Interpreting the Results of Patient-reported Outcomes

Gordon Guyatt
McMaster University

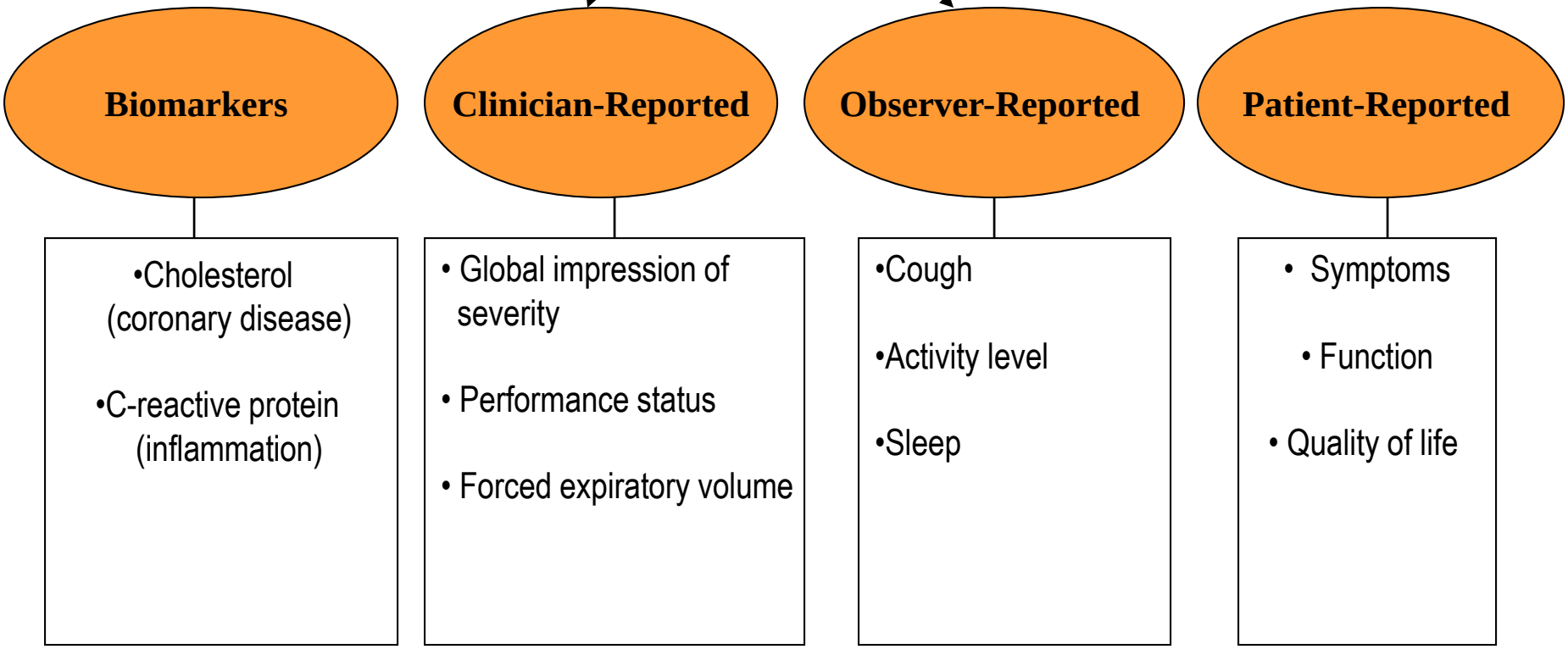
Plan

- Something about you
- What is a patient-important outcome?
- The problem of interpretability
- How to interpret results in individual studies
  - The minimal important difference
- How to interpret results in systematic reviews and meta-analyses

Something about you

- Experience conducting a systematic review
  - If not, plans
- Experience leading a systematic review
  - If not, plans
- Experience continuous outcomes in SR?
- Experience different measures for some outcome
  - Came to find out

Clinical Outcomes Assessment -- Sources and Examples



Survival and Major Morbid Events

Patient-Reported Outcomes (PRO)

- **PRO**: Any report directly from patients, without interpretation by anyone else, about how they function or feel in relation to health condition and therapy (from diaries, questionnaires, interviews, etc.)
- Term PRO requires construct be specified, i.e., respiratory symptoms, physical function, reduction in pain severity
- Almost invariably important to patients
- What PROs have you seen in the literature?

Plan

- What is a patient-important outcome?
- *The problem of interpretability*
- How to interpret results in individual studies
  - The minimal important difference
- How to interpret results in systematic reviews and meta-analyses

Interpretability: The problem

- Mean score for treatment group improves 5 points on the PRO measure, no change in control
- Is this trivial, large, or somewhere between?
- Statistically significant – does that help?

Br J Dermatology, 2004

- Effect of alefacept on quality of life in 553 patients with psoriasis
- Alefacept improved mean Dermatology Quality of Life Scale scores compared with placebo: 4.4 vs. 1.8 at 2 weeks after the last dose ($P<0.0001$)
- Magnitude of Effect?
  - trivial, small but important, large?

Plan

- What is a patient-important outcome?
- The problem of interpretability
- ***How to interpret results in individual studies***
  - ***The minimal important difference***
- How to interpret results in systematic reviews and meta-analyses

Randomized trial of lung volume reduction surgery

- Severe emphysema over inflated
- Reducing lung volume may improve mechanical properties
- RCT of 55 pts followed for 1 year
- Key QOL CRQ
  - Dyspnea, fatigue, motional function
- 1.5 point difference: recommend surgery?
- What could investigators do to help?

Minimally important difference

- Smallest change that patients would consider important
- Approaches
  - Patient scenarios
  - Between-patient ratings
  - Within-patient ratings
    - ***Global ratings of change***
      - ***Are you the same, a little better, a lot better***

Establishing Interpretability

- Chronic Respiratory Questionnaire (CRQ)
- Chronic Heart Failure Questionnaire (CHQ)
- 20 Items
  - Dyspnea
  - Fatigue
  - Emotional Function
- Evidence for validity, responsiveness

1 EXTREMELY SHORT OF BREATH
2 VERY SHORT OF BREATH
3 QUITE A BIT SHORT OF BREATH
4 MODERATE SHORTNESS OF BREATH
5 SOME SHORTNESS OF BREATH
6 A LITTLE SHORTNESS OF BREATH
7 NOT AT ALL SHORT OF BREATH

Establishing Interpretability

- Clinical impression
  - MID 0.5 per question
- 31 patients respiratory rehab program
  - before, 2, 6, 12, and 24 weeks after
- 24 CAL patients in bronchodilator trial
- 20 patient with CHF in digoxin trial

Global rating of change

Overall, has there been any change in your shortness of breath since the last time you saw us?

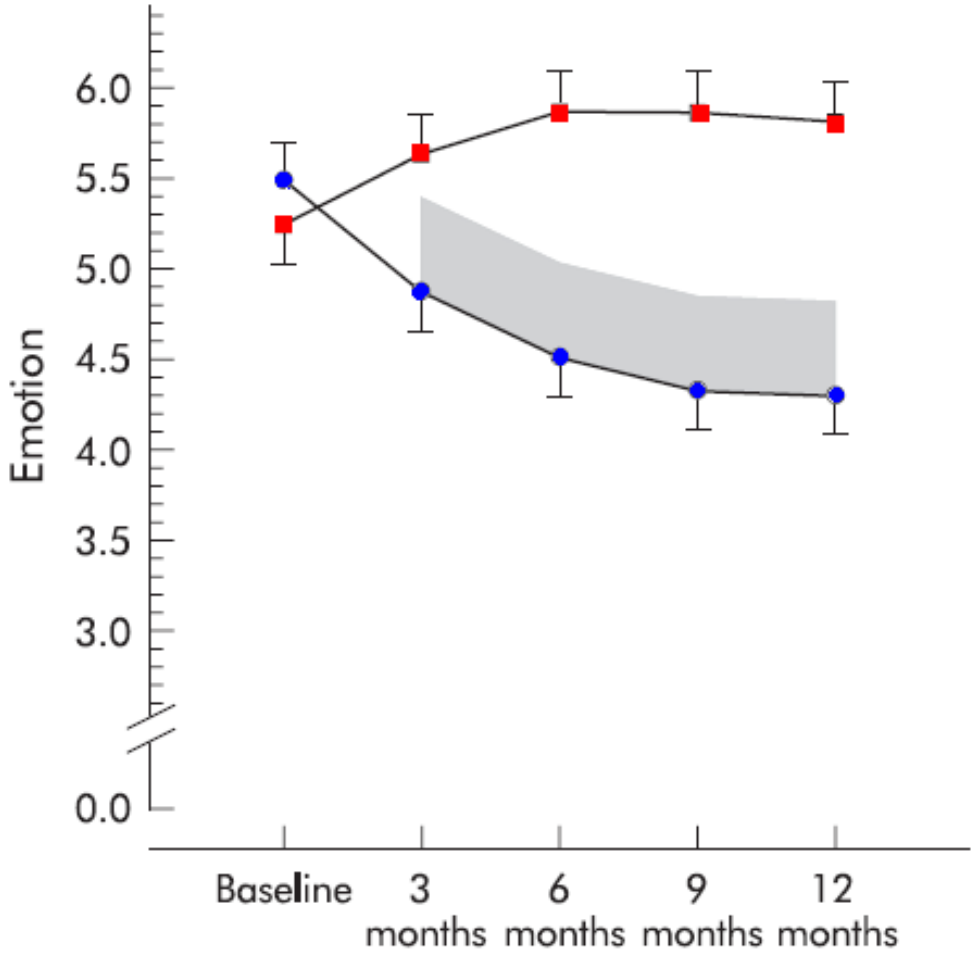
1 WORSE
2 ABOUT THE SAME
3 BETTER

1 ALMOST THE SAME, HARDLY ANY WORSE AT ALL
2 A LITTLE WORSE
3 SOMEWHAT WORSE
4 MODERATELY WORSE
5 A GOOD DEAL WORSE
6 A GREAT DEAL WORSE
7 A VERY GREAT DEAL WORSE

All Trials Combined Mean Change per Question

Global Rating of Change in CRQ and CHQ	Unchanged	**Small Important**	Moderate	Large
Dyspnea	0.10	**0.43**	0.96	1.47
Fatigue	0.03	**0.46**	0.88	0.96
Emotional Function	0.02	**0.49**	0.81	0.86

Effect of Surgery and Medical Control Treatment

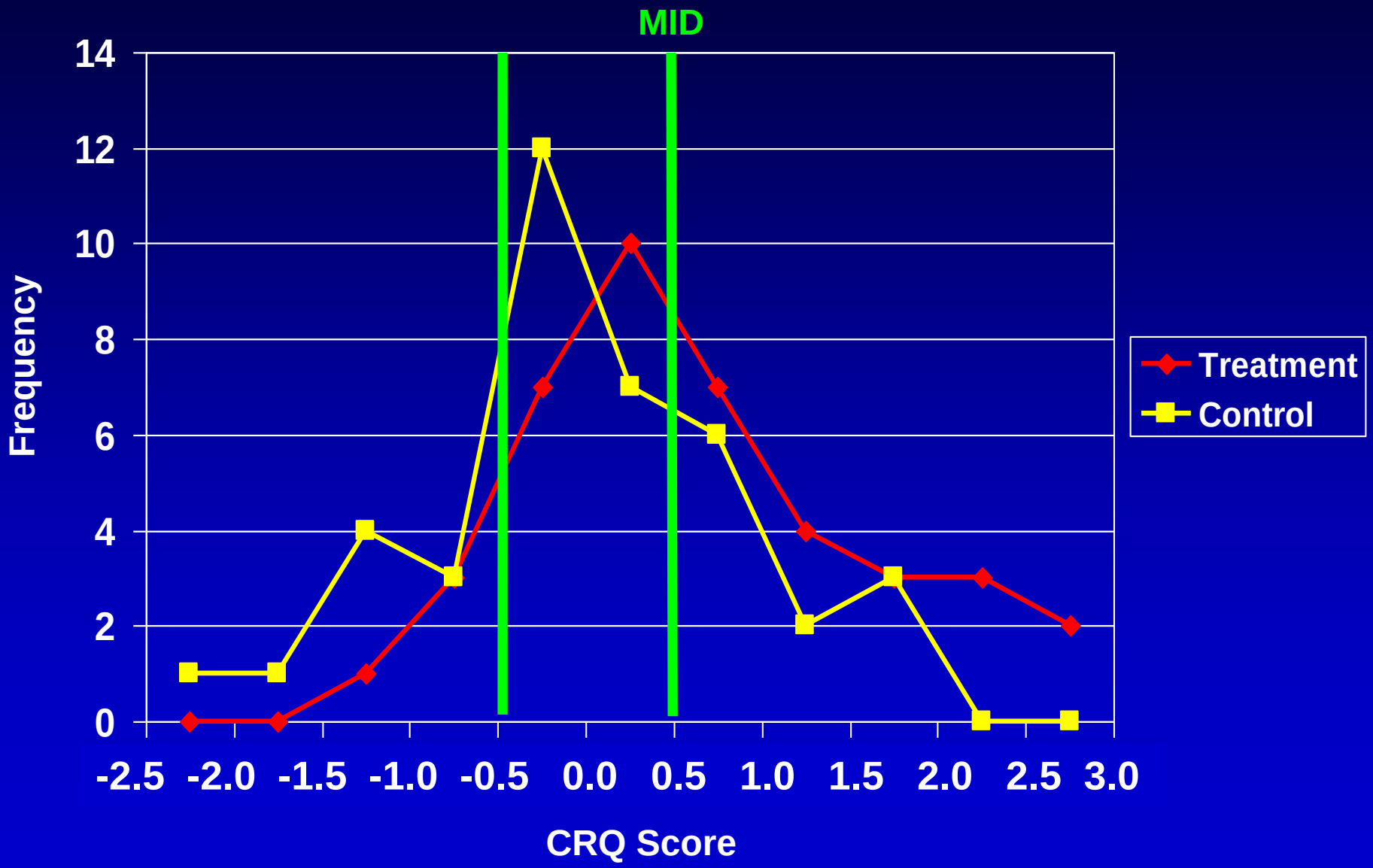


Would you recommend surgery to your patients on the basis of these results?

Interpreting MID Results

- RCT respiratory rehabilitation in COPD
- Assume MID is 0.50 and patients mean improvement vs control is 0.25
- What is your conclusion about rehabilitation?
- Does this mean no one benefits?
- What if 0.6 – everyone benefits?
- If 0.25 mean change could mean:
  - 75% have 0 improvement
  - 25% have 1.0
  - NNT of 4

CRQ Emotion Change Scores



Differences between rehabilitation and conventional care in CAL

CRQ domain	Difference between groups		Estimated proportion better on rehabilitation	Estimated proportion better on conventional care	Proportion benefiting from rehabilitation	NNT for a single patient to benefit
	Mean	P value				
Dyspnoea	0.60	0.0003	0.47	0.28	0.19	5.2
Fatigue	0.45	0.06	0.45	0.23	0.23	4.4
Emotional function	0.40	0.001	0.47	0.17	0.30	3.3

Plan

- What is a patient-important outcome?
- The problem of interpretability
- How to interpret results in individual studies
  - The minimal important difference
- *How to interpret results in systematic reviews and meta-analyses*
  - ***When all studies use the same PROM***

Meta-analysis

- Studies all use same or similar outcome
- Could give weighted mean difference in natural units
- Not intuitively interpretable to the audience
- Solution
  - MID if available
  - range of possible results if not

Systematic review respiratory rehabilitation

CRQ	Point estimate (95% Confidence Interval)
Dyspnea	1.06 (0.85, 1.26)
Emotional Function	0.76 (0.52, 1.00)
Fatigue	0.92 (0.71, 1.13)
Mastery	0.97 (0.74, 1.20)
Overall	0.94 (0.57, 1.32)

Would you recommend respiratory rehabilitation to your patients?

Alternative: dichotomize

- Rankin Stroke Scale
- Five levels
  - no symptoms
  - minor handicap
    - restriction in life style, can look after self
  - moderate handicap
    - restrict life style, prevent independent existence
  - moderately severe handicap
    - clearly prevent independence, no constant attention
  - severe handicap, require constant attention

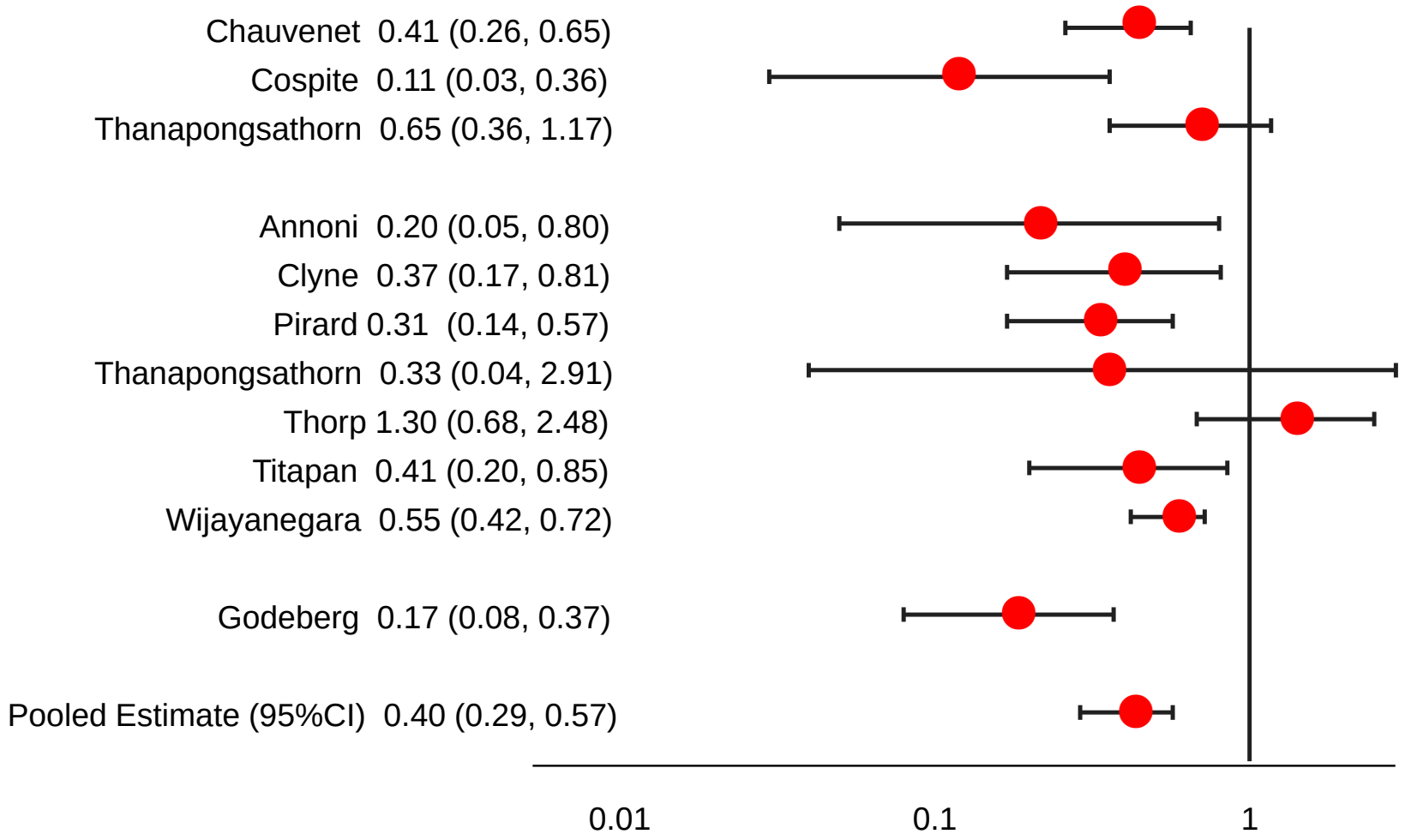
Systematic review of RCTs of thrombolysis in acute stroke

- Use Rankin threshold 2 to 3
  - 2 minor handicap
  - 3 moderate handicap
  - proportion “dead or disabled”
- “Death or dependency”
  - odds ratio 0.84 (95% CI 0.75 to 0.95)
  - 4% absolute risk reduction
  - NNT 25

Flavanoids for Hemorrhoids

- Venotonic agents
  - Increase venous return
- Popularity
  - 90 venotonics commercialized in France
  - None in Sweden and Norway
  - France 70% of world market
- Possibilities
  - French misguided, rest of world missing out
- Key outcome
  - Risk not improving/persistent symptoms
  - 11 studies, 1002 patients, 375 events

Phlebotonics for Hemorrhoids (Venotonics vs. Placebo)
Relative Risk (95%CI)



Plan

- What is a patient-important outcome?
- The problem of interpretability
- How to interpret results in individual studies
  - The minimal important difference
- *How to interpret results in systematic reviews and meta-analyses*
  - ***When studies use different PROMs for the same construct***

Plan

- When studies use different PROMS
  - Standardized mean difference
  - Natural units
  - Dichotomize
  - Ratio of means
  - MID units

Do clinicians understand treatment effects?

- Cross-sectional, paper-based survey
  - Academic centers in 8 countries,
  - Internal and family medicine, 531/610 (87%)
- Summary estimates hypothetical interventions vs placebo chronic pain

Clinicians' understanding

In units of the pooled standard deviation of all the pain scores in the opioid and control groups, expressed as a standardized mean difference, a meta-analysis finds the effect of intervention A vs placebo control for patient-reported pain is 0.20 standard deviation units in favor of intervention A. The magnitude of this difference is:

- ☐ trivial difference, probably not important
- ☐ small difference, but probably important
- ☐ moderate difference, surely important
- ☐ large difference, very important

In units of the pooled standard deviation of all the pain scores in the opioid and control groups, expressed as a standardized mean difference, a meta-analysis finds the effect of intervention A vs placebo control for patient-reported pain is 0.80 standard deviation units in favor of intervention A. The magnitude of this difference is:

- ☐ trivial difference, probably not important
- ☐ small difference, but probably important
- ☐ moderate difference, surely important
- ☐ large difference, very important

In units of the pain scale, where 0 represents no pain and 10 represents the worst pain ever on numeric rating scale, a meta-analysis finds the effect of intervention C vs placebo control for patient-reported pain to be 2.0 in favor of intervention C. The magnitude of difference is:

- ☐ trivial difference, probably not important
- ☐ small difference, but probably important
- ☐ moderate difference, surely important
- ☐ large difference, very important

In units of the pain scale, where 0 represents no pain and 10 represents the worst pain ever on numeric rating scale, a meta-analysis finds the effect of intervention C vs placebo control for patient-reported pain to be 0.60 in favor of intervention C. The magnitude of difference is:

- ☐ trivial difference, probably not important
- ☐ small difference, but probably important
- ☐ moderate difference, surely important
- ☐ large difference, very important

As a relative risk (or risk ratio), where the ratio of the risk of the event occurring in the treatment group is divided by the risk in the placebo control group, a meta-analysis finds the effect of intervention D vs placebo control for patient-reported pain to be 0.80 (80%) in favor of intervention D, representing a relative risk reduction of 20% (relative to the control group, 20% fewer patients experience moderate to severe pain). The magnitude of difference is:

- ☐ trivial difference, probably not important
- ☐ small difference, but probably important
- ☐ moderate difference, surely important
- ☐ large difference, very important

As a relative risk (or risk ratio), where the ratio of the risk of the event occurring in the treatment group is divided by the risk in the placebo control group, a meta-analysis finds the effect of intervention D vs placebo control for patient-reported pain to be 0.50 (50%) in favor of intervention D, representing a relative risk reduction of 50% (relative to the control group, 50% fewer patients experience moderate to severe pain). The magnitude of difference is:

- ☐ trivial difference, probably not important
- ☐ small difference, but probably important
- ☐ moderate difference, surely important
- ☐ large difference, very important

As a risk difference, where the risk of in the treatment group is subtracted from the risk in the placebo control group, a meta-analysis finds the effect of intervention E vs placebo control for patient-reported pain to be 0.20 (20%) in favor of intervention E, representing 20 fewer patients per 100 experiencing moderate to severe pain. The magnitude of difference is:

- ☐ trivial difference, probably not important
- ☐ small difference, but probably important
- ☐ moderate difference, surely important
- ☐ large difference, very important

Plan

- When studies use different outcomes
  - ***Standardized mean difference***
  - Natural units,
  - Dichotomize,
  - ratio of means,
  - MID units

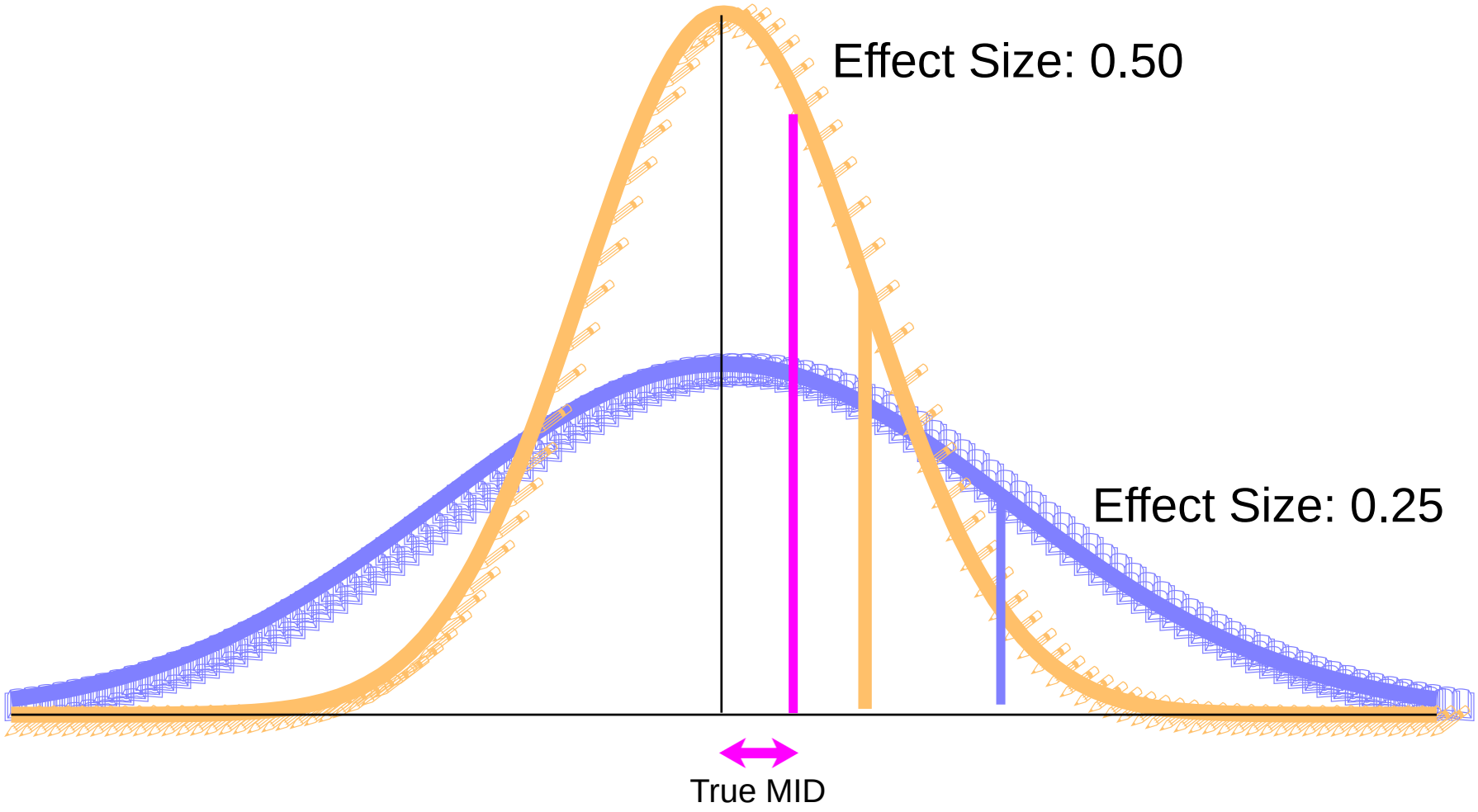
Studies different instruments

- CRQ is one QoL measure for CAL
- St. George's respiratory questionnaire another
- Some studies use one and some other?
  - What now?
- Divide each effect by standard deviation
- Ultimate result in SD units
- "Effect size" or SMD
- Study shows effect size of 0.4
  - Trivial, small but important, medium or large effect?

Cohen:
small effect 0.2 SD units
moderate effect 0.5
large effect 0.8

More recent suggestions in terms of MID
across all instruments 0.5 or 0.35

Rules of thumb likely to be limited


Effect Size: 0.50
Effect Size: 0.25
True MID

Results – SD Units

Study or Subgroup	Experimental Mean	SD	Total	Control Mean	SD	Total	Weight	Std. Mean Difference IV, Random, 95% CI
2.1.1 SGRQ								
Boxall 2005	5.8	11.8	23	1.4	13.3	24	6.8%	0.34 [-0.23, 0.92]
Chlumsky 2001	4.07	19.76	13	4.22	19.2	6	3.9%	-0.01 [-0.97, 0.96]
Engstrom 1999	-0.3	17.3	26	-0.5	16.2	24	7.0%	0.01 [-0.54, 0.57]
Finnerty 2001	9.3	12.2	24	2.2	15	25	6.9%	0.51 [-0.06, 1.08]
Ringbaek 2000	2.1	19	17	2.2	17	19	6.1%	-0.01 [-0.66, 0.65]
2.1.2 CRQ								
Behnke 2000	1.9	0.7	15	-0.07	1.1	15	4.2%	2.08 [1.17, 2.99]
Cambach 2004	1.04	0.91	15	0.01	0.75	8	4.1%	1.15 [0.22, 2.09]
Goldstein 2004	0.43	0.92	40	-0.13	0.75	40	8.1%	0.66 [0.21, 1.11]
Gosselink 2000	0.67	1.02	34	-0.1	1.11	28	7.4%	0.72 [0.20, 1.23]
Griffiths 2000	0.97	1	93	-0.15	0.9	91	9.6%	1.17 [0.86, 1.49]
Guell 1995	0.98	1.01	29	-0.18	1.05	27	6.9%	1.11 [0.55, 1.68]
Guell 1998	0.45	0.89	18	-0.3	0.97	17	5.8%	0.79 [0.10, 1.48]
Hernandez 2000	0.86	1	20	0.14	1.03	17	6.0%	0.69 [0.03, 1.36]
Simpson 1992	0.86	1.26	14	0.13	1.11	14	5.2%	0.60 [-0.16, 1.36]
Singh 2003	0.91	0.75	20	0.1	0.68	20	6.0%	1.11 [0.44, 1.78]
Wijkstra 1994	0.8	0.83	28	0.07	0.82	15	6.1%	0.87 [0.21, 1.52]
Total (95% CI)			**429**			**390**	**100.0%**	**0.73 [0.49, 0.96]**

Heterogeneity: $Tau^2 = 0.13$; $Chi^2 = 35.82$, df = 15 (P = 0.002); $I^2 = 58\%$

Test for overall effect: Z = 6.04 (P < 0.00001)

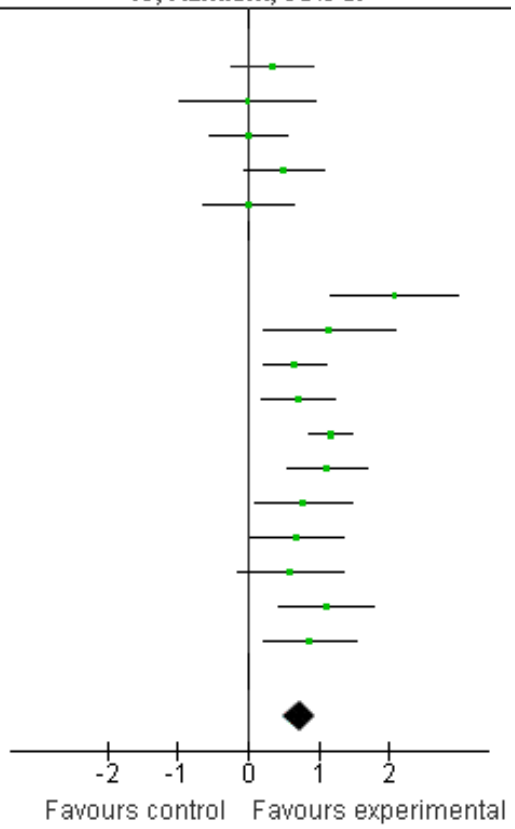


Table 5: Application of approaches to chronic respiratory rehabilitation for health-related quality of life impairment in patients with chronic airflow limitation

Outcomes	Estimated baseline score/proportion improving in control patients	Absolute increase in proportion improving in patients receiving respiratory rehabilitation	Relative Effect (95% CI)	Number of Participants (studies)	Confidence in effect estimate[1]	Comments
(A) Health-related quality of life (HRQL) Investigators measured HRQL using different instruments. Higher scores mean better HRQL.	The HRQL score in the respiratory rehabilitation group improved on average **0.72 (95% CI 0.48 to 0.96) SDs** more in the respiratory rehabilitation patients than in control patients		---	818 (16)	⊕⊕⊕⊕ High	As a rule of thumb, 0.2 SD represents a small difference, 0.5 moderate, and 0.8 large

Plan

- When studies use different outcomes
  - Standardized mean difference
  - ***Natural units***
  - Dichotomize
  - Ratio of means
  - MID units

Conversion to familiar units

- All instruments into most familiar
  - Two statistical approaches
- Rescale to units of most familiar
  - St. George's 0 to 100
  - Divide by 7 to go to CRQ units

(B) Health-related quality of life (HRQL) measured on a scale of 1 to 7	Control group baseline 4.5[1] Average improvement in control 0.04	HRQL improved on average **0.71 (95% CI 0.48 to 0.94)** more in the respiratory rehabilitation patients than in the control patients	---	818 (16)	⨁⨁⨁⨁ High	Calculated by transforming all scores to the Chronic Respiratory Questionnaire in which the minimal important difference is 0.5

- Confident encourage
- Possibly encourage
- Probably discourage
- Certainly discourage

What if mean difference 0.4

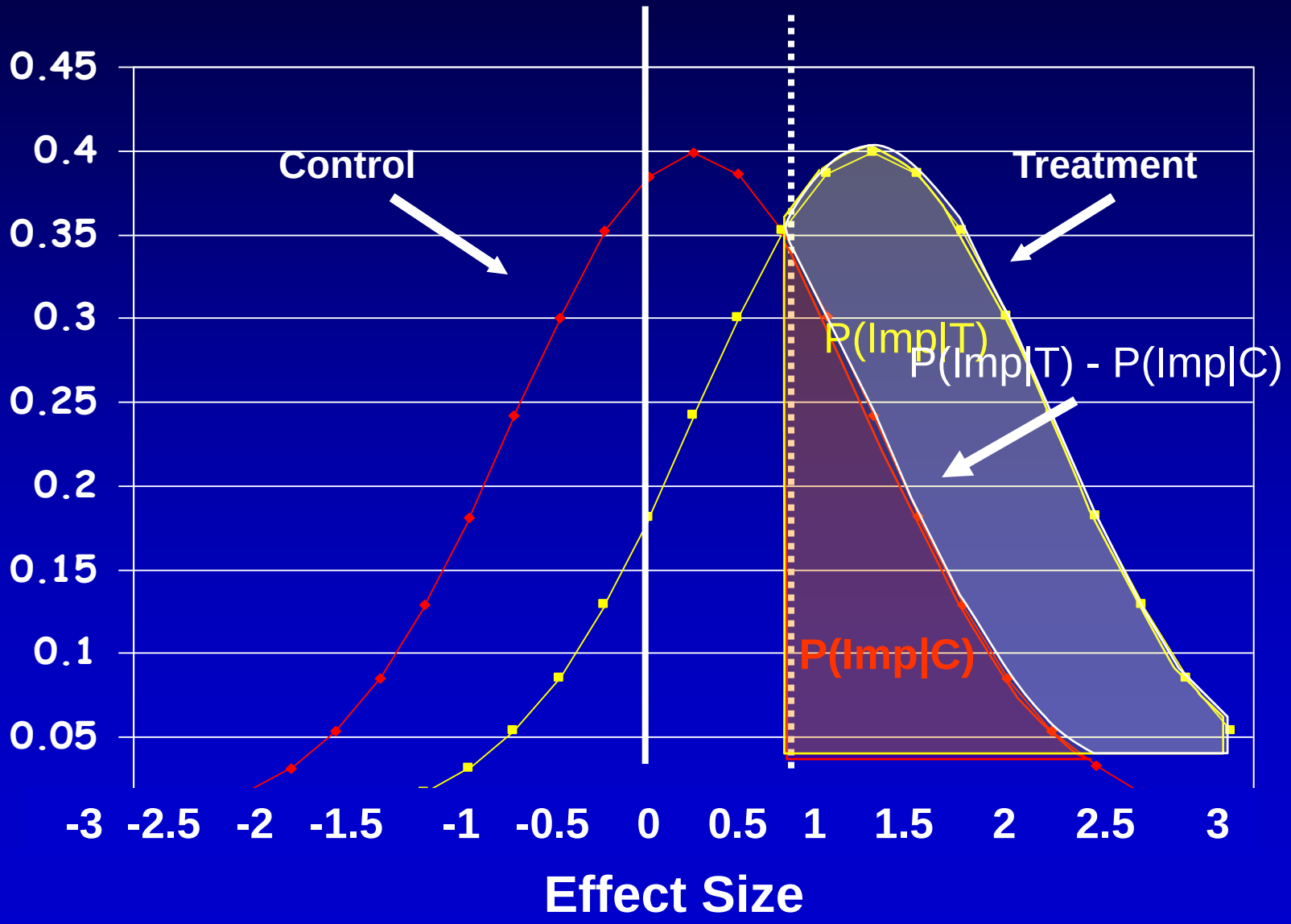
Vulnerable to no one benefits/everyone benefits

Plan

- When studies use different outcomes
  - Standardized mean difference
  - Natural units
  - ***Dichotomize***
  - Ratio of means
  - MID units

Dichotomize

Assume standard symmetrical distribution
Assume equal variance in intervention and control groups



Dichotomize

- Yields relative and absolute effects
- 50% RRR in number of patients severe pain
  - Big or small effect?
- Can't tell
  - Could be reduction from 2% to 1%
  - Or reduction from 40% to 20%

(C) Proportion of patients with important improvement in health-related quality of life (HRQL)	**0.30**[2]	Differences in proportion achieving important improvement **0.31 (95% CI 0.22 to 0.40)** in favor of rehabilitation	**OR=3.36 (95% CI 2.31 to 4.86)**	818 (16)	⊕⊕⊕⊕ High	Calculation uses established minimal important difference of 0.5 units on the CRQ and 4 units on the St. George's Respiratory Questionnaire

- Confident encourage
- Possibly encourage
- Probably discourage
- Certainly discourage

Plan

- When studies use different outcomes
  - Standardized mean difference
  - Natural units
  - Dichotomize
  - ***Ratio of means***
  - MID units

Ratio of means

$$\text{RoM} = \frac{\text{mean}_{exp}}{\text{mean}_{control}}.$$

- Can tell us for instance:
  - Treatment had 30% less pain than control
- Analogous to relative risk
  - Greater absolute difference with greater control risk

Plan

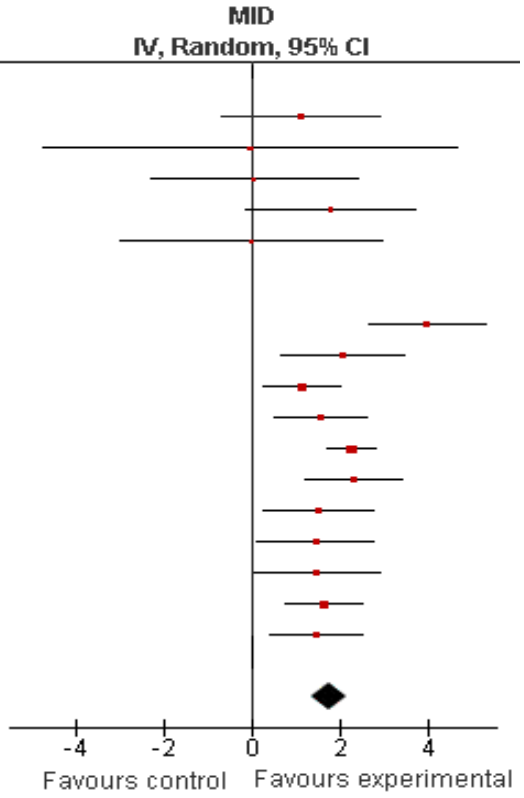
- When studies use different outcomes
  - Standardized mean difference
  - Natural units
  - Dichotomize
  - Ratio of means
  - ***MID units***

Results – MID Units

Study or Subgroup	MID	SE	Experimental Total	Control Total	Weight	MID IV, Random, 95% CI
1.3.1 SGRQ						
Boxall 2005	1.1	0.926	23	23	3.7%	1.10 [-0.71, 2.91]
Chlumsky 2001	-0.0375	2.391	13	6	0.6%	-0.04 [-4.72, 4.65]
Engstrom 1999	0.05	1.184	26	24	2.4%	0.05 [-2.27, 2.37]
Finnerty 2001	1.775	0.974	24	25	3.4%	1.77 [-0.13, 3.68]
Ringbaek 2000	-0.025	1.509	17	17	1.5%	-0.03 [-2.98, 2.93]
1.3.2 CRQ						
Behnke 2000	3.96	0.683	15	15	5.9%	3.96 [2.62, 5.30]
Cambach 2004	2.06	0.713	15	8	5.5%	2.06 [0.66, 3.46]
Goldstein 2004	1.12	0.445	40	40	10.1%	1.12 [0.25, 1.99]
Gosselink 2000	1.545	0.545	34	28	8.0%	1.54 [0.48, 2.61]
Griffiths 2000	2.25	0.281	93	91	14.9%	2.25 [1.70, 2.80]
Guell 1995	2.3	0.553	29	27	7.9%	2.30 [1.22, 3.38]
Guell 1998	1.5	0.63	18	17	6.6%	1.50 [0.27, 2.73]
Hernandez 2000	1.445	0.674	20	17	6.0%	1.45 [0.12, 2.77]
Simpson 1992	1.465	0.73	14	14	5.3%	1.47 [0.03, 2.90]
Singh 2003	1.63	0.452	20	20	10.0%	1.63 [0.74, 2.52]
Wijkstra 1994	1.45	0.537	28	15	8.2%	1.45 [0.40, 2.50]
Total (95% CI)			**429**	**387**	**100.0%**	**1.75 [1.37, 2.13]**

Heterogeneity: $Tau^2 = 0.17$; $Chi^2 = 22.15$, df = 15 (P = 0.10); $I^2 = 32\%$

Test for overall effect: Z = 9.00 (P < 0.00001)



Do clinicians understand treatment effects?

- Cross-sectional, paper-based survey
  - Academic centers in 8 countries,
  - Internal and family medicine, 531/610 (87%)
- Summary estimates hypothetical interventions vs placebo chronic pain

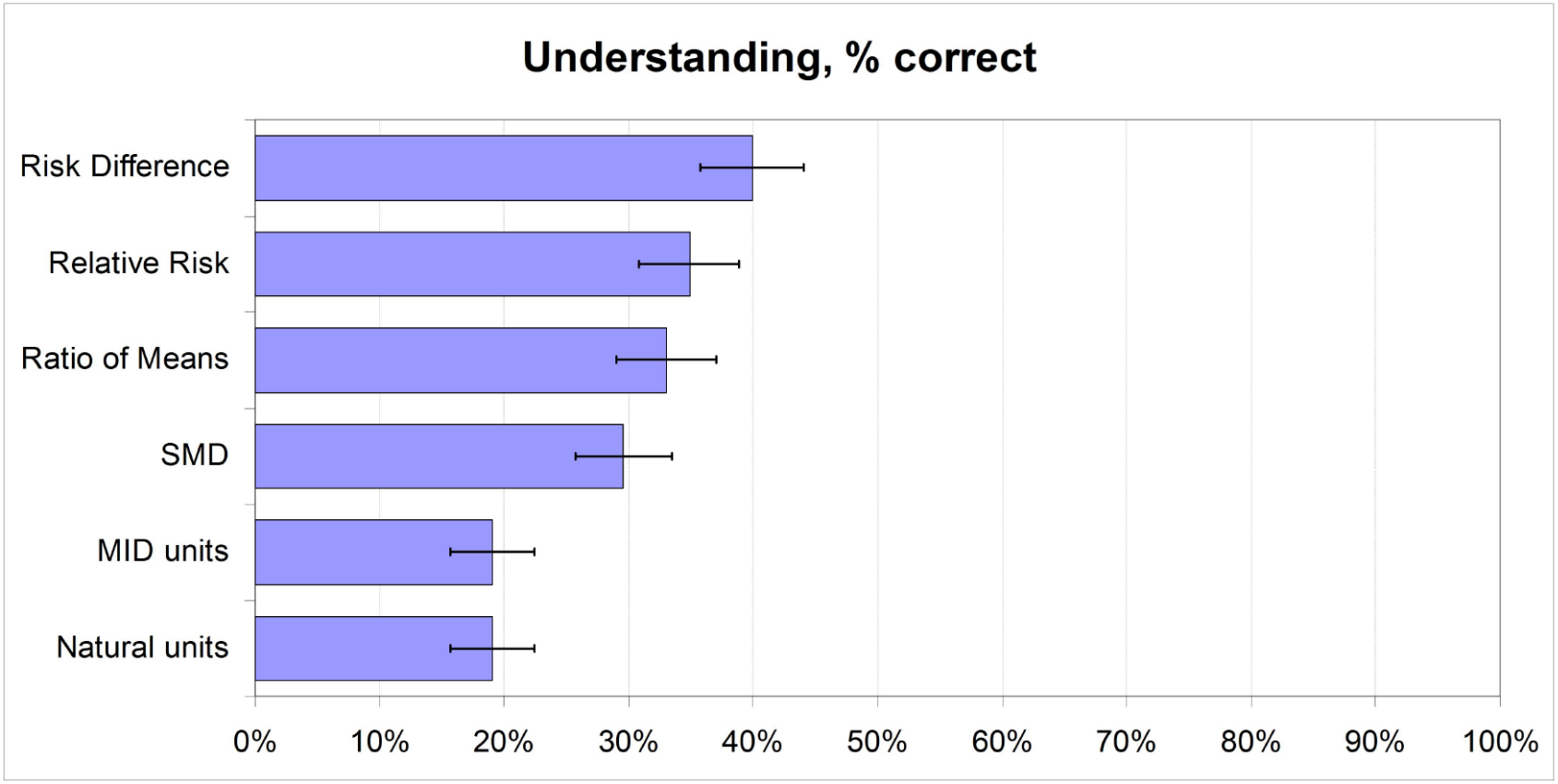
Do clinicians understand treatment effects?

- Objective: determine clinicians understanding and perspective of 6 approaches to presenting continuous outcomes
  - Standardized Mean Difference
  - Natural units
  - Ratio of Means
  - Relative Risk
  - Absolute Risk
  - Ratio of Means
  - MID Units
- Random assign 1 of 4 versions, differing magnitude of effect and presentation order

Presentation method	Small effect	Large effect
SMD	**0.20**	**0.80**
MD in natural units	**0.60**	**2.00**
RR	**0.80**	**0.50**
RD	**0.04**	**0.20**
RoM	**0.92** (8% less)	**0.63** (37% less)
MID units	**0.60**	**2.00**

Results: Correct answers

Figure 3: Understanding of the presentation approaches, *n* = 531

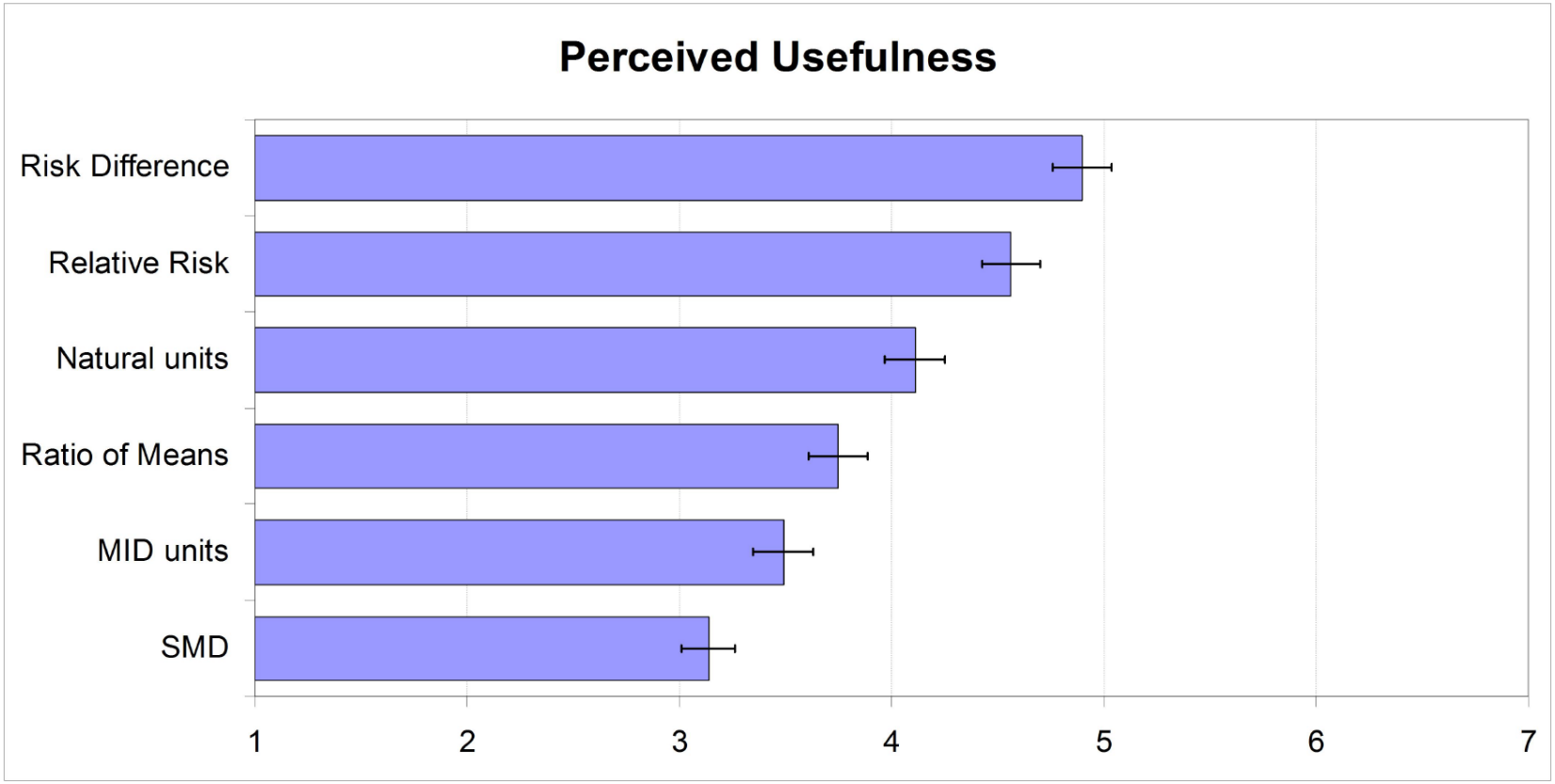


18. In pooled standard deviation units of all pain scores in the treatment and control groups, a meta-analysis finds the effect of intervention A vs placebo control for patient-reported pain to be 0.20 standard deviation units in favoring of intervention A. Please clearly indicate whether this presentation approach is useful:

1	2	3	4	5	6	7
Not useful in understanding size and importance of the effect						Extremely useful in understanding the size and importance of the effect

Results: Usefulness

Figure 4: Perceived Usefulness, *n* = 531



Higher scores represent higher perceived usefulness

Conclusions

- Patient-reported outcomes often critical
  - Almost always patient-important
  - Symptoms, function, quality of life
- Interpretations can be challenging
- Single study, SR all same instrument
  - Fine if everyone familiar with units
  - If not, need MID or dichotomize
- SR different instruments
  - SMD most used – limitations
  - Look for natural units, conversion RR and RD

More conclusions

- Use more than one method
  - Decreases selection bias
  - If similar reassuring
  - If not, need to explain, appropriate doubt
- If very familiar instrument, use as approach
- Use comments in SoF, especially MID
- One of approaches should be dichotomy

For copies of the slides

- Contact

guyatt@mcmaster.ca

Ebcp.mcmaster.ca

Evidence-Based Clinical Practice Workshop

@EBCPMcMaster

Follow us on Twitter