Clinical Outcome Assessments: Establishing and Interpreting Meaningful Within-Patient Change

4/4/17
CLINICAL OUTCOME ASSESSMENTS: INTERPRETING MEANINGFUL CHANGE

Duke Margolis Expert Workshop
April 4, 2017

Elektra J. Papadopoulos, MD, MPH
Clinical Outcome Assessments Staff
Office of New Drugs
Center for Drug Evaluation and Research
U.S. Food and Drug Administration
Dr. Janet Woodcock:

• "It turns out that what is really bothering the patient and what is really bothering the doctor can be radically different things....patients are true experts in their disease”.

• “It's clear you have to start with an understanding of the impact of the disease on the people who have it, and what they value most in terms of alleviation before you set up a measurement and go forward with truly patient-focused drug development."
**Basic Framing of FDA Drug Benefit-Risk Assessment**

<table>
<thead>
<tr>
<th>Decision Factor</th>
<th>Evidence and Uncertainties</th>
<th>Conclusions and Reasons</th>
</tr>
</thead>
</table>
| Analysis of Condition            | **Sets the context for the weighing of benefits and risks:**  
  • How serious is this indicated condition, and why?  
  • How well is the patient population’s medical need being met by currently available therapies? |                                                                                         |
| Current Treatment Options        | **Characterize and assess the evidence of benefit:**  
  • How meaningful is the benefit, and for whom?  
  • How compelling is the expected benefit in the post-market setting?                                                                            |                                                                                         |
| Benefit                          | **Characterize and assess the safety concerns:**  
  • How serious are the safety signals identified in the submitted data?  
  • What potential risks could emerge in the post-market setting?                                                                              |                                                                                         |
| Risk                             | **Assess what risk management (e.g., labeling, REMS) may be necessary to address the identified safety concerns**                                                                                                         |                                                                                         |
| Risk Management                  |                                                                                                                                                            |                                                                                         |
FDA’s Patient-Focused Drug Development (PFDD) Initiative

• Patients are uniquely positioned to inform understanding of the therapeutic context for drug development and evaluation
  – There is a need for more systematic ways of gathering patient perspective on their condition and treatment options

• Patient-Focused Drug Development (PFDD) is part of FDA commitments under PDUFA V*
  – FDA is convening 24 meetings on specific disease areas in FY 2013-17
  – Meetings can help advance a systematic approach to gathering input

*The fifth authorization of the Prescription Drug User Fee Act, enacted in 2012
PFDD in Chronic Disease

- PFDD meetings routinely ask for patients’ perspectives on what an ideal treatment would look like and what clinical benefit would be the most meaningful to them.
- Concepts such as emotional impact of disease, ability to perform activities are often cited by patients as important.
  - E.g., in Parkinson’s disease patients want to know functional status over time.
  - Depending on the stage of disease even small amounts of deterioration can make the difference between being able to perform basic activities (e.g., feeding oneself) independently or not.
**PFDD Next Steps**

- **Advance science of patient input**
  - Engage wider community to discuss methodologically sound approaches that:
    - Bridge from initial PFDD meetings to more systematic collection of patients’ input
    - Generate meaningful input on patients’ experiences and perspectives to inform drug development and B-R assessment
    - Are “fit for purpose” in drug development and regulatory context

- **Provide guidance**
  - To: patient communities, researchers, and drug developers
  - On: pragmatic and methodologically sound strategies, pathways, and methods to gather and use patient input
Interpretation of Clinically Meaningful

• Statistical significance alone is not sufficient

• Clinical benefit: a positive clinically meaningful effect of an intervention, i.e., a positive effect on how an individual feels, functions, or survives.

• To establish clinical benefit we consider two questions:
  – 1) Does the assessment measure or reflect something of significance to patients?
    • Relies on patient, caregiver and expert input/engagement
  
  – 2) Is the magnitude of change at the individual level sufficiently large to affect how patients feel or function in daily life?
Triangulation of Evidence

• **Multiple methods** used to select a benchmark for meaningful change
  – Often result in a range of values for what is a clinically meaningful benchmark
  – Triangulation of evidence consists of examining these values to converge on an appropriate value or range of values likely to represent meaningful change in the outcome of interest
Guidance for Industry

Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims


- FDA PRO Guidance defines good measurement principles to consider for “well-defined and reliable” (21 CFR 314.126) PRO measures
- All COAs can benefit from the good measurement principles described within the guidance
- But, judgment and flexibility are needed!
Final PRO Guidance (2009)

• Clinically meaningful thresholds may vary by target population:
  – “we will evaluate an instrument’s responder definition in the context of each specific clinical trial.”

• Anchor-based methods emphasized:
  – “Empiric evidence for any responder definition is derived using anchor-based methods”
  – “…explore the associations between the targeted concept of the PRO instrument and the concept measured by the anchors”
  – Multiple anchors recommended

• Distribution-based methods:
  – “…should be considered as supportive and are not appropriate as the sole basis for determining a responder definition”
Final PRO Guidance (2009)

• Emphasizes the display of individual responses to treatment:
  – “...it is possible to present the entire distribution of responses for treatment and control group, avoiding the need to pick a responder criterion. Whether the individual responses are meaningful represents a judgment...”
  – “...cumulative distribution displays show a continuous plot of the percent change from baseline on the X-axis and the percent of patients experiencing that change on the Y-axis.
  – “A variety of responder definitions can be identified along the cumulative distribution of response curve.”
But...

• The presentation of all possible response level cut-off points does not eliminate the need to identify the level of change that is clinically important (or at least to state our uncertainty about that level)
Cumulative Distribution Function (CDF) (DB4, pooled across treatment arms)

Change from Baseline in Nocturia Episodes

Cumulative Percentage (Responder Rate)

- Much Better (n=298)
- Somewhat Better (n=288)
- Not Changed (n=185)

Source: Dr. Jia Guo; Bone, Reproductive and Urologic Drugs Advisory Committee 10/19/2016
CDF Plot by Treatment Arms (DB4)

Change from Baseline in Nocturia Episodes

SER 120 1.5 mcg (n=260)
Placebo (n=260)

Source: Dr. Jia Guo; Bone, Reproductive and Urologic Drugs Advisory Committee 10/19/2016
Establishing Meaningful Change: Examples from FDA Guidance

• Alzheimer’s Disease: Developing Drugs for the Treatment of Early Stage Disease (2013)
  – Co-primary endpoint of cognitive test and a functional or global assessment
    • “The intent of this dual measurement is to ensure the clinical meaningfulness of a cognitive benefit that may be observed”

• Irritable Bowel Syndrome: Clinical Evaluation of Drugs for Treatment (2012)
  – Patient global assessments
    • Example: “How would you rate your IBS signs or symptoms overall over the past 7 days?”

• Analgesic Indications: Developing Drugs and Biological Products (2014)
  – Allows the use of a responder analysis (e.g., 30% reduction in pain with early discontinuation counted as failure) in addition to differences in group means
  – Encourages use of cumulative distribution functions in the package insert
A word about “MID” and “MCID”

• Minimum important difference (MID) was removed from the Final PRO Guidance 2009
  – Confusion resulted from the term being used interchangeably to indicate either group-level mean differences as well as individual level change
  – Use of the term “minimal” is problematic:
    • While a minimal amount of change may be noticeable, it does not necessarily imply the change is meaningful to patients
• Minimal clinically important difference” (MCID)-
  – i.e., the smallest difference in score...patients perceive as beneficial and which would mandate a change in the patient’s management
  – Again, smallest change does not imply meaningful change
Beyond Anchor-based and Distribution-based methods: Examples of Emerging Methods

• Bookmarking/Standard Setting
  – Patients and experts are presented with clinical vignettes of a disease in order to reach a consensus on thresholds for severity levels
  – Designed for measures that have been calibrated using an IRT model

• Scale judgment
  – Panels of judges evaluate pairs of completed tests to determine whether the amount of change specified by the responses before and after treatment is meaningful

• Exit interviews
  – Interviews of patients who recently completed a clinical trial
  – Can be used to collect qualitative and quantitative data about patients’ experience of disease or treatment burden and changes during the course of the clinical trial

• Others
Today’s Goals

• Advance the discussion on methods to identify meaningful within-patient change in COAs by discussing key issues and major challenges, including:
  – What are the advantages and disadvantages of each of the methods?
  – How might threshold determinations differ across the four types of COAs?
  – What are special considerations for establishing meaningful change in small and heterogeneous study populations?
  – How and when could these methods be most feasibly used in drug development?
Clinical Outcome Assessments: Establishing and Interpreting Meaningful Within-Patient Change

4/4/17
Exploring the Use of Emerging Methods to Derive and Interpret Meaningful Within-Patient Change Using Idiosyncratic Scale-Judgment (Bookmarking/Standard-Setting)

Karon F. Cook
Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL

April 4, 2017
Washington, DC
Background

How much we know

Building State of the Art Measures

How to Interpret Scores on State of the Art Measures
Meaningful Change
Methods for Defining Meaningful Score Differences

Statistical

Global Ratings of Change

External Anchor
How often are you too tired to socialize with family?
Item Response Samples

T=40

- never was too tired to do household chores.
- never needed to sleep during the day.
- rarely had trouble finishing things because she was too tired.
- rarely was so tired that she needed to rest during the day.
- rarely felt that she had no energy.
Item Response Samples

- sometimes was too tired to eat.
- often had trouble finishing things because she was too tired.
- often was too tired to do her household chores.
- often needed to sleep during the day.
- always frustrated by being too tired to do the things she wanted to do.
Grant #H00145
Deborah Miller, PI
Developed 18, 5-item sample response sets, 2 pts apart.

Online panel of 500 participants with Multiple Sclerosis

Responded to NeuroQoL Fatigue Short Form

Branched into 7 fatigue levels

Presented with 7 response samples

Ms. Butler

e.g. 48-51
In PART B, you will

• Look at the fatigue reports of 7 people who have MS

• Compare each person’s fatigue to your own fatigue. For example, your fatigue might be greater.

• Or, you might decide your fatigue is the SAME or LESS than that other person’s.
If you decide your fatigue is DIFFERENT from the other person’s, you will then

- Consider what it would be like to have this person’s fatigue, and
- Decide if the difference would matter to you in your daily life.
Depending on your own fatigue, you may decide that none, some, or all of these people have more, less, or the same amount fatigue.

There are no “right” answers—just your own thoughtful judgments.
This is what Ms. Anderson said about her fatigue over the last 7 days. She reported that she:
• sometimes felt weak all over.
• often had to limit social activity because she was tired.
• sometimes had trouble starting things because she was too tired.
• often was too tired to take a short walk.
• often had trouble finishing things because she was too tired.

Compared to Ms. Anderson’s, has YOUR FATIGUE been:
☐ Greater than Ms. Anderson’s
☐ The same as Ms. Anderson’s
☐ Less than Ms. Anderson’s
You said YOUR FATIGUE over the past week was Greater

If your fatigue IMPROVED to Ms. Anderson’s level, would it make a difference in your daily life?

- It wouldn’t really make a difference in my daily life.
- It would make a difference in my daily life (things I do day-to-day would be easier).

This is what Ms. Anderson said about his fatigue over the last 7 days. She reported that she:
- sometimes felt weak all over.
- often or always had to limit social activity because she was tired.
- sometimes had trouble starting things because she was too tired.
- often was too tired to take a short walk.
- often had trouble finishing things because she was too tired.
If your fatigue WORSENED to Ms. Anderson’s level, would it make a difference in your daily life?

- It wouldn’t really make a difference in my daily life.
- It would make a difference in my daily life (many of the things I do day-to-day would be harder).

This is what Ms. Anderson said about her fatigue over the last 7 days. She reported that she:

• sometimes felt weak all over.
• often or always had to limit social activity because she was tired.
• sometimes had trouble starting things because she was too tired.
• often was too tired to take a short walk.
• often had trouble finishing things because she was too tired.
Results
Analyses to Estimate Thresholds for Interpreting Change

- Calculate minimum distance endorsed by respondent as meaningful improvement/decrement
- Identify thresholds that would capture different percentages of respondents minimums.
Threshold locations for capturing 50, 75, and 95% of distances endorsed as important worsening

Thresholds for Worsening
Mean of Individual Thresholds ≤ 3.2
Threshold locations for capturing 50, 75, and 95% of distances endorsed as important improvement

Thresholds for Improvement
Mean of Individual Thresholds ≤ -3.5
98% Reversals
Judgments Consistent with Prior Judgment

7 points >
5 points >
3 points >
1 pt

Butler
Richardson
Woods
Anderson
7.6% of all judgments were inconsistent with the one prior (of 3000 opportunities).
When global rating of change contradicts observed change: Examining appraisal processes underlying paradoxical responses over time

Carolyn E. Schwartz1,2 • Victoria E. Powell1 • Bruce D. Rapkin3

Accepted: 14 September 2016 / Published online: 8 October 2016
© Springer International Publishing Switzerland 2016

Scores of N=525
- Declined
- Unchanged

On GROC, 48.6% made a paradoxical judgment
- reporting worse status when observed score was unchanged or
- endorsing the same status when observed scores had declined.
• We were able to estimate plausible responder thresholds for consequential change.

• Participants evaluated range of IRT-Vs close to their own fatigue levels.

• Participants reported high confidence that their judgments (3.3 between moderately and highly confident).
• Judgments contextualized in a patient-relevant context—“make a difference in daily life.”

• Design allowed large samples.

• Judgment errors existed, but were within range of other methods.

• Judgement errors not strongly associated with demographics (e.g., education)
How can we do better

• Qualitative research to understand what is important to people in assessing change.

• Cognitive debriefing to understand what people are attending to. Are they attending to different things.
  • Could we selection of concepts of the vignettes, should they be standardized.

• Frame vignettes by what is important to people.

• Try to understand the variation in levels of change that people believe is important.

• Can set a threshold that is most representative, but it will not catch everyone.

• Change study design so that everyone is getting same distances. Branch on every score.
How can we do better?
An Alternative Model:
Clinical Outcome Assessments: Establishing and Interpreting Meaningful Within-Patient Change

4/4/17
Some Remarks about
(Educational) Standard Setting,
Characterizing Meaningful Change,
and the Scale-Judgment Method

David Thissen
L.L. Thurstone Psychometric Laboratory
The University of North Carolina at Chapel Hill
“Minimally Important Difference” (MID) estimation is not like answering the question

“What is the ratio of the circumference of a circle to its diameter?”
“Minimally Important Difference” (MID) estimation is not like answering the question

“What is the ratio of the circumference of a circle to its diameter?”
“Minimally Important Difference” (MID) estimation is like a speed limit:

At least for regulatory purposes, MID is like a speed limit: A policy decision informed by data (And likely between 0.2 and 0.5 standard units)
“Most authorities on standard setting (e.g., Green, Trimble, and Lewis, 2003; Hambleton, 1980; Jaeger, 1989; Shepard, 1980; Zieky, 2001) suggest that, when setting cut scores, it is prudent to use and compare results from different standard setting methods.” (p. 155)


Section II: Standard Setting Methods

Chapter 4. The Nedelsky Method
Chapter 5. The Ebel Method
Chapter 6. The Angoff Method and Angoff Variations
Chapter 7. The Direct Consensus Method
Chapter 8. The Contrasting Groups and Borderline Group Methods
Chapter 9. The Body of Work and Other Holistic Methods

The Body of Work Method
The Judgmental Policy Capturing Method
The Dominant Profile Method
The Analytic Judgment Method

Chapter 10. The Bookmark Method
Chapter 11. The Item-Descriptor Matching Method
Chapter 12. The Hofstee and Beuk Methods
National Assessment of Educational Progress (NAEP)
Achievement Levels
Grade 4 Mathematics, 1996
A Ranking Procedure to Find Score Ranges Associated with “Mild”, “Moderate”, and “Severe” Conditions

(Cella et al., 2008, 2014)


### Sample Fatigue “Vignette”—T-score 40

#### FATIGUE – PINK

<table>
<thead>
<tr>
<th>Rating ______</th>
</tr>
</thead>
</table>

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>How often did your fatigue make you feel less alert?</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Always</td>
</tr>
<tr>
<td>2</td>
<td>How often did you have trouble starting things because of your fatigue?</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Always</td>
</tr>
<tr>
<td>3</td>
<td>How often did you feel run-down?</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Always</td>
</tr>
<tr>
<td>4</td>
<td>How often were you energetic?</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Always</td>
</tr>
<tr>
<td>5</td>
<td>How easily did you find yourself getting tired on average?</td>
<td>Not at all</td>
<td>A little bit</td>
<td>Some-what</td>
<td>Quite a bit</td>
<td>Very much</td>
</tr>
</tbody>
</table>

---

Sample Depression “Vignette”—T-score 60

<table>
<thead>
<tr>
<th>Rating ______</th>
</tr>
</thead>
</table>

DEPRESSION - MINT

<table>
<thead>
<tr>
<th></th>
<th>I felt that I had nothing to look forward to</th>
<th>Never</th>
<th>Rarely</th>
<th>Sometimes</th>
<th>Often</th>
<th>Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I felt that I wanted to give up on everything</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Always</td>
</tr>
<tr>
<td>2</td>
<td>I felt disappointed in myself</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Always</td>
</tr>
<tr>
<td>3</td>
<td>I felt lonely</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Always</td>
</tr>
<tr>
<td>4</td>
<td>I felt I had no reason for living</td>
<td>Never</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>Often</td>
<td>Always</td>
</tr>
</tbody>
</table>

**Step 1:** Please review the ten different cards in the "Anxiety" envelope. Each card represents a patient who falls along a different place on the anxiety continuum. Sort the cards in order from least severe to most severe, giving each color a ranking ("1" being least severe).

Please enter the card color (e.g. "Pink", "Blue", etc.) below the number ranking that you have assigned it. You are encouraged to give each card a unique ranking, but this is not required. If you believe two patients are tied, for example, at rank "6", then write both color names under the number "6."

<table>
<thead>
<tr>
<th>Least severe</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Most severe</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLOR:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COLOR:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(if applicable)</td>
</tr>
<tr>
<td>COLOR:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(if applicable)</td>
</tr>
</tbody>
</table>

**Step 2:** Now please draw three vertical lines between ranks (e.g. between "3" and "4"); one delineating each of the following:

1. A separation between those cards (i.e. patients) that you believe represent a normal level of anxiety and a mild level of anxiety
2. A separation between those cards representing a mild level of anxiety and a moderate level of anxiety
3. A separation between those cards representing a moderate level of anxiety and a severe level of anxiety

---

**Expert Ranking Sheet: Anxiety, Step 2**

**ANXIETY – Case Examples Exercise**

**Step 1:** Please review the ten different cards in the “Anxiety” envelope. Each card represents a patient who falls along a different place on the anxiety continuum. Sort the cards in order from least severe to most severe, giving each color a ranking (“1” being least severe).

Please enter the card color (e.g. “Pink”, “Blue”, etc.) below the number ranking that you have assigned it. You are encouraged to give each card a unique ranking, but this is not required. If you believe two patients are tied, for example, at rank “6,” then write both color names under the number “6.”

<table>
<thead>
<tr>
<th>Least severe</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Most severe</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLOR:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COLOR:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(if applicable)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>Mild</td>
<td>Moderate</td>
<td>Severe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Step 2:** Now please draw three vertical lines between ranks (e.g. between “3” and “4”); one delineating each of the following:

1. A separation between those cards (i.e. patients) that you believe represent a **normal** level of anxiety and a **mild** level of anxiety
2. A separation between those cards representing a **mild** level of anxiety and a **moderate** level of anxiety
3. A separation between those cards representing a **moderate** level of anxiety and a **severe** level of anxiety

After a consensus-building process …

A Bookmarked-Vignettes Procedure to Find Score Ranges Associated with “Mild”, “Moderate”, and “Severe” Conditions

(Cook et al., 2014; Morgan et al., 2017)
Judges do not know this vignette is for a $T$-score of 47.5

Ms. Miller’s Fatigue

In the last 7 days, Ms. Miller rarely felt weak all over and rarely was so tired she couldn’t take a short walk. However, she sometimes felt tired, which got in the way of her doing her household chores. Feeling too tired to do the things she wanted to do was sometimes frustrating for her.

In summary, Ms. Miller reports being:

- Rarely weak all over.
- Rarely too tired to take a short walk.
- Sometimes tired.
- Sometimes too tired to do household chores.
- Sometimes frustrated by being too tired to do the things she wanted to do.

Mild Problems

Anna’s Pain
In the last 7 days, Mr. Turner has rarely felt tired. Often, he has felt alert when he woke up and ready to start the day.

In summary, Mr. Turner has rarely felt tired.

Julia’s Pain
In the last 7 days, Mr. Turner has never had trouble sleeping because of bad dreams.

In summary, Mr. Turner has never had trouble sleeping because of bad dreams.

Andrea’s Pain
Mr. Turner has rarely felt red.

In summary, Mr. Turner has rarely felt red.

Jacob’s Pain
Mr. Turner has never had trouble controlling his emotions because of poor sleep.

In summary, Mr. Turner has never had trouble controlling his emotions because of poor sleep.

Chloe’s Pain
Mr. Turner has never had trouble sleeping because of bad dreams.

In summary, Mr. Turner has never had trouble sleeping because of bad dreams.

Kristen’s Pain

In summary, Mr. Turner has never had trouble sleeping because of bad dreams.

Addison’s Pain
In the last 7 days, Addison has rarely felt tired. Often, he has felt alert when he woke up and ready to start the day.

In summary, Mr. Turner reported:

Establishing clinical meaning and defining important differences for PROMIS measures in Juvenile Idiopathic Arthritis. Presentation at UNC PROMIS Pediatric Investigators Meeting, Chapel Hill, NC.
The “Scale-Judgment Method” to Estimate the “Minimally Important Difference” (MID) between Scores

(Thissen et al., 2016)

A minimally important difference (MID) has been defined as the “smallest difference in score … that patients perceive as important, … and which would lead the clinician to consider a change in the patient’s management”

—Guyatt et al. (2002)

Existing methods:

• Distribution-based indices (not an empirical method; merely expresses change in standard units)

• Anchor-based methods (“contrasting groups” in educational standard setting)

Earlier judgment-based methods

Delphi Method, Delphi plus anchor, physician survey, expert panels using visual analog scales or changes to item responses


One month ago

- I felt alone.
  - never
  - almost never
  - sometimes
  - often
  - almost always
- I felt like I couldn’t do anything right.
  - never
  - almost never
  - sometimes
  - often
  - almost always
- I felt everything in my life went wrong.
  - never
  - almost never
  - sometimes
  - often
  - almost always
- I felt sad.
  - never
  - almost never
  - sometimes
  - often
  - almost always
- I thought that my life was bad.
  - never
  - almost never
  - sometimes
  - often
  - almost always
- I could not stop feeling sad.
  - never
  - almost never
  - sometimes
  - often
  - almost always
- I felt lonely.
  - never
  - almost never
  - sometimes
  - often
  - almost always
- I felt unhappy.
  - never
  - almost never
  - sometimes
  - often
  - almost always

Today

- I felt alone.
  - never
  - almost never
  - sometimes
  - often
  - almost always
- I felt like I couldn’t do anything right.
  - never
  - almost never
  - sometimes
  - often
  - almost always
- I felt everything in my life went wrong.
  - never
  - almost never
  - sometimes
  - often
  - almost always
- I felt sad.
  - never
  - almost never
  - sometimes
  - often
  - almost always
- I thought that my life was bad.
  - never
  - almost never
  - sometimes
  - often
  - almost always
- I could not stop feeling sad.
  - never
  - almost never
  - sometimes
  - often
  - almost always
- I felt lonely.
  - never
  - almost never
  - sometimes
  - often
  - almost always
- I felt unhappy.
  - never
  - almost never
  - sometimes
  - often
  - almost always

**T-score 62.1**

**T-score 58.9**

The *scale-judgment method* presents judges with pairs of questionnaires, artificially completed using IRT, with scores known to the experimenter but not the judges …

---

One month ago

<table>
<thead>
<tr>
<th>Question</th>
<th>Never</th>
<th>Almost Never</th>
<th>Sometimes</th>
<th>Often</th>
<th>Almost Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>I felt alone.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt like I couldn’t do anything right.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt everything in my life went wrong.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt sad.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I thought that my life was bad.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I could not stop feeling sad.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt lonely.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt unhappy.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Today

<table>
<thead>
<tr>
<th>Question</th>
<th>Never</th>
<th>Almost Never</th>
<th>Sometimes</th>
<th>Often</th>
<th>Almost Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>I felt alone.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt like I couldn’t do anything right.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt everything in my life went wrong.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt sad.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I thought that my life was bad.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I could not stop feeling sad.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt lonely.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt unhappy.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**T-score 62.1**

**T-score 58.9**

The judges (clinicians, adolescents, parents) judge for each pair whether the (imaginary) respondent is “doing or feeling better, worse, or about the same.”...

For the Depressive Symptoms example, this process yields data with summary statistics like these …

<table>
<thead>
<tr>
<th>Pair</th>
<th>1 month ago</th>
<th>Today</th>
<th>Difference</th>
<th>Better</th>
<th>Frequency</th>
<th>No difference</th>
<th>Worse</th>
<th>Proportion Wrong Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>49.5</td>
<td>57.9</td>
<td>8.4</td>
<td>23</td>
<td>19</td>
<td>185</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>56.7</td>
<td>62.1</td>
<td>5.4</td>
<td>32</td>
<td>18</td>
<td>176</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>43.5</td>
<td>45.9</td>
<td>2.4</td>
<td>15</td>
<td>151</td>
<td>61</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>64.3</td>
<td>62.1</td>
<td>-2.2</td>
<td>133</td>
<td>66</td>
<td>27</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>62.1</td>
<td>58.9</td>
<td>-3.2</td>
<td>179</td>
<td>33</td>
<td>15</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>73.4</td>
<td>66.0</td>
<td>-7.4</td>
<td>189</td>
<td>21</td>
<td>17</td>
<td>0.07</td>
<td></td>
</tr>
</tbody>
</table>

There were more data for Fatigue, Mobility, and Pain.

If the judges were homogeneous, data analysis could be logistic regression of the probability “different” on the scale score difference, with the 50-50 point the MID:

But the judges were not homogeneous.

So we treated the pairs of questionnaires as items, the same-different judgments as item responses, and fitted the data with the 1PL IRT model:

\[ P(\text{"different"}) \]

\[ \theta \text{ (Propensity to respond "different")} \]

Then we interpolated the scale-score difference for a hypothetical item that would be judged “different” 50% of the time by an average respondent:

\[ \theta \] (Propensity to respond "different")

---

We used quadratic regression to interpolate the scale-score difference associated with a pair of questionnaires that would have a 1PL \( b \) of zero:

![Wrong Direction Omitted](image1)

![Wrong Direction Reversed](image2)

### Table: Minimally Important Difference (MID) in Health Outcomes Measures

<table>
<thead>
<tr>
<th></th>
<th>Wrong Direction</th>
<th>Wrong Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Omitted</td>
<td>Omitted</td>
</tr>
<tr>
<td><strong>MID</strong></td>
<td><strong>s.e.</strong></td>
<td><strong>MID</strong></td>
</tr>
<tr>
<td>Clinicians</td>
<td>2.1</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>Adolescents</td>
<td>2.2</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>Parents</td>
<td>2.4</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.7</td>
</tr>
</tbody>
</table>

MID is about two points on the T-score scale for these health outcomes measures, with no clear difference among the domains.

---

A Free-Response Method to Estimate the “Minimally Important Difference” (MID) between Scores

(Morgan et al., 2017)
DeWitt, Cook, and their colleagues also used something like the scaled-judgment method, but with the judges filling out the responses to the “after” protocol to make it “minimally different” from the (given) “pre” protocol.

This can be conceptualized as a “free response” variant of the scaled-judgment method.


Grace’s Fatigue Report

<table>
<thead>
<tr>
<th></th>
<th>Never</th>
<th>Almost Never</th>
<th>Sometimes</th>
<th>Often</th>
<th>Almost Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>Being tired kept me from having fun.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Being tired made it hard for me to keep up with my schoolwork.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Being tired made it hard for me to play or go out with my friends as much as I like.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt more tired than usual when I woke up in the morning.</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>I felt tired.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt too tired to spend time with my friends.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I felt weak.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I got tired easily.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I had trouble finishing things because I was too tired.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I had trouble starting things because I was too tired.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I needed to rest during the day.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was tired it was hard for me to focus on my work.</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was tired it was hard for me to pay attention.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was tired to do sports or exercises.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was tired doing housework.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was tired to eat.</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>I was tired to go out with my family.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was tired to go up and down a lot of stairs.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“Just enough improvement to make a difference.”
Future Research

Do these different methods of data collection yield consistent results? Or are there predictable differences?

How do results from these methods compare to results obtained with anchor-based methods, when anchors are available?

Everyone finds differences between groups of judges... adolescents, parents, clinicians; what is to be made of that?
Acknowledgments

This work was funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant 1U01AR052181-01.

Thanks to Dave Cella, Karon Cook, and Esi Morgan for their graphics used in this presentation, and my collaborators Yang Liu, Brooke Magnus, Hally Quinn, Debbie S. Gipson, Carlton Dampier, I-Chan Huang, Pamela S. Hinds, Bryce B. Reeve, Heather E. Gross, and Darren A. DeWalt at UNC and across the rest of the PROMIS pediatric multi-site project.
Clinical Outcome Assessments: Establishing and Interpreting Meaningful Within-Patient Change

4/4/17
Clinical Trial Exit Interviews

Presented at the Clinical Outcome Assessments: Establishing and Interpreting Meaningful Within-Patient Change Meeting

The Duke-Margolis Center for Health Policy, Washington, DC, April 4, 2017

Dana DiBenedetti, PhD
Executive Director, Patient-Centered Outcomes Assessment
Acknowledgments

• T. Michelle Brown
• Carla (DeMuro) Romano
• Lynda Doward
• Claire Ervin
• Sheri Fehnel
• Sandy Lewis
• Diane Whalley
What is an “Exit Interview?”

• The collection of (mostly) qualitative data from clinical trial participants
  – Most commonly, interviews are conducted soon after participants complete the treatment period
  – However, patients’ (and/or caregivers’) experiences and perspectives regarding treatment benefit may not be fully captured with traditional COAs.

• Interviews with clinical trial participants provide the opportunity to more fully explore the impacts of investigational products
  – Describe the meaningfulness of treatment-related changes (positive and negative)
  – Identify unanticipated treatment benefits

• Information regarding pre-study experiences, as well as treatment-related expectations and unmet needs can also be collected.
Why Do Exit Interviews?

To identify

- Characteristics of (sometimes new or rare) patient populations
- What symptoms/impacts are **most important** to patients
  - Allows participants to articulate concepts that may be important to them but that are not obtained (or fully obtained) in the trial, thus
    - Enriching researchers’ and sponsors’ understanding of the patient experience
    - Aiding in interpretation of other clinical data
- Full impact of treatment (meaningful changes)
- Unmet needs of treatment
- Expectations for and experiences with disease and of treatment
- Thematic information used to inform future COA strategies and clinical trial designs
- Potential treatment differentiators
Exit Interviews

• Supplement, support, and facilitate the interpretation of data from traditional PRO, PerfO, ObsRO and/or clinical measures
  – Provide greater depth and rationale for data from traditional measures
  – Describe treatment effects
  – Explore the relevance and clinical meaningfulness of specific treatment changes beyond clinical indices and side effects
  – Explain anomalous results
Sample Interview Concepts

Patients’ (and Caregivers’) Experiences With and Attitudes About Treatment

- Symptoms/impact prior to study start
- Expectations of changes/outcomes
  - Can compare pre-study expectations with clinical outcomes
- Anticipated or unanticipated benefits, impact of those benefits
  - Impact of treatment on daily life/functioning
  - Impact of treatment on most important/bothersome symptoms
  - Onset of benefits/changes
- Treatment experiences
  - Convenience of visits, monitoring
  - Managing treatment schedule (e.g., regimen schedule, infusions, monitoring)
  - Most challenging aspect of study treatment
  - Managing adverse events
- How well treatment addresses most important/bothersome symptoms
- Impact of treatment on daily life/functioning, quality of life
- Satisfaction levels with treatment
  - Reasons for satisfaction
Potential Applications

When to conduct interviews

• Both within and outside the context of a clinical trial
  – Implementing as part of a clinical trial is generally more efficient and maximizes participation as compared with a separate or subsequent study

• At various time points (not just at the end of a study)
  – Baseline, at key time point(s) during the study, at the end of a randomized treatment phase, at the end of open-label extension, etc.

• With all participants or select samples of study participants
  – Participants can be selected by site, country, experience of a particular side effect, patient-reported data
Approaches to Conducting Patient Interviews

**Approach 1: Experienced, trained qualitative researchers conduct interviews**
- Interviews conducted via telephone or in-person at designated time(s)
- Can be prospectively planned into the CT protocol or done as a substudy
- Interviews follow a semi-structured guide
- Values of this approach
  - Richest source of data, robust methodologically
  - Level of granularity from experienced interviewers
  - Limits the variability in data quality (vs large number of individuals with varying degrees of qualitative experience)…
  - Qualitative analysis usually done by interviewers themselves

**Approach 2: Study coordinators (SCs) conduct interviews**
- Qualitative interviewers would develop interview guide/related materials, and provide training to SCs
  - Certify, demonstrate proficiency
- Use a more standardized and heavily scripted interview guide
- SCs provide field notes, audio recording etc. to qualitative researchers who analyze qualitative results
- Values of this approach
  - Although data may be less in-depth than Approach 1
  - Particularly effective in global trials in which interview process needs to be scaled to allow for maximal participation
  - Allows for interview to be conducted by a someone familiar to patient
Issues to Consider in Operationalizing

• What questions are you trying to answer with the interviews?
  – Exploratory, looking for a signal vs providing data/support for primary endpoint?
  – Do you need patients from all countries to answer your questions or sample of participants?

• Population
• Sample size
• Who is going to conduct interviews?
• Method
• Timelines
• Budget
• Senior-management buy in
Potential Methodological Considerations / Limitations

• How, if at all, exit interview activities influence CT data
• Self-selection bias of exit interview volunteers (site and patient level)
• Sample
  – All patients, subsample(s), size
• How data will be analyzed
  – How interview data relate to CT data
• Potential for additional adverse event reporting
Factors Contributing to a More Successful Interview Study

• General rule of thumb: the more sites and patients, the easier and less expensive it is to recruit

• Include prospectively in clinical trial (vs. relying on sites and patients to volunteer their participation)
  – Increases site and patient willingness and compliance
  – Increases patient sample size
  – Interview substudy can be included as a component of a clinical trial for select countries (does not have to be for the entire study)
  – Additional protocol amendments and IRB reviews would not be needed
  – Does not significantly add to site burden
  – Training for interview substudy adds ~ 30 minutes to site initiation visits
Factors Contributing to a More Successful Interview Study

• Adequate time to design interview substudy and materials
• Target an adequate sample size (e.g., 30-50 interviewed patients)
  – More likely to identify themes/signals (vs. 10-15 patients)
• Larger site and patient pool increases likelihood of success
  – Easier and more efficient to recruit
  – More “buy-in” from sites and patients
• Include in phase 1B or phase 2 study
  – Increases chances of early identification of signals (e.g., treatment benefits, impacts)
  – Learn what is important to patients that may not be included in protocols
  – Early signals can help inform future study design, PRO measurement strategy, selection of other study endpoints, systematic measurement of new endpoints
Exit Interview Study Examples
Example 1: Exit Interviews with COPD and Asthma Patients in Prospective, Real World Clinical Studies

- RTI-HS designed and is implementing an exploratory study to capture patient-centered information in the context of two real-world studies being conducted in chronic obstructive pulmonary disease (COPD) and asthma.

- The study is investigating the impact and management of COPD and asthma from the patients’ perspective and highlighting the potential relationship between treatment and both behavioral and psychological factors on patients’ experiences.
  - Goal is to identify key risk factors for exacerbations and treatment adherence.

- A mixed methods approach is being used:
  - Quantitative data is being collected through the administration of structured, closed-ended questions administered to all patients via telephone interviews.
  - Qualitative data is also being collected through semi-structured, open-ended questions on key topic areas administered to a subset of patients via face to face interviews.
Example 2: Interviews with Patients with Diabetic Gastroparesis Before and After Treatment

• RTI-HS recently collaborated with a pharmaceutical client developing a new treatment for diabetic gastroparesis (DG)

• Participation in qualitative interviews at both the beginning (pre-treatment) and end (post-treatment) of a phase 2 study was offered to all clinical trial participants

• Primary objective of the pre-treatment interviews was to inform the development of a new PRO measure or modification of an existing PRO measure by:
  – Identifying a comprehensive set of DG symptoms
  – Learning how patients describe the burden and natural variation in these symptoms
  – Understanding the relative bothersomeness of the symptoms
  – Describing expectations related to successful treatment

• Primary objective of the post-treatment interviews was to gather in-depth information about participants’ experience with the study drug, including the magnitude and relative importance of both positive and negative changes

• A manuscript describing the methods and results of this study have just been submitted for publication
Example 3: Exit Interviews with Clinical Trial Participants with Carcinoid Syndrome (CS)

- Task: Regulatory requirement that client assess and document the relevance and clinical meaningfulness of specific CS-related symptoms and their impacts
- Designed and implemented a qualitative study to explore perceptions and experiences of patients following their participation in a clinical trial.
  - Conducted telephone exit interviews with 35 patients across 16 sites in 5 countries enrolled in a phase 3 clinical trial investigating a new treatment for carcinoid syndrome to assess:
    - Participants’ experiences (symptoms and impacts) with their disease
    - Perceived benefits of the study treatment
    - The clinical meaningfulness of specific symptom improvements and their associated impact to the patients
- Mixed methods (qualitative and quantitative data)
- Data analyzed
  - Qualitative
  - Quantitative
  - Compared with selected clinical trial data
Example 3: Exit Interviews with Clinical Trial Participants with Carcinoid Syndrome (CS): Results

- Supported the primary endpoint of decrease in diarrhea
- The 3 most important symptoms to treat and the most bothersome symptoms were diarrhea, BM frequency, and urgency.
  - BM frequency was reported as being more important to treat than stool form/consistency.
- Meaningfulness of changes with treatment
- 95% of participants who reported reductions in BM frequency noted that this was meaningful to them, allowing them to better enjoy life, leave the house, and participate in social and other activities.
  - “I definitely feel like I'm not a prisoner in my house, staying 10 feet to the nearest bathroom. I can go out to activities…”
  - “But the biggest change is not having to run to the toilet constantly…You can't live going 20 times a day. I was able to go out more often…”
- Most participants reported that a BM frequency reduction of at least 30% would be considered meaningful.
Clinical Outcome Assessments:
Establishing and Interpreting Meaningful Within-Patient Change

4/4/17
Exploring the Use of Anchor-Based Methods to Derive and Interpret Meaningful Within-Patient Change

April 4, 2017
ANCHOR-BASED METHODS

- “The anchor-based approaches use an external indicator, either clinical or patient-based, to assign subjects into several groupings reflecting no change, small positive changes, large positive changes, small negative changes, or large negative changes in clinical or health status” (Revicki 2008; p. 104)
- “[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” (PRO Consortium 2015)
- Meaningful within-person change = Change on the target COA measure for patients who experience meaningful improvement or worsening on the anchor
- Gold standard for estimating meaningful within-person change (FDA 2009)
TYPES OF ANCHORS

• Global Impression Change
  • Patient, Caregiver, Clinician Reported

• Global Impression of Symptoms
  • Patient Reported

• Disease Severity Categories
  • e.g., New York Heart Association Classification among heart failure patients

• Occurrence of a Meaningful Event
  • e.g., Hospitalization, disease relapse

• Experience of certain degree of change on a disease-related variable
  • e.g., Loss of 5% body fat in obese patients (Crosby 2003)
Please choose the response below that best describes the overall change in your <SYMPTOM/OVERALL STATUS/ETC> since you started taking the study medication.

- Very much Better
- Moderately Better
- A Little Better
- No Change
- A Little Worse
- Moderately Worse
- Very much Worse
ESTIMATING WITHIN-PATIENT MEANINGFUL CHANGE

Farrar 2001; Pain
TYPES OF ANCHORS

• Global Impression Change
  • Patient, Caregiver, Clinician Reported

• Global Impression of Symptoms
  • Patient Reported

• Disease Severity Categories
  • e.g., New York Heart Association Classification among heart failure patients

• Occurrence of a Meaningful Event
  • e.g., Hospitalization, disease relapse

• Experience of certain degree of change on a disease-related variable
  • e.g., Loss of 5% body fat in obese patients (Crosby 2003)
Please choose the response below that best describes the severity of your <SYMPTOM/OVERALL STATUS/ETC> over the past week.

- None
- Mild
- Moderate
- Severe
- Very Severe
TYPES OF ANCHORS

• Global Impression Change
  • Patient, Caregiver, Clinician Reported

• Global Impression of Symptoms
  • Patient Reported

• Disease Severity Categories
  • e.g., New York Heart Association Classification among heart failure patients

• Occurrence of a Meaningful Event
  • e.g., Hospitalization, disease relapse

• Experience of certain degree of change on a disease-related variable
  • e.g., Loss of 5% body fat in obese patients (Crosby 2003)
CONSIDERATIONS WHEN SELECTING ANCHORS

- Anchors should be easier to interpret than the PRO measure itself (FDA, 2009)
- Correlation between anchor and target COA should be greater than 0.30-0.40 (Hays 2005; Revicki 2008)
- Should anchor assess change in a specific symptom/function or a more global assessment of health?
- Recall bias with impression of change items
- Most appropriate anchor type for different types of COAs?
- Recommended to use multiple independent anchors and to examine and confirm responsiveness across multiple samples (Revicki 2008)
CONSIDERATIONS WHEN USING ANCHOR-BASED METHODS

- Type of analysis to determine meaningful change?
  - Descriptive: Average COA score at each level of PGIC
  - Formal: Regression analysis, ROC curve
- What level of change should be considered as the marker for meaningful change? Minimal? Moderate? Large?
- Only use estimate from group that has changed? Difference between changed and stable groups?
- Non-linear relationship between anchor and COA score
NON-LINEAR RELATIONSHIP BETWEEN PGIC AND PRO

Distribution of Change Scores by PGIC Category

Farrar 2001; Pain

Gwaltney Consulting Confidential
NON-LINEAR RELATIONSHIP BETWEEN PGIC AND PRO

Farrar 2001; Pain

Gwaltney Consulting  Confidential
FIG. 1. Boxplot of absolute change from baseline to week 12 in the number of moderate to severe hot flushes by answer to CGI question. CGI, Clinical Global Impression; v., very; minim., minimally; abs., absolute.
CONSIDERATIONS WHEN USING ANCHOR-BASED METHODS

- Use of cross-sectional approaches?
  - e.g., Difference between disease severity categories at single point in time
- Effect of unblinding on PGIC rating?
- False sense of precision
  - Clinical trials are less likely to acknowledge the error associated with estimates
- Different anchors can lead to substantially different findings – How integrate findings?
REFERENCES

- PRO Consortium 2015. Interpreting Change in Scores on COA Endpoint Measures
Clinical Outcome Assessments: Establishing and Interpreting Meaningful Within-Patient Change

4/4/17