when the only available quantitative data are from a non-interventional study?
- How do we ensure that we are linking change to the patient perspective?

Alternative Approaches: Bookmarking/Standard Setting

- A group of patients and experts reviews the items and reaches a consensus on the location of thresholds for interpreting scores and change scores

- Example: Cook et al. *Qual Life Res*. 2015;24(3):575-89.

Advantages	Disadvantages
➢ Does not require an interventional study ➢ Directly ties thresholds to the patient perspective ➢ Results can be available quickly after data collection	➢ Requires a stand-alone study ➢ Can be cognitively difficult for participants ➢ Can be inconclusive if the group fails to reach a consensus

Alternative Approaches: Exit Interviews

- Subjects whose condition changed over time are recruited from a longitudinal study, and interviews are conducted to understand how the change that was experienced influenced their scores on the COA measure

- Example: No published applications to date?

Advantages	Disadvantages
➢ Recruitment uses the pool of subjects from an already-recruited study ➢ Qualitative data is used to give greater insight into quantitative results ➢ Patients are available to directly ask if their scores are reflective of the change they experienced	➢ The qualitative data could contradict the quantitative data and could be difficult to synthesize ➢ The study is sensitive to recall bias ➢ Timing can be challenging in recruiting soon after the longitudinal study is completed

Alternative Approaches: Conjoint Analysis

- Patient preference data on hypothetical item response profiles are used to the calculate utilities of actual COA scores in the clinical trial, and the conjoint model identifies responders based on their utilities pre- and post-treatment

- Example: Coon et al. *17th Annual ISOQOL Conference*. 2010.

Advantages	Disadvantages
➢ Does not require an interventional study ➢ Directly ties responder definitions to the patient perspective ➢ Responder definitions are sensitive to pre-treatment scores	➢ Requires a stand-alone study ➢ Can be cognitively difficult for participants ➢ Needs a limited number of items and response categories to be able to fit the model ➢ Does not result in one CIR threshold on the same metric as the COA

Summary

- Researchers in our field are considering innovative approaches for interpreting change on COAs
  - These approaches can complement mainstream approaches, providing further insight into the meaning of change

- There are pros and cons to each approach, and it is important to acknowledge limitations while also being open to original or supplementary methods

- Regardless of which methods are implemented, justification for and the value of the selected methods should be documented

Topics for Discussion

PRO CONSORTIUM
CRITICAL PATH INSTITUTE

1. Should CIR be used as the primary metric for evaluating COA endpoints, or should it be used to supplement statistical methods that treat the COA as a continuous measure?

2. When is group-level interpretation (i.e., CID) preferred over individual-level interpretation (i.e., CIR)?

3. The same methodology has been used to identify one threshold that defines both CID and CIR. Is that appropriate?

4. Are there other examples where innovative methods have been used successfully to assess CIR (or CID) for supporting endpoints?

5. Is it preferable to include standardized effect sizes (i.e., difference within or between groups, divided by variability) to accompany CID and CIR?

6. What approaches would YOU (the audience) recommend?

Discussion and/or Questions?

Session Participants

- Cheryl D. Coon, PhD
  - Director, Healthcare Analytics, Adelphi Values
- Joseph C. Cappelleri, PhD, MPH
  - Senior Director of Biostatistics, Pfizer Inc
- Scott Komo, DrPH
  - Senior Statistical Reviewer, Office of Biostatistics, FDA
- Laura Lee Johnson, PhD
  - Associate Director, Division of Biometrics III, Office of Biostatistics, FDA