Methods for Interpretation of Patient-Reported Outcomes

Joseph C. Cappelleri, PhD, MPH
Pfizer Inc
joseph.c.cappelleri@pfizer.com

Invited one-hour tutorial at BASS XXV – 25th Annual Biopharmaceutical Applied Statistics Symposium, Savannah, Georgia, October 15-19, 2018

Learning Objective

- To understand the methods for interpretation of patient-reported outcomes

Outline

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

- Distribution-based approaches
  - Standardized effect size
  - Probability of relative benefit
  - Cumulative distribution function

- Mediation analysis

Importance of Interpretation

- PRO results must be interpreted by attaching meaning to them
- Patients and other stakeholders benefit
- Applying methods to enrich interpretation of PRO scores

Anchor-based Approaches

What is an Anchor?

- Anchor measure is external 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

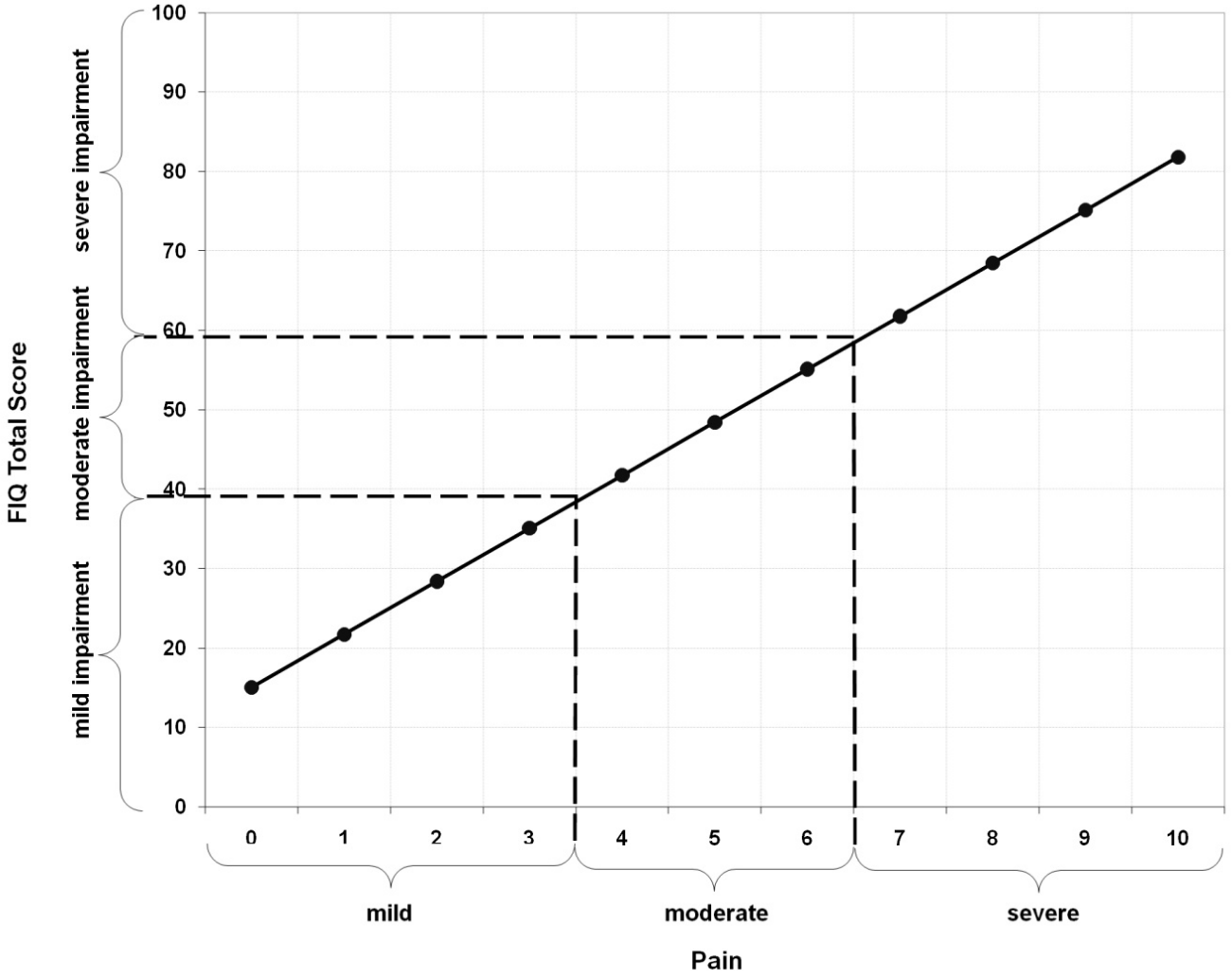
Percentage Based on Thresholds

- Show percentage of patients above and below some specified value, which is an anchored value with a meaningful criterion.

- Example: Erectile function domain of International Index of Erectile Function

- Example: Severity categorization on Fibromyalgia Impact Questionnaire (FIQ)

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)

VIEWTABLE: Work._mixed_2

	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

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;
```

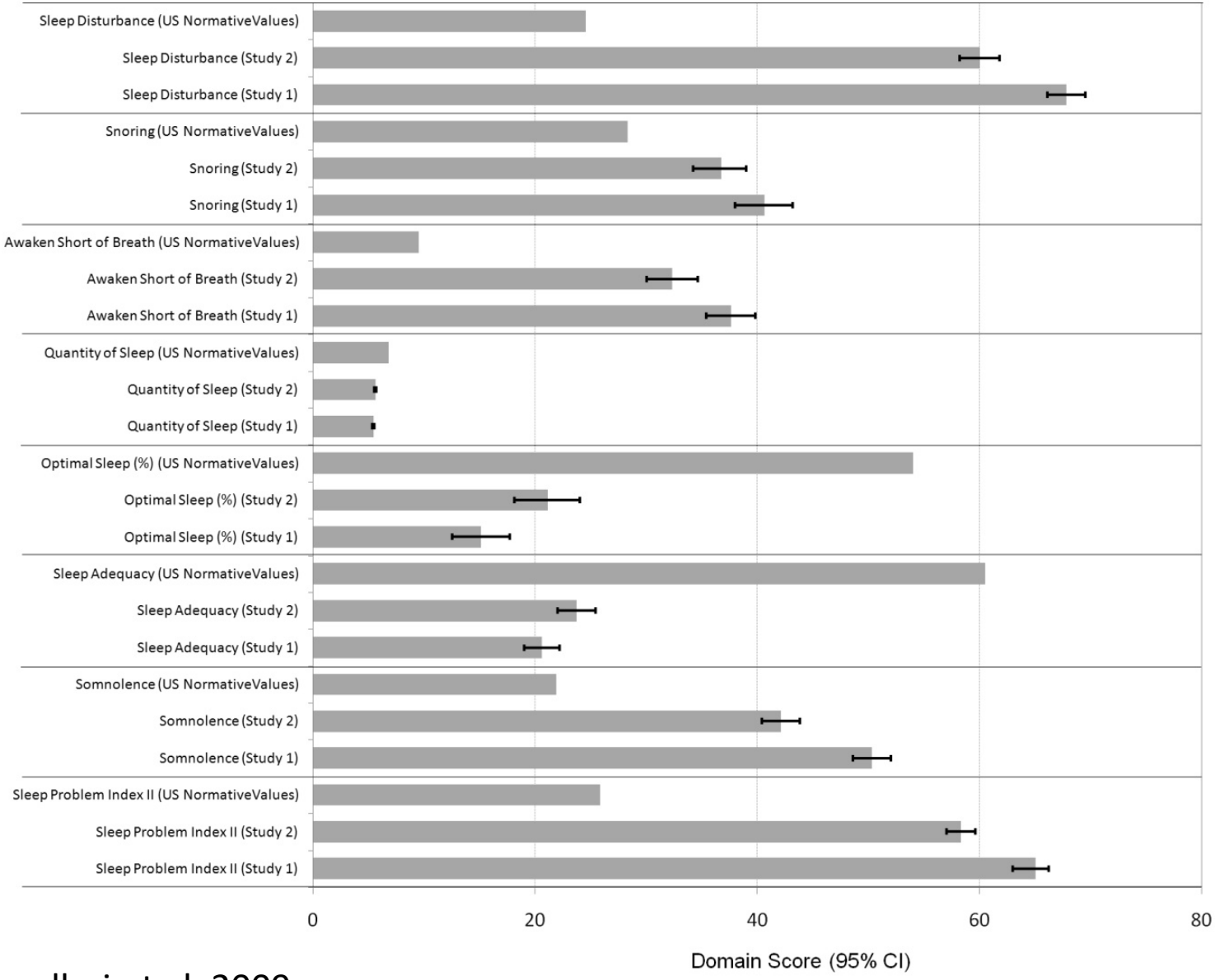
Results from Simulated Example

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

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

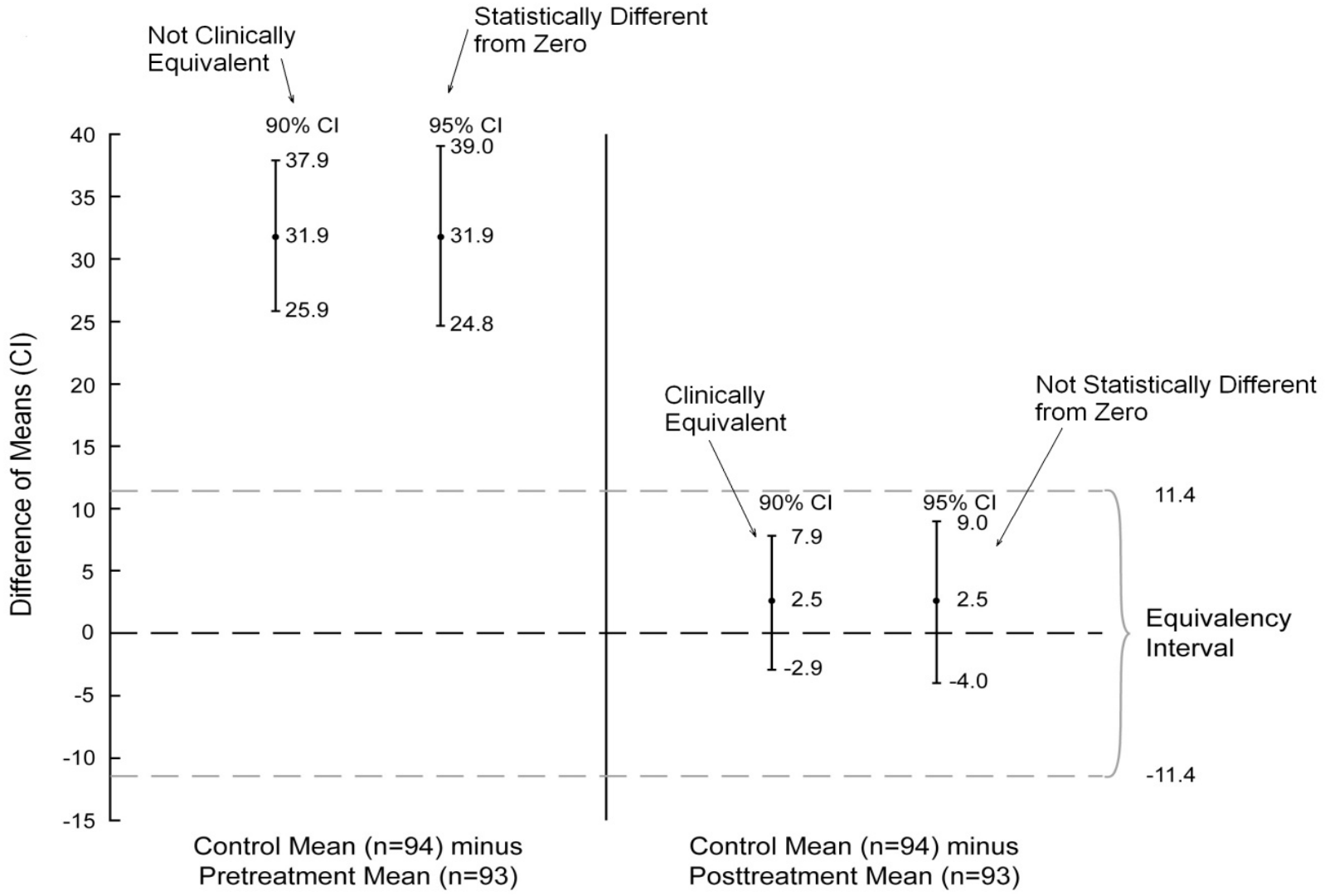


Source: Cappelleri et al. 2009

Classification of Tests on Statistical Significance and Clinical Equivalence

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

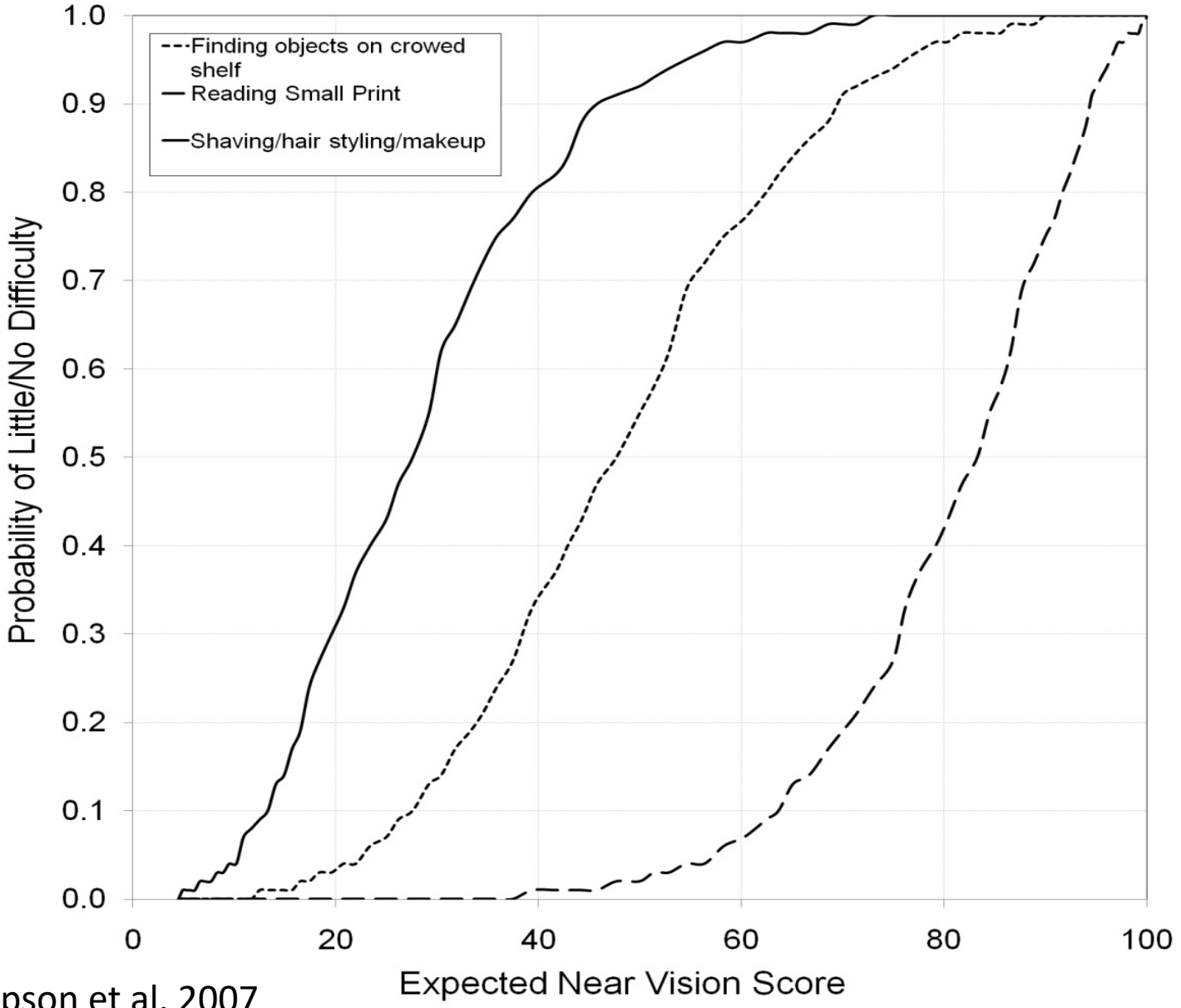


Source: Cappelleri et al. 2006

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, ordinal logistic regression, and binary logistic regression

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



Source: Thompson et al. 2007

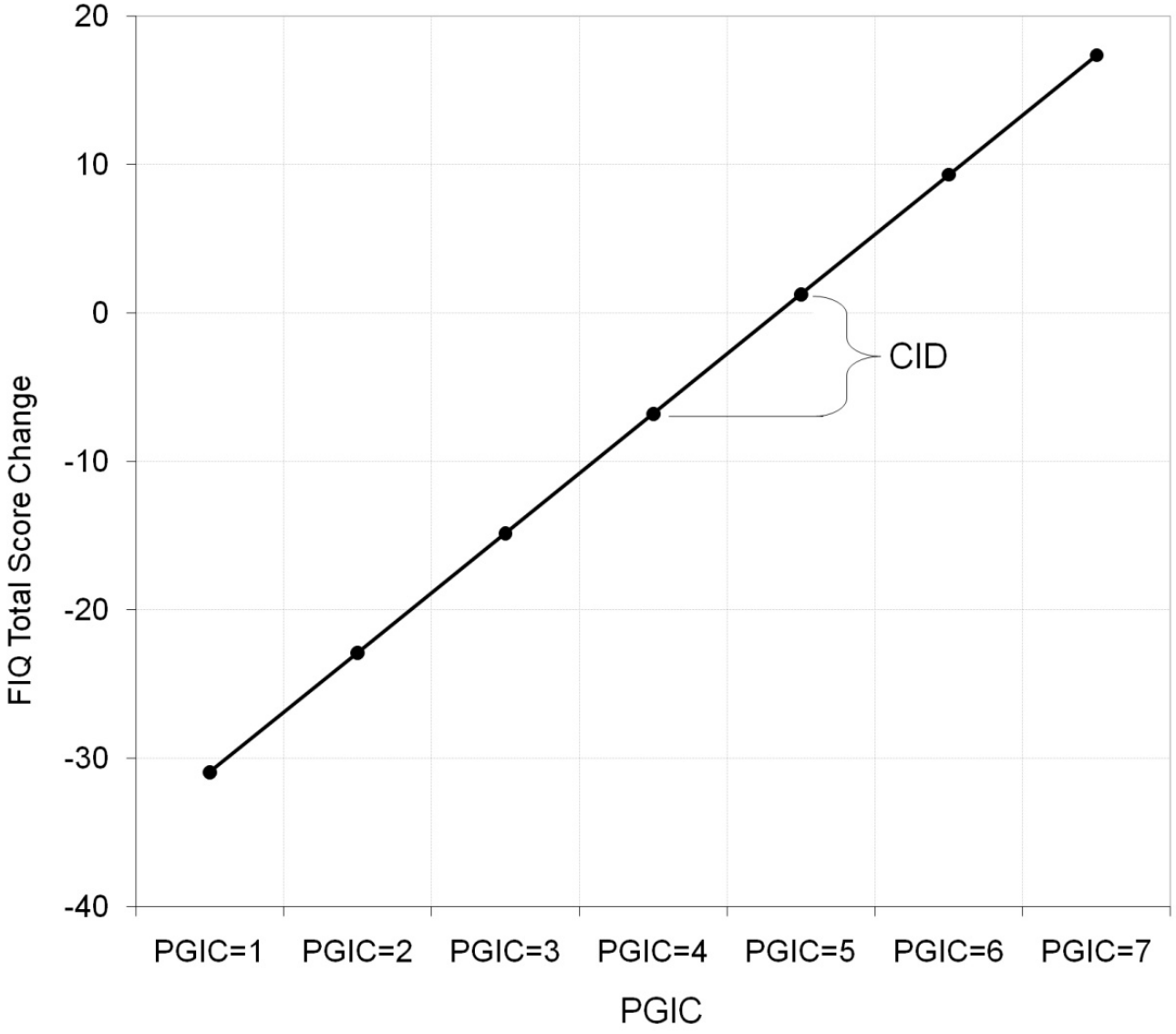
Clinical Important Difference (CID)

- Statistical significance does not imply clinical significance
- PRO score (or change in PRO score) as outcome regressed on an anchor 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
- Anchor: Clinical Global Impression–Severity (CGIC, serial)

CID on FIQ Using PGIC as Continuous Anchor



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

Source: Bennett et al. 2009

Dataset Structure in Simulated Example

VIEWTABLE: Work._mixed_2

	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

Proc Mixed Longitudinal Modeling: CID 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 CID

Label	Estimate	Standard Error	Pr > |t|	Lower	Upper
CID (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

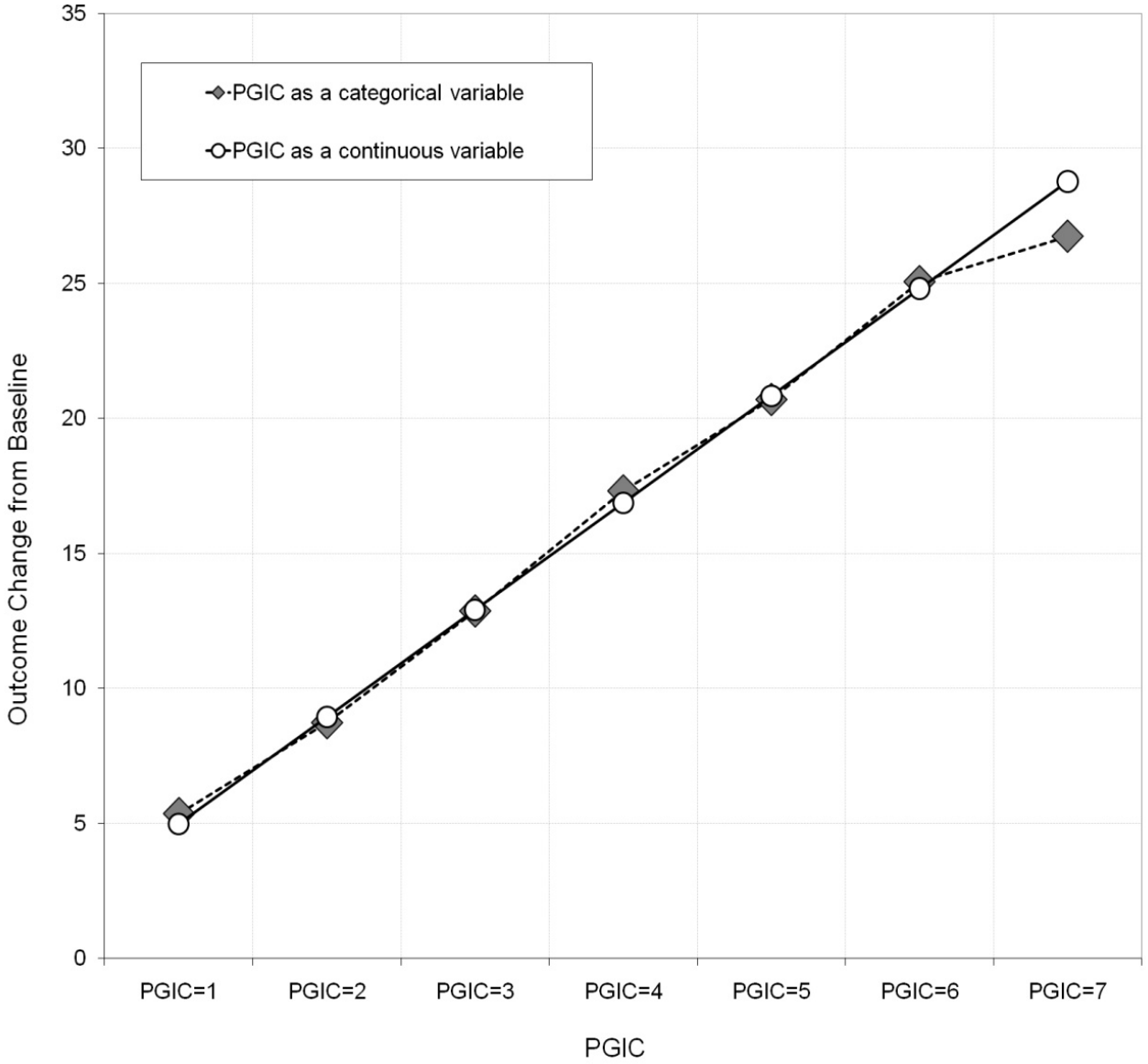
Proc Mixed Longitudinal Modeling: CID Estimation (Categorical Anchor) – Sensitivity Analysis

```
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 CID: Sensitivity Analysis (Same Simulated Data)

Effect	*PGIC*	*Estimate*	*Standard 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 Difference in PRO Measure as Function of PGIC



Frequencies on PGIC

PGIC	*Frequency*	*Cumulative Percent*	*Cumulative 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

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 whether change is *clinically* significant to patients
- Different types
  - Standardized Effect Size
  - Probability of Relative Benefit
  - Cumulative Distribution Function

Standardized Effect Size

- (Standardized) Effect size = magnitude of effect relative to variability
  - 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

- Effect size for all subjects in single intervention study

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

Example: Effect Size

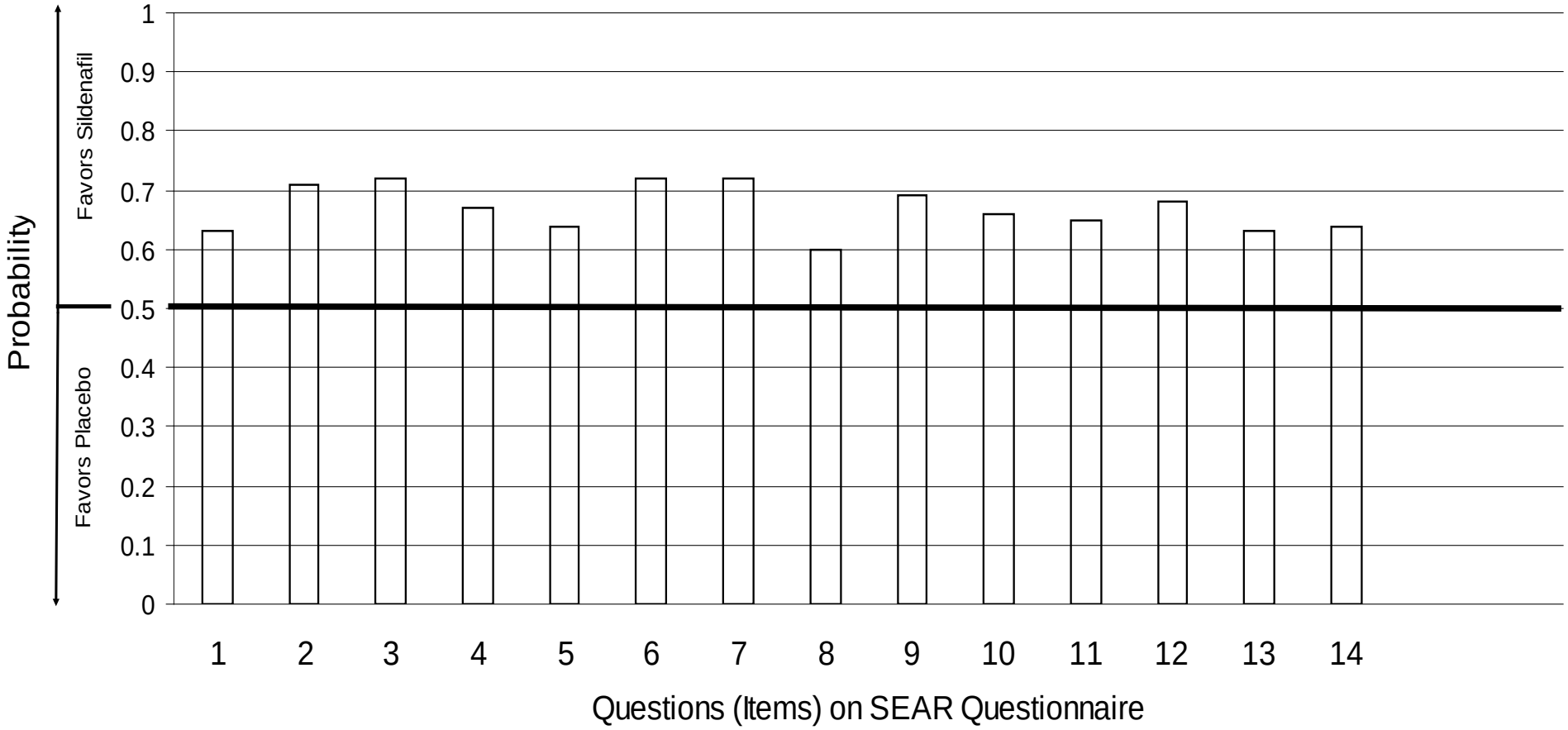
SEAR Component	Baseline Mean ± SD	End Mean ± SD	Difference	Effect Size
Sexual Relationship	42 ± 22	78 ± 21	36 ± 23	1.6
Confidence	55 ± 26	81 ± 21	26 ± 26	1.0
Self-esteem	52 ± 27	81 ± 22	29 ± 28	1.1
Overall Relationship	62 ± 30	80 ± 24	18 ± 32	0.6
Overall	48 ± 22	79 ± 20	31 ± 22	1.4

Source: Althof et al. 2003

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

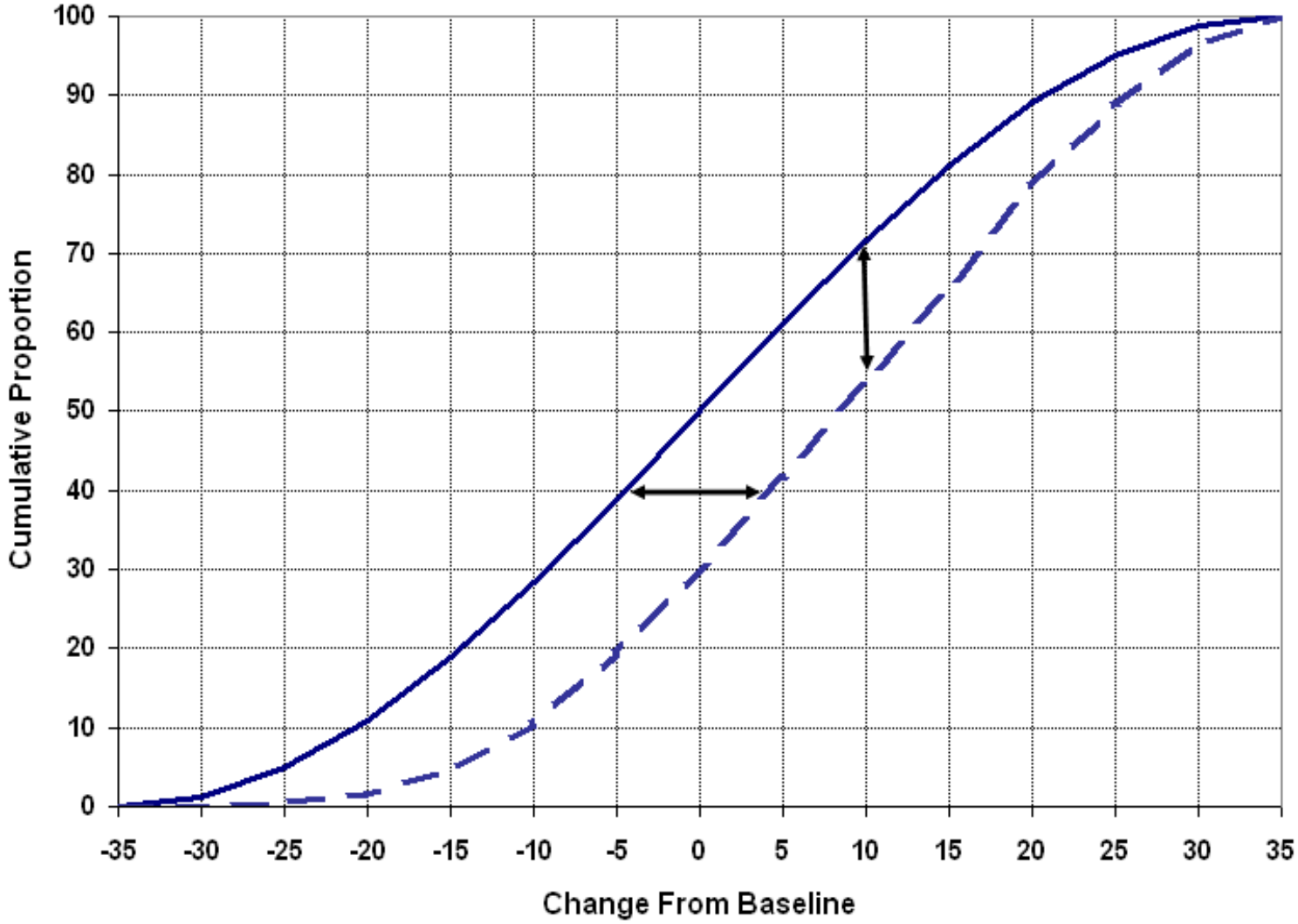


Source: Cappelleri et al. 2007

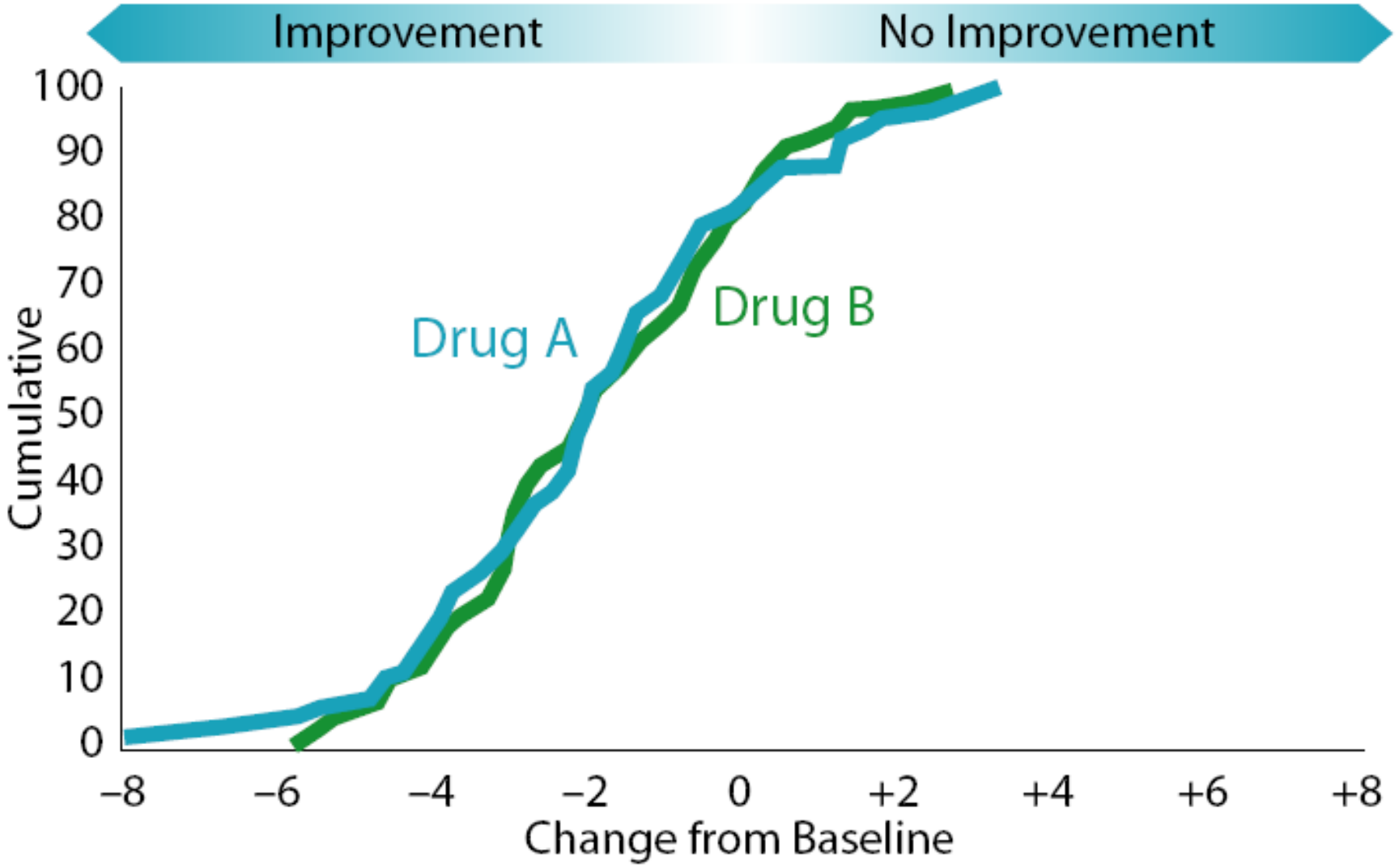
Cumulative Distribution Function

- An alternative or supplement to responder analysis
- Display a continuous plot of the percent change (or absolute 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

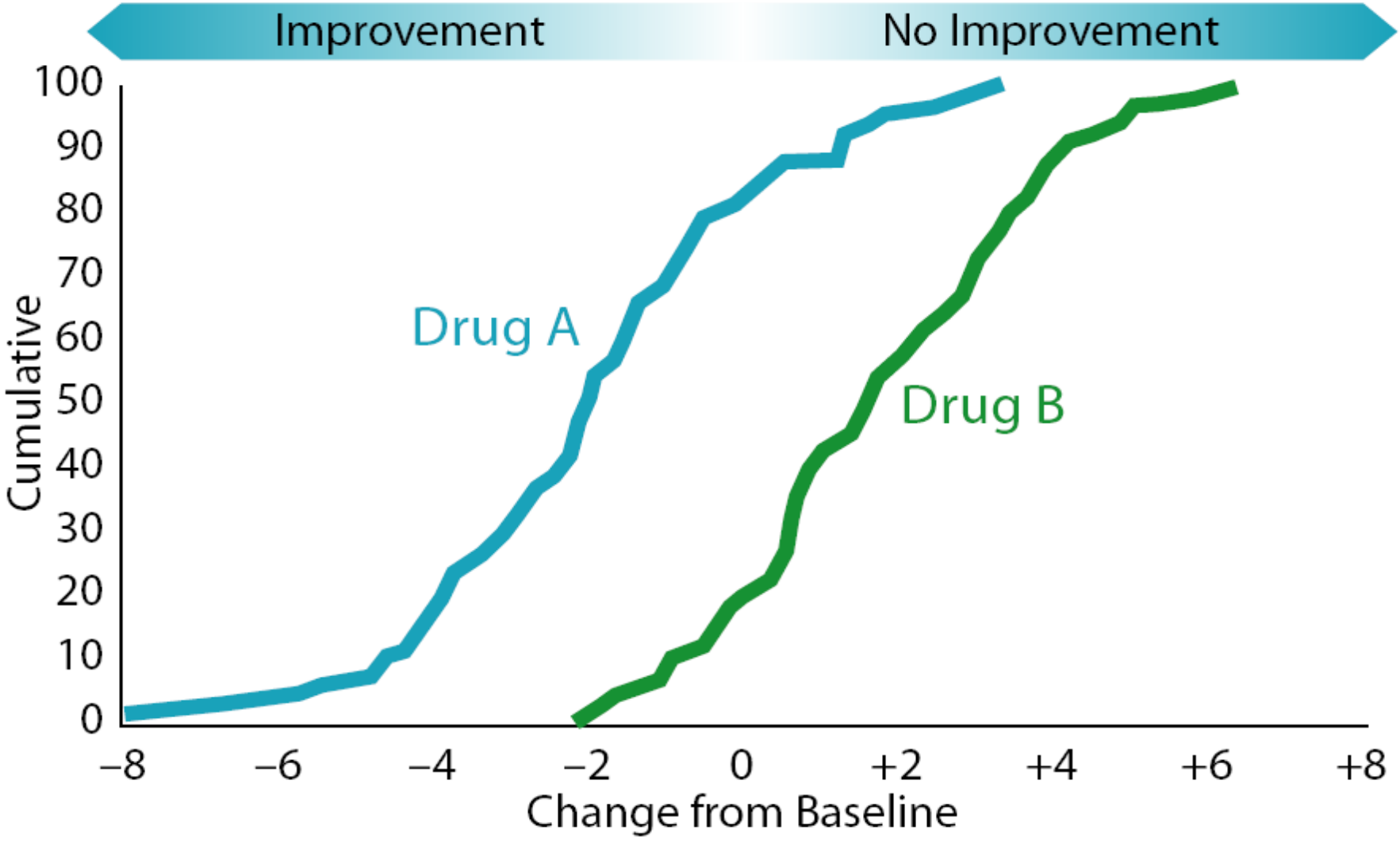
Illustrative Cumulative Distribution Function: Experimental Treatment (solid line) better than Control Treatment (dash line) -- Negative changes indicate improvement



Results showing no comparative efficacy of Drug A or Drug B



Results showing the efficacy of Drug A over Drug B



Aricept® label from 10/13/2006

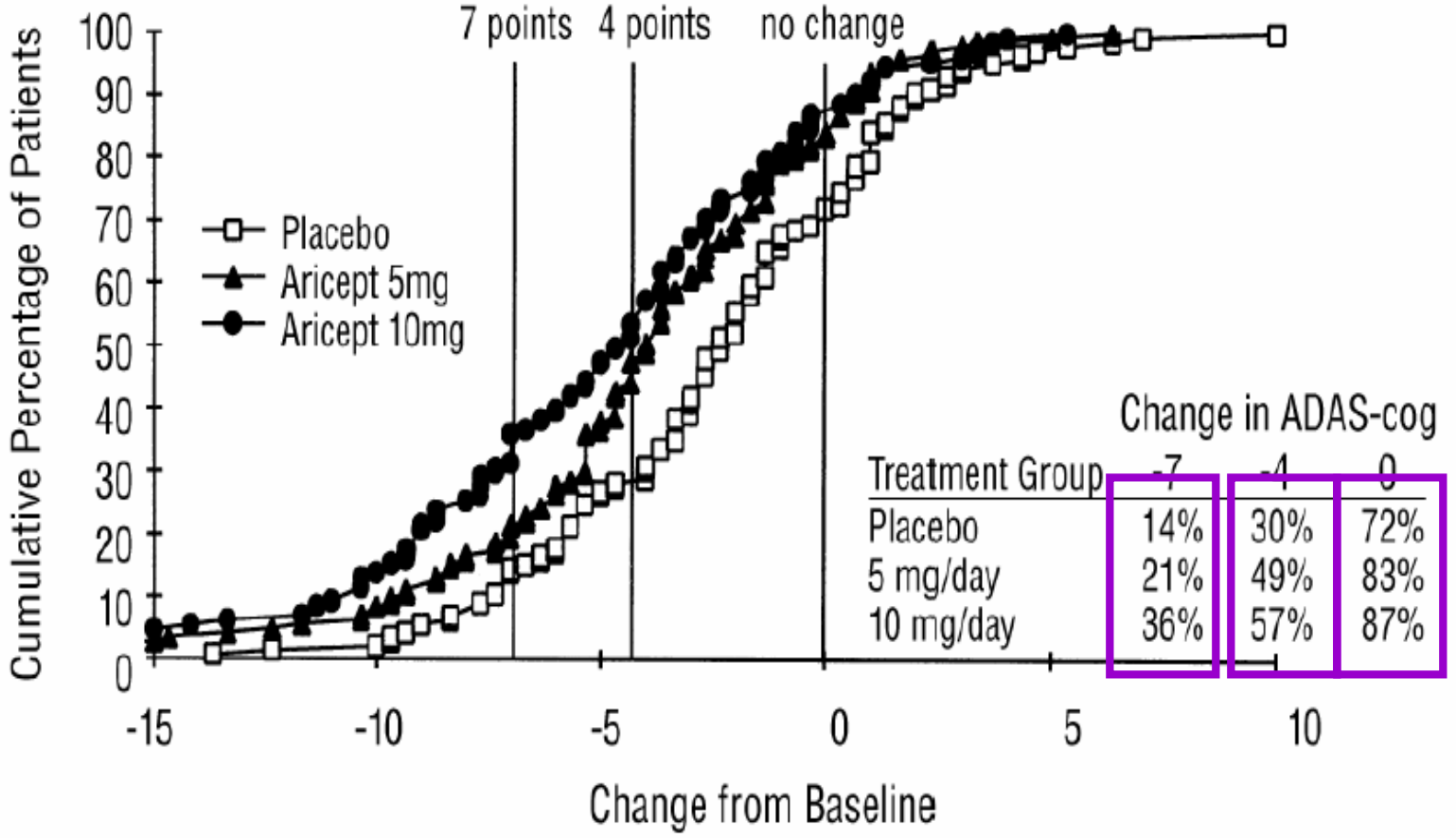


Figure 5. Cumulative Percentage of Patients with Specified Changes from Baseline ADAS-cog Scores. The Percentages of Randomized Patients Within Each Treatment Group Who Completed the Study Were: Placebo 93%, 5 mg/day 90% and 10 mg/day 82%.

Cymbalta® label from 11/19/2009 (x-axis reversed)

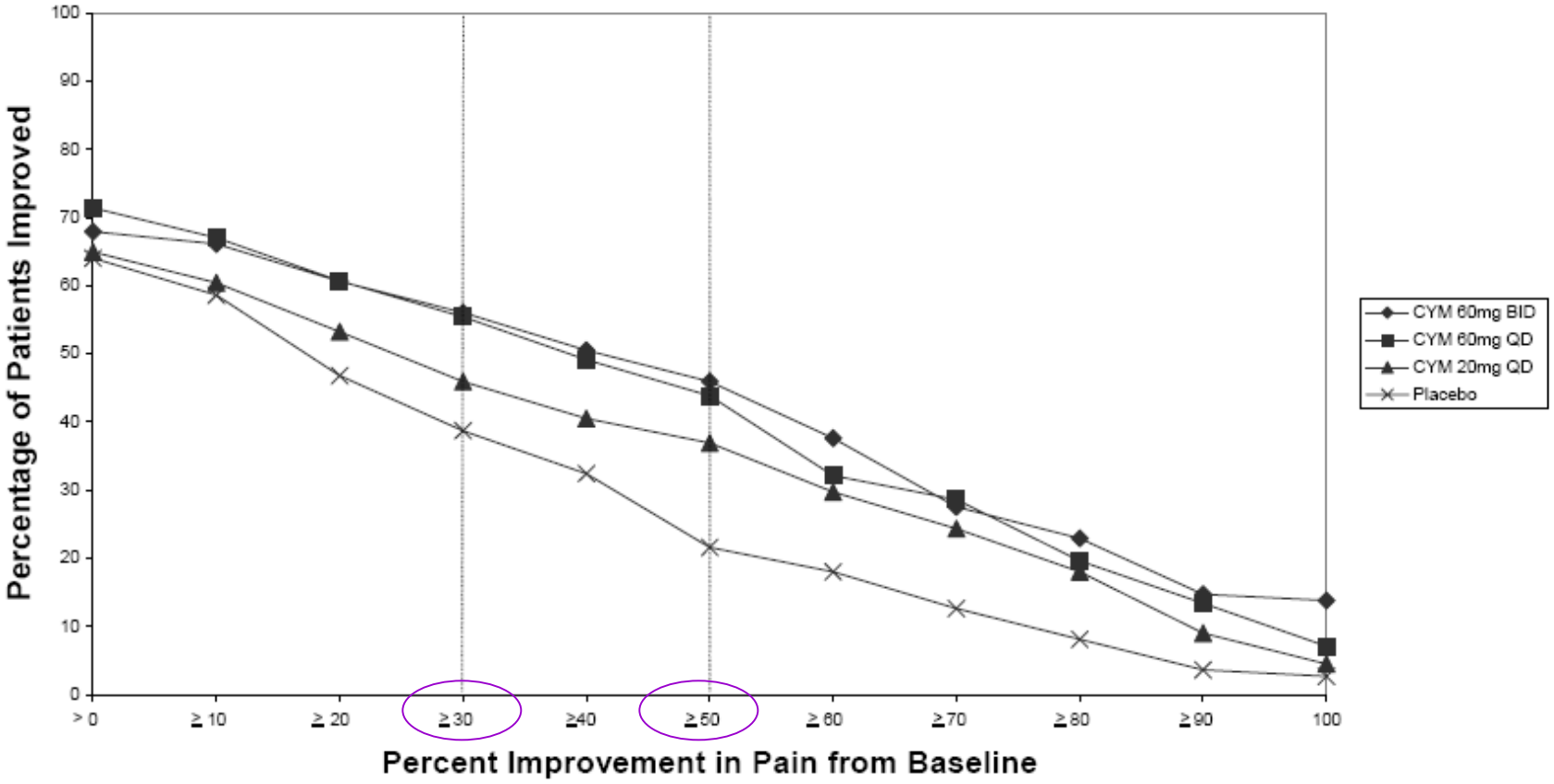
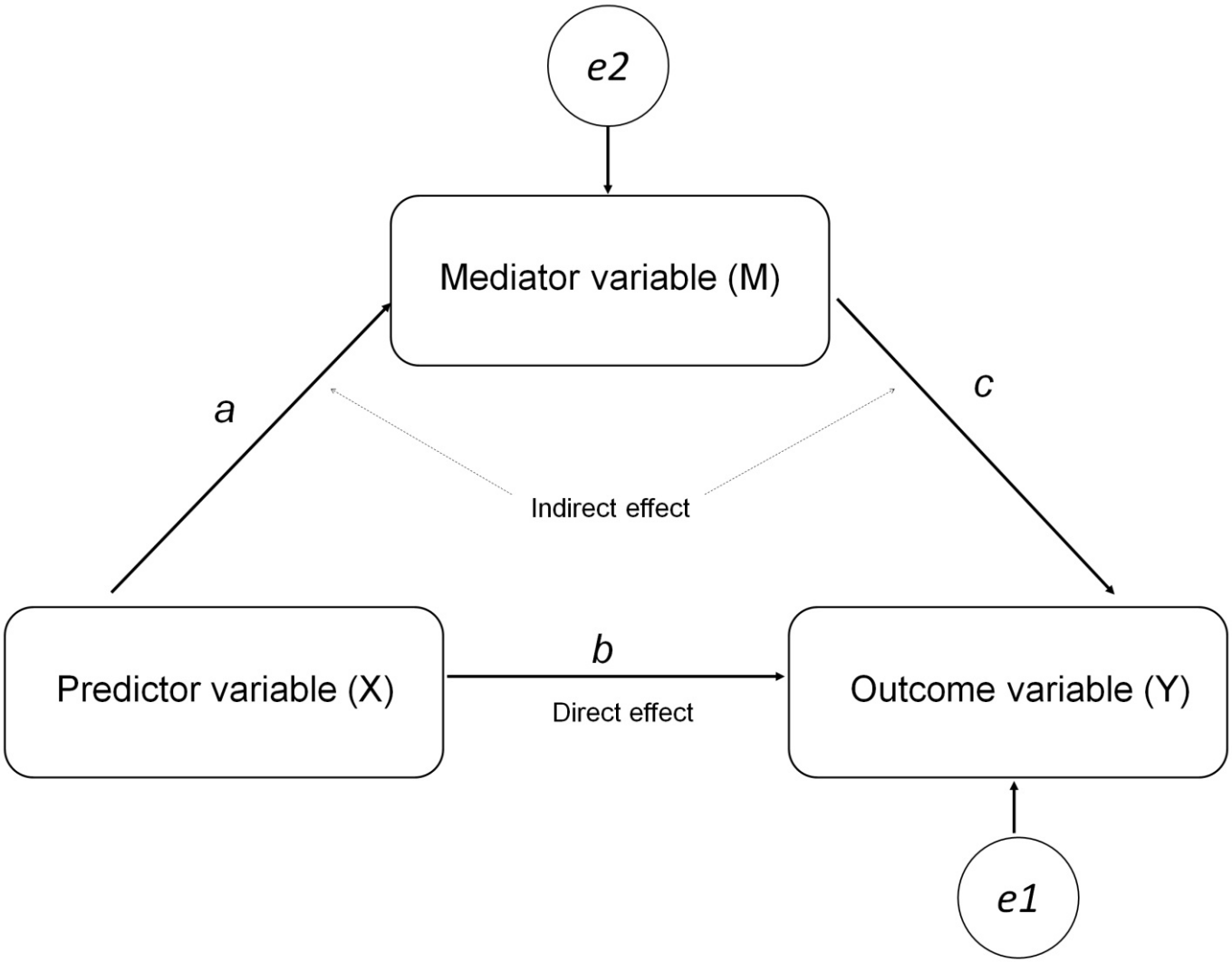


Figure 1: Percentage of Patients Achieving Various Levels of Pain Relief as Measured by 24-Hour Average Pain Severity - Study 1

Mediation Analysis

Basic Mediation Model



A Few Equations

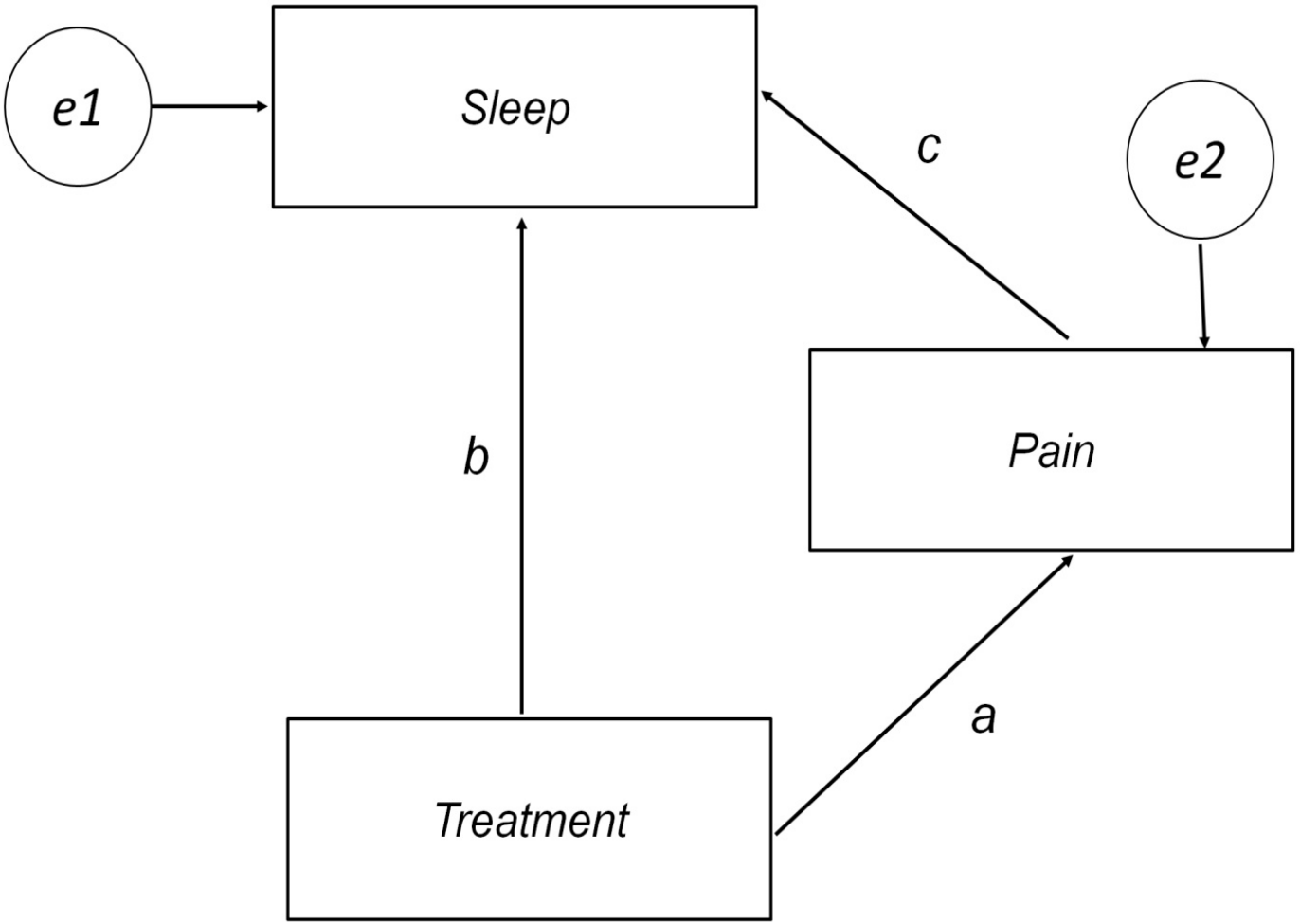
- $Y_j = i_1 + b \times X_j + c \times M_j + e_{1j}$
- $M_j = i_2 + a \times X_j + e_{2j}$

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

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

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

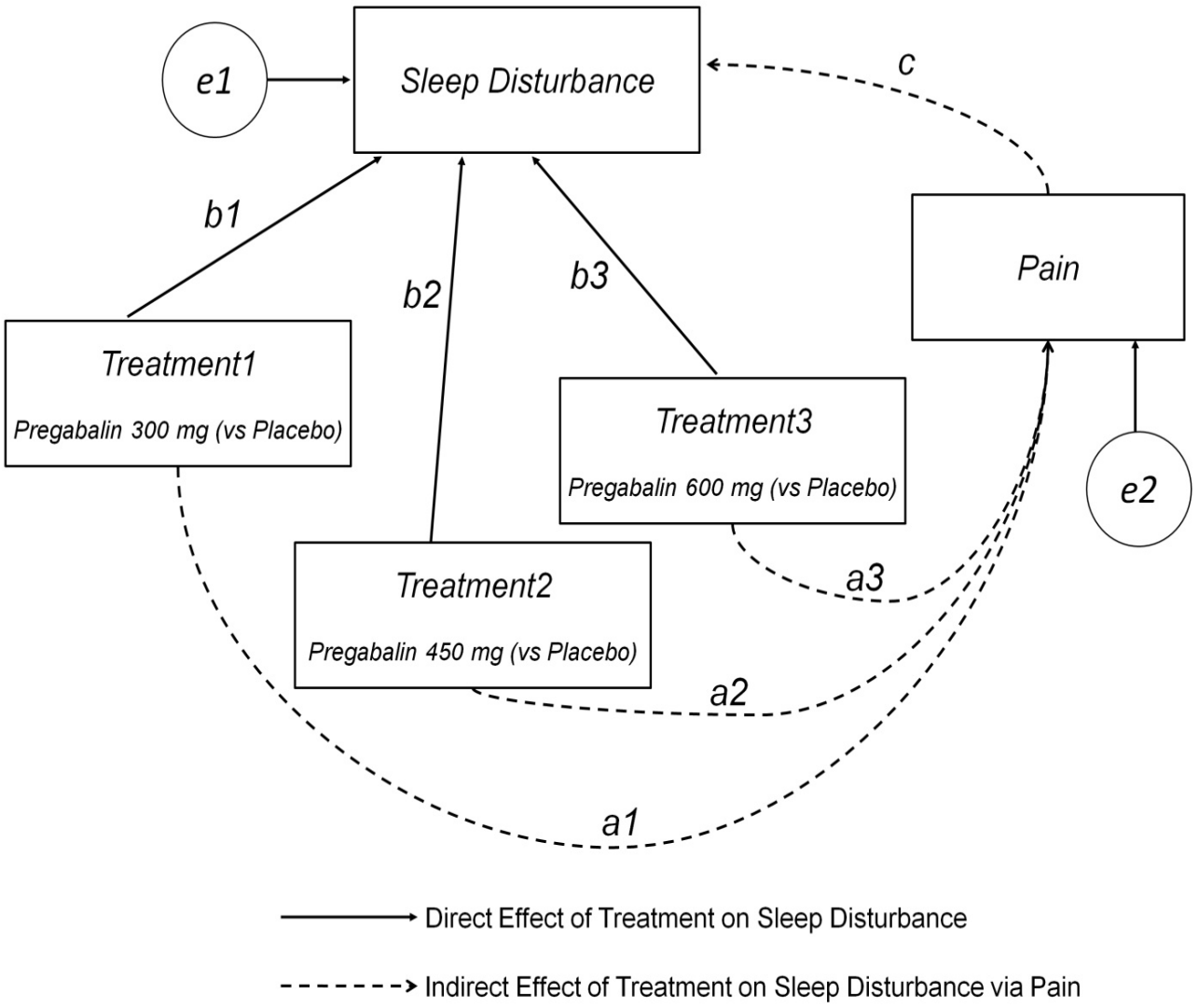
Treatment Affects *Sleep* Directly and Indirectly via *Pain*



Assumptions

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

Published Example



Source: Russell et al. 2009

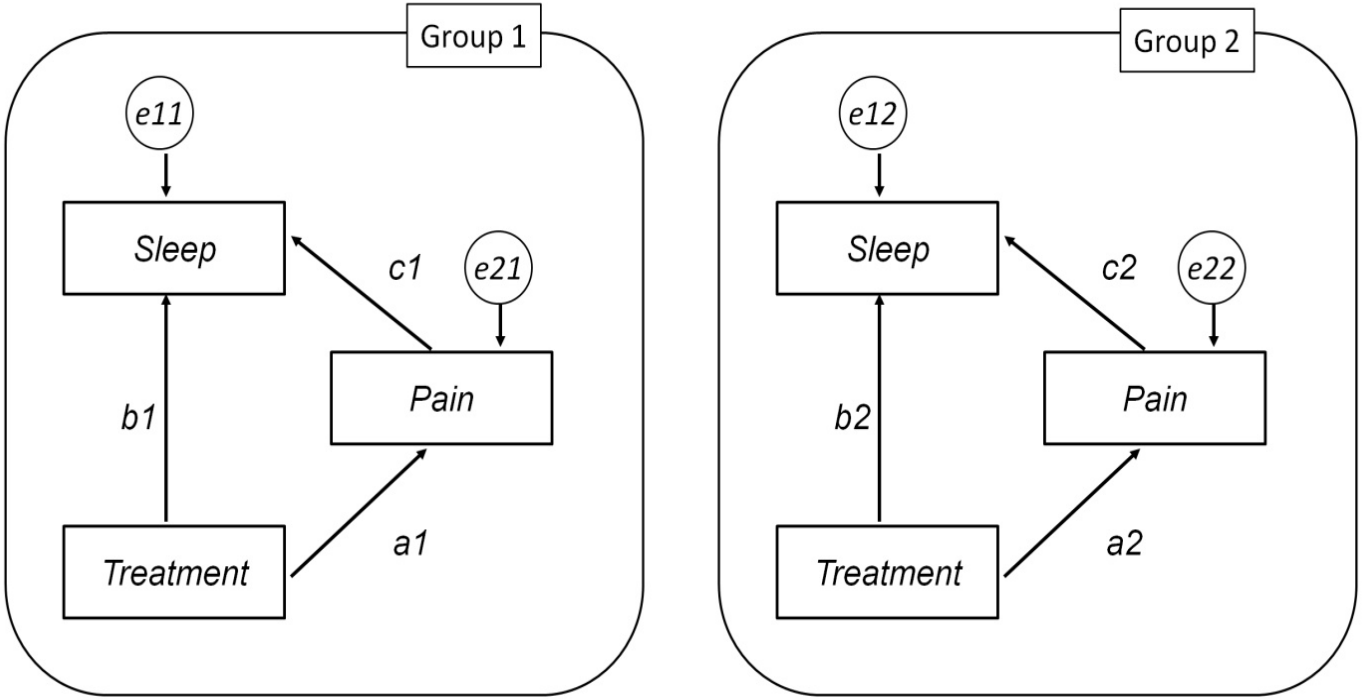
Results

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

Testing for Model Invariance between Groups



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

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

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

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

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
  - Standardized effect size
  - Probability of relative benefit
  - Cumulative distribution function

- Mediation analysis

Journal References: Illustrations Cited

- 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.

- 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.

- 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.

- 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.

- 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.

- Thompson JR, Cappelleri JC, Getter C, Pleil A, Reichel M, Wolf S. 2007. Enhanced interpretation of instrument scales using the Rasch model. *Drug Information Journal* 41:541-550**.**

General References

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

- 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.

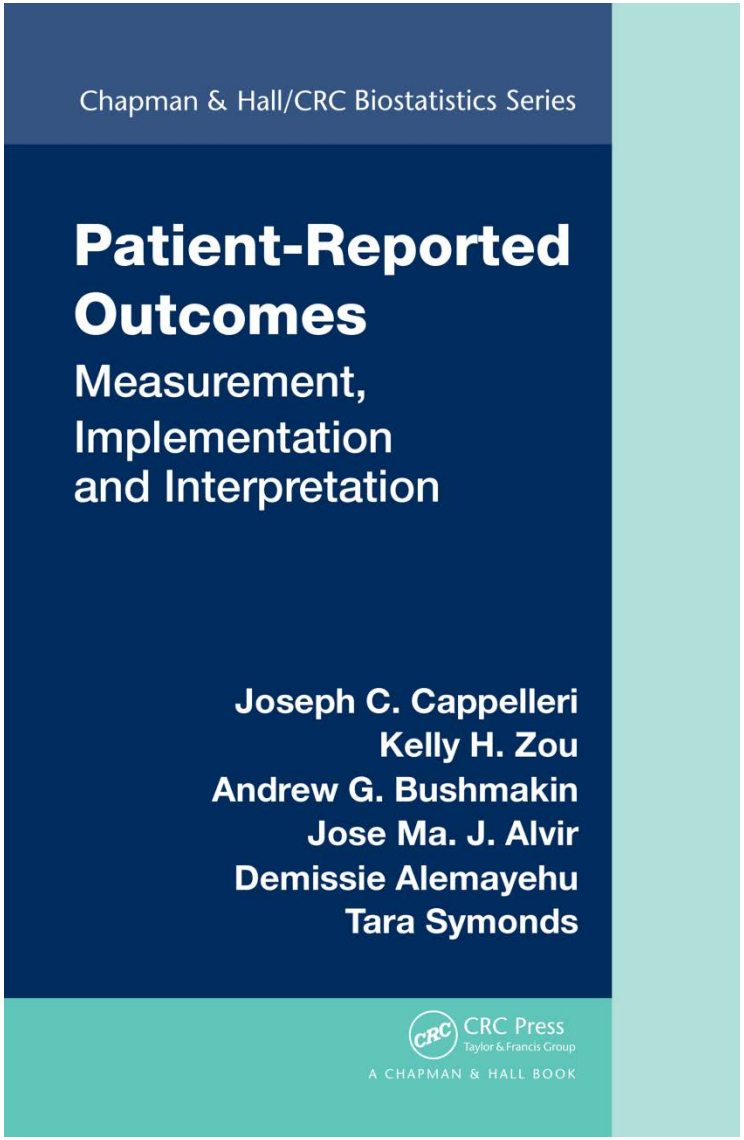
- Coon CD, Cook KF. 2018. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Quality of Life Research* 27:33-40.

- McLeod LD, Cappelleri JC, Hays RD. 2016. Best (but of forgotten) practices: Expressing and interpreting meaning and effect sizes in clinical outcome assessments. *American Journal of Clinical Nutritio*n 103:685-693 (with erratum).

- Revicki D, Erickson PA, Sloan JA, Dueck A, Guess H, Santanello NC and the Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. 2007. Interpreting and reporting results based on patient-reported outcomes. *Value in Health* 10:S116-S124.

- Revicki, D, Hays RD, Cella D, Sloan J. 2008. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology* 61:102-109.

- Sloan JA, Cella D, Hays. 2005. Clinical significance of patient-reported questionnaire data: another step toward consensus. *Journal of Clinical Epidemiology* 58:1217-1219.


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