# Advances in Interpretation of Patient-Reported Outcomes

Joseph C. Cappelleri, PhD, MPH, MS
Executive Director of Biostatistics
Pfizer Inc
joseph.c.cappelleri@pfizer.com

Invited 90-minute presentation tutorial at BASS XXX – 30th Annual Biopharmaceutical Applied Statistics Symposium, October 23-25, 2023, Savannah, Georgia

Forsyth Park - 30 acres in the historic district of Savannah

# Welcome to Historic Savannah, Georgia!

**Bonaventure Cemetery**
**(*Midnight in the Garden of Good and Evil*)**

**Wormsloe Historic Site**
**(1.5-mile-long tree-lined driveway)**

**River Street (Waterfront)**

**Tybee Island Lighthouse (1736, 145 ft.)**

# Happy Birthday BASS!

Flashback to 1993

## The Party Times

SPECIAL EDITION
Read all about the year 1993!

Vol. I No. 1 NEWS ★ COST OF LIVING ★ SPORTS ★ TECHNOLOGY ★ & MORE! PRICE 60 cents

# 30 YEARS AGO BACK IN 1993

360 months ★ 10,957 days ★ 262,968 hours ★ 15,778,080 minutes ★ 946,684,800 seconds

42rd U.S. PRESIDENT
*Bill Clinton*
U.S. POPULATION
*260.3 million*

### What happened in 1993

IBM announces a $4,970,000,000 loss for 1992, the largest single-year corporate loss in United States history to date. * $7,400,000 USD is stolen from Brinks Armored Car Depot in Rochester, NY in the fifth-largest robbery in U.S. history. * The Intel Corporation ships the first Pentium chips. * The Great Blizzard strikes the eastern U.S., bringing record snowfall and other severe weather all the way from Cuba to Quebec. * NASA loses contact with the Mars Observer spsocraft. * Windows NT 3.1, the first version of Microsoft's line of Windows NT operating systems, is released to manufacturing.

*Happy 30th Birthday*
1993
YEAR OF THE ROOSTER

THE 65th ACADEMY AWARDS
*ACTOR*
Al Pacino
*Scent of a Woman*
*ACTRESS*
Emma Thompson
*Howards End*
*DIRECTING*
Clint Eastwood
*Unforgiven*
*BEST PICTURE*
*Unforgiven*

POPULAR BABY NAMES
*GIRLS*
Jessica, Ashley, Sarah, Samantha, Emily, Brittany
*BOYS*
Michael, Christopher, Matthew, Joshua, Tyler, Brandon

### WHAT THINGS COST

Minimum wage: $4.25/ hour
New house: $113,000.00
Gallon of gas: $1.16
Gallon of milk: $2.85
Loaf of bread: $1.57
Dozen eggs: $1.05
Postage stamp: $0.29
Movie ticket: $4.15

*Average Income Per Year*
***$31,200.00***

### *POPULAR GAMES & TOYS*

GAMES Doom, Myst, Secret of Mana, Mortal Kombat II, Star Wars: X-Wing

TOYS Aladdin Game, Beanie Babies, Sesame Street Big Bird Story Magic

## *WORLD POPULATION HIT 5.5 BILLION*

### ★ STARS BORN IN 1993 ★

Ariana Grande, KSI, Savannah LaBrant, Zayn Malik, Liam Payne, Sofia Carson, Debby Ryan, Alisha Marie, Niall Horan

*GENERATION Y (GEN Y)*

### ON THE BIG SCREEN

Jurassic Park, Sleepless in Seattle, The Fugitive, Mrs. Doubtfire, Indecent Proposal, Schindler's List, The Firm, Philadelphia, Cliffhanger, The Pelican Brief, Gettysburg

### ON TELEVISION

Seinfeld, Frasier, Coach, 60 Minutes, Grace Under Fire, Home Improvement, Roseanne, 20/20, Murphy Brown, Love & War

### IN *Style...*

Plaid flannel shirts
Ripped jeans
Denim overalls
Timberland boots
Doc Martens
JanSport backpacks
CK 1 Fragrance
Rollerblades
The Rachel haircut

### ON THE RADIO

"I Will Always Love You" -Whitney Houston
"Whoomp! (There It Is)"-Tag Team
"Can't Help Falling in Love" - UB40
"That's the Way Love Goes"-Janet Jackson
"What's Up?" - 4 Non Blondes

### 90's SLANG

Chillin' - Taking it easy
Diss - Lack of respect
Da Bomb - Really Cool
Boo Ya! - Excitement
Dope - Something great

### *SPORTS HIGHLIGHTS*

*NBA FINALS:*
Chicago Bulls win 4 games to 2 over the Phoenix Suns to complete their first three-peat of the decade.

*WORLD SERIES:*
The Toronto Blue Jays win 4 games to 2 over the Philadelphia Phillies.

*NFL SUPER BOWL XXVII:*
Dallas Cowboys win 52-17 over the Buffalo Bills.

*STANLEY CUP WINNER:*
Montreal Canadiens win 4 games to 1 over the Los Angeles Kings.

HAPPY BIRTHDAY! ALL THE BEST TO YOU!

# Learning Objectives

- To understand methods for interpretation of patient-reported outcomes

- To understand specific applications of the methods

# Outline

- Anchor-based approaches
  - Percentages based on thresholds
  - Cutoff scores based on severity
  - Criterion-group interpretation
  - Statistical significance and clinical equivalence
  - Content-based interpretation
  - Clinically meaningful change and difference

- Distribution-based approaches
  - Effect size, % of range, reliability change index
  - Probability of relative benefit
  - Cumulative distribution function

- Mediation analysis

# Importance of Interpretation

- Results on a patient-reported outcome (PRO) scores should be interpreted in a meaningful way
- Benefit to patients and other stakeholders
- Methods are needed to enrich interpretation of PRO scores

SCENE of ENDURING LOVE

*Life* magazine cover, February 25, 1909

# Anchor-Based Approaches

# What is an Anchor?

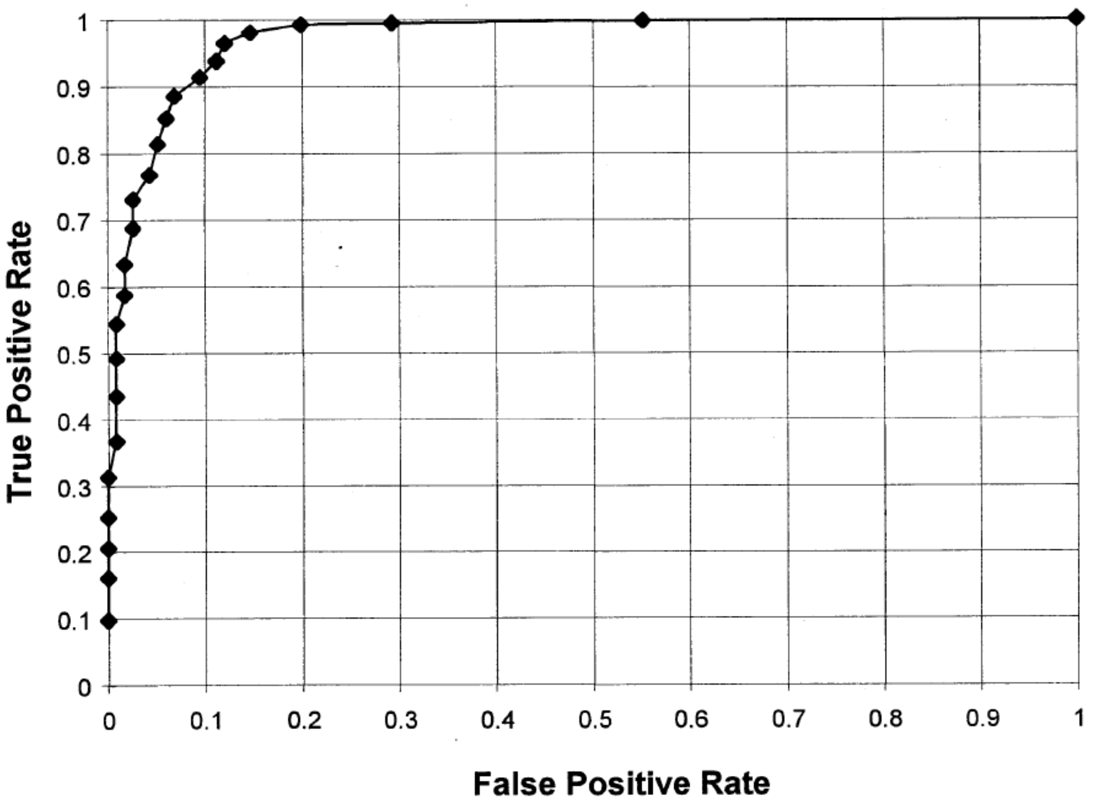
- Anchor measure is external or adjunct to the target PRO measure of interest
- Anchor measure should bear an appreciable correlation with the PRO measure
- Anchor measure should itself be clearly interpretable

shutterstock.com · 1818321527

# Percentages Based on Thresholds or Cutoff Scores

- Show percentage of patients above and below some specified value, which is an anchored value with a meaningful criterion
- Useful for inclusion criterion or treatment efficacy
- Erectile function domain of International Index of Erectile Function (IIEF)
- European Organization for Research and Treatment of Cancer Quality-of-Life Questionnaire 30 (EORTC QLQ-C30)
- Severity categorization on Fibromyalgia Impact Questionnaire (FIQ)
- Fatigue measure from National Institutes of Neurological Quality of Life measurement initiative (Neuro-QOL)
- Anxiety measure in oncology from Patient-Reported Outcomes Measurement Information System (PROMIS)

# Self-Reported Diagnosis of Erectile Function (IIEF): Receiver Operating Characteristic Curve



- Outcome: Clinically diagnosed erectile dysfunction (ED) vs. no such ED
- Predictor: Erectile function domain on IIEF (range 1-30, higher scores better)
- Model: Logistic regression
- Optimal cutoff (Youden's index):
  <= 25, ED; > 25, ED

- 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

# Patients with Functional Scale Scores Below (Worse) the Clinical Problem Threshold: POLARIS

| EORTC QLC-C30 Scale | Clinical problem (threshold scores) | Baseline n (%) | Month 6 n (%) | Month 12 n (%) |
|---|---|---|---|---|
| **Functioning Scales** | | | | |
| Physical functioning | <83 | 629 (54.6) | 365 (50.2) | 239 (50.4) |
| Role functioning | <58 | 330 (28.6) | 146 (20.1) | 88 (18.6) |
| Social functioning | <58 | 278 (24.2) | 109 (15.0) | 71 (15.0) |
| Emotional functioning | <71 | 427 (37.1) | 215 (29.6) | 139 (29.4) |
| Cognitive functioning | <75 | 394 (34.2) | 251 (34.5) | 150 (31.7) |

POLARIS = Palbociclib in Hormone-Receptor-Positive Advanced Breast Cancer

Source: Karaturi et al. 2021

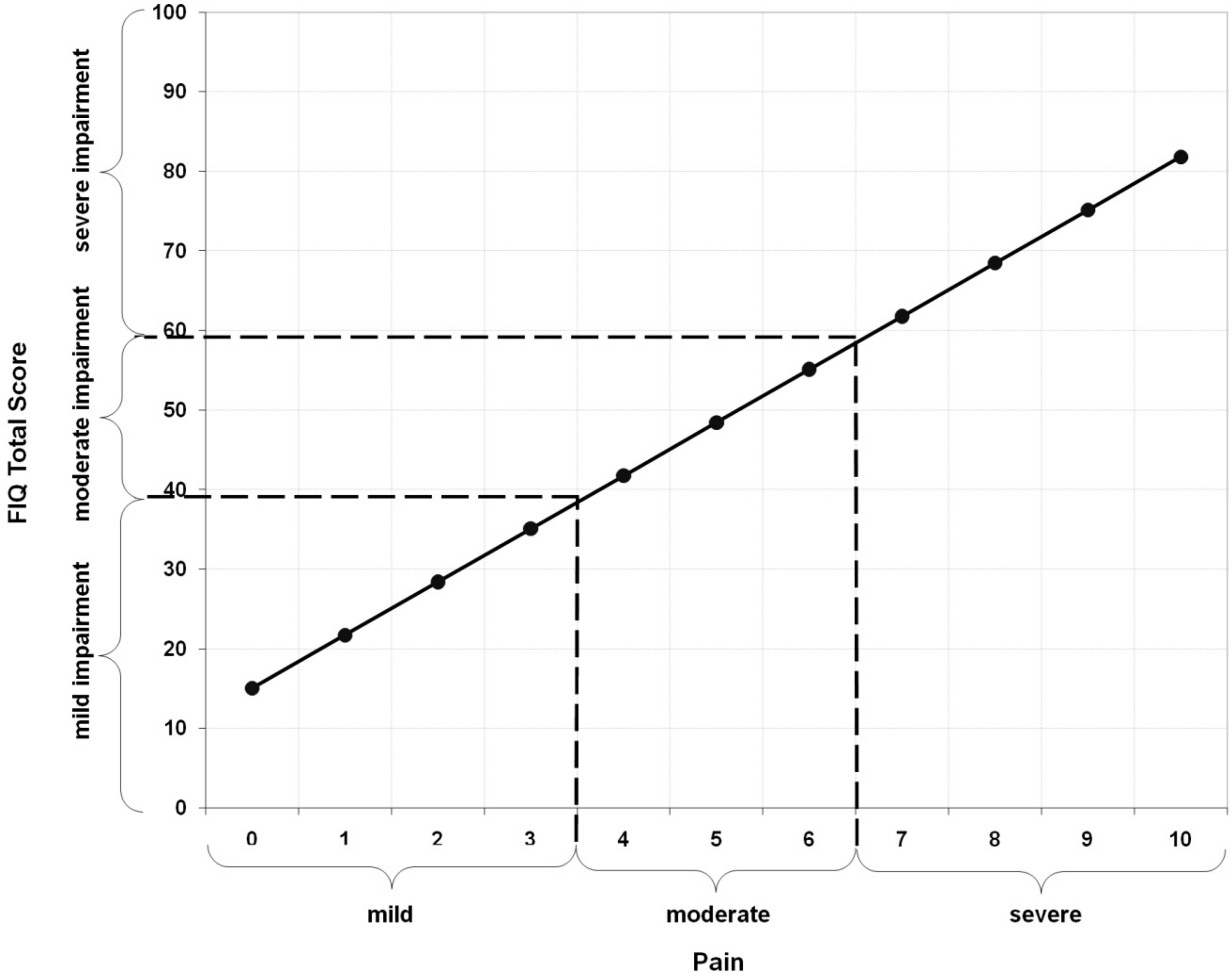
# Patients with Functional Scale Scores Above (Worse) the Clinical Problem Threshold: POLARIS

| | Clinical problem | Baseline | Month 6 | Month 12 |
|---|---|---|---|---|
| EORTC QLQ-C30 Scale | (threshold score) | n (%) | n (%) | n (%) |
| Symptom Scales | | | | |
| Fatigue | >39 | 421 (36.5) | 256 (35.2) | 144 (30.4) |
| Pain | >25 | 614 (53.3) | 341 (46.9) | 221 (46.6) |
| Nausea and vomiting | >8 | 423 (36.7) | 233 (32.0) | 153 (32.3) |
| Insomnia | >50 | 279 (24.2) | 156 (21.5) | 89 (18.8) |
| Appetite loss | >50 | 205 (17.8) | 72 (9.9) | 48 (10.1) |
| Constipation | >50 | 142 (12.3) | 71 (9.8) | 46 (9.7) |
| Dyspnea | >17 | 547 (47.6) | 323 (44.5) | 207 (43.7) |
| Diarrhea | >17 | 327 (28.5) | 209 (28.7) | 120 (25.4) |
| Financial impact of disease | >17 | 563 (49.0) | 322 (44.3) | 198 (42.0) |

POLARIS = Palbociclib in Hormone-Receptor-Positive Advanced Breast Cancer

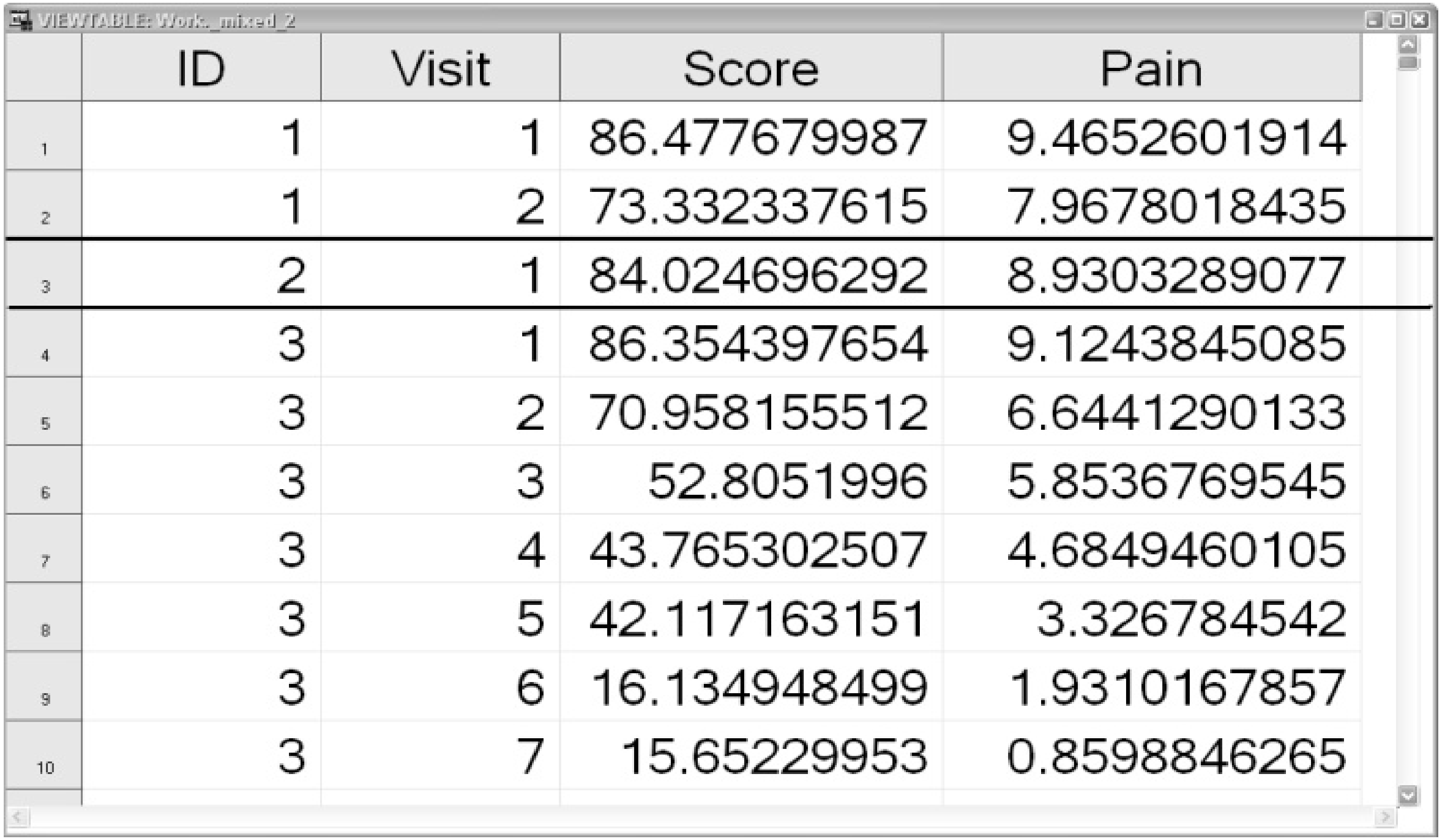
Source: Karaturi et al. 2021

# Severity Categorization of FIQ Total Score Using Pain Severity as an Anchor



Source: Bennett et al. 2009

# Simulated Example in SAS: FIQ Severity Categorization (first 3 subjects)



|  | ID | Visit | Score | Pain |
|---|---|---|---|---|
| 1 | 1 | 1 | 86.477679987 | 9.4652601914 |
| 2 | 1 | 2 | 73.332337615 | 7.9678018435 |
| 3 | 2 | 1 | 84.024696292 | 8.9303289077 |
| 4 | 3 | 1 | 86.354397654 | 9.1243845085 |
| 5 | 3 | 2 | 70.958155512 | 6.6441290133 |
| 6 | 3 | 3 | 52.8051996 | 5.8536769545 |
| 7 | 3 | 4 | 43.765302507 | 4.6849460105 |
| 8 | 3 | 5 | 42.117163151 | 3.326784542 |
| 9 | 3 | 6 | 16.134948499 | 1.9310167857 |
| 10 | 3 | 7 | 15.65229953 | 0.8598846265 |

Source: Cappelleri, Zou, Bushmakin et al. 2013

# SAS Code:
# FIQ Severity Categorization

```
Proc Mixed data=_mixed_2;
 Class ID Visit ;
 Model Score = Pain / ddfm=kr s;
 Repeated Visit / Type=UN Subject=ID;
 Estimate " Pain =0 " Intercept 1 Pain 0 /cl;
 Estimate " Pain =1 " Intercept 1 Pain 1 /cl;
 Estimate " Pain =2 " Intercept 1 Pain 2 /cl;
 Estimate " Pain =3 " Intercept 1 Pain 3 /cl;
 Estimate " Pain =4 " Intercept 1 Pain 4 /cl;
 Estimate " Pain =5 " Intercept 1 Pain 5 /cl;
 Estimate " Pain =6 " Intercept 1 Pain 6 /cl;
 Estimate " Pain =7 " Intercept 1 Pain 7 /cl;
 Estimate " Pain =8 " Intercept 1 Pain 8 /cl;
 Estimate " Pain =9 " Intercept 1 Pain 9 /cl;
 Estimate " Pain =10" Intercept 1 Pain 10 /cl;
 Estimate " Pain =3.5 " Intercept 1 Pain 3.5 /cl;
 Estimate " Pain =6.5 " Intercept 1 Pain 6.5 /cl;
Run;
```

# Results from Simulated Example

| *Label* | *Estimate* | *Standard Error* | *Pr > \|t\|* | *Alpha* | *Lower* | *Upper* |
|---|---|---|---|---|---|---|
| Pain =0 | 6.5523 | 1.8715 | 0.0024 | 0.05 | 2.6299 | 10.4746 |
| Pain =1 | 15.5845 | 1.5984 | <.0001 | 0.05 | 12.2173 | 18.9517 |
| Pain =2 | 24.6168 | 1.3292 | <.0001 | 0.05 | 21.7971 | 27.4364 |
| Pain =3 | 33.6490 | 1.0668 | <.0001 | 0.05 | 31.3650 | 35.9330 |
| Pain =4 | 42.6812 | 0.8179 | <.0001 | 0.05 | 40.9150 | 44.4475 |
| Pain =5 | 51.7135 | 0.5995 | <.0001 | 0.05 | 50.4335 | 52.9935 |
| Pain =6 | 60.7457 | 0.4576 | <.0001 | 0.05 | 59.8182 | 61.6733 |
| Pain =7 | 69.7780 | 0.4679 | <.0001 | 0.05 | 68.8473 | 70.7087 |
| Pain =8 | 78.8102 | 0.6229 | <.0001 | 0.05 | 77.5709 | 80.0495 |
| Pain =9 | 87.8425 | 0.8465 | <.0001 | 0.05 | 86.1555 | 89.5294 |
| Pain =10 | 96.8747 | 1.0976 | <.0001 | 0.05 | 94.6826 | 99.0669 |
| Pain =3.5 | 38.1651 | 0.9400 | <.0001 | 0.05 | 36.1427 | 40.1876 |
| Pain =6.5 | 65.2619 | 0.4408 | <.0001 | 0.05 | 64.3820 | 66.1417 |

# Bookmarking

- Bookmarking is designed for measured calibrated using Item Response Theory (IRT)

- IRT is probability-based, psychometric method that estimates the probability of a particular response to a scale item based on patient severity levels and characteristics of the item (e.g., item difficulty or severity)
  - For example, items with higher levels of difficulty or severity may be those that require higher levels of health to endorse
  - For instance, running as opposed to walking for physical functioning

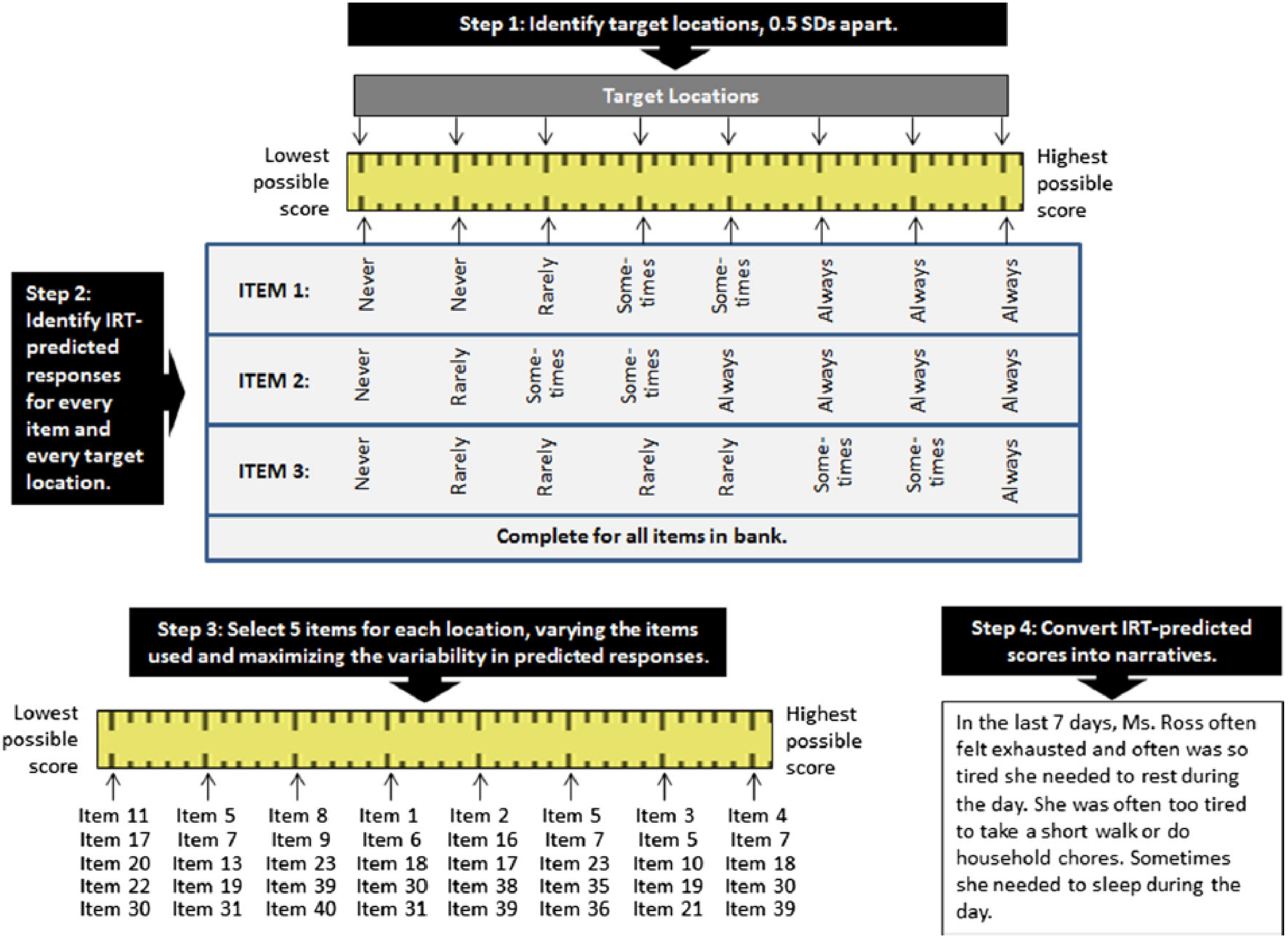
- Performed with a facilitator & set of panelists (experts, patients)

# Bookmarking with Fatigue Measure (Neuro-QOL)

- Based on IRT, clinical vignettes were developed to represent graduated levels of symptom severity on fatigue in persons with multiple sclerosis
- Panelists identified adjacent vignettes they judged to present the threshold between two levels of severity
  - For example, threshold between a vignette that indicated “no problem” with fatigue and an adjacent one that represent “mild problem”
- Cutoff scores on the multi-item fatigue measure were defined as the mean location for each pair of threshold vignettes

Source: Cook et al. 2015

# Steps for Developing Clinical Vignettes from an IRT-Calibrated Item Bank



**Step 1: Identify target locations, 0.5 SDs apart.**

**Step 2: Identify IRT-predicted responses for every item and every target location.**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ITEM 1: | Never | Never | Rarely | Some-times | Some-times | Always | Always | Always |
| ITEM 2: | Never | Rarely | Some-times | Some-times | Always | Always | Always | Always |
| ITEM 3: | Never | Rarely | Rarely | Rarely | Rarely | Some-times | Some-times | Always |

Complete for all items in bank.

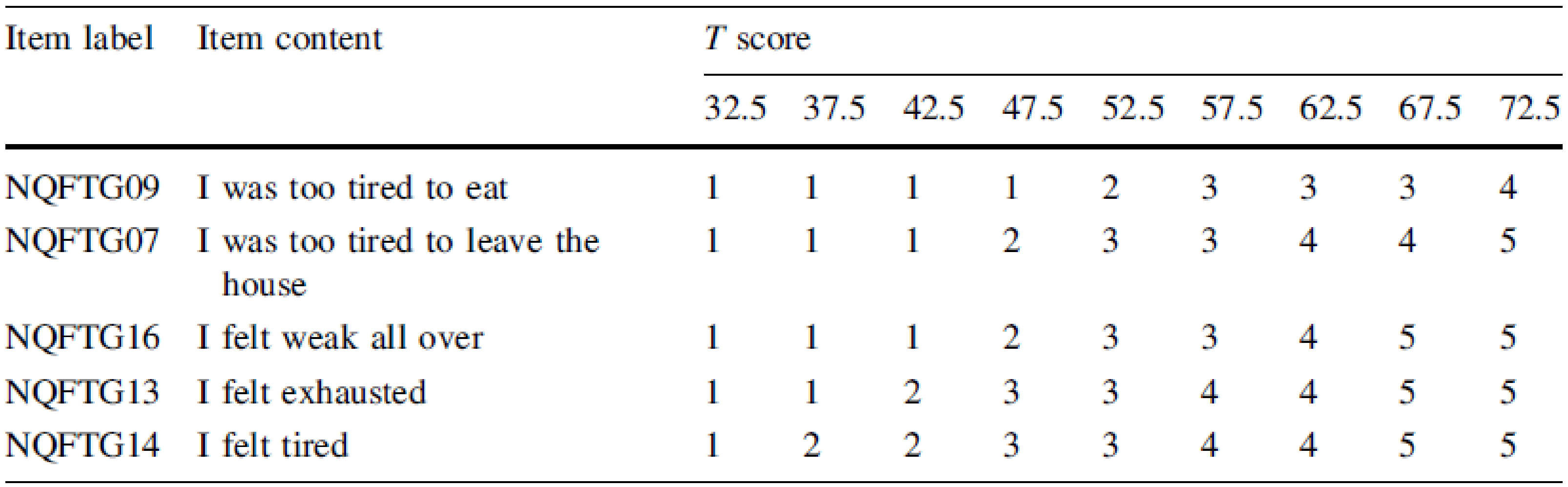
**Step 3: Select 5 items for each location, varying the items used and maximizing the variability in predicted responses.**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Item 11 | Item 5 | Item 8 | Item 1 | Item 2 | Item 5 | Item 3 | Item 4 |
| Item 17 | Item 7 | Item 9 | Item 6 | Item 16 | Item 7 | Item 5 | Item 7 |
| Item 20 | Item 13 | Item 23 | Item 18 | Item 17 | Item 23 | Item 10 | Item 18 |
| Item 22 | Item 19 | Item 39 | Item 30 | Item 38 | Item 35 | Item 19 | Item 30 |
| Item 30 | Item 31 | Item 40 | Item 31 | Item 39 | Item 36 | Item 21 | Item 39 |

**Step 4: Convert IRT-predicted scores into narratives.**

In the last 7 days, Ms. Ross often felt exhausted and often was so tired she needed to rest during the day. She was often too tired to take a short walk or do household chores. Sometimes she needed to sleep during the day.

# IRT-Derived Most Probable Response by *T* Score for Five Neuro-QoL Fatigue Items



| Item label | Item content | *T* score | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 32.5 | 37.5 | 42.5 | 47.5 | 52.5 | 57.5 | 62.5 | 67.5 | 72.5 |
| NQFTG09 | I was too tired to eat | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 3 | 4 |
| NQFTG07 | I was too tired to leave the house | 1 | 1 | 1 | 2 | 3 | 3 | 4 | 4 | 5 |
| NQFTG16 | I felt weak all over | 1 | 1 | 1 | 2 | 3 | 3 | 4 | 5 | 5 |
| NQFTG13 | I felt exhausted | 1 | 1 | 2 | 3 | 3 | 4 | 4 | 5 | 5 |
| NQFTG14 | I felt tired | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 5 |

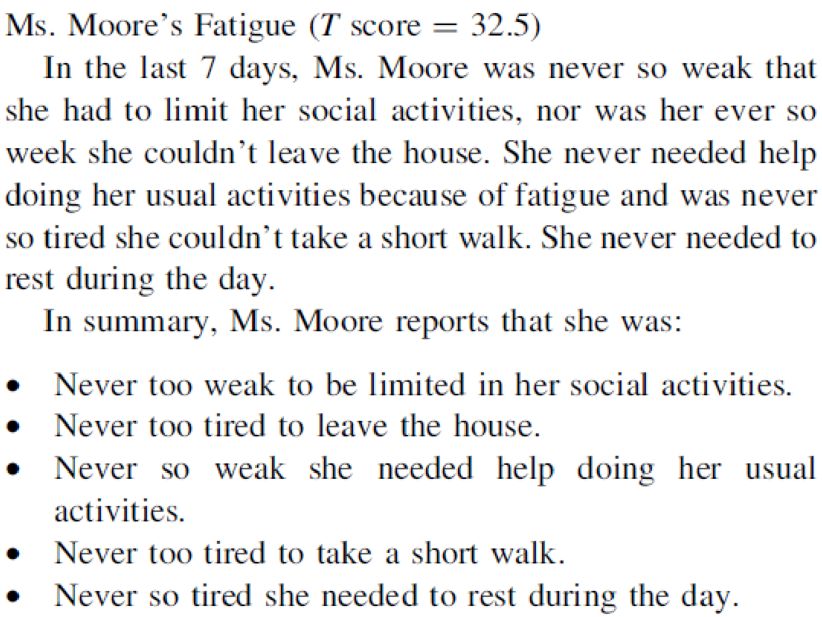
1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Often, 5 = Always

- Multi-item scores are normed by *T* scores with a mean of 50 and standard deviation of 10 based on a reference population (e.g., U.S. general population sample)

- Thus, a *T* score of 60 indicates a score 1 standard deviation above reference mean

- Higher scores reflect more of the domain measured (higher level of fatigue)

# Clinical Vignettes on Fatigue Neuro-QOL: Examples

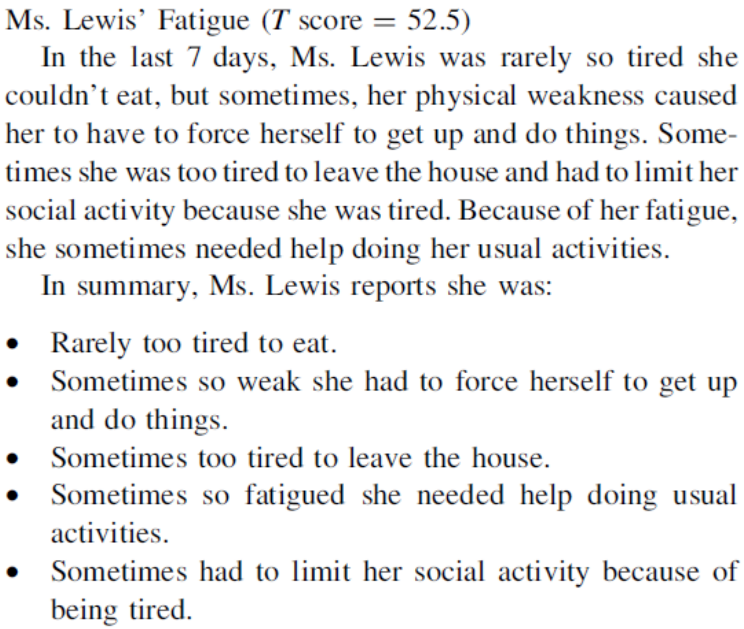


Ms. Moore's Fatigue (*T* score = 32.5)

In the last 7 days, Ms. Moore was never so weak that she had to limit her social activities, nor was her ever so week she couldn't leave the house. She never needed help doing her usual activities because of fatigue and was never so tired she couldn't take a short walk. She never needed to rest during the day.

In summary, Ms. Moore reports that she was:

- Never too weak to be limited in her social activities.
- Never too tired to leave the house.
- Never so weak she needed help doing her usual activities.
- Never too tired to take a short walk.
- Never so tired she needed to rest during the day.

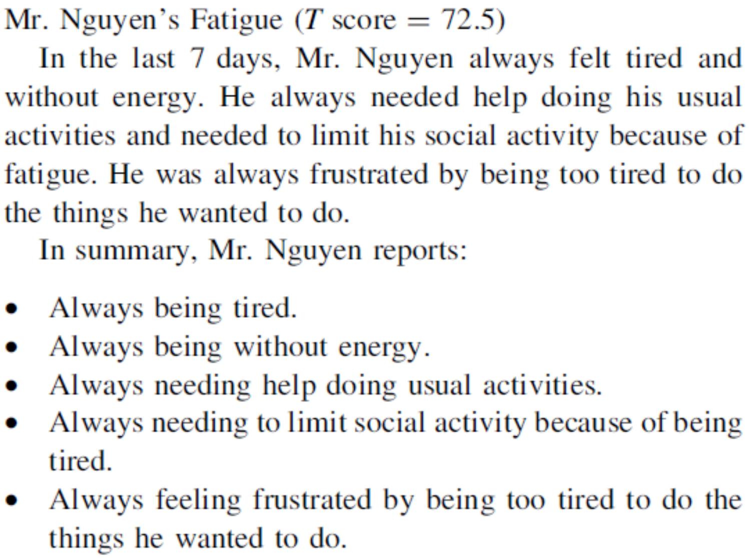


Ms. Lewis' Fatigue (*T* score = 52.5)

In the last 7 days, Ms. Lewis was rarely so tired she couldn't eat, but sometimes, her physical weakness caused her to have to force herself to get up and do things. Sometimes she was too tired to leave the house and had to limit her social activity because she was tired. Because of her fatigue, she sometimes needed help doing her usual activities.

In summary, Ms. Lewis reports she was:

- Rarely too tired to eat.
- Sometimes so weak she had to force herself to get up and do things.
- Sometimes too tired to leave the house.
- Sometimes so fatigued she needed help doing usual activities.
- Sometimes had to limit her social activity because of being tired.



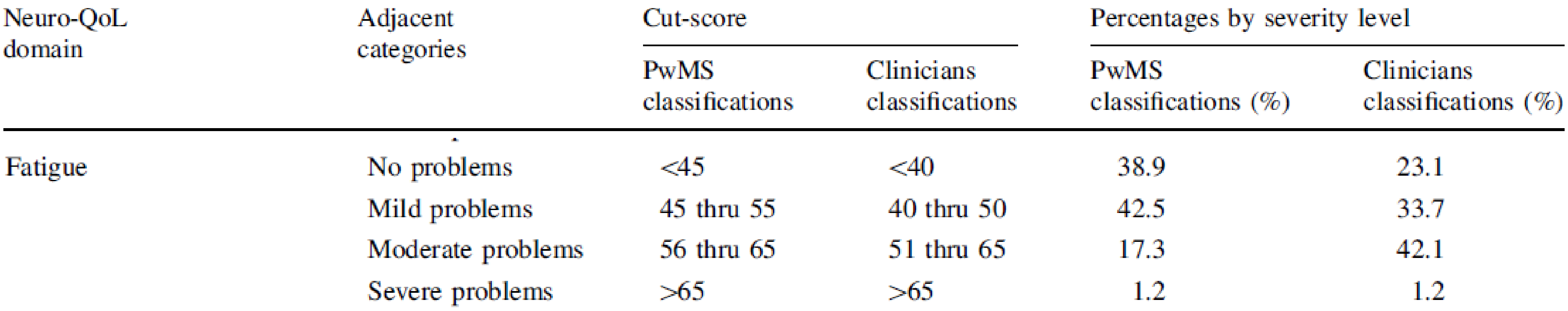
Mr. Nguyen's Fatigue (*T* score = 72.5)

In the last 7 days, Mr. Nguyen always felt tired and without energy. He always needed help doing his usual activities and needed to limit his social activity because of fatigue. He was always frustrated by being too tired to do the things he wanted to do.

In summary, Mr. Nguyen reports:

- Always being tired.
- Always being without energy.
- Always needing help doing usual activities.
- Always needing to limit social activity because of being tired.
- Always feeling frustrated by being too tired to do the things he wanted to do.

# Consensus Cutoff by Expert Panel

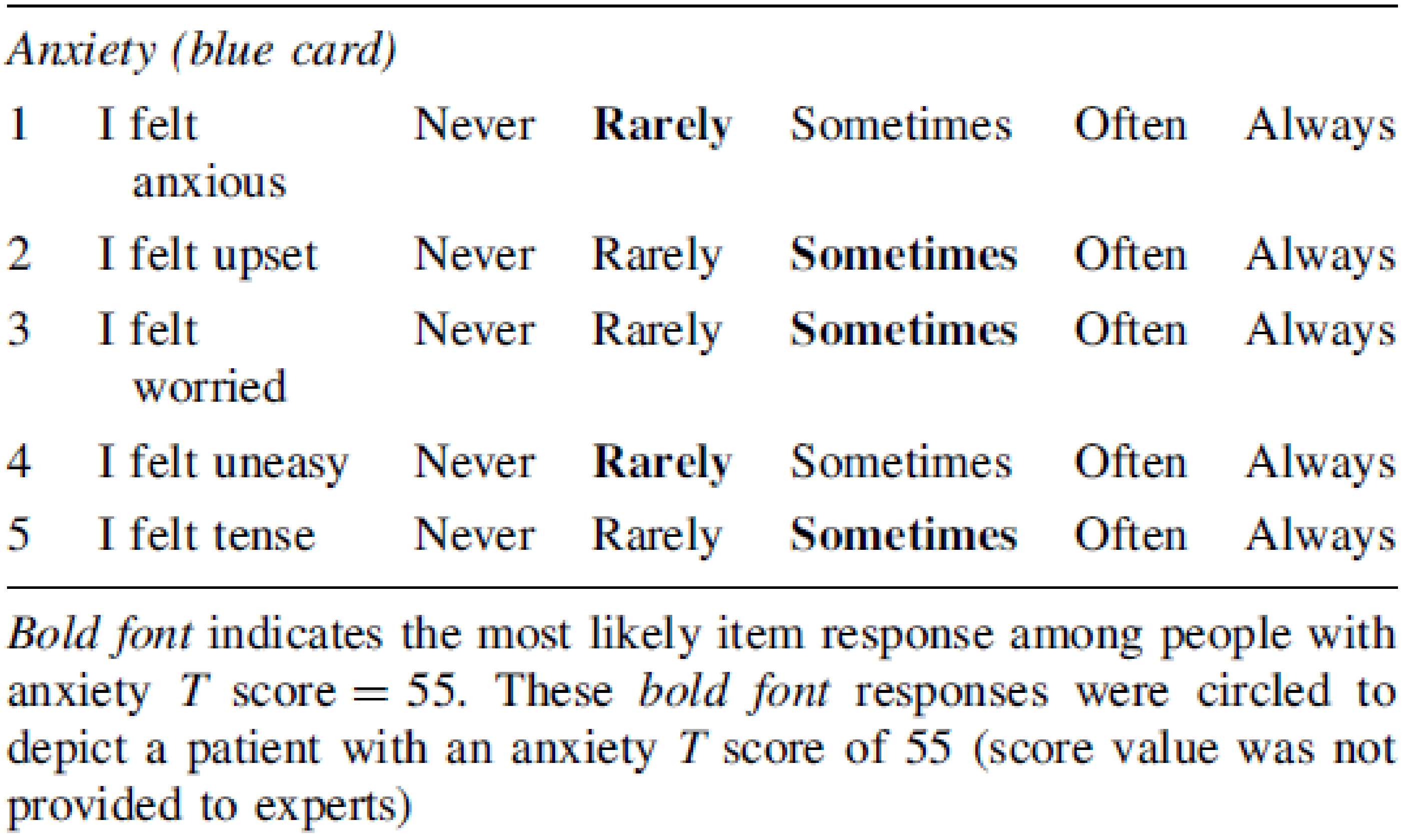


| Neuro-QoL domain | Adjacent categories | Cut-score | | Percentages by severity level | |
|---|---|---|---|---|---|
| | | PwMS classifications | Clinicians classifications | PwMS classifications (%) | Clinicians classifications (%) |
| Fatigue | No problems | <45 | <40 | 38.9 | 23.1 |
| | Mild problems | 45 thru 55 | 40 thru 50 | 42.5 | 33.7 |
| | Moderate problems | 56 thru 65 | 51 thru 65 | 17.3 | 42.1 |
| | Severe problems | >65 | >65 | 1.2 | 1.2 |

PwMS = Persons with multiple sclerosis

# Setting Standards for Severity of Common Symptoms in Oncology: Example – Anxiety (PROMIS)

Sample Vignette Card Presented to Experts for Ranking & Bookmarking



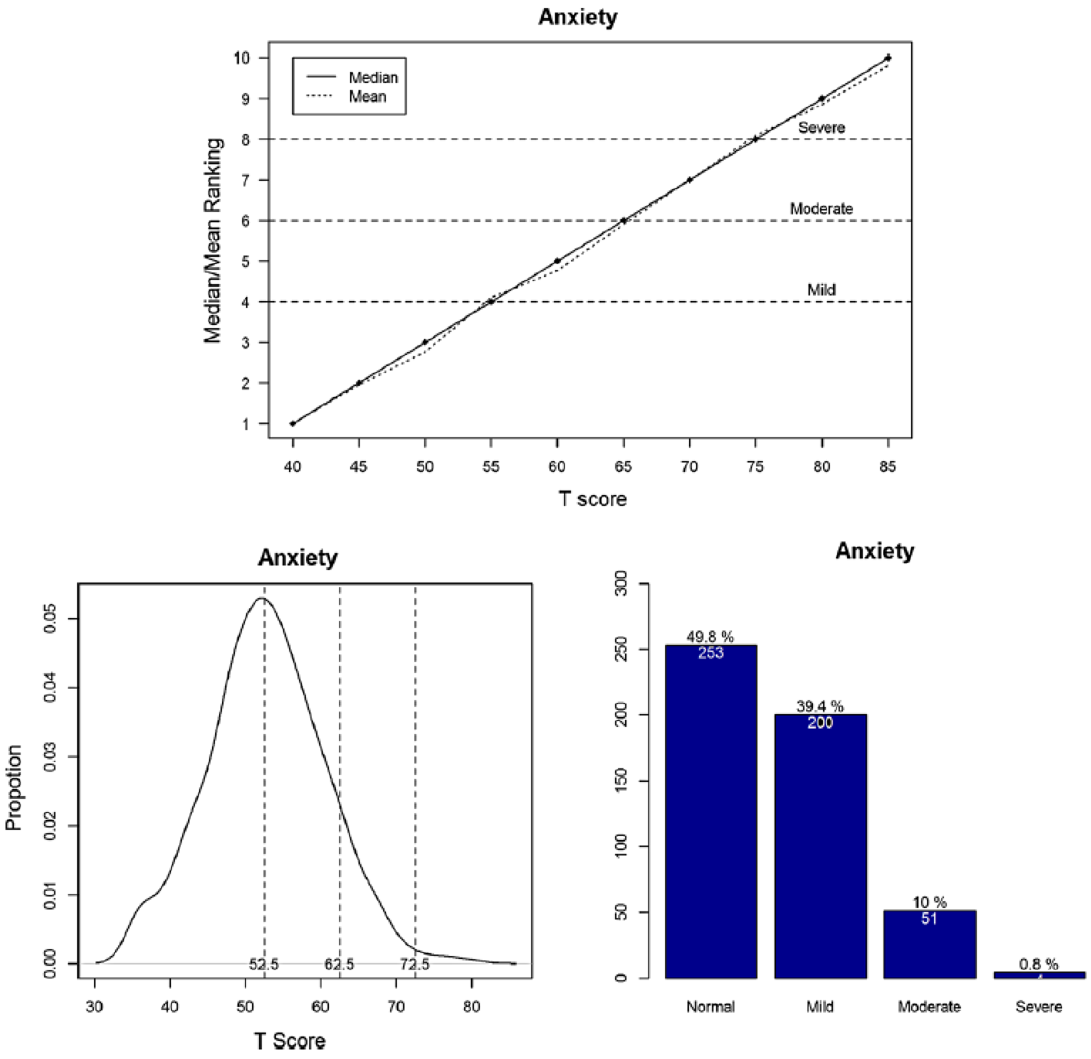
*Anxiety (blue card)*

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | I felt anxious | Never | **Rarely** | Sometimes | Often | Always |
| 2 | I felt upset | Never | Rarely | **Sometimes** | Often | Always |
| 3 | I felt worried | Never | Rarely | **Sometimes** | Often | Always |
| 4 | I felt uneasy | Never | **Rarely** | Sometimes | Often | Always |
| 5 | I felt tense | Never | Rarely | **Sometimes** | Often | Always |

*Bold font* indicates the most likely item response among people with anxiety *T* score = 55. These *bold font* responses were circled to depict a patient with an anxiety *T* score of 55 (score value was not provided to experts)

Cella et al. 2014

# Anxiety: < 55 normal; 55-64 mild; 65-74 moderate; >=75 severe



*Top panel* plots the vignette *T* score (*x* axis) against the median and mean card rankings according to expert consensus (*y* axis). *Dotted horizontal lines* reflect the expert consensus on bookmarks separating the severity of symptom vignettes (mild; moderate; severe). Experts were blind to vignette *T* score values throughout the exercise. *Lower left panel* displays the distribution of anxiety scores (*y* axis) by *T* score (*x* axis), with *vertical lines* separating clinical categories (none; mild; moderate; severe). *Lower right panel* indicates the number and proportion of patients in each of the four clinical categories

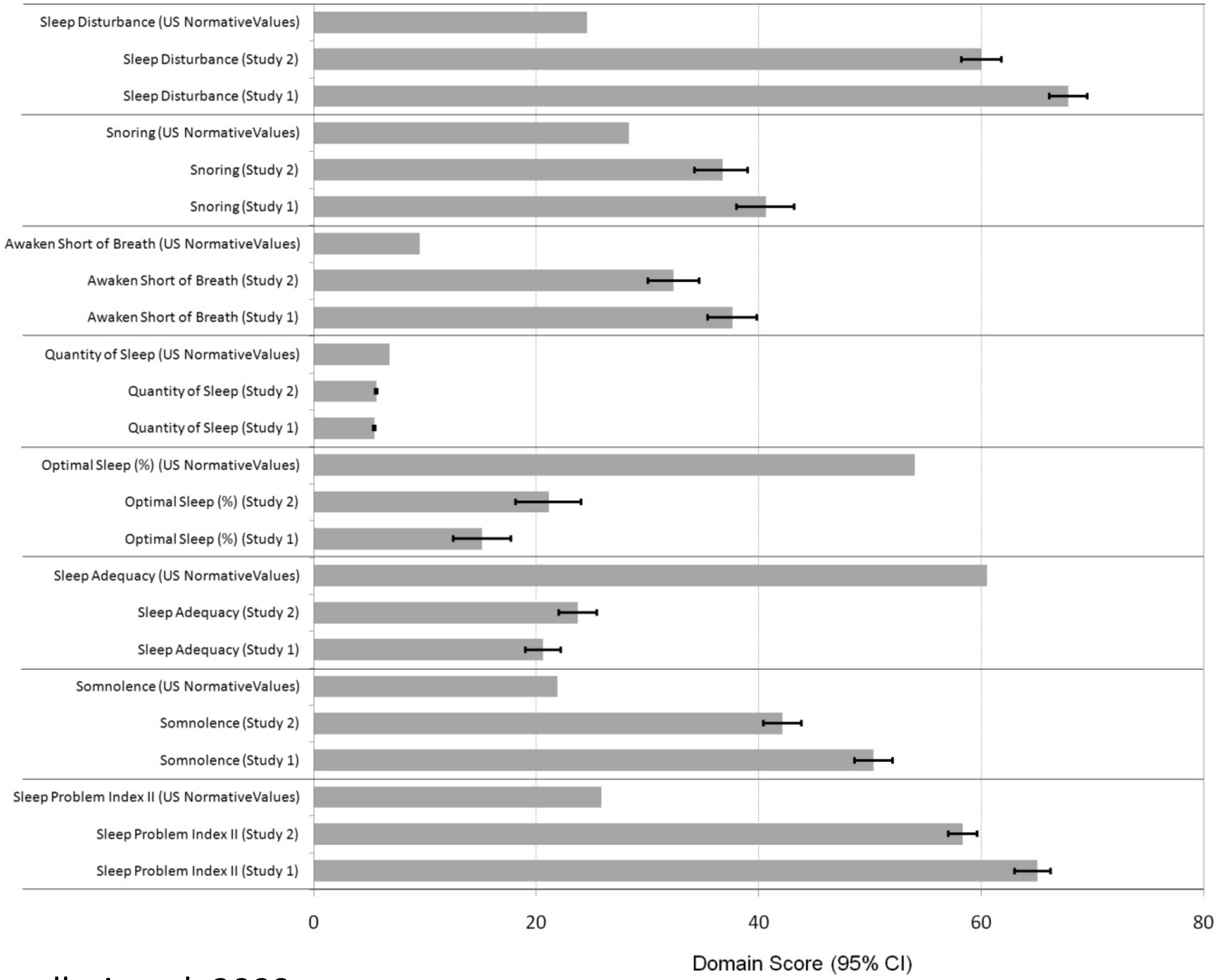
# Criterion-Group Interpretation

- Involves a comparison of scores from the particular group of interest to a criterion group

- Criterion group is a known group worthy of comparison which can serve as a yardstick

- For example, criterion group can be a healthy group, general population, or clinical group

# Baseline Mean Scores on the Medical Outcomes Study Sleep Scale: Patients with Fibromyalgia vs. Values from the U.S. General Population

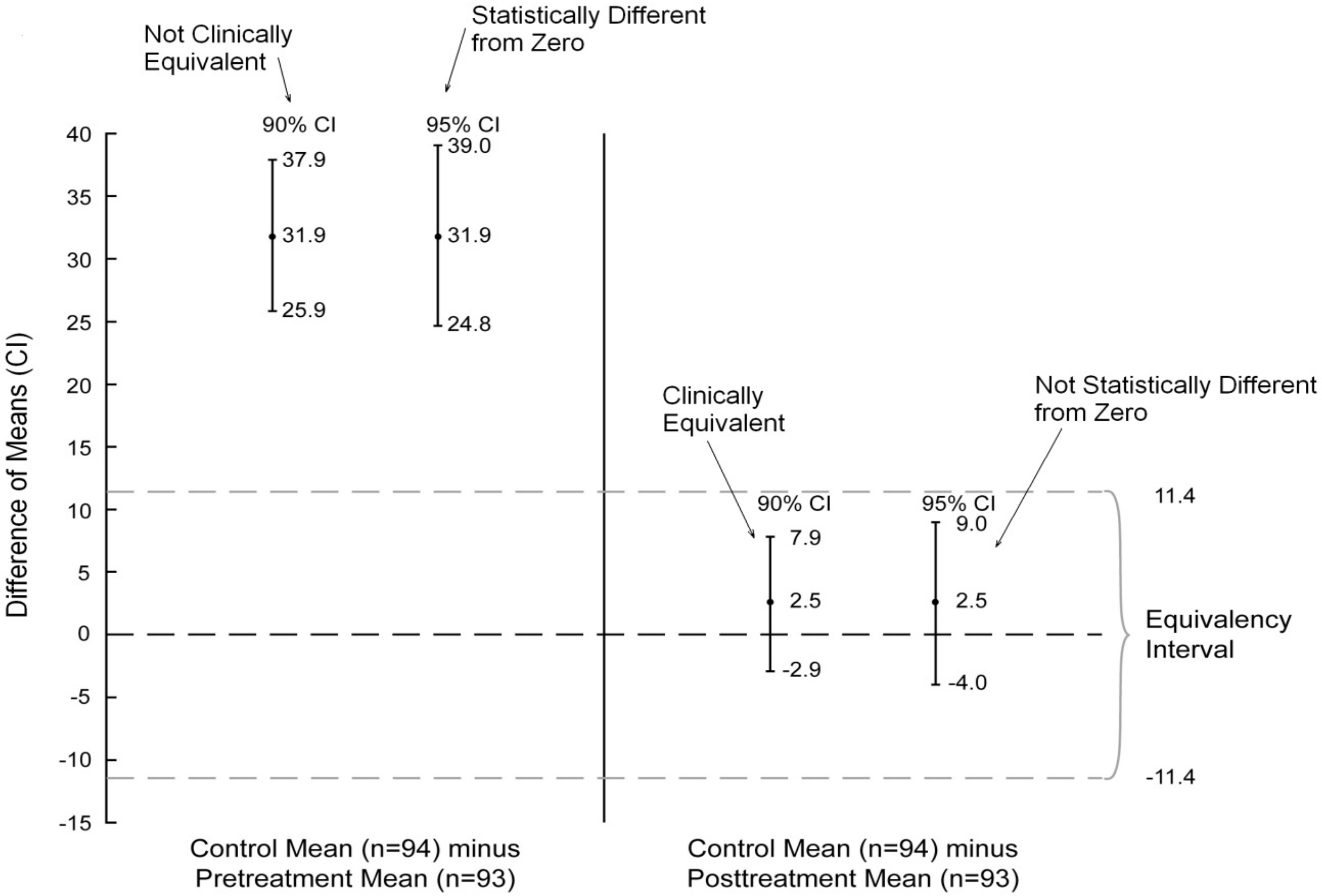


Source: Cappelleri et al. 2009

# Classification of Tests on Statistical Significance and Clinical Equivalence

| | | **Statistical Significance Test** | |
|---|---|---|---|
| | | *Statistically Significant from 0 (95% CI excludes 0)* | *Not Statistically Significant from 0 (95% CI includes 0)* |
| **Clinical Equivalence Test** | *Clinically Equivalent (entire 90% CI within region of equivalence)* | **Cell I**<br>Clinically Equivalent<br>and<br>Statistically Significant | **Cell II**<br>Clinically Equivalent<br>and<br>Not Statistically Significant |
| | *Not Clinically Equivalent (entire 90% CI not within region of equivalence)* | **Cell III**<br>Not Clinically Equivalent<br>and<br>Statistically Significant | **Cell IV**<br>Not Clinically Equivalent<br>and<br>Not Statistically Significant |

# Difference of Control (No ED) Mean versus Pre-treatment and Post-treatment Means on the Self-Esteem Subscale of the Self-Esteem And Relationship Questionnaire



Source: Cappelleri et al. 2006

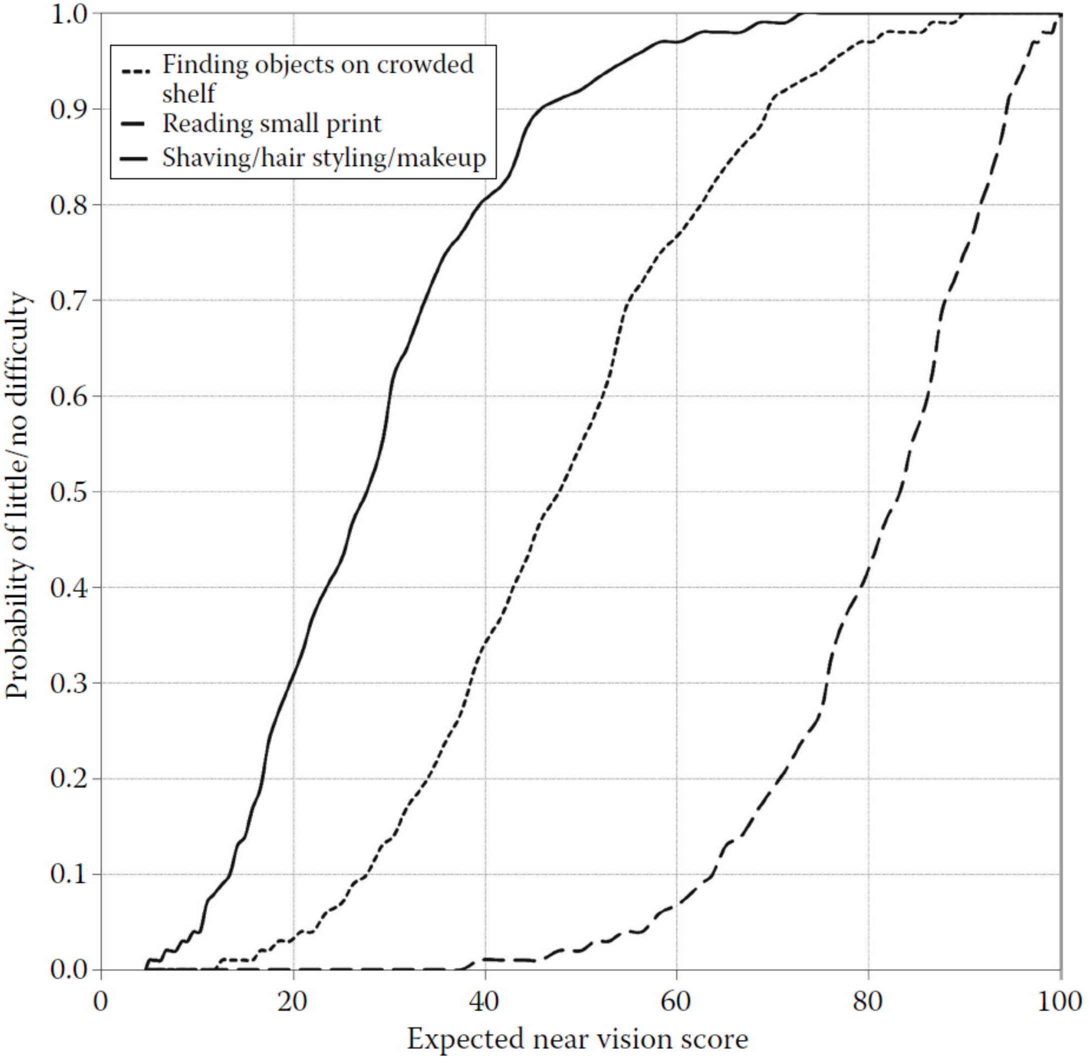
# Content-based Interpretation

- Considered for a multi-item PRO measure

- Uses a representative item, along with its response categories, internal to the measure itself

- Mapping can be obtained using descriptive statistics, item response theory (IRT), ordinal logistic regression, and binary logistic regression

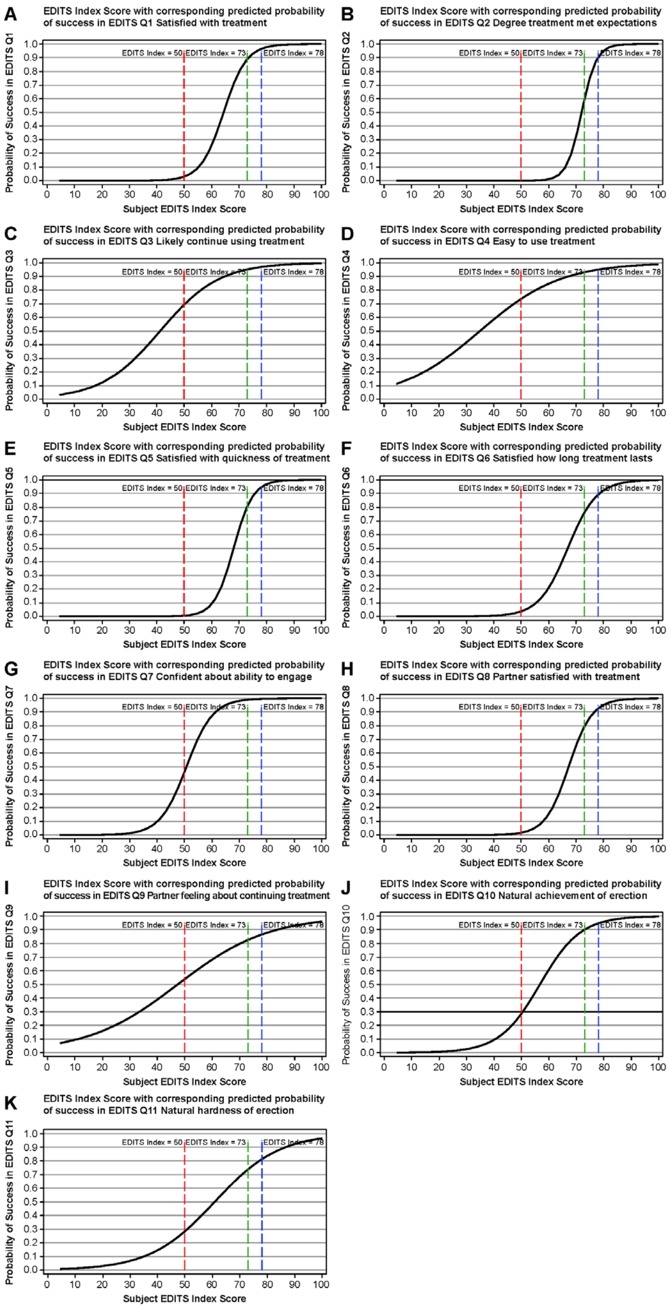
# Probability of Little or No Difficulty: Near-Vision Subscale of the NEI-VFQ



Fitted with a Rasch (IRT) model

Source: Thompson et al. 2007

# Probability of "Success": Erectile Dysfunction Inventory of Treatment Satisfaction (EDITS)



- Item responses: 0 (no or low satisfaction or dissatisfaction) to 4 (high satisfaction)

- Outcome – "Success" is two most satisfied responses to a question (item)
- Predictor – 11-item EDITS total score: 0 to 100, higher treatment satisfaction (mean satisfaction value multiplied by 25)
- Model – Logistic regression

- EDITS index scores of 50 (red line), 73 (green line), and 78 (blue line) represent mean scores observed after 8 weeks of treatment in the placebo, sildenafil 50-mg, and sildenafil 100-mg groups, respectively.

Source: Cappelleri et al. 2018

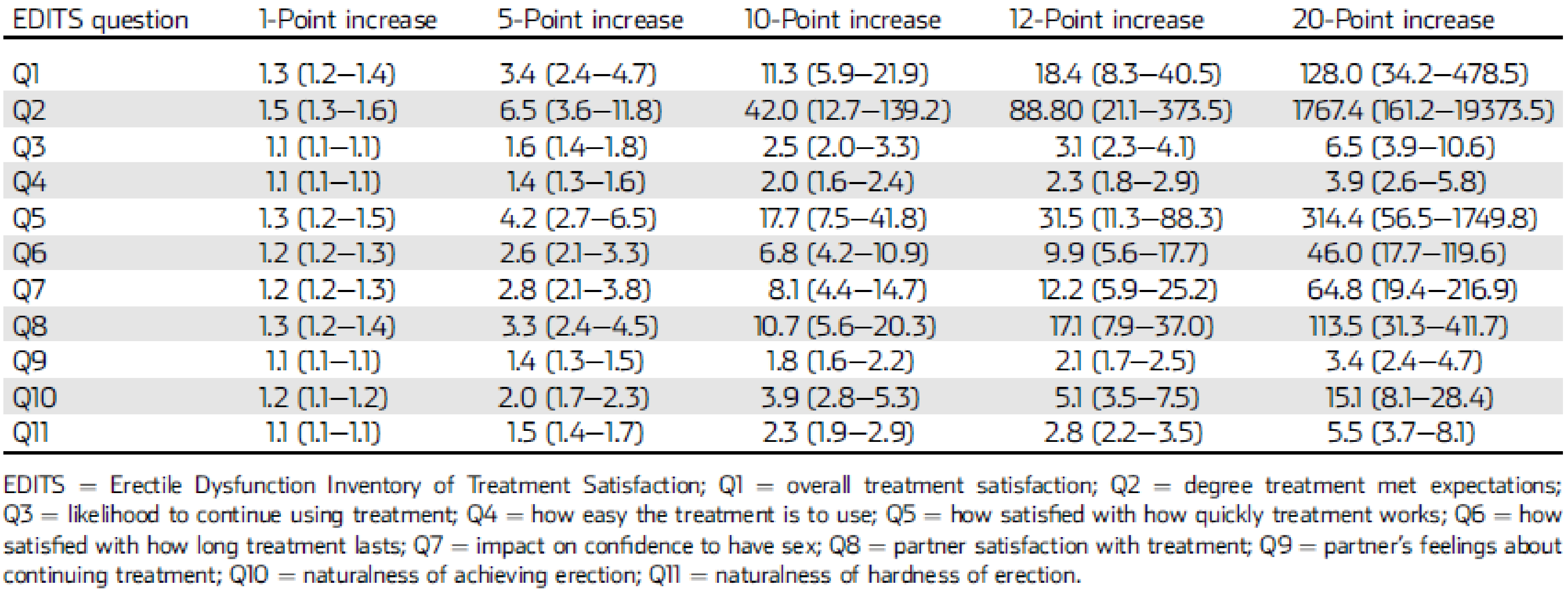
# Profiles of Success: EDITS

- At the end of the double-blinded treatment phase (week 8), the sildenafil 100-mg group had a mean EDITS index score of 78.3, which translates to a probability of success of
  - 96% for satisfaction with treatment(Figure A)
  - 88% for degree of meeting expectations (Figure B)
  - 94% for satisfaction with treatment quickness (Figure E)
  - 88% for satisfaction with how long the treatment lasts (Figure F)

- The sildenafil 50-mg group reported a week 8 mean EDITS index score of 72.7, which is associated with corresponding probabilities of success of 88%, 57%, 80%, and 75%

- The placebo group reported a week 8 mean EDITS index score of 50.1, which is associated with the corresponding probabilities of success of 3%, less than 0.1%, 1%, and 4%

# Estimated Odds Ratio (95% CI) of Success for Each EDITS Item According to Increase in Overall EDITS Index



| EDITS question | 1-Point increase | 5-Point increase | 10-Point increase | 12-Point increase | 20-Point increase |
|---|---|---|---|---|---|
| Q1 | 1.3 (1.2–1.4) | 3.4 (2.4–4.7) | 11.3 (5.9–21.9) | 18.4 (8.3–40.5) | 128.0 (34.2–478.5) |
| Q2 | 1.5 (1.3–1.6) | 6.5 (3.6–11.8) | 42.0 (12.7–139.2) | 88.80 (21.1–373.5) | 1767.4 (161.2–19373.5) |
| Q3 | 1.1 (1.1–1.1) | 1.6 (1.4–1.8) | 2.5 (2.0–3.3) | 3.1 (2.3–4.1) | 6.5 (3.9–10.6) |
| Q4 | 1.1 (1.1–1.1) | 1.4 (1.3–1.6) | 2.0 (1.6–2.4) | 2.3 (1.8–2.9) | 3.9 (2.6–5.8) |
| Q5 | 1.3 (1.2–1.5) | 4.2 (2.7–6.5) | 17.7 (7.5–41.8) | 31.5 (11.3–88.3) | 314.4 (56.5–1749.8) |
| Q6 | 1.2 (1.2–1.3) | 2.6 (2.1–3.3) | 6.8 (4.2–10.9) | 9.9 (5.6–17.7) | 46.0 (17.7–119.6) |
| Q7 | 1.2 (1.2–1.3) | 2.8 (2.1–3.8) | 8.1 (4.4–14.7) | 12.2 (5.9–25.2) | 64.8 (19.4–216.9) |
| Q8 | 1.3 (1.2–1.4) | 3.3 (2.4–4.5) | 10.7 (5.6–20.3) | 17.1 (7.9–37.0) | 113.5 (31.3–411.7) |
| Q9 | 1.1 (1.1–1.1) | 1.4 (1.3–1.5) | 1.8 (1.6–2.2) | 2.1 (1.7–2.5) | 3.4 (2.4–4.7) |
| Q10 | 1.2 (1.1–1.2) | 2.0 (1.7–2.3) | 3.9 (2.8–5.3) | 5.1 (3.5–7.5) | 15.1 (8.1–28.4) |
| Q11 | 1.1 (1.1–1.1) | 1.5 (1.4–1.7) | 2.3 (1.9–2.9) | 2.8 (2.2–3.5) | 5.5 (3.7–8.1) |

EDITS = Erectile Dysfunction Inventory of Treatment Satisfaction; Q1 = overall treatment satisfaction; Q2 = degree treatment met expectations; Q3 = likelihood to continue using treatment; Q4 = how easy the treatment is to use; Q5 = how satisfied with how quickly treatment works; Q6 = how satisfied with how long treatment lasts; Q7 = impact on confidence to have sex; Q8 = partner satisfaction with treatment; Q9 = partner's feelings about continuing treatment; Q10 = naturalness of achieving erection; Q11 = naturalness of hardness of erection.

CI = confidence interval

# Clinically Meaningful Change and Difference

# Clinically Meaningful Within-Patient Change (MWPC)

- Statistical significance does not imply clinical significance

- Regress change in PRO measure as outcome on change in anchor measure as predictor

- Anchor: Patient Global Impression of Change (PGIC, retrospective)

  1=very much improved, 2=much improved, 3=minimally improved, 4=no change, 5=minimally worse, 6=much worse, 7=very much worse

- Anchor: Patient Global Impression–Severity (PGIS, serial)

  1=none, 2=mild, 3=moderate, 4=severe

# Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC): MWPC - Study Design

- WOMAC with 24 items using 0–10 numerical rating scale
  - 0 = no pain/stiffness/ difficulty to 10 = extreme pain/stiffness/difficulty
  - Pain (5 items), Stiffness (2 items), Physical Function (17 items)
  - 48-hour recall period, scores averaged

- Here MWPC illustrated for WOMAC Pain domain

- Anchor: Patient Global Assessment of Osteoarthritis (PGA-OA)
  - "Considering all the ways your OA in your hip/knee affects you, how are you doing today?" 1 = very good to 5 = very poor

- Randomized, double-blind, active-controlled trial (tanezumab)
  - Measurements on WOMAC and PGA-OA
  - Completed at baseline and weeks 2, 4, 8, 16, 24, 32, 40, 48, 56, and 64

Source: Conaghan et al. 2022

# Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC): MWPC – Statistical Analysis

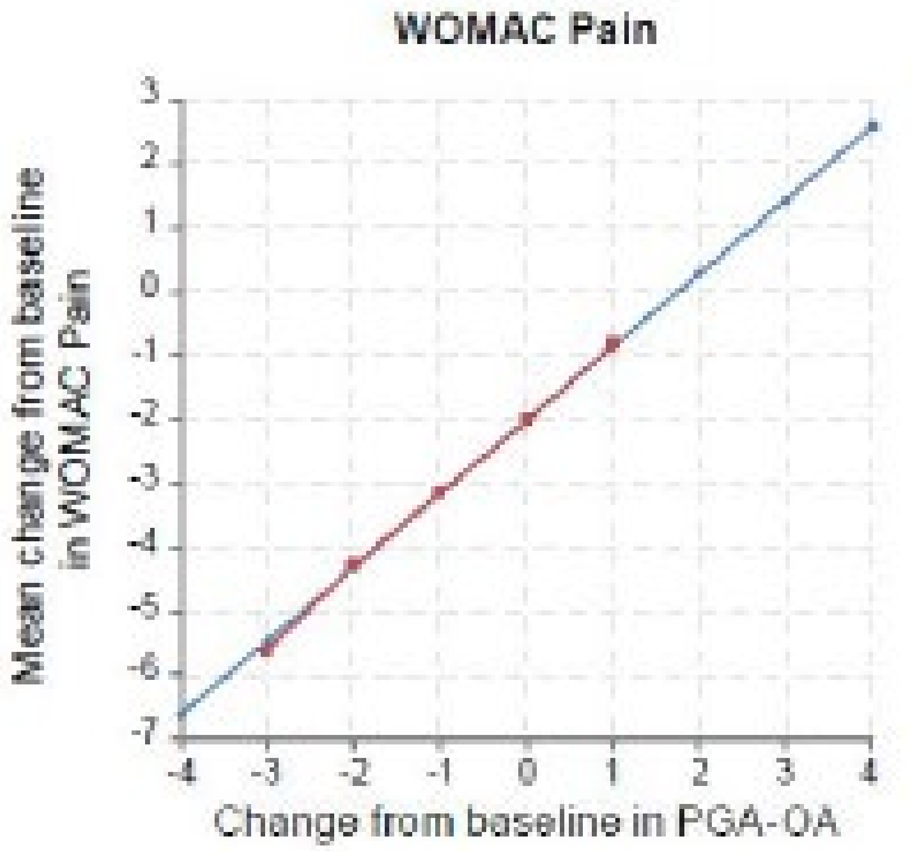
- Outcome: Changes from baseline in WOMAC Pain
- Predictor: Changes from baseline in PGA-OA
- Negative changes reflect improvement

- Repeated measures longitudinal measures model
  - Difference in mean change in outcome scores between adjacent categories of predictor (one-category change, two-category change)
  - Used all available data
  - Combined data across treatment groups

- Predictor taken as continuous and, separately, as categorical

Source: Conaghan et al. 2022

# Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC): MWPC – Results



Blue line reflects predictor as continuous

Red line reflects predictor as categorical

Correlation between changes at the primary time point (week 16) = 0.60

1-category change on WOMAC Pain in PGA-OA = 1.15 [95% confidence interval (CI), 1.12 to 1.18]

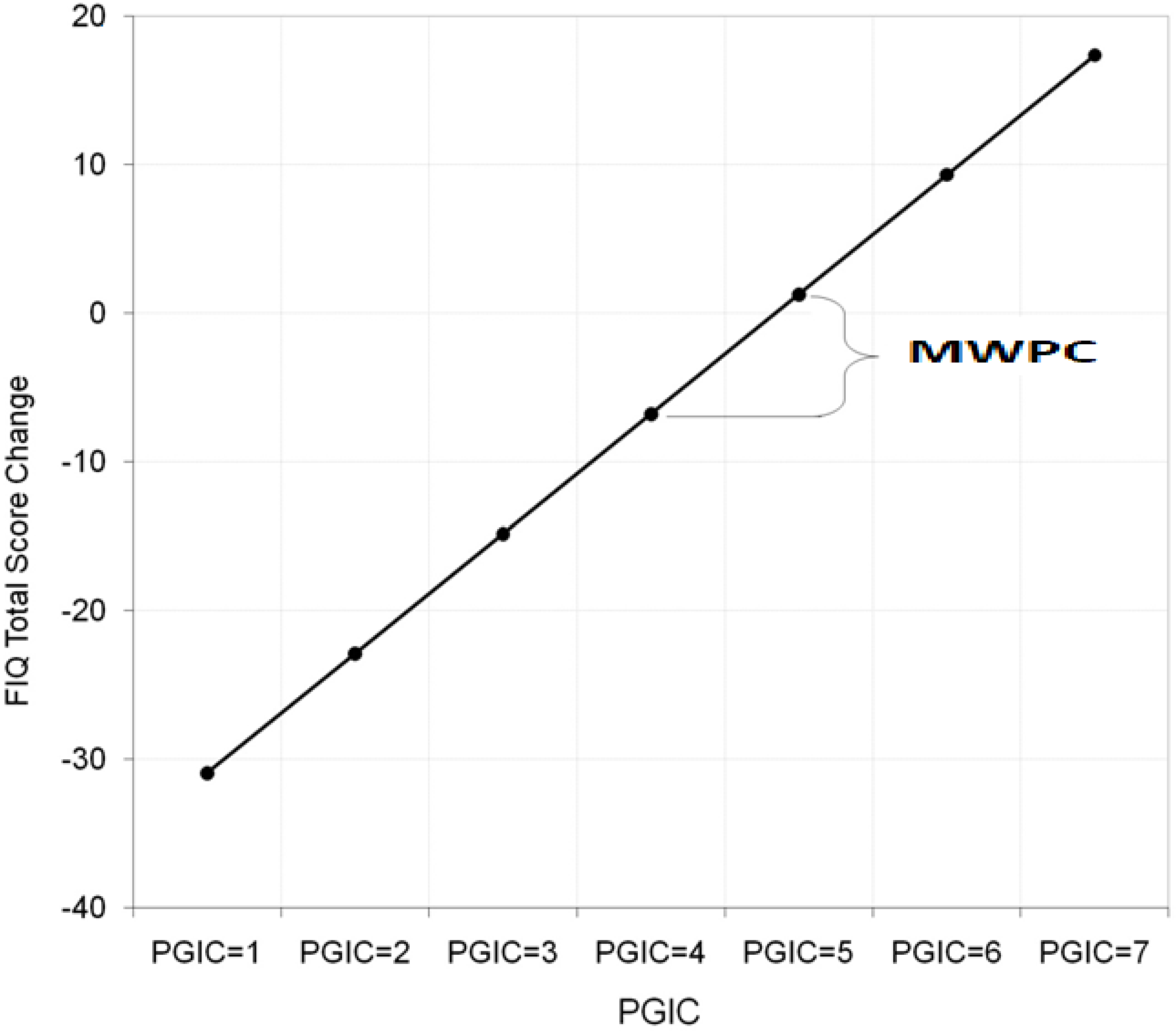
2-category change on WOMAC in PGA-OA = 2.30 (95% CI, 2.25 to 2.35)

Qualitative research can be sought on whether a 1-category change on PGA-OA is meaningfully acceptable change (1.15)

Otherwise, go with the 2-category change on PGA-OA (2.30)

Source: Conaghan et al. 2022

# MWPC on Fibromyalgia Impact Questionnaire



MWPC = 8.1
(95% CI: 7.6 to  8.5)

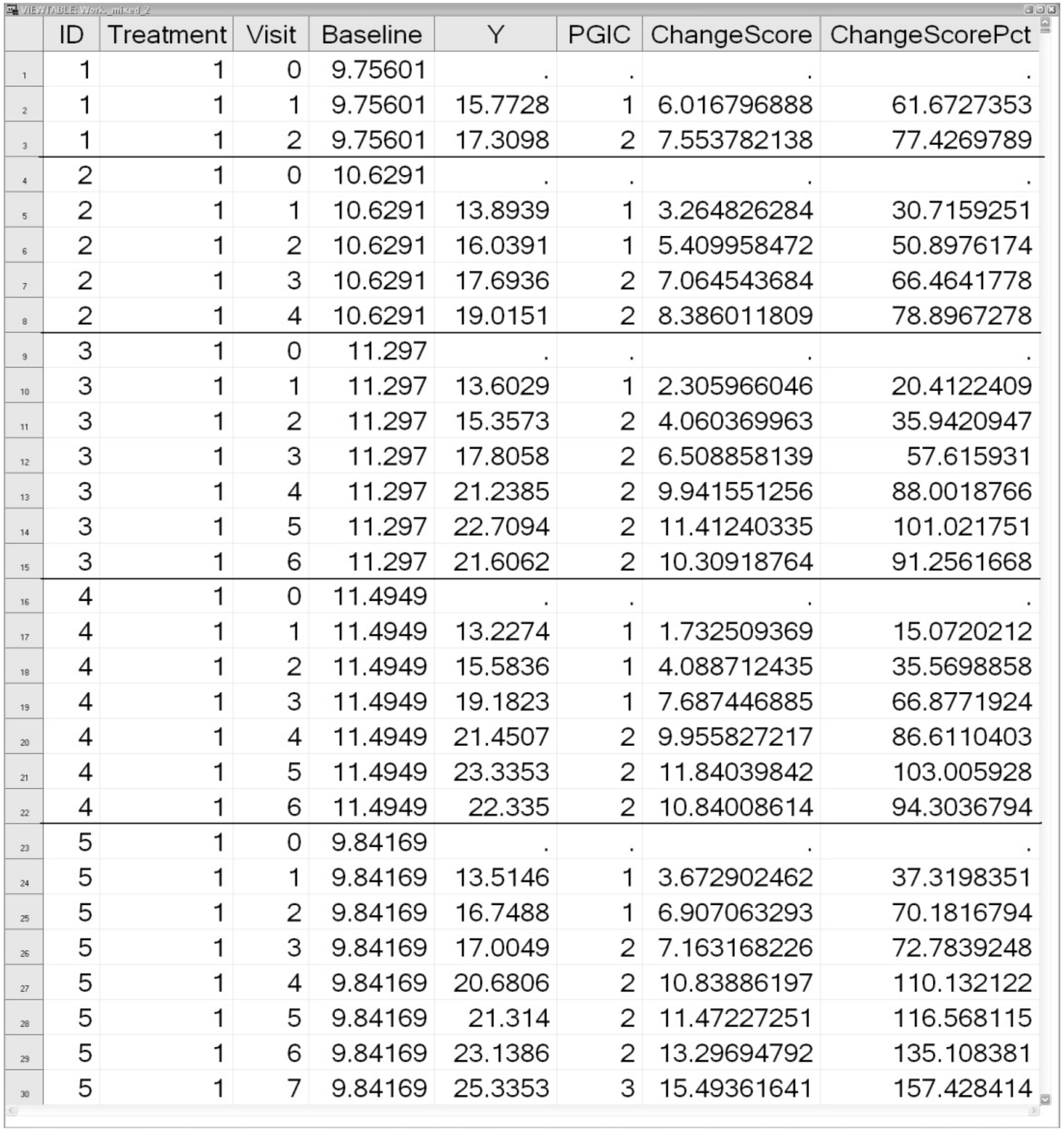
Higher FIQ scores – greater impact of fibromyalgia

Higher FIQ Change (follow-up minus baseline) is less favorable

PGIC is continuous anchor predictor

Source: Bennett et al. 2009

# Dataset Structure in Simulated Example



| | ID | Treatment | Visit | Baseline | Y | PGIC | ChangeScore | ChangeScorePct |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 9.75601 | . | . | . | . |
| 2 | 1 | 1 | 1 | 9.75601 | 15.7728 | 1 | 6.016796888 | 61.6727353 |
| 3 | 1 | 1 | 2 | 9.75601 | 17.3098 | 2 | 7.553782138 | 77.4269789 |
| 4 | 2 | 1 | 0 | 10.6291 | . | . | . | . |
| 5 | 2 | 1 | 1 | 10.6291 | 13.8939 | 1 | 3.264826284 | 30.7159251 |
| 6 | 2 | 1 | 2 | 10.6291 | 16.0391 | 1 | 5.409958472 | 50.8976174 |
| 7 | 2 | 1 | 3 | 10.6291 | 17.6936 | 2 | 7.064543684 | 66.4641778 |
| 8 | 2 | 1 | 4 | 10.6291 | 19.0151 | 2 | 8.386011809 | 78.8967278 |
| 9 | 3 | 1 | 0 | 11.297 | . | . | . | . |
| 10 | 3 | 1 | 1 | 11.297 | 13.6029 | 1 | 2.305966046 | 20.4122409 |
| 11 | 3 | 1 | 2 | 11.297 | 15.3573 | 2 | 4.060369963 | 35.9420947 |
| 12 | 3 | 1 | 3 | 11.297 | 17.8058 | 2 | 6.508858139 | 57.615931 |
| 13 | 3 | 1 | 4 | 11.297 | 21.2385 | 2 | 9.941551256 | 88.0018766 |
| 14 | 3 | 1 | 5 | 11.297 | 22.7094 | 2 | 11.41240335 | 101.021751 |
| 15 | 3 | 1 | 6 | 11.297 | 21.6062 | 2 | 10.30918764 | 91.2561668 |
| 16 | 4 | 1 | 0 | 11.4949 | . | . | . | . |
| 17 | 4 | 1 | 1 | 11.4949 | 13.2274 | 1 | 1.732509369 | 15.0720212 |
| 18 | 4 | 1 | 2 | 11.4949 | 15.5836 | 1 | 4.088712435 | 35.5698858 |
| 19 | 4 | 1 | 3 | 11.4949 | 19.1823 | 1 | 7.687446885 | 66.8771924 |
| 20 | 4 | 1 | 4 | 11.4949 | 21.4507 | 2 | 9.955827217 | 86.6110403 |
| 21 | 4 | 1 | 5 | 11.4949 | 23.3353 | 2 | 11.84039842 | 103.005928 |
| 22 | 4 | 1 | 6 | 11.4949 | 22.335 | 2 | 10.84008614 | 94.3036794 |
| 23 | 5 | 1 | 0 | 9.84169 | . | . | . | . |
| 24 | 5 | 1 | 1 | 9.84169 | 13.5146 | 1 | 3.672902462 | 37.3198351 |
| 25 | 5 | 1 | 2 | 9.84169 | 16.7488 | 1 | 6.907063293 | 70.1816794 |
| 26 | 5 | 1 | 3 | 9.84169 | 17.0049 | 2 | 7.163168226 | 72.7839248 |
| 27 | 5 | 1 | 4 | 9.84169 | 20.6806 | 2 | 10.83886197 | 110.132122 |
| 28 | 5 | 1 | 5 | 9.84169 | 21.314 | 2 | 11.47227251 | 116.568115 |
| 29 | 5 | 1 | 6 | 9.84169 | 23.1386 | 2 | 13.29694792 | 135.108381 |
| 30 | 5 | 1 | 7 | 9.84169 | 25.3353 | 3 | 15.49361641 | 157.428414 |

Source: Cappelleri, Zou, Bushmakin et al. 2013

# Proc Mixed Longitudinal Modeling: MWPC Estimation (Continuous Anchor)

```
Data _mixed_3;
 Set  _mixed_2;
 Where Visit In (1 2 3 4 5 6 7);
 Run;
 Proc Mixed data=_mixed_3;
 Class ID Visit ;
 Model ChangeScore = PGIC / ddfm=kr s;
 Repeated Visit / Type=AR(1) /*UN*/  Subject=ID;
 Estimate "CID(One Category Change) = "  PGIC 1 /cl;
 Estimate " PGIC=1 " Intercept 1 PGIC 1 /cl;
 Estimate " PGIC=2 " Intercept 1 PGIC 2 /cl;
 Estimate " PGIC=3 " Intercept 1 PGIC 3 /cl;
 Estimate " PGIC=4 " Intercept 1 PGIC 4 /cl;
 Estimate " PGIC=5 " Intercept 1 PGIC 5 /cl;
 Estimate " PGIC=6 " Intercept 1 PGIC 6 /cl;
 Estimate " PGIC=7 " Intercept 1 PGIC 7 /cl;
 Run;
```

# Estimated Mean Changes and MWPC

| | | Standard | | | |
|---|---|---|---|---|---|
| Label | Estimate | Error | Pr > \|t\| | Lower | Upper |
| MWPC (one-category change) | 3.9665 | 0.0724 | <.0001 | 3.8242 | 4.1088 |
| PGIC=1 | 4.9722 | 0.1417 | <.0001 | 4.6939 | 5.2504 |
| PGIC=2 | 8.9387 | 0.0987 | <.0001 | 8.7445 | 9.1328 |
| PGIC=3 | 12.9052 | 0.0997 | <.0001 | 12.7090 | 13.1013 |
| PGIC=4 | 16.8717 | 0.1437 | <.0001 | 16.5893 | 17.1540 |
| PGIC=5 | 20.8381 | 0.2046 | <.0001 | 20.4363 | 21.2400 |
| PGIC=6 | 24.8046 | 0.2712 | <.0001 | 24.2719 | 25.3374 |
| PGIC=7 | 28.7711 | 0.3403 | <.0001 | 28.1028 | 29.4394 |

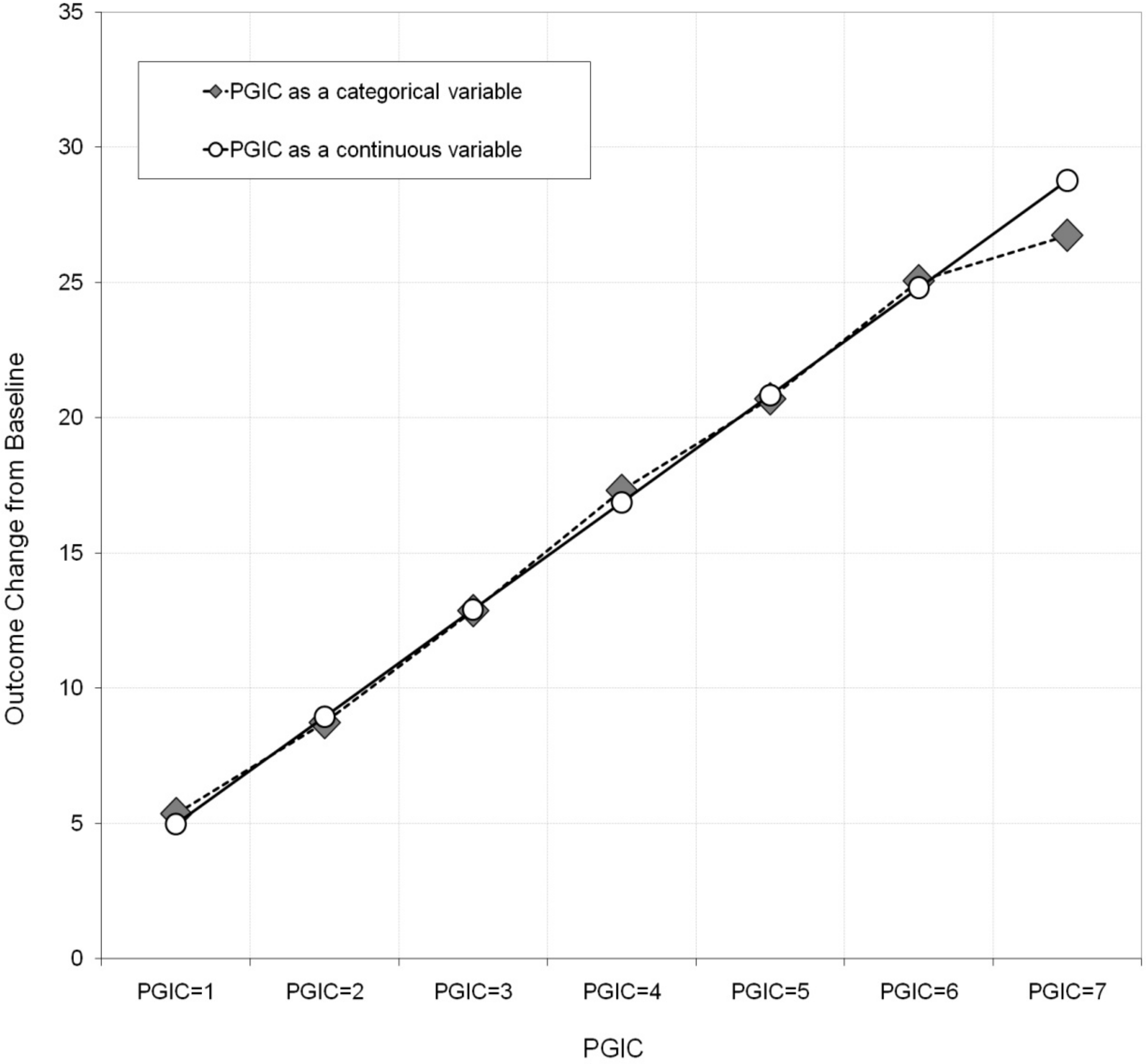
# Proc Mixed Longitudinal Modeling: MWPC Estimation (Categorical Anchor) – Sensitivity Analysis

```
Proc Mixed data=_mixed_3;
 Class ID Visit PGIC ;
 Model ChangeScore = PGIC / ddfm=kr s;
 Repeated Visit / Type=AR(1) Subject=ID;
 Lsmeans PGIC /cl;
 Run;
```

# Estimated Mean Changes and MWPC: Sensitivity Analysis (Same Simulated Data)

| | | | *Standard* | | | |
|---|---|---|---|---|---|---|
| *Effect* | *PGIC* | *Estimate* | *Error* | *Pr > \|t\|* | *Lower* | *Upper* |
| PGIC | 1 | 5.3561 | 0.1939 | <.0001 | 4.9757 | 5.7365 |
| PGIC | 2 | 8.7256 | 0.1233 | <.0001 | 8.4836 | 8.9677 |
| PGIC | 3 | 12.8642 | 0.1564 | <.0001 | 12.5572 | 13.1713 |
| PGIC | 4 | 17.3115 | 0.2384 | <.0001 | 16.8438 | 17.7792 |
| PGIC | 5 | 20.6988 | 0.3406 | <.0001 | 20.0305 | 21.3672 |
| PGIC | 6 | 25.0653 | 0.5040 | <.0001 | 24.0764 | 26.0542 |
| PGIC | 7 | 26.7490 | 2.3192 | <.0001 | 22.1987 | 31.2993 |

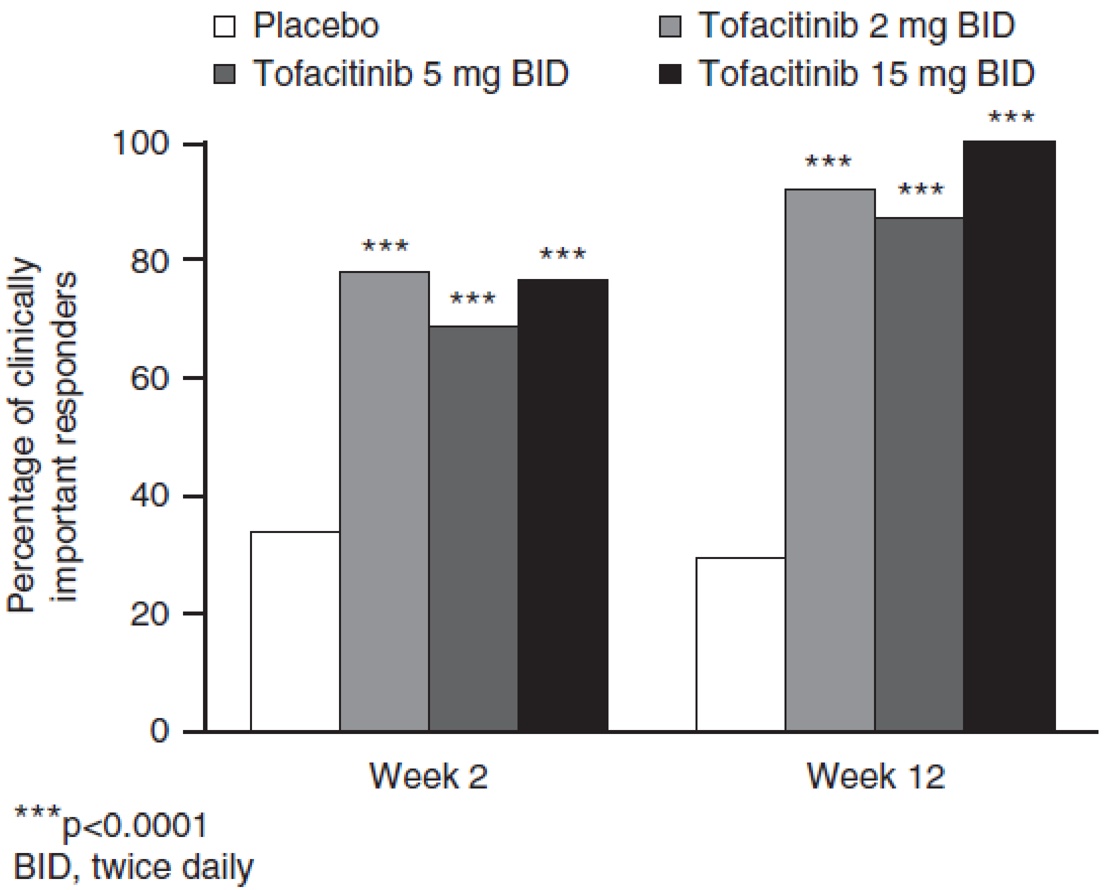
# Mean Change in PRO Measure as Function of PGIC



# Frequencies on PGIC

| | | *Cumulative* | *Cumulative* | |
|---|---|---|---|---|
| *PGIC* | *Frequency* | *Percent* | *Frequency* | *Percent* |
| 1 | 179 | 14.98 | 179 | 14.98 |
| 2 | 518 | 43.35 | 697 | 58.33 |
| 3 | 300 | 25.10 | 997 | 83.43 |
| 4 | 114 | 9.54 | 1111 | 92.97 |
| 5 | 57 | 4.77 | 1168 | 97.74 |
| 6 | 26 | 2.18 | 1194 | 99.92 |
| 7 | 1 | 0.08 | 1195 | 100.00 |

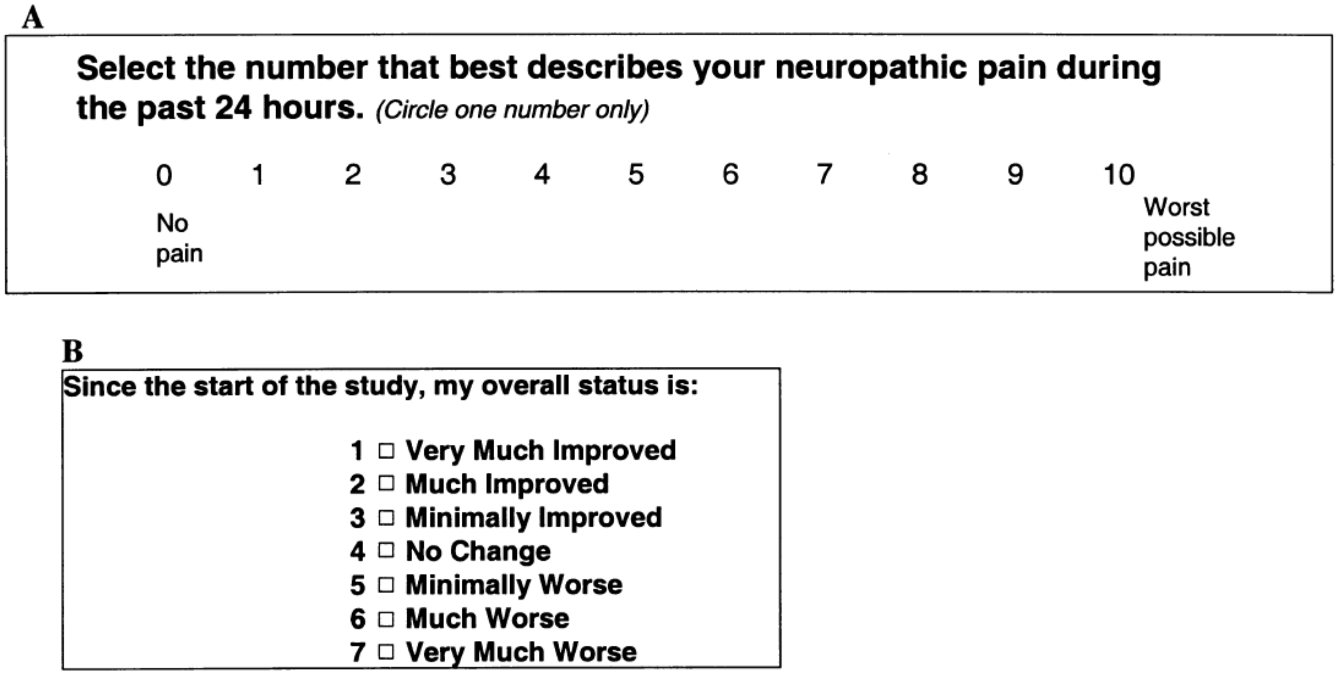
# MWPC on Itch Severity Score (ISS): Responder Analysis



- Patients with moderate-to-sever plaque psorasis
- ISS is 0-10 numeric rating scale (lower values less itching)
- Outcome: % change in ISS
- Predictor: Subject Global Impression of Change
- Repeated measures model
- MWPC on % change in ISS = 30% (95% CI, 23.3-36.4%)

Source: Mamolo et al. 2015

# MWPC on Pain Intensity Numerical Rating Scale (PI-NRS)



A

**Select the number that best describes your neuropathic pain during the past 24 hours.** *(Circle one number only)*

0 1 2 3 4 5 6 7 8 9 10

No pain — Worst possible pain

B

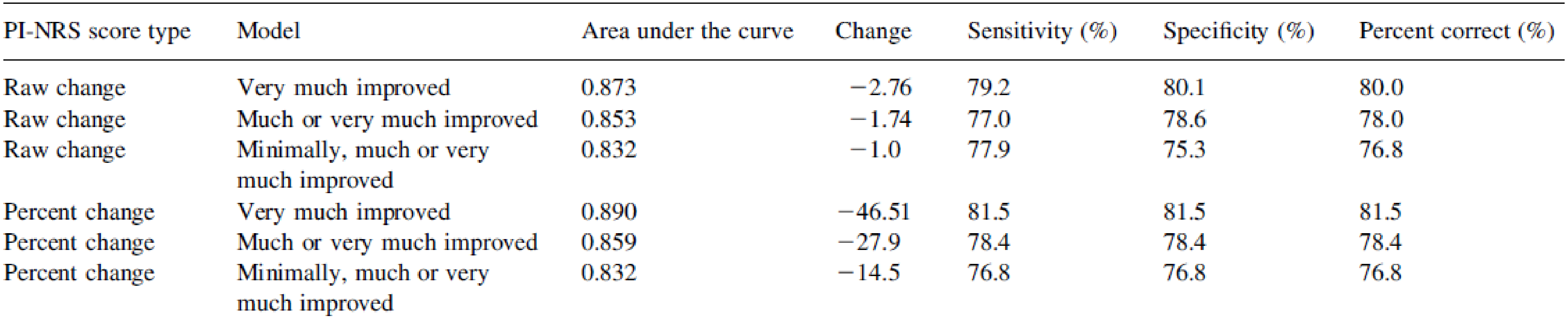
**Since the start of the study, my overall status is:**

**1 □ Very Much Improved**
**2 □ Much Improved**
**3 □ Minimally Improved**
**4 □ No Change**
**5 □ Minimally Worse**
**6 □ Much Worse**
**7 □ Very Much Worse**

- 2,724 subjects from 10 placebo-controlled trials of pregabalin
- Different chronic diseases (e.g., diabetic neuropathy, fibromyalgia)

Source: Farrar et al. 2001

# MWPC on PI-NRS: ROC Results

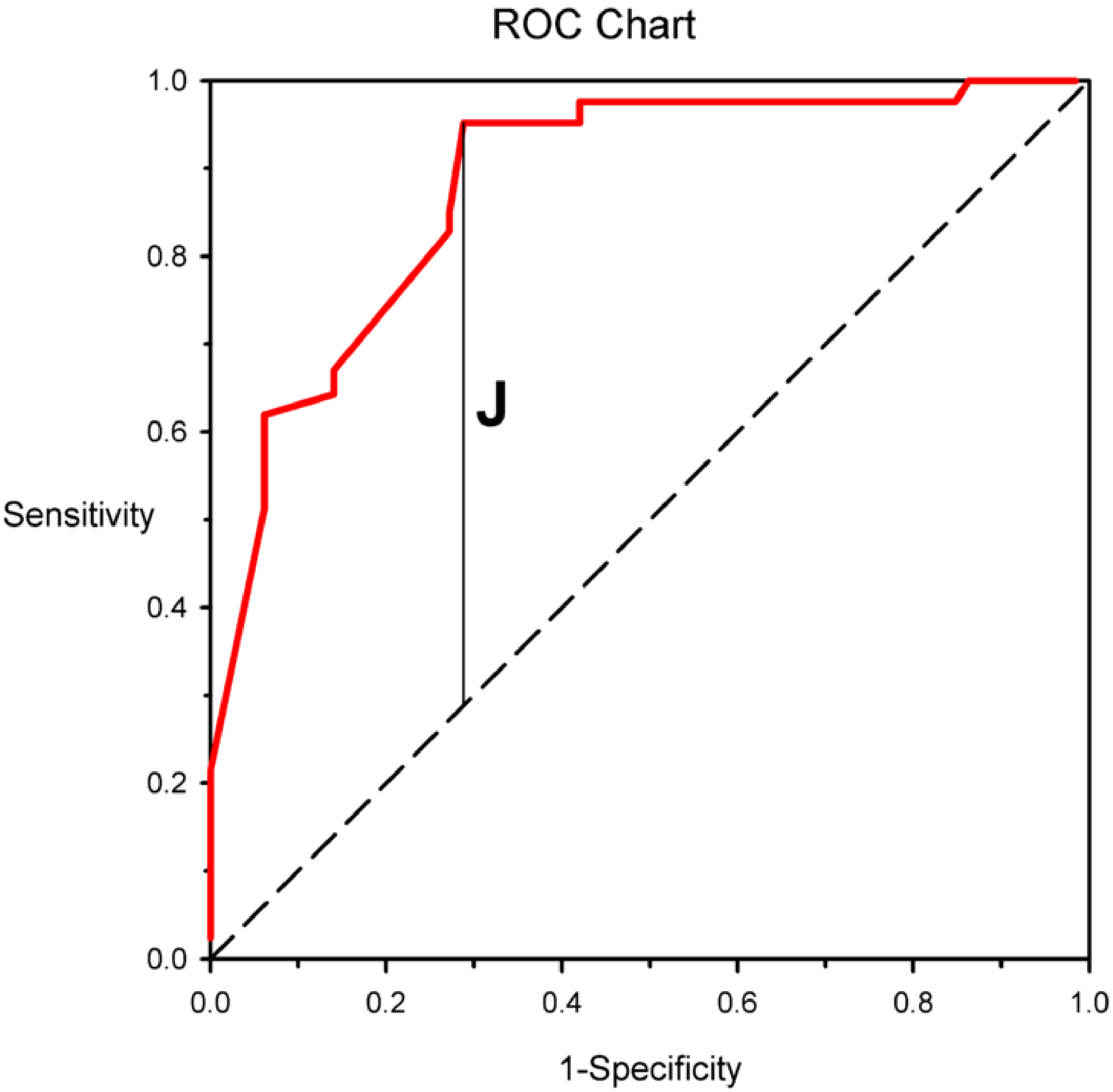


| PI-NRS score type | Model | Area under the curve | Change | Sensitivity (%) | Specificity (%) | Percent correct (%) |
|---|---|---|---|---|---|---|
| Raw change | Very much improved | 0.873 | −2.76 | 79.2 | 80.1 | 80.0 |
| Raw change | Much or very much improved | 0.853 | −1.74 | 77.0 | 78.6 | 78.0 |
| Raw change | Minimally, much or very much improved | 0.832 | −1.0 | 77.9 | 75.3 | 76.8 |
| Percent change | Very much improved | 0.890 | −46.51 | 81.5 | 81.5 | 81.5 |
| Percent change | Much or very much improved | 0.859 | −27.9 | 78.4 | 78.4 | 78.4 |
| Percent change | Minimally, much or very much improved | 0.832 | −14.5 | 76.8 | 76.8 | 76.8 |

- Receiver Operating Characteristic (ROC) curve via logistic regression
- Primary Outcome: PGIC much improved or better
- Predictor: PI-NRS changes (raw or percent)
- MWPC = cutoff on PI-NRS where sensitivity & specificity are closest

- MWPC (improvement) on PI-NRS (raw change) at least 2 points
- MWPC (improvement) on PI-NRS (percent change) at least 30%

# General Note on ROC Curve for MWPC



Selection of optimal cutoff:

J = Youden's Index = maximum of (sensitivity + specificity – 1)

# Clinically Meaningful Between-Group Difference (MBGD)

- Also referred to as clinically important difference (CID)
- Similar to MWPC approach
- But now use absolute scores instead of change scores
- Regress PRO measure on anchor measure
- Anchor: Patient Global Impression of Change (PGIC)

  1=very much improved, 2=much improved, 3=minimally improved,
  4= no change, 5=minimally worse, 6=much worse, 7=very much worse

- Anchor: Patient Global Impression–Severity (PGIS)

  1=none, 2=mild, 3=moderate, 4=severe

# Itch Severity Score (ISS): Methodology - MBGD

- Trial of patients with moderate-to-severe plaque psoriasis

- Repeat measures longitudinal model using all available data
  - Assessments at baseline and week 4, 8, 12, and 16

- Outcome: ISS score

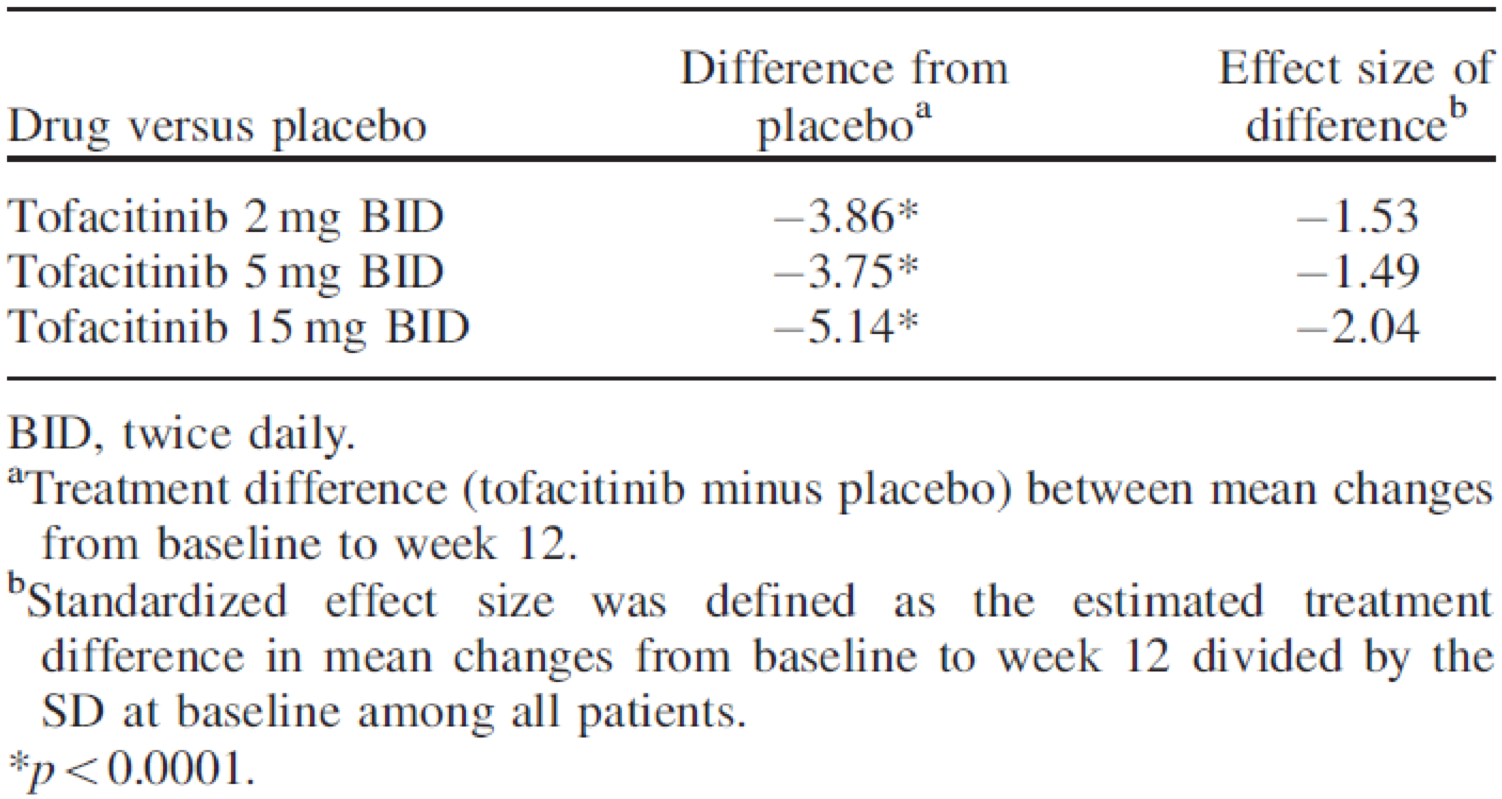
- Anchor: Patient Global Assessment
  - Evaluated the overall extent of cutaneous disease at a given time
  - "Clear," "Almost Clear," "Mild," "Moderate," "Severe"

- MBGD (or CID) on ISS was defined as the difference between one-category change on the Patient Global Assessment

Source: Mamolo et al. 2015

# ISS: Results - MBGD

- ISS MBGD = 1.64 (95% CI, 1.50-1.78)



| Drug versus placebo | Difference from placebo[a] | Effect size of difference[b] |
|---|---|---|
| Tofacitinib 2 mg BID | −3.86* | −1.53 |
| Tofacitinib 5 mg BID | −3.75* | −1.49 |
| Tofacitinib 15 mg BID | −5.14* | −2.04 |

BID, twice daily.

[a] Treatment difference (tofacitinib minus placebo) between mean changes from baseline to week 12.

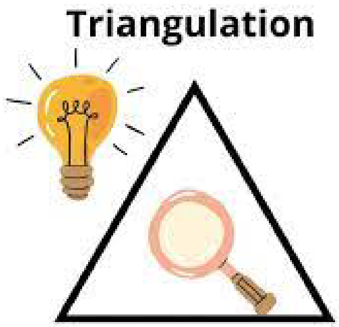
[b] Standardized effect size was defined as the estimated treatment difference in mean changes from baseline to week 12 divided by the SD at baseline among all patients.

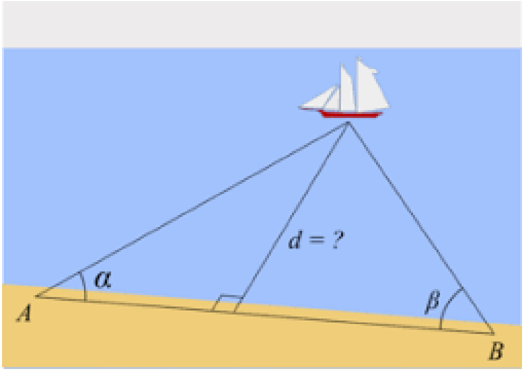
$^{*}p < 0.0001$.

Source: Mamolo et al. 2015

# Triangulation in Clinically Meaningful Change and Difference

- Several anchor-based, distribution-based, and qualitative methods have been used to determine MWPC or MBGD
- Given uncertainty in them, a typical strategy is to triangulate
- Examining multiple values from different approaches and converging on a single value or small range of values



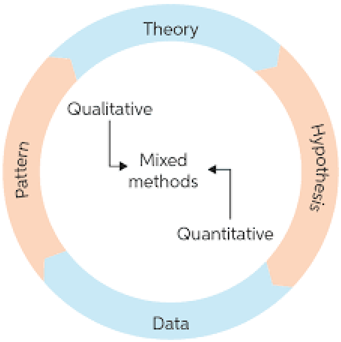


# Methods to Estimate Meaningful Change or Difference

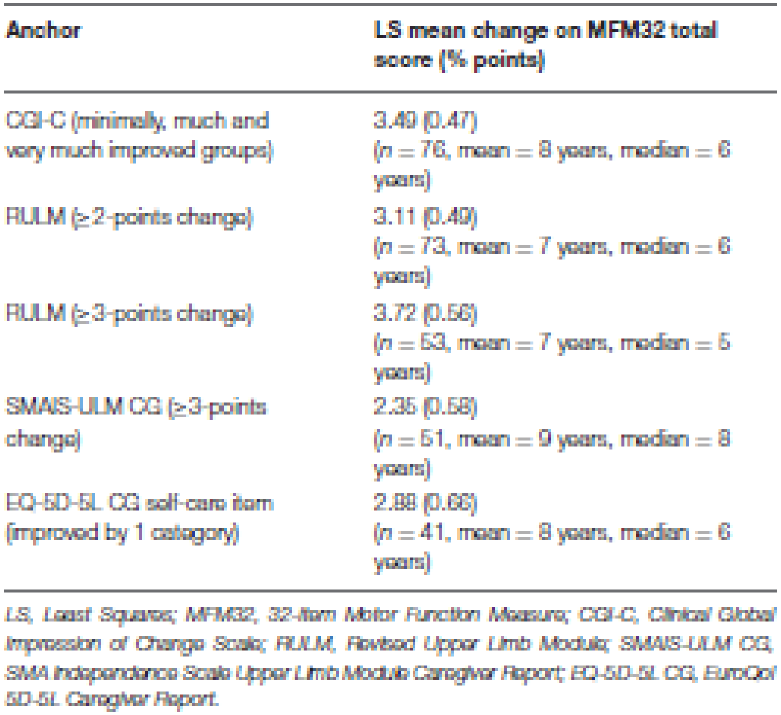
- Anchor-based – external indicator that classifies patients to score changes considered meaningful and maps this relation to score changes in the target PRO measure

- Distribution-based – distribution, variation, and reliability of values in target PRO in the sample

- Qualitative-based – semi-structured interviews (clinical trial exit interviews, focus groups, vignettes, Delphi panel)



# Triangulating MWPC in a Single Study – 32-Item Motor Function Measure (MFM32) in Spinal Muscular Trophy: Anchor-Based Findings



| Anchor | LS mean change on MFM32 total score (% points) |
|---|---|
| CGI-C (minimally, much and very much improved groups) | 3.49 (0.47)<br>(*n* = 76, mean = 8 years, median = 6 years) |
| RULM (≥2-points change) | 3.11 (0.49)<br>(*n* = 73, mean = 7 years, median = 6 years) |
| RULM (≥3-points change) | 3.72 (0.56)<br>(*n* = 53, mean = 7 years, median = 5 years) |
| SMAIS-ULM CG (≥3-points change) | 2.35 (0.58)<br>(*n* = 51, mean = 9 years, median = 8 years) |
| EQ-5D-5L CG self-care item (improved by 1 category) | 2.88 (0.66)<br>(*n* = 41, mean = 8 years, median = 6 years) |

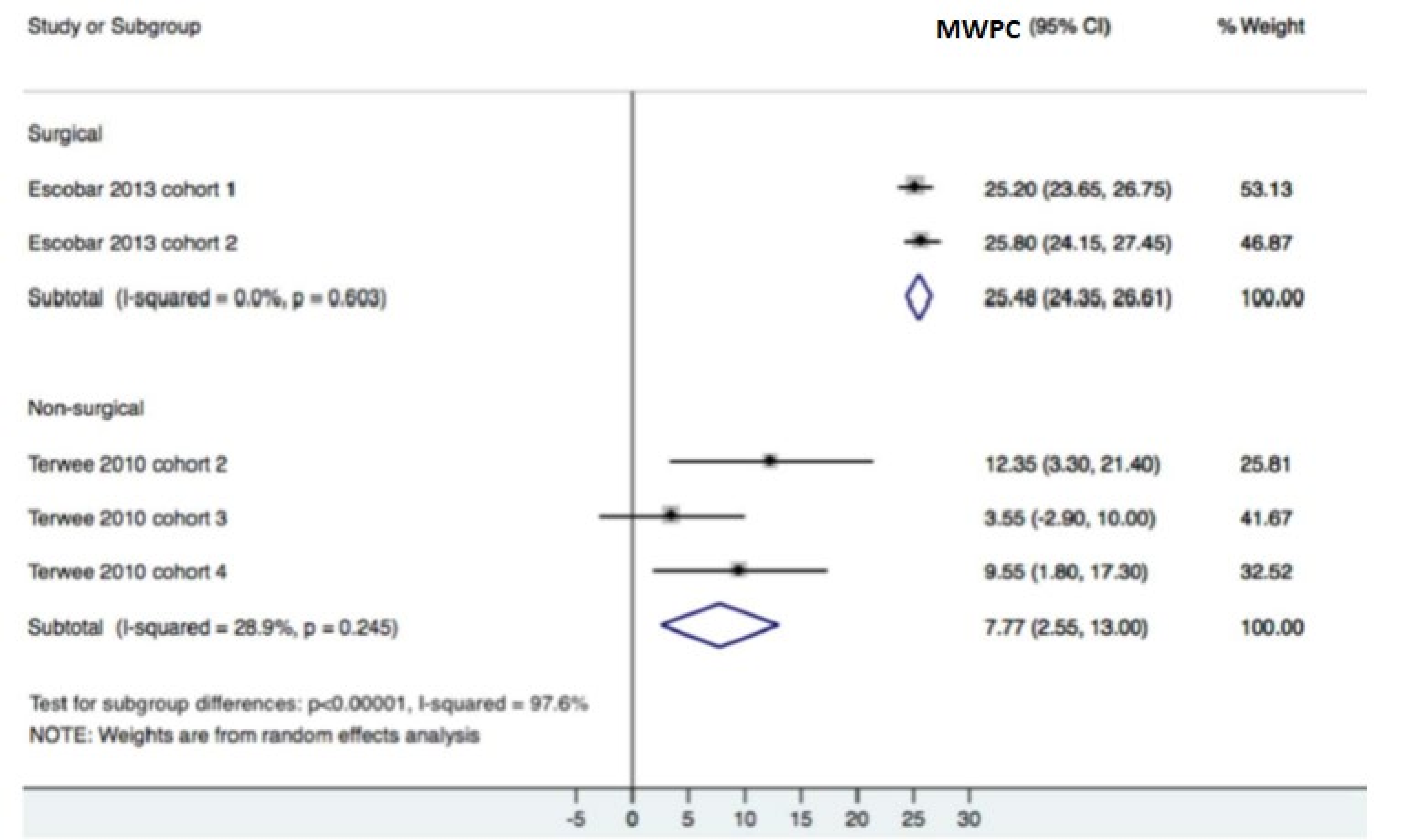
*LS, Least Squares; MFM32, 32-item Motor Function Measure; CGI-C, Clinical Global Impression of Change Scale; RULM, Revised Upper Limb Module; SMAIS-ULM CG, SMA Independence Scale Upper Limb Module Caregiver Report; EQ-5D-5L CG, EuroQol 5D-5L Caregiver Report.*

| Spearman's Rank Correlation | Fisher's Z Transformation |
|---|---|
| 0.48 | 0.52 |
| 0.50 | 0.55 |
| 0.50 | 0.55 |
| 0.22 | 0.22 |
| 0.20 | 0.20 |

Sources: Triggs and Griffiths 2021, Duong et al. 2022

$$\mathbf{MWPC_{weighted} = \{[3.49(0.52) + 3.11(0.55) + 2.35(0.22) + 2.88(0.20)] / (0.52 + 0.55 + 0.22 + 0.20)\} = 3.10}$$

# Triangulating MWPC Across Studies – WOMAC Pain Domain in Degenerative Knee Disease



| Study or Subgroup | MWPC (95% CI) | % Weight |
|---|---|---|
| Surgical | | |
| Escobar 2013 cohort 1 | 25.20 (23.65, 26.75) | 53.13 |
| Escobar 2013 cohort 2 | 25.80 (24.15, 27.45) | 46.87 |
| Subtotal (I-squared = 0.0%, p = 0.603) | 25.48 (24.35, 26.61) | 100.00 |
| Non-surgical | | |
| Terwee 2010 cohort 2 | 12.35 (3.30, 21.40) | 25.81 |
| Terwee 2010 cohort 3 | 3.55 (-2.90, 10.00) | 41.67 |
| Terwee 2010 cohort 4 | 9.55 (1.80, 17.30) | 32.52 |
| Subtotal (I-squared = 28.9%, p = 0.245) | 7.77 (2.55, 13.00) | 100.00 |

Test for subgroup differences: p<0.00001, I-squared = 97.6%

NOTE: Weights are from random effects analysis

WOMAC = Western Ontario and McMaster University Osteoarthritis Index, CI = Confidence Interval. Source: Devji et al. 2017

**MWPC$_{weighted}$ using inverse variance weights from a random effects model = 25 (95% CI 24 to 27) for Surgical Intervention; 8 (95% CI 3 to 13) for Nonsurgical**

# Distribution-Based Methods

# Distribution-Based Methods

- Based on empirical distribution and characteristics of the data
- Adjunct to, not substitute for, anchor-based methods
- Informs on meaning of difference or change in PRO measure but not necessarily whether change is *clinically* significant to patients
- Different types
  - Standardized Effect Size for group difference
  - Change Indexes for individual change
  - Probability of Relative Benefit
  - Cumulative Distribution Function

# Standardized Effect Size

- (Standardized) Effect Size = magnitude of effect relative to variability
  - Measured in standard deviation (SD) units
  - 0.2, ‘small’; 0.5, ‘medium’; 0.8, ‘large’
- Within group: before vs. after therapy
- Between groups: treatments A vs. B

# (Standardized) Effect Size

- Within group
  - Effect = average change score on PRO
  - Variability = baseline standard deviation (SD)
  - Or variability = SD of individual changes

- Between groups
  - Effect = average difference between groups at follow-up
  - Or effect = average difference between groups from baseline to follow-up

  - Variability = pooled between-group SD at baseline
  - Or variability = pooled between-group SD at follow-up
  - Or variability = pooled SD of individual changes

# Example: Effect Size for Within Group

- Effect size for all subjects in single intervention study

- $$\text{Effect size} = \frac{\text{Mean difference score}}{\text{SD at baseline}}$$

# Example: Effect Size for Within Group

| SEAR<br>Component | Baseline<br>Mean ± SD | End<br>Mean ± SD | Difference | Effect<br>Size |
|---|---|---|---|---|
| Sexual Relationship | 42 ± 22 | 78 ± 21 | 36 ± 23 | 1.6 |
| Confidence | 55 ± 26 | 81 ± 21 | 26 ± 26 | 1.0 |
| Self-esteem | 52 ± 27 | 81 ± 22 | 29 ± 28 | 1.1 |
| Overall Relationship | 62 ± 30 | 80 ± 24 | 18 ± 32 | 0.6 |
| Overall (Total) | 48 ± 22 | 79 ± 20 | 31 ± 22 | 1.4 |

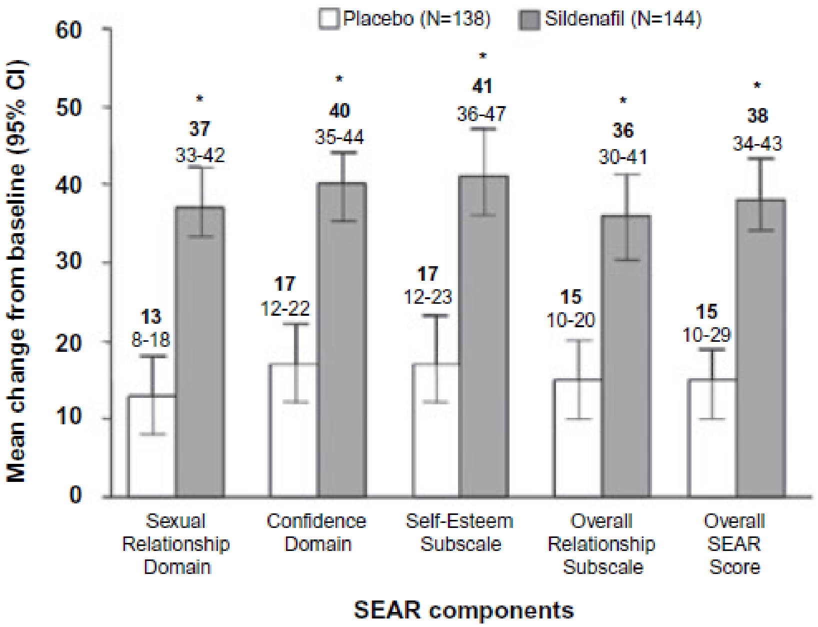
Source: Althof et al. 2003

# Example: Effect Size for Between Groups

- Effect size between two treatment groups
- Mean changes from baseline
- Effect size =

$$\frac{\text{Difference in mean changes between treatment groups}}{\text{Pooled SD at baseline}}$$

# Example: Effect Size for Between Groups



Mean change scores and 95% CI

| SEAR Component | Difference in Mean Change | Baseline SD | Effect Size |
|---|---|---|---|
| Sexual Relationship | 37-13=24 | 18.5 | 1.3 |
| Confidence | 40-17=23 | 18.0 | 1.3 |
| Self-Esteem | 41-17=24 | 18.0 | 1.3 |
| Overall Relationship | 36-15=21 | 26.0 | 0.8 |
| Overall (Total) | 38-15=23 | 16.5 | 1.4 |

Source: Althof et al. 2006

# Effect Sizes on EQ-5D-5L Index Scores (U.S. Weights) for U.S. Adult Outpatients with COVID-19

| | BNT162b2 Cohort | | | | Unvaccinated Cohort | | | | Between-Cohort Difference | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Score | Change from Baseline | | | Score | Change from Baseline | | | | | |
| Time | LSE (95% CI) | LSE (95% CI) | P | ES$_w$ | LSE (95% CI) | LSE (95% CI) | P | ES$_w$ | LSE (95% CI) | P | ES |
| | | | | | | | | | | | |
| Day 3 | 0.84 (0.81, 0.88) | -0.08 (-0.12, -0.04) | <.01 | -0.49 | 0.77 (0.74, 0.81) | -0.15 (-0.18, -0.11) | <.01 | -0.64 | 0.07 (0.03, 0.11) | <.01 | 0.36 |
| Week 4 | 0.90 (0.87, 0.94) | -0.02 (-0.05, 0.02) | 0.37 | -0.13 | 0.86 (0.83, 0.89) | -0.06 (-0.09, -0.03) | <.01 | -0.38 | 0.04 (0.01, 0.08) | <.01 | 0.32 |

Source: Di Fusco et al. 2022. Abbreviations: LSE = Least-Square Mean Estimate; CI = Confidence Interval, P = P-value

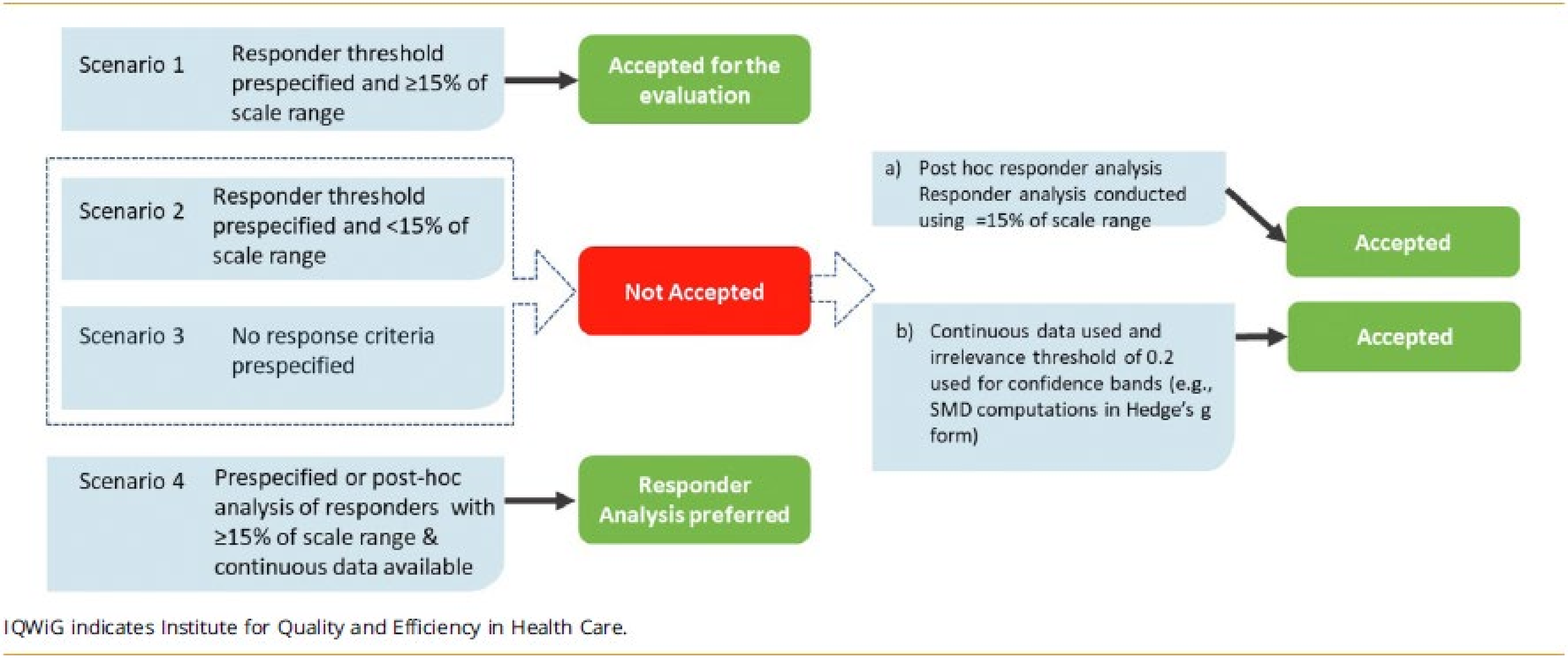
Multivariate models include variables for time, vaccination status and interaction of time by vaccination status, as well as covariates of participant pre-COVID-19 symptom onset score, sociodemographic characteristics (age, sex, regions, social vulnerability, race/ethnicity, high risk occupations), previously tested positive for COVID-19, severity of acute illness (number of symptoms reported on index date), and immunocompromised status.

**ES$_w$, within-cohort effect size, was calculated as the least square estimate of mean change from divided by the observed standard deviation of change scores from baseline to follow-up.**

**ES$_b$, between-cohort effect size, was calculated as the difference in least square estimates of mean changes from baseline between cohorts, divided by the observed pooled standard deviation of change scores**

# IQWiG's 15% Threshold for Meaningful Change: Percent of Range

At least 15% of scale range on a PRO measure, universally applied



Scenario 1 — Responder threshold prespecified and ≥15% of scale range → Accepted for the evaluation

Scenario 2 — Responder threshold prespecified and <15% of scale range

Scenario 3 — No response criteria prespecified

→ Not Accepted →

a) Post hoc responder analysis Responder analysis conducted using =15% of scale range → Accepted

b) Continuous data used and irrelevance threshold of 0.2 used for confidence bands (e.g., SMD computations in Hedge's g form) → Accepted

Scenario 4 — Prespecified or post-hoc analysis of responders with ≥15% of scale range & continuous data available → Responder Analysis preferred

IQWiG indicates Institute for Quality and Efficiency in Health Care.

Source: Schlichting et al. 2022, IQWiG 2020

# Reliable Change Index for Meaningful Within-Patient Change

- Reliable Change Index (RCI)= $(Y - X) / \sqrt{2} * SEM$
- $Y$ = individual patient's value on PRO at follow-up
- $X$ = individual patient's value on PRO at baseline
- $SEM$=Standard Error of Measurement = $SD_X * \sqrt{1 - reliabilityx}$
- SEM estimates how repeated measures of a person's scores are distributed around his or her "true" score

- RCI categorizes patients as significantly changed
  - Deteriorated or improved
  - RCI 95% confidence: |RCI| > 1.96
  - RCI 68% confidence: |RCI| > 0.994 (likely change index)
  - RCI 50% confidence: |RCI| > 0.674 (likely change index)
  - Confidence reflects the likelihood that change is or is not due to chance

# Indexes for Meaningful Within-Patient Change: Example

- PROMIS Physical Function 10a (PF10a) Measure
  - Each item from 1 to 5, higher scores better
  - Summed score range of 10-50

- Sample of 1120 adult cancer patients

- Coefficient alpha for PF10a at baseline = 0.90
- SEM for PF10a at baseline = 2.29

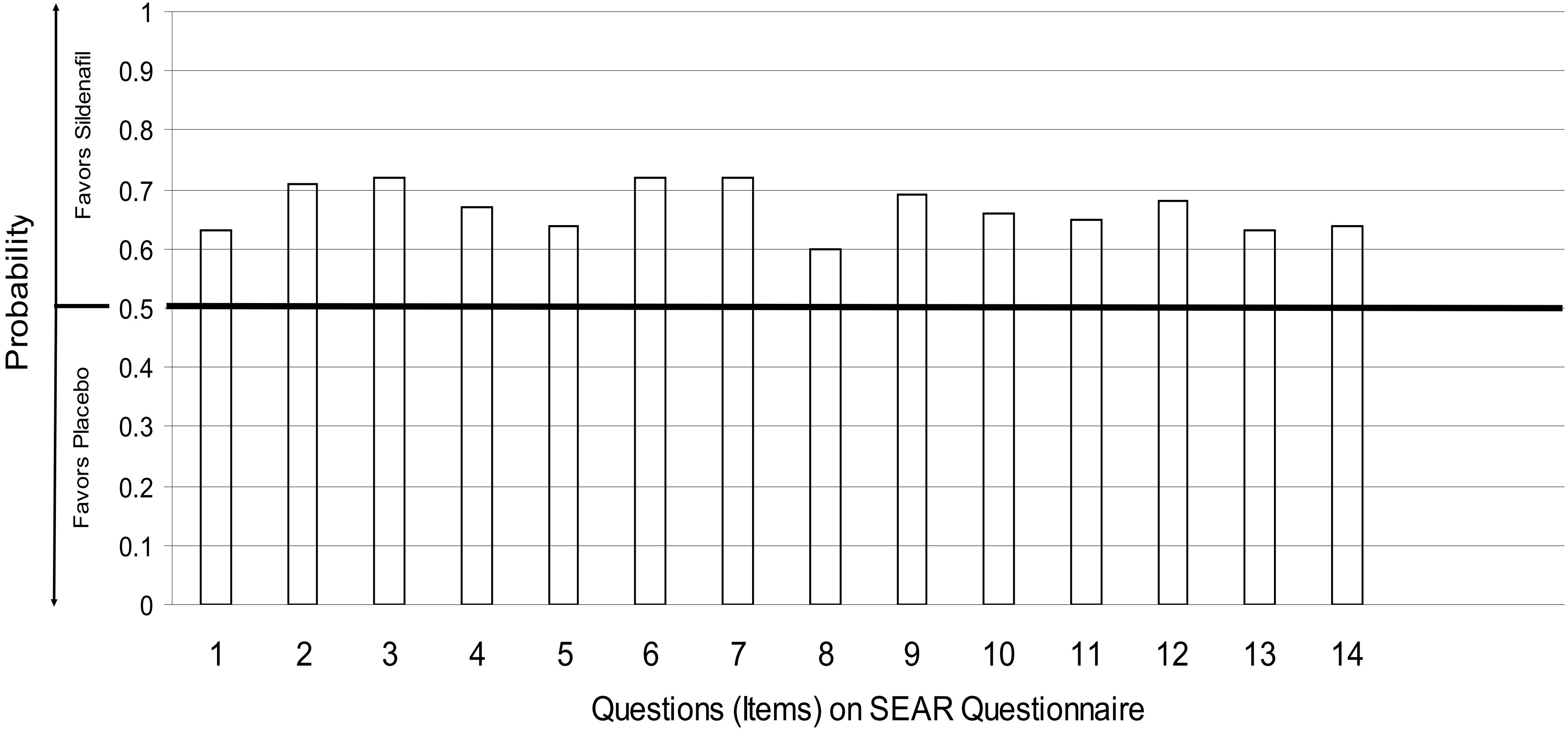
- RCI 95% confidence = 6.35, threshold of 7 points
- RCI 68% confidence = 3.22, threshold of 4 points
- RCI 50% confidence = 2.18, threshold of 3 points

Source: Peipert et al. 2023

# Probability of Relative Benefit

- Based on Wilcoxon rank-sum test using ridit analysis
- Convert Mann-Whitney *U* statistic to a probability
- Probability represents the chance that a randomly selected patient from the treatment group has a more favorable response than a randomly selected patient from the control group

# Example: Probability of Relative Benefit



Source: Cappelleri et al. 2007

# Cumulative Distribution Function (CDF) Curves

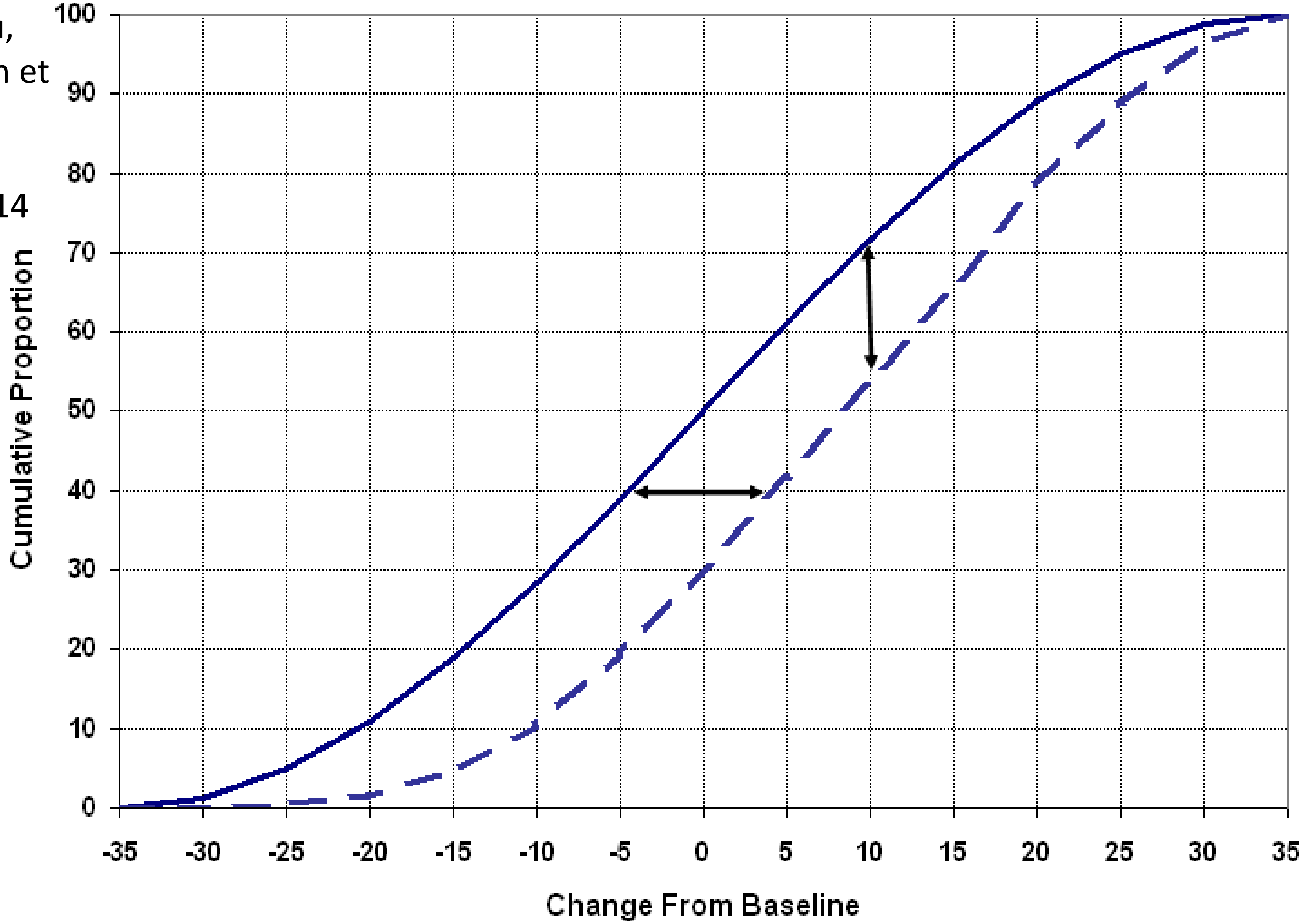
- An alternative or supplement to responder analysis

- Display a continuous plot of change from baseline on the horizontal axis and the cumulative percent of patients experiencing up to that change on the vertical axis

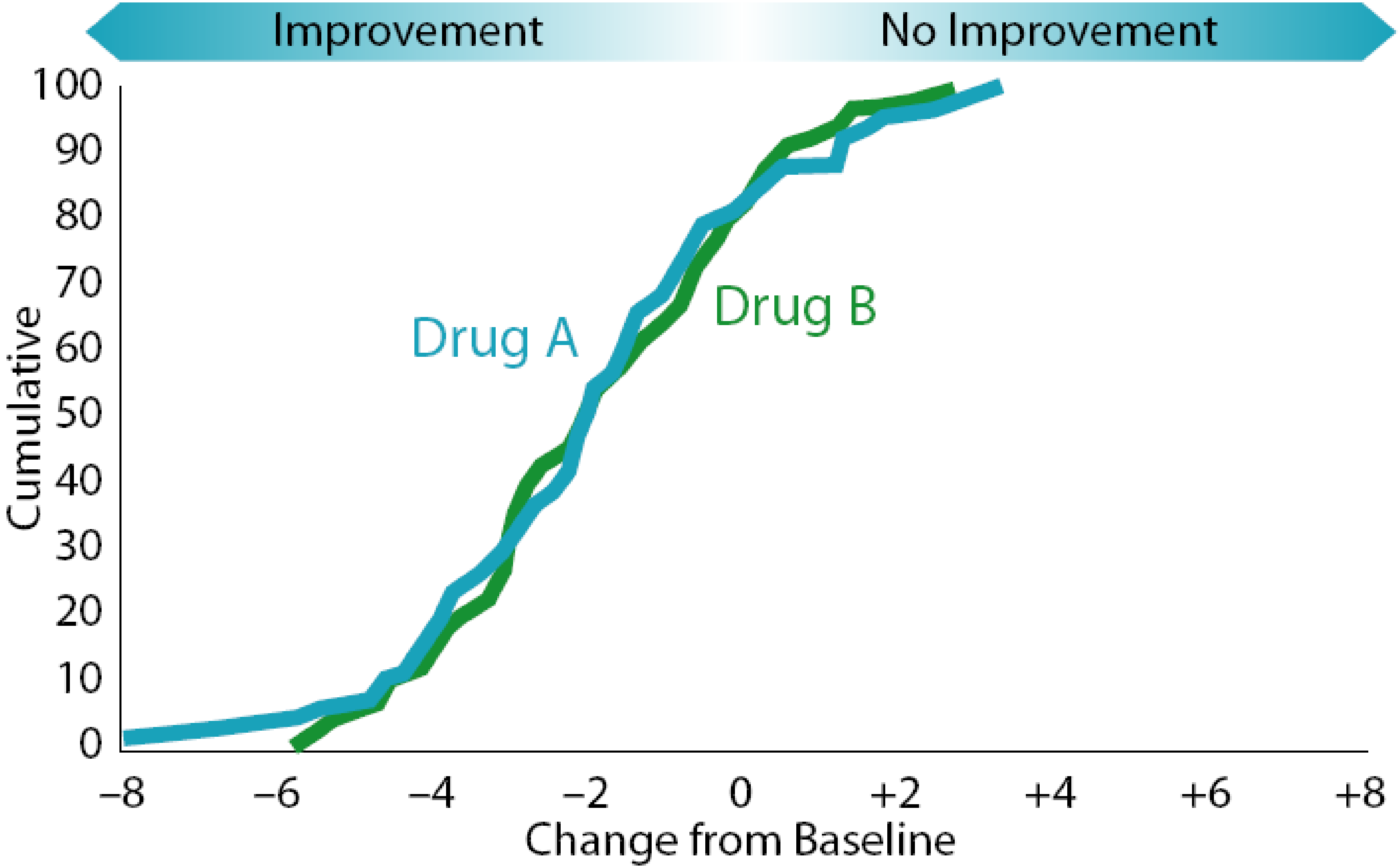
- Such a cumulative distribution of response curve – one for each treatment group – would allow a variety of response thresholds to be examined simultaneously and collectively, encompassing all available data

# Illustrative CDF Curves: Experimental Treatment (solid line) Better Than Control Treatment (dash line) – Negative Changes Indicate Improvement

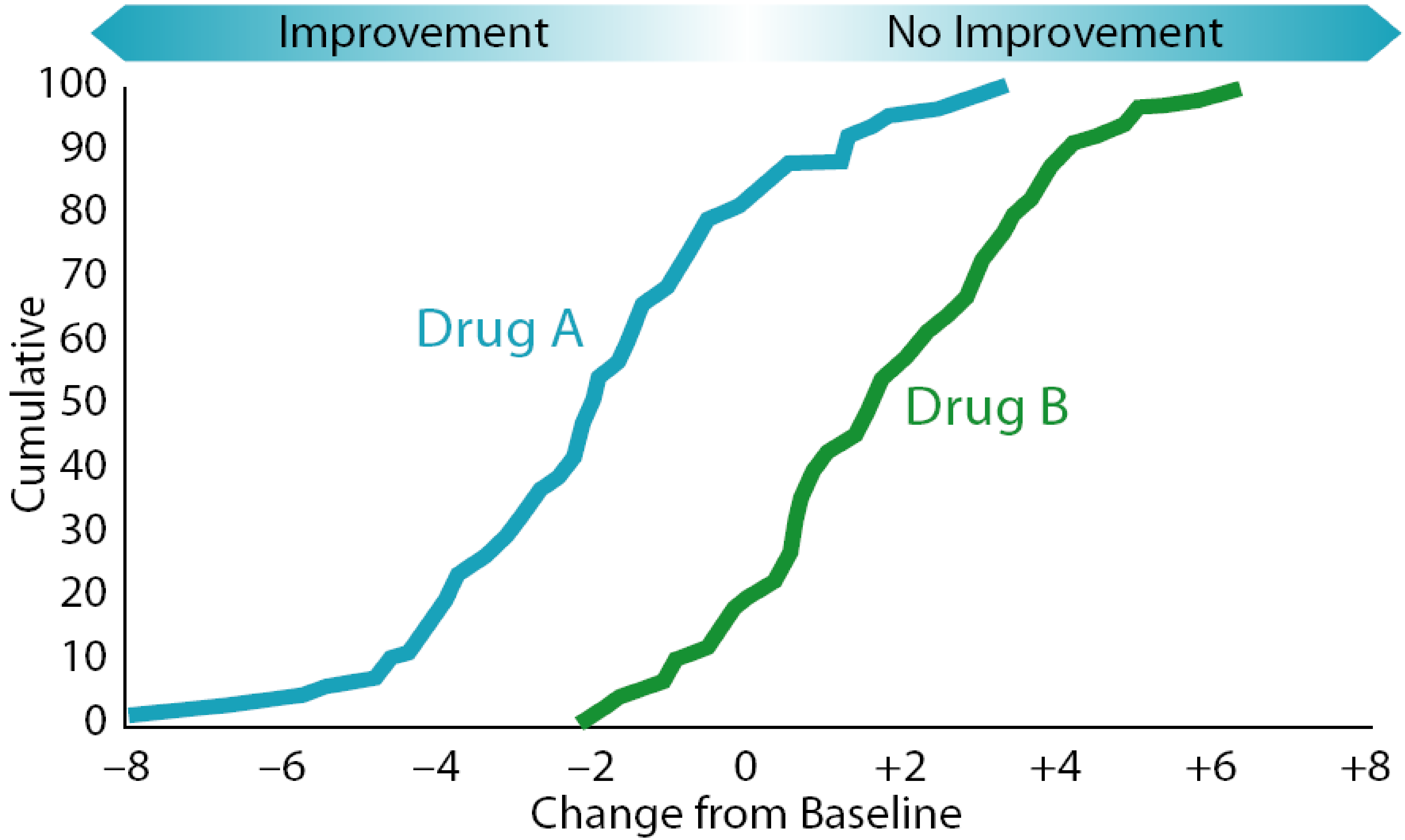
Source: Cappelleri, Zou, and Bushmakin et al. 2013, Cappelleri and Bushmakin 2014



# *Results Showing No comparative Efficacy of Drug A or Drug B*

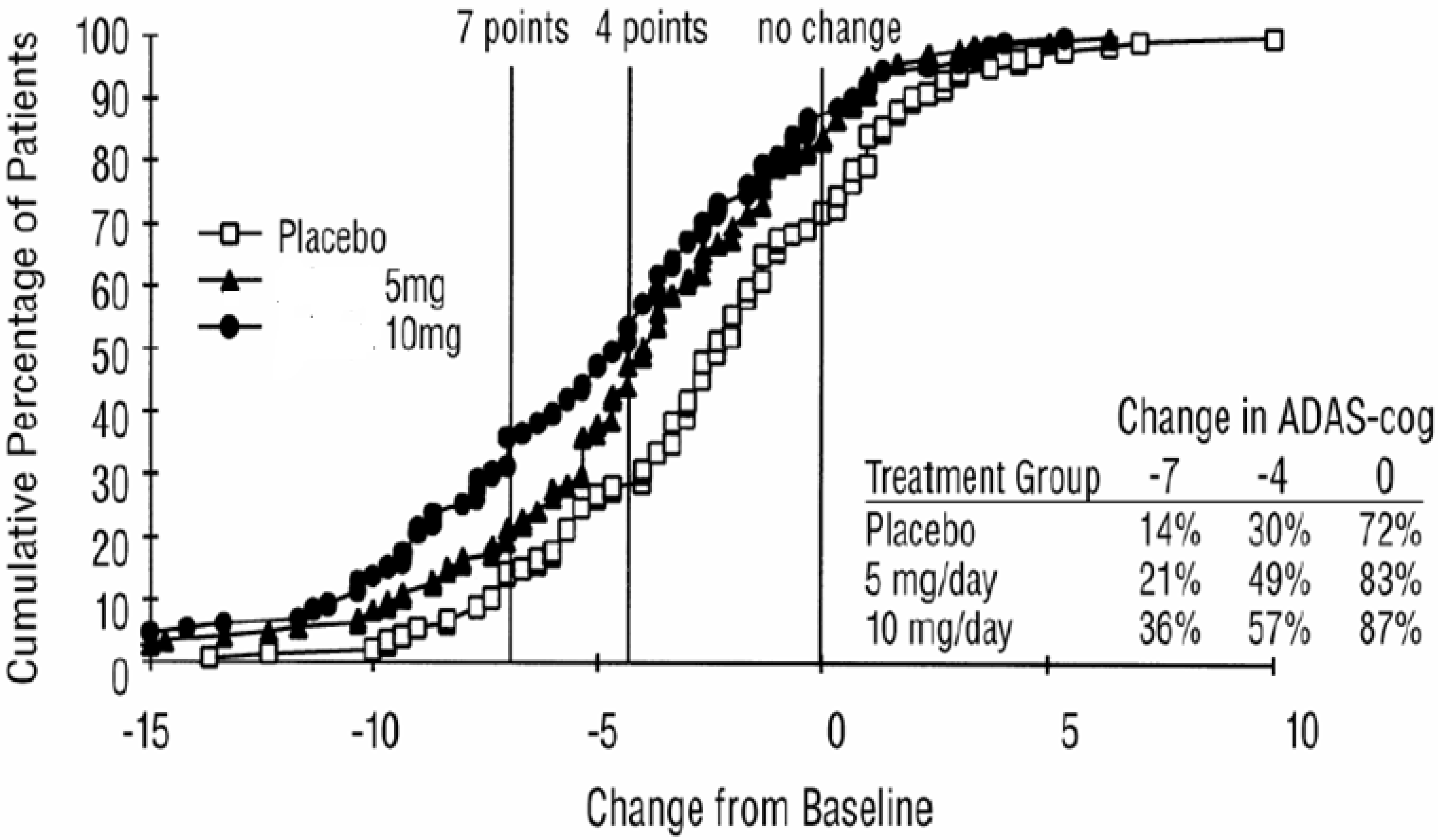


# *Results Showing the Efficacy of Drug A over Drug B*



# *Example with the Alzheimer's Disease Assessment Scale-Cognitive (ADAS-Cog)*

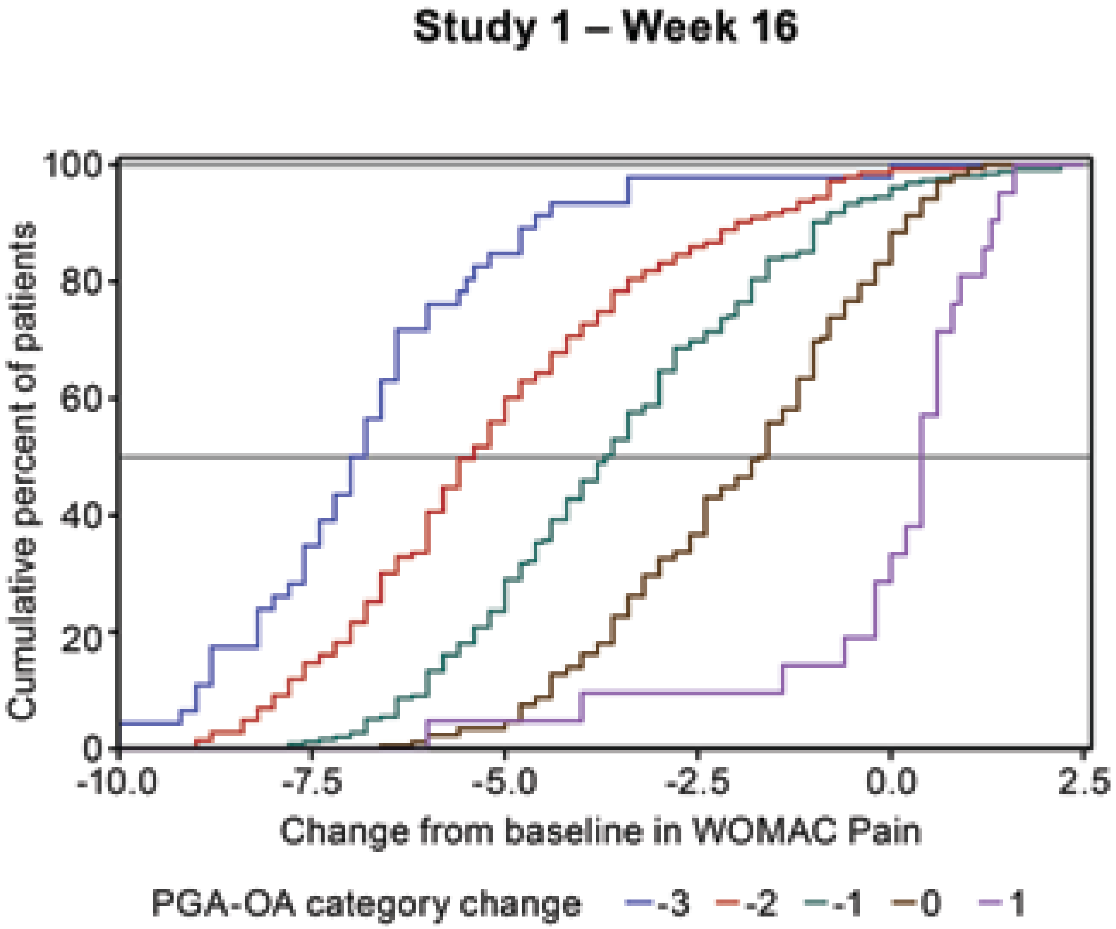
Aricept® Label from 10/13/2006



| | Change in ADAS-cog | | |
|---|---|---|---|
| Treatment Group | -7 | -4 | 0 |
| Placebo | 14% | 30% | 72% |
| 5 mg/day | 21% | 49% | 83% |
| 10 mg/day | 36% | 57% | 87% |

Cumulative Percentage of Patients with Specified Changes from Baseline ADAS-cog Scores. The Percentages of Randomized Patients Within Each Treatment Group Who Completed the Study Were: Placebo 93%, 5 mg/day 90% and 10 mg/day 82%.

# *Example with WOMAC Pain in Osteoarthritis*



Source:
Conaghan et al.
2022

Patient Global Assessment of Osteoarthritis (PGA-OA) was a single question: “Considering all the ways your OA in your hip/knee affects you, how are you doing today?”

PGA-OA was measured on a 5-point Likert scale, with higher scores indicating worse symptoms (1 = very good [asymptomatic and no limitation of normal activities] to 5 = very poor [very severe symptoms that are intolerable and inability to carry out all normal activities]

# CDF Curves to Support Clinical Relevance of Treatment Effect

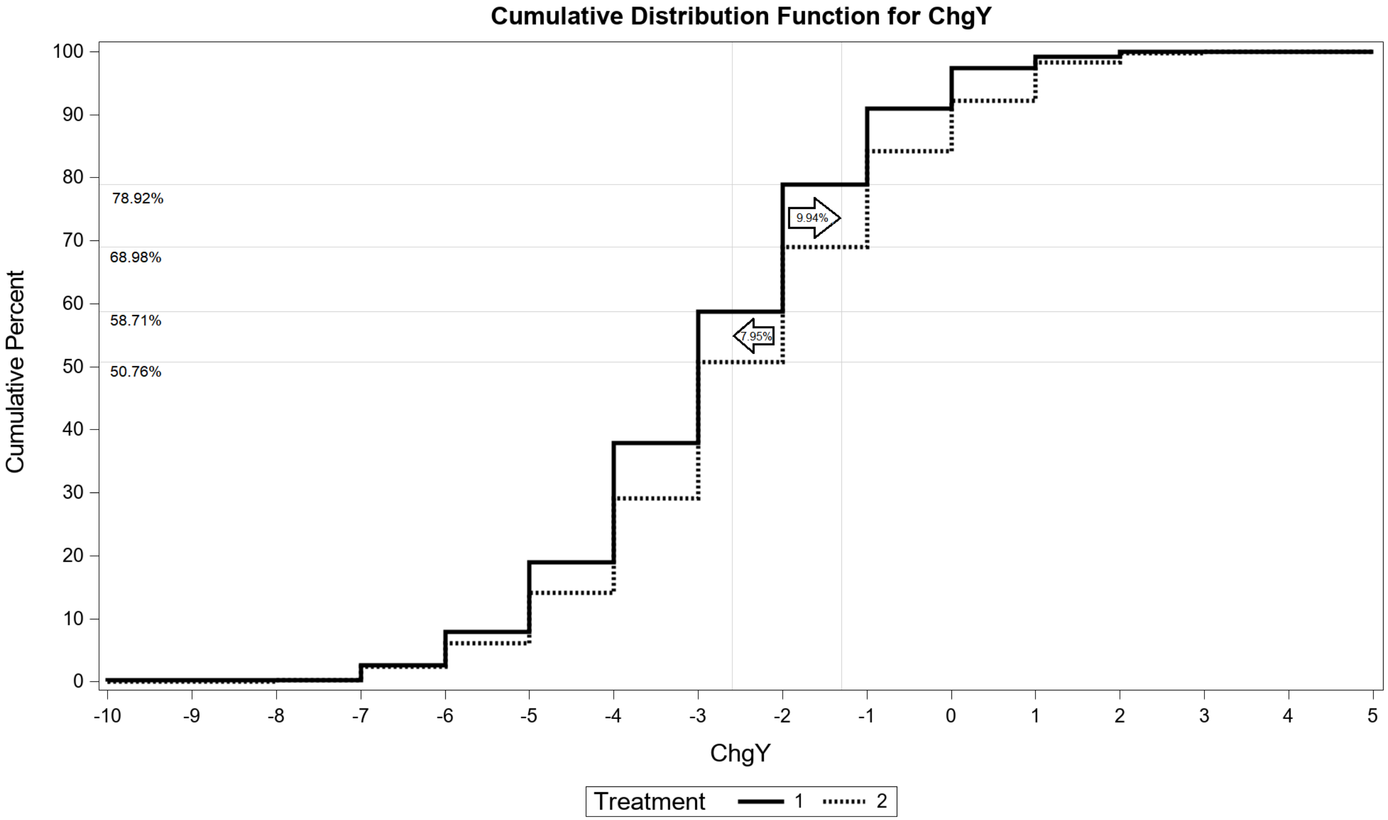
- To support clinical relevance of the estimated treatment effect, CDF curves must meet two criteria:

1) Consistent separation between treatment arms

2) Treatment effect occurs in the range patients consider to be clinically meaningful

Source: FDA 2019

# eCDF Curves by Treatment Arm



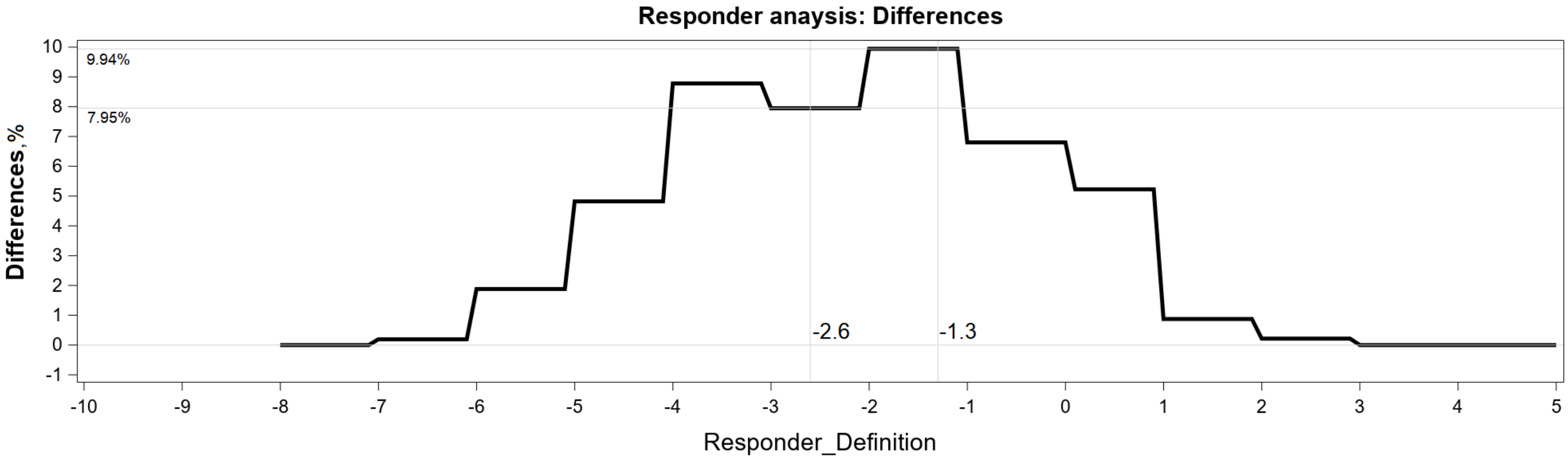
Source: Bushmakin and Cappelleri 2022

More negative change scores are more favorable

“Consistent separation between treatment arms”

Suppose MWPC estimates were -1.3 (for a one-category difference in the anchor) and -2.6 (for a two-category difference in the anchor):
Response Rate for Threshold of -1.3: Treatment 1, 78.92%; Treatment 2, 68.98%
Response Rate for Threshold of -2.6: Treatment 1, 58.71%; Treatment 2, 50.76%

# Difference in Percentages of Responders Between Treatment Arms



More negative change scores are more favorable

“Treatment effect occurs in the range patients consider to be clinically meaningful”

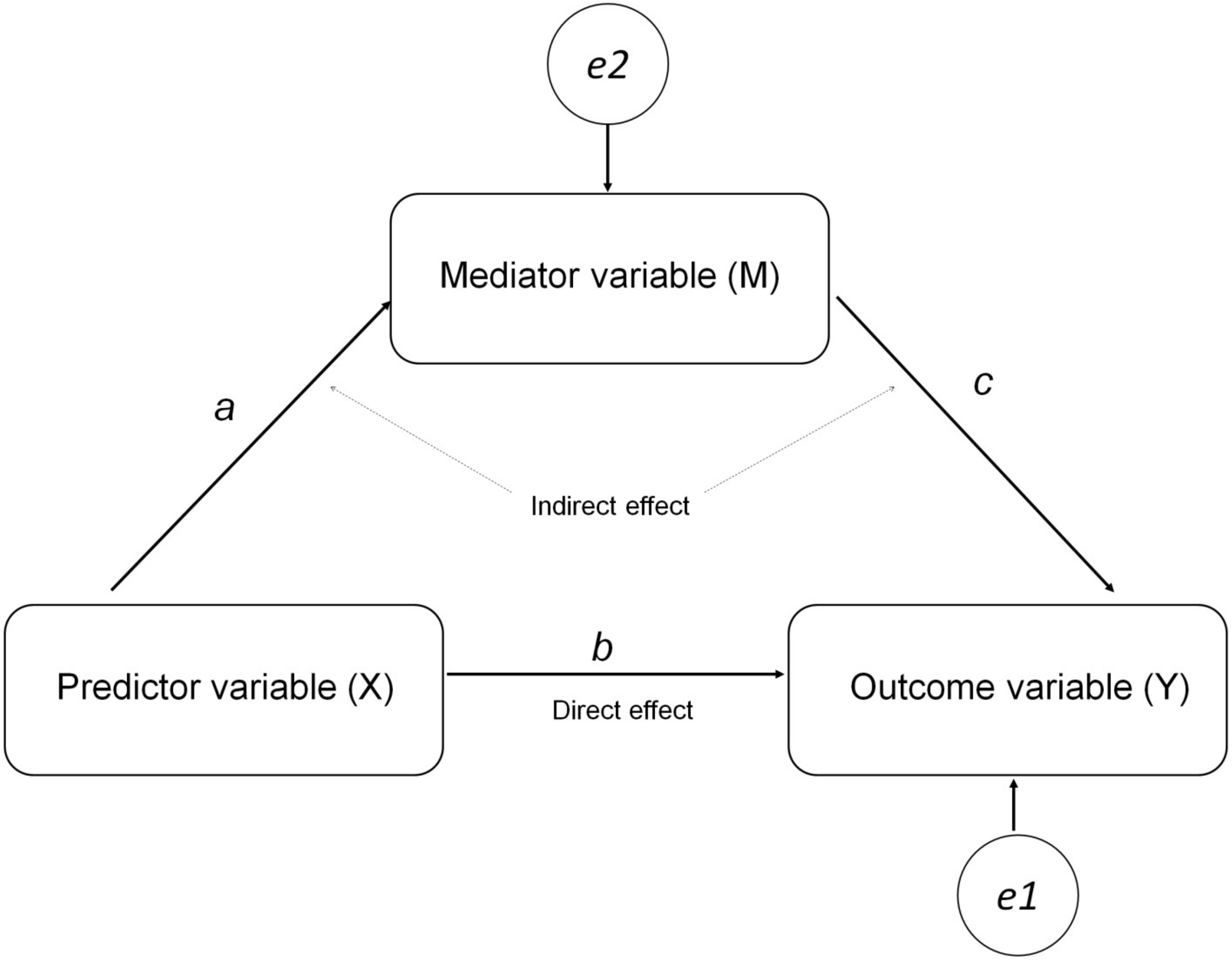
Suppose MWPC estimates were -1.3  (for a one-category difference in the anchor) and -2.6 (for a two-category difference in the anchor).

Largest differences between treatment arm occurs in the regions of MWPC threshold: Threshold of -1.3: 9.94%, Threshold of -2.3: 7.95%

Source: Bushmakin and Cappelleri 2022

# Mediation Analysis

# Basic Mediation Model



# A Few Equations

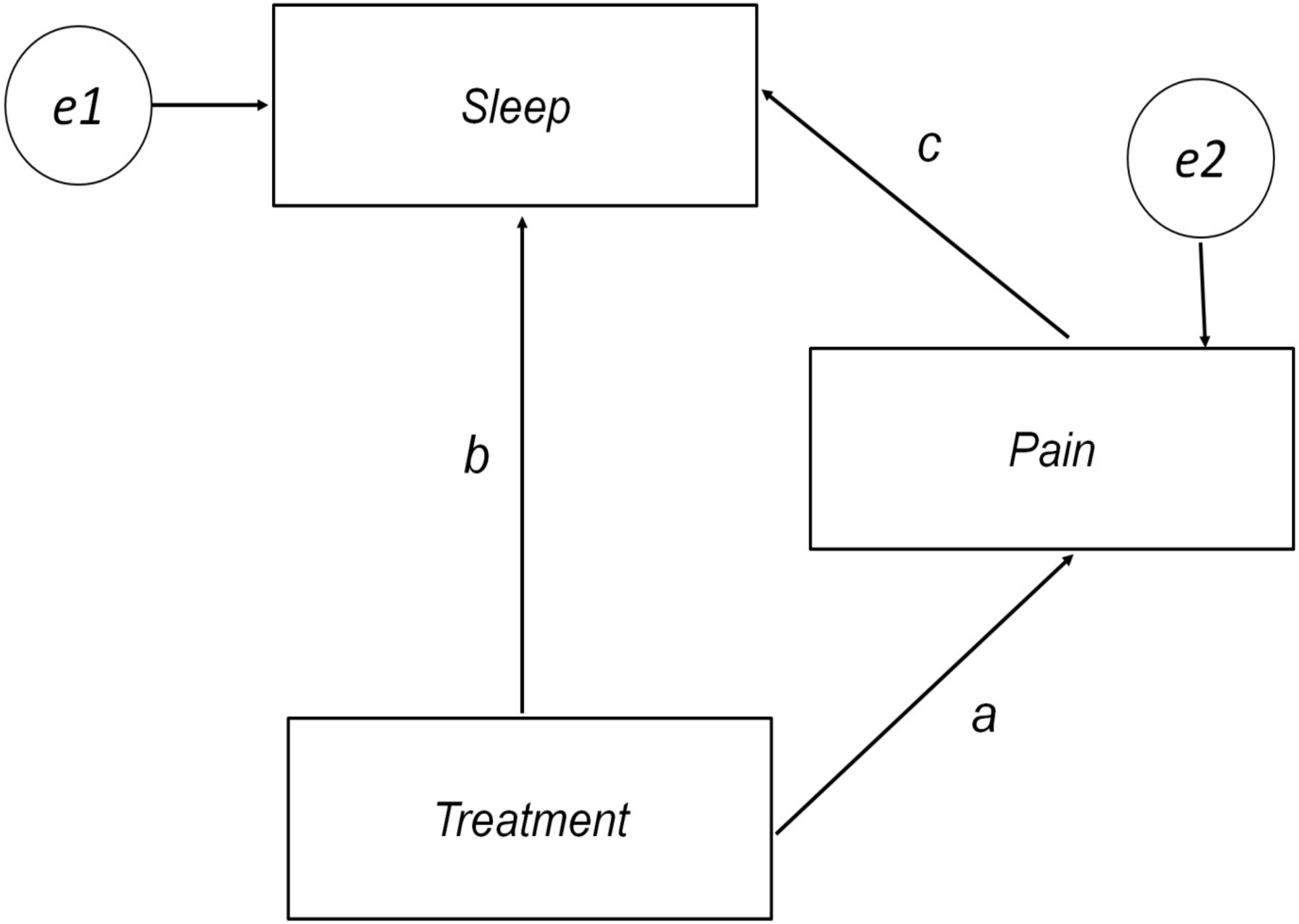
- $Y_j = i_1 + b \times X_j + c \times M_j + e_{1j}$
- $M_j = i_2 + a \times X_j + e_{2j}$

- $Y_j = (i_1 + c \times i_2) + (b + c \times a) \times X_j + (c \times e_{2j} + e_{1j})$

$$direct\ effect = 100\left(\frac{b}{b + c \times a}\right)$$

$$indirect\ effect = 100\left(\frac{c \times a}{b + c \times a}\right)$$

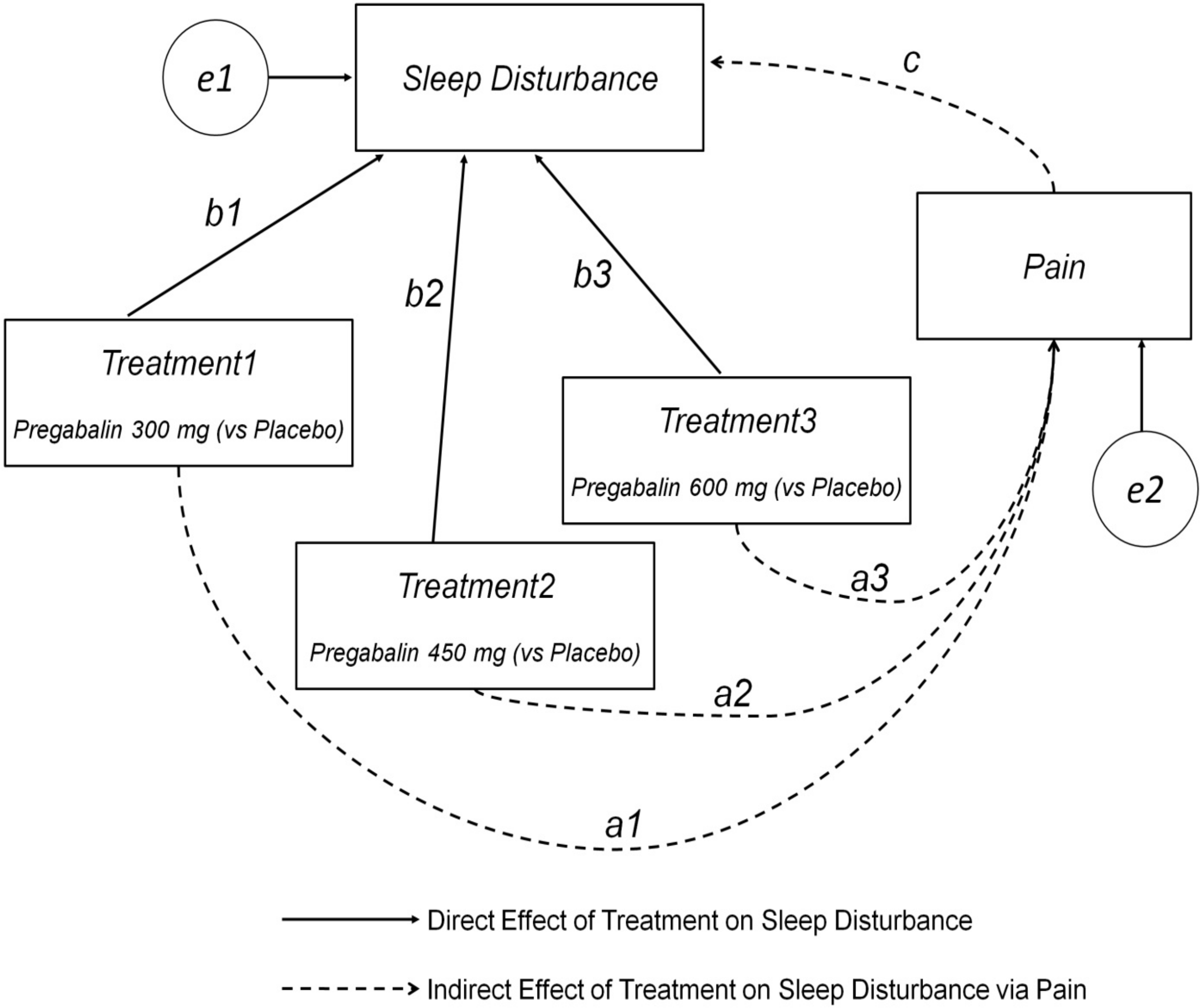
# *Treatment* Affects *Sleep* Directly and Indirectly via *Pain*



# Assumptions

- No unmeasured confounding
  - Predictor-outcome
  - Predictor-mediator
  - Mediator-outcome
- Model with no interaction is correctly specified
  - Predictor and mediator on outcome

# Published Example


e1
Sleep Disturbance
c
b1
b2
b3
Pain
Treatment1
Pregabalin 300 mg (vs Placebo)
Treatment3
Pregabalin 600 mg (vs Placebo)
e2
Treatment2
Pregabalin 450 mg (vs Placebo)
a3
a2
a1
Direct Effect of Treatment on Sleep Disturbance
Indirect Effect of Treatment on Sleep Disturbance via Pain

Source: Russell et al. 2009

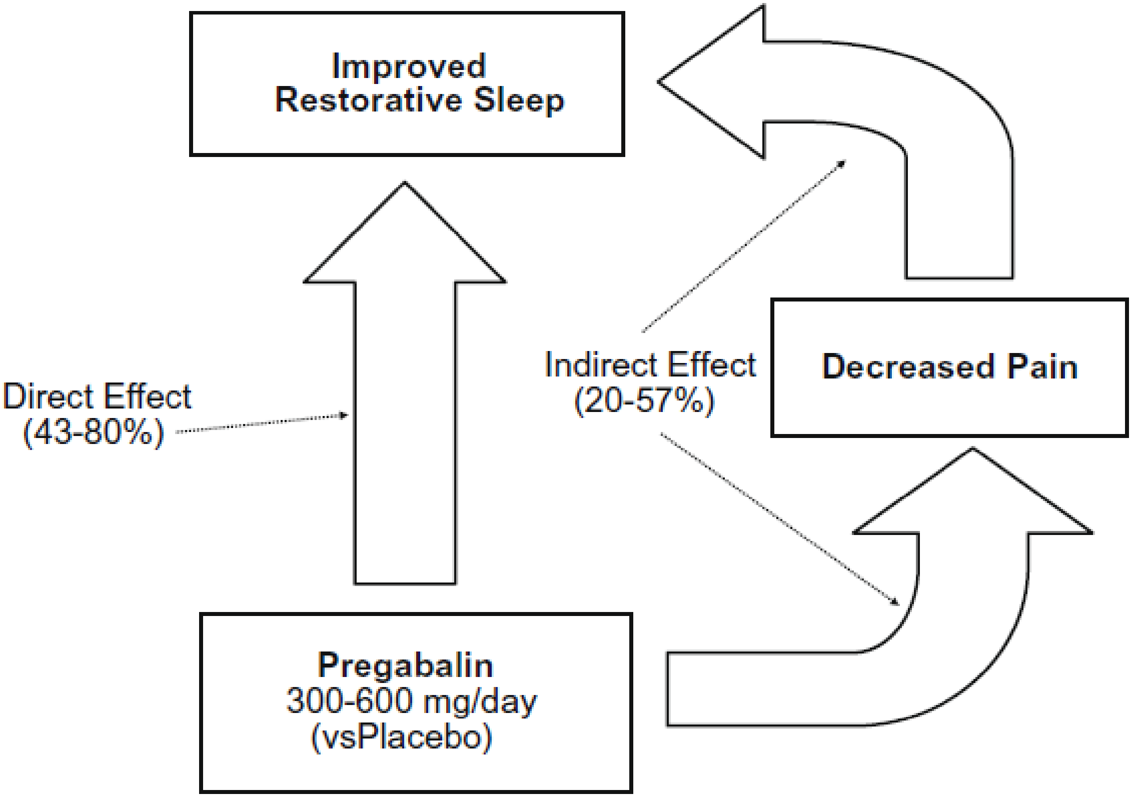
# Results on Sleep Disturbance (One Study)

| Effect | Effects from TRT300 to SLEEP | Effects from TRT450 to SLEEP | Effects from TRT600 to SLEEP |
|---|---|---|---|
| Total | -9.94 | -12.73 | -17.79 |
| Indirect | -1.95(*) | -3.44 | -4.35 |
| (Indirect / Total) x 100% | 19.6%(*) | 27% | 24.4% |
| (Direct / Total) x 100% | 80.4% | 73% | 75.6% |

(*) indicates not statistically significant result, p-value > 0.05

Source: Russell et al. 2009

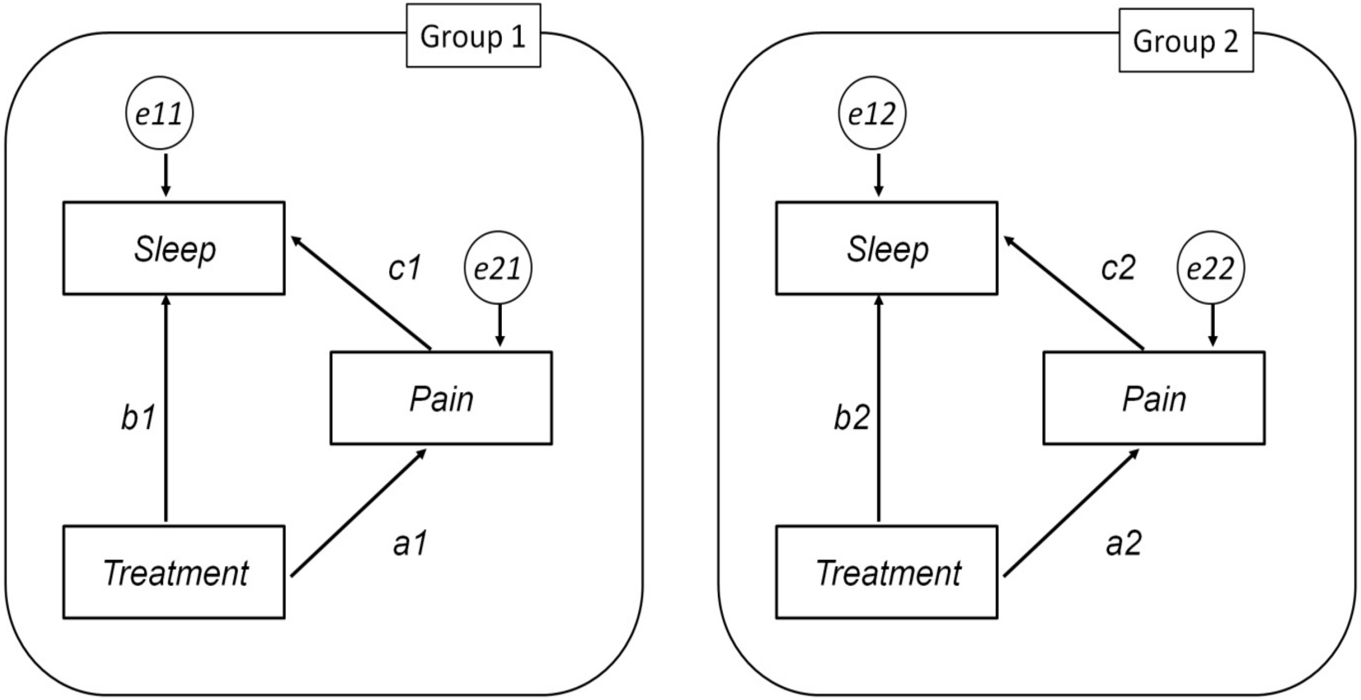
# Results on Sleep Quality and Sleep Disturbance (Each from the Same Two Studies)



Organizational diagram illustrating the direct and indirect effects of pregabalin on restorative sleep determined by mediation analysis. While the direct effect is independent of the effect of pregabalin on pain, the indirect effect is mediated through the analgesic effects of pregabalin. For the Sleep Quality Diary, direct effects range from 43% to 61% and indirect effects range from 40% to 57%. For the Medical Outcomes Study Sleep Disturbance subscale, direct effects range from 66% to 80% and indirect effects range from 20% to 34%.

Source: Russell et al. 2009, Cappelleri and Bushmakin 2014

# Testing for Model Invariance Between Groups



$$difference\ of\ direct\ effects\ (Group\ 1\ vs\ Group\ 2):$$

$$= 100\ (\frac{b1}{b1+c1\times\ a1} - \frac{b2}{b2+c2\times\ a2})$$

$$difference\ of\ indirect\ effects\ (Group\ 1\ vs\ Group\ 2):$$

$$= 100\ (\frac{c1\times\ a1}{b1+c1\times\ a1} - \frac{c2\times\ a2}{b2+c2\times\ a2})$$

Source: Cappelleri, Zou, Bushmakin et al. 2013

# Summary

- Anchor-based approaches
  - Percentage based on thresholds
  - Criterion-group interpretation
  - Statistical significance and clinical equivalance
  - Content-based interpretation
  - Clinical important difference

- Distribution-based approaches
  - Effect size, % of range, reliability change index
  - Probability of relative benefit
  - Cumulative distribution function

- Mediation analysis

# References

- Althof SE. Cappelleri JC, Shpilsky A, Stecher V, Diuguid C, Sweeney M, Duttagupta S. 2003. Treatment responsiveness of the Self-Esteem And Relationship (SEAR) questionnaire in erectile dysfunction. *Urology* 61:888-893.

- Althof SE, O'Leary MP, Cappelleri JC, Hvidsten K, Stecher VJ, Glina S, King R, RL Siegel on behalf of the International SEAR Study Group. 2006. Sildenafil citrate improves self-esteem, confidence, and relationships in men with erectile dysfunction: Results from an international, multi-center, double-blind, placebo-controlled trial. *Journal of Sexual Medicine* 3:521-529.

- Bennett RM, Bushmakin AG, Cappelleri JC, Zlateva G, Sadosky AB. 2009. Minimally clinically important difference in the Fibromyalgia Impact Questionnaire (FIQ). *Journal of Rheumatology* 36:1304-1311.

- Bushmakin AG, Cappelleri JC. 2022. *A Practical Approach to Quantitative Validation of Patient-Reported Outcomes: A Simulation-Based Guide Using SAS*. Hoboken, New Jersey: John Wiley & Sons.

- Cappelleri JC, Rosen RC, Smith MD, Quirk F, Maytom MC, Mishra A, Osterloh IH. 1999. Some developments on the International Index of Erectile Function (IIEF). *Drug Information Journal* 33:179-190.

- Cappelleri JC, Bell SS, Althof SE, Siegel RL, Stecher VJ. 2006. Comparison between sildenafil-treated subjects with erectile dysfunction and control subjects on the Self-Esteem And Relationship questionnaire. *Journal of Sexual Medicine* 3:274-282.

- Cappelleri JC, Althof SE, O'Leary MP, Tseng L-J, on behalf of the US and International SEAR Study Group. 2007. Analysis of single items on the Self-Esteem And Relationship questionnaire in men treated with sildenafil citrate for erectile dysfunction: Results of two double-blind placebo-controlled trials. *BJU International* 101:861-866.

- Cappelleri JC, Bushmakin AG, McDermott, A, Dukes E, Sadosky A, Petrie CD, Martin S. 2009. Measurement properties of the Medical Outcomes Study Sleep Scale in patients with fibromyalgia. *Sleep Medicine* 10:766-770.

# References

- Cappelleri JC, Zou KH, Bushmakin AG, Alvir JMJ, Alemayehu D, Symonds T. 2013. *Patient-Reported Outcomes: Measurement, Implementation and Interpretation*. Boca Raton, Florida: Chapman & Hall/CRC Press.

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

- Cappelleri JC, Tseng L-J, Stecher V, Goldstein I. 2018. Enriching the interpretation of the Erectile Dysfunction Inventory of Treatment Satisfaction: Characterizing success in treatment satisfaction. *Journal of Sexual Medicine* 15:732-740.

- Cella D, Choi S, Garcia S, Cook KF, Rosenbloom S, Lai J-Sh, Tatum Surges D, Gershon R. 2014. Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment. *Quality of Life Research* 23:2651-2661.

- Conaghan PG, Dworkin RH, Schnitzer TJ, Berenbaum F, Bushmakin AG, Cappelleri JC, Viktrup L, Abraham L. 2022. WOMAC meaningful within-patient change: Results from three studies of tanezumab in patients with moderate-to-severe osteoarthritis of the hip or knee. *The Journal of Rheumatology* 49:615-621.

- Cook KF, Victorson DE, Cella D, Schalet BD, Miller D. 2015. Creating meaningful cut-scores for Neuro-QOL measures of fatigue, physical functioning, and sleep disturbance using standard setting with patients and providers. *Quality of Life Research* 24:575-589.

- Devji T, Guyatt GH, Lytvyn L, Brignardello-Petersen R, Foroutan F, Sadeghirad B, Buchbinder R, Poolman RW, Harris IA, Carrasco-Labra A, Siemieniuk RAC, Vandvik PO. 2017. Application of minimal important differences in degenerative knee disease outcomes: a systematic review and case study to inform BMJ Rapid recommendations. *BMJ Open* 7:e-15587.

- Di Fusco M, Sun X, Moran MM, Coetzer H, Zamparo JM, Puzniak L, Alvarez MB, Tabak YP, Cappelleri JC. 2022. Impact of COVID-19 and effects of BNT162b2 on patient-reported outcomes: Quality of life, symptoms and work productivity among US adult outpatients. *Journal of Patient-Reported Outcomes* 6:123. https://doi.org/10.1186/s41687-022-00528-w.

- Duong T, Staunton H, Braid J, Barriere A, Trzaskoma B, Gao L, Willgoss T, Cruz R, Gusset N, Gorni K, Randhawa S, Yang L, Vuillerot C.2022. A patient-centered evaluation of meaningful change on the 32-item motor function measure in spina muscular atrophy using qualitative and quantitative data. *Frontiers in Neurology* 12:770423.

# References

- Farrar JT, Young JP Jr, LaMoreaux L, Werth JL, Poole RM. 2001. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain* 94:149-148.

- FDA (Food and Drug Administration). 2019. Patient-focused drug development guidance series for enhancing the incorporation of the patient's voice in medical product development and regulatory decision making. Draft guidance documents. https://www.fda.gov/drugs/development-approval-process-drugs/fda-patient-focused-drug-development-guidance-series-enhancing-incorporation-patients-voice-medical. (Accessed September 10, 2023.).

- IQWiG General Methods: Version 6.0 of 5 November 2020. 2020. https://www.iqwig.de/methoden/general-methods_version-6-0.pdf. (Access September 10, 2023.)

- Karuturi M, Rocque GB, Cappelleri JC, Blum J, McCune S, Bijoy Telivala B, Kurian S, Anderson D, Pluard T, Migas J, Wang Y, Montelongo MZ, Tripathy D. 2021. Real-world quality of life (QoL) in patients with hormone receptor-positive (HR+), human epidermal growth factor receptor 2-negative (HER2-), advanced breast cancer (ABC) treated with palbociclib: A patient-reported outcome (PRO) analysis from POLARIS. Poster presentation at the 2021 San Antonio Breast Cancer Symposium (SABCS) – 44th Annual Meeting, San Antonio, Texas, December 7–10.

- Mamolo CM, Bushmakin AG, Cappelleri JC. 2015.Application of the Itch Severity Score in patients with moderate-to-severe plaque psoriasis: Clinically important difference and responder analysis. *Journal of Dermatological Treatment* 26:121-123.

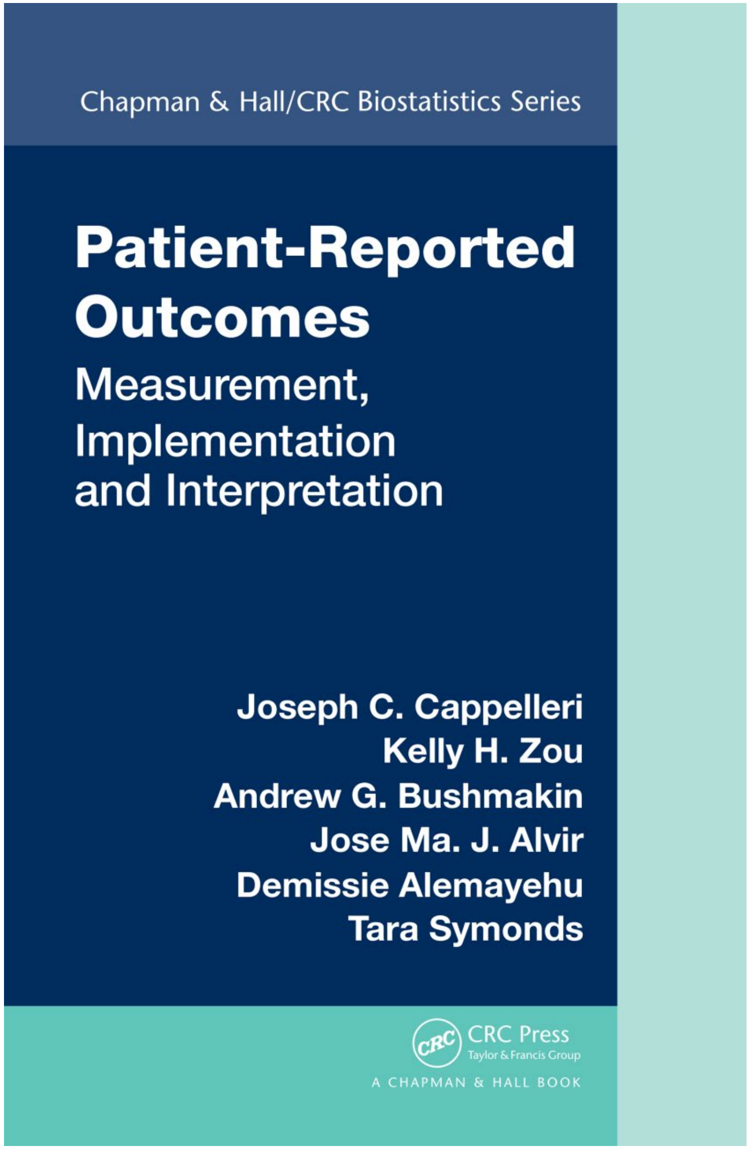
- Peipert JD, Hays RD, Cella D. 2023. Likely change indexes improve estimates of individual change on patient-reported outcomes. *Quality of Life Research* 32:1341-1352.

- Russell IJ, Crofford LJ, Leon T, Cappelleri JC, Bushmakin AG, Whalen E, Barrett JA, Sadosky A. 2009. The effects of pregabalin on sleep disturbance symptoms among individuals with fibromyalgia syndrome. *Sleep Medicine* 10:604-610.

# References

- Schlichting M, Hennig M, Rudell K, McLeod L, Bennett B, Shaw J, Doward L, Molsen-David E, Chassany O. 2022. Is IQWiG's 15% threshold universally applicable in assessing the clinical relevance of patient-reported outcomes changes? An ISPOR special interest group report. *Value in Health* 25:1463-1468.

- Trigg A, Griffiths P. 2021. Triangulation of multiple meaningful change thresholds for patient-reported outcome scores. *Quality of Life Research* 30:2755-2764.



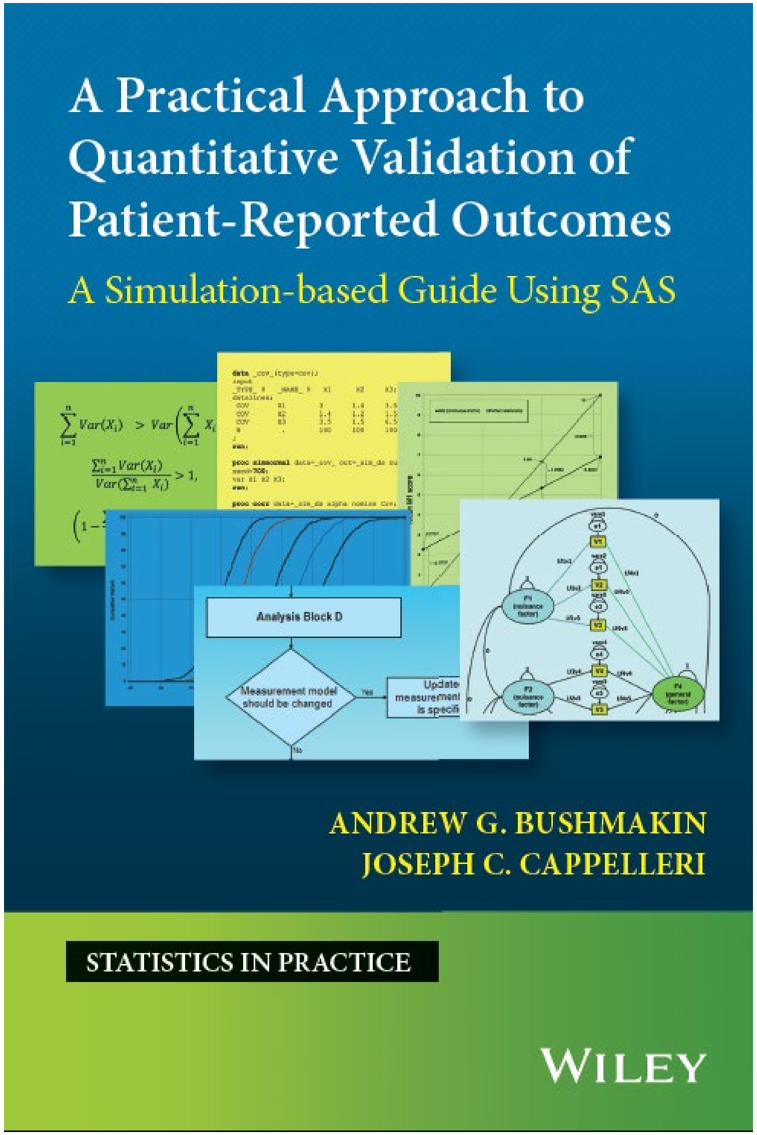
Chapman & Hall/CRC Biostatistics Series

**Patient-Reported Outcomes**

Measurement, Implementation and Interpretation

Joseph C. Cappelleri
Kelly H. Zou
Andrew G. Bushmakin
Jose Ma. J. Alvir
Demissie Alemayehu
Tara Symonds

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK



A Practical Approach to Quantitative Validation of Patient-Reported Outcomes

A Simulation-based Guide Using SAS

ANDREW G. BUSHMAKIN
JOSEPH C. CAPPELLERI

STATISTICS IN PRACTICE

WILEY