Making results of patient-reported outcomes interpretable

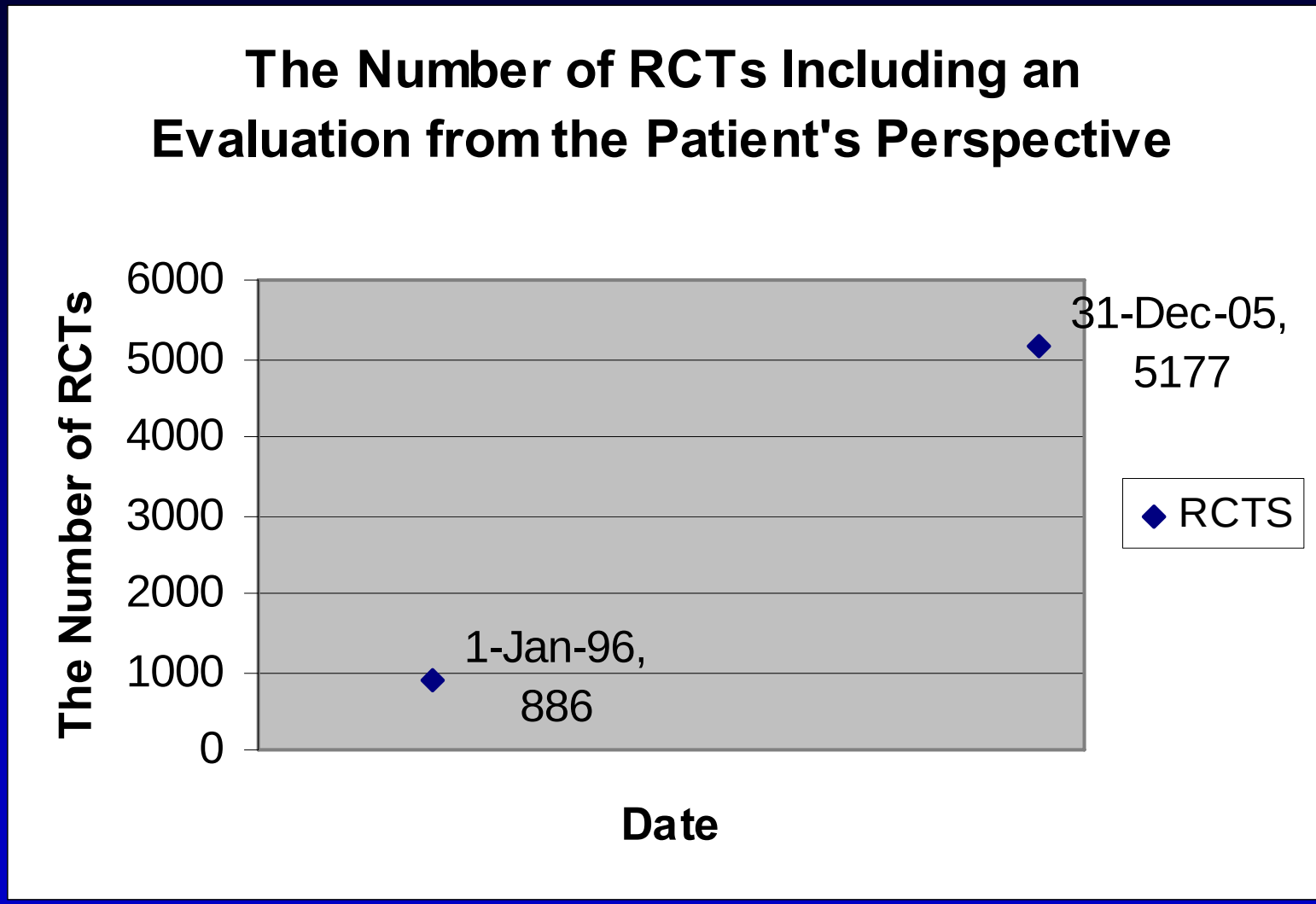
Gordon Guyatt, MD, MSc
Donald L. Patrick, PhD, MSPH

Plan

- what are patient-reported outcomes?
- the problem of interpretability
- strategies for making results interpretable
  - effect sizes
  - minimal important differences
- systematic reviews and meta-analyses
  - options for summarizing effects

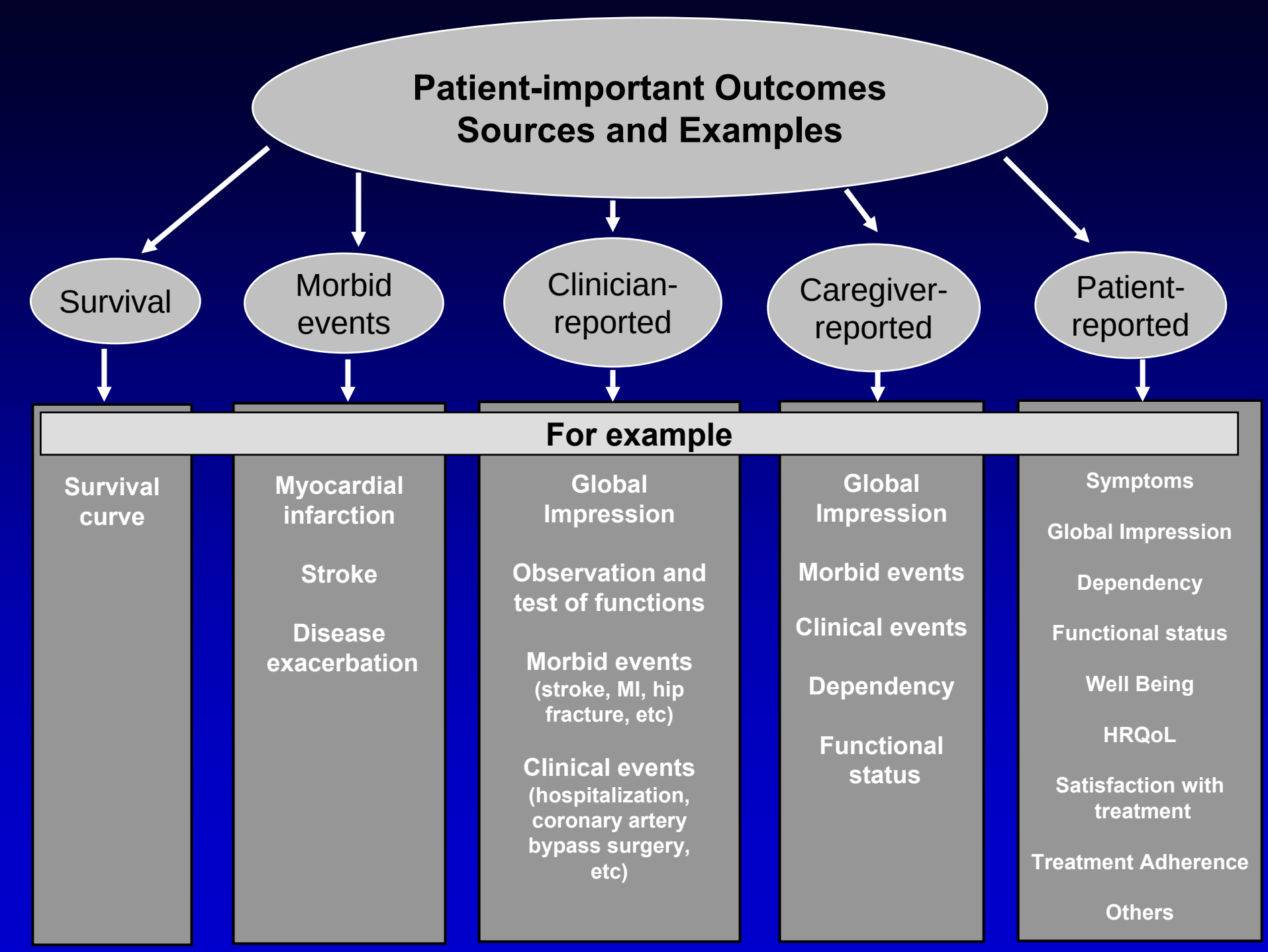
What is a Patient-Reported Outcome (PRO)?

- **PRO:** Any report directly from patients, without interpretation by physicians or anyone else, about how they function or feel in relation to a health condition and its therapy (from diaries, questionnaires, interviews, etc.)
- **PROs developed with patient input using qualitative methods a guiding principle**
- **PRO term requires concept purported to be measured be specified**
- **PRO ≠QoL≠HrQoL**


The Number of RCTs Including an Evaluation from the Patient's Perspective
The Number of RCTs
6000
5000
4000
3000
2000
1000
0
1-Jan-96, 886
31-Dec-05, 5177
RCTS
Date

Why PROs?

- Some treatment effects known only to the patient, i.e. pain, symptoms, feelings
- Small changes in survival further informed by symptoms, function, and feelings
- Survival not only outcome of interest for many interventions
- Physiologic measures often do reflect how patient functions or feels
- Well-developed assessment by patients is as reliable if not more reliable than ratings of patient's condition by clinicians


Patient-important Outcomes
Sources and Examples
Survival
Morbid events
Clinician-reported
Caregiver-reported
Patient-reported
For example
Survival curve
Myocardial infarction
Stroke
Disease exacerbation
Global Impression
Observation and test of functions
Morbid events (stroke, MI, hip fracture, etc)
Clinical events (hospitalization, coronary artery bypass surgery, etc)
Global Impression
Morbid events
Clinical events
Dependency
Functional status
Symptoms
Global Impression
Dependency
Functional status
Well Being
HRQoL
Satisfaction with treatment
Treatment Adherence
Others

Interpretability

- 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 significantly reduced (improved) mean Dermatology Quality of Life Scale scores compared with placebo: 4.4 vs. 1.8 at 2 weeks after the last dose ($P<0.0001$) and 3.4 vs. 1.4 at 12 weeks after the last dose ($P<0.001$).
- effect size?
  - trivial, small but important, large?

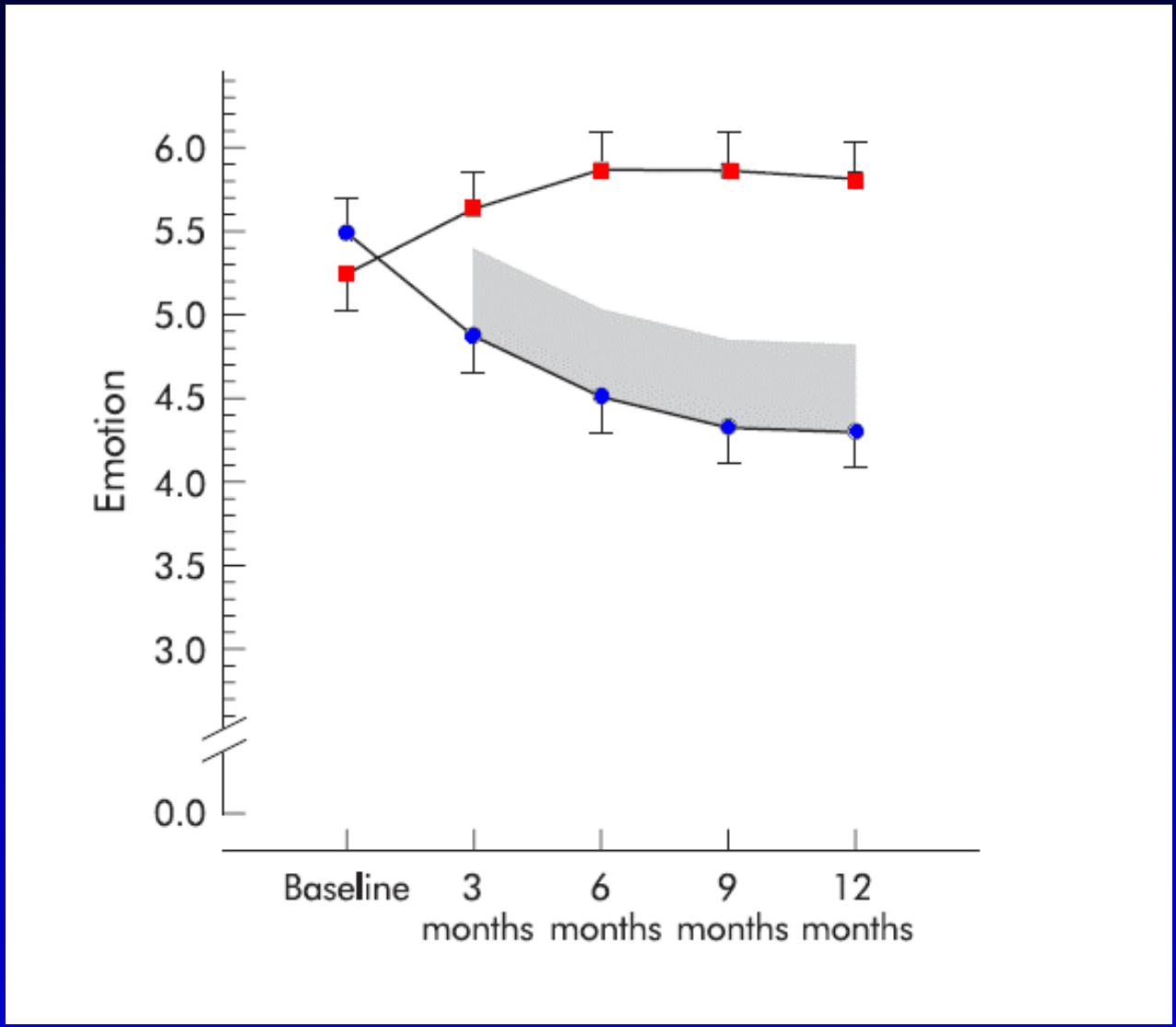
Minimally important difference

- smallest change that patients would consider important
- global ratings of change
  - are you the same, a little better, a lot better
- instruments on 1 to 7 scale 0.5 often represents MID

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, emotional function

Effect of Surgery and Medical Control Treatment

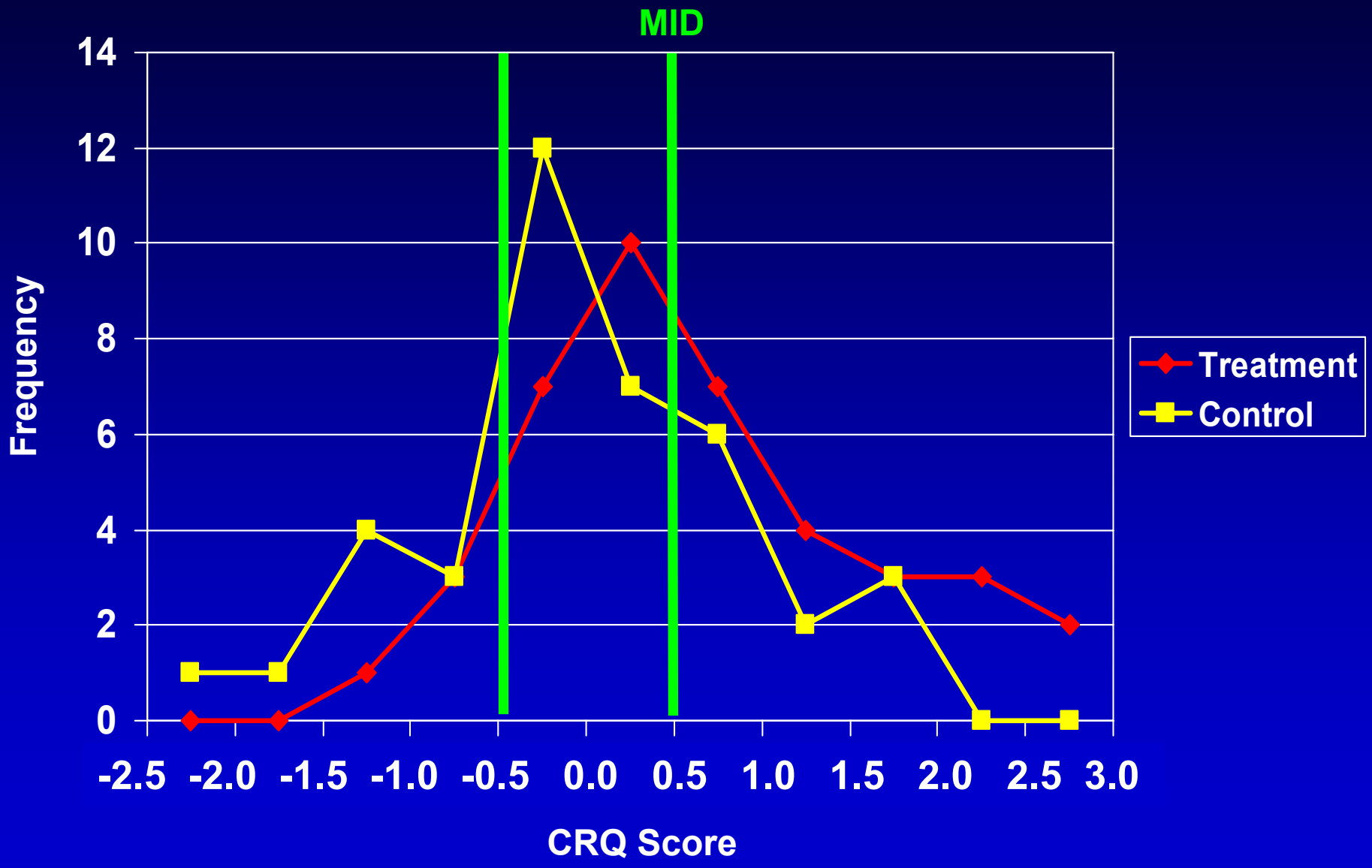


Would you recommend surgery to your patients on the basis of these results?

What if effect smaller

- randomized trial respiratory rehabilitation in COPD
- effect on emotional function 0.4
- important? how important?

CRQ Emotion Change Scores



Number Needed to Treat

- Number needed to treat (NNT) for 1 person to achieve a specified change in a PRO (responder criteria)
- *NNT = 100/(pT – pC)*
- pT is the percentage of patients who improved in the treatment group, and
- pC is the proportion of patients who improved in the control group

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	No 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

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?

Systematic review

- CRQ overall pooled
  - mean difference 0.7, CI 0.2 to 1.2
- St. George's MID 4
  - mean difference 6, CI 2 to 8
- what is your conclusion on size/importance of effect?

Solution for RCT interpretation

- 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

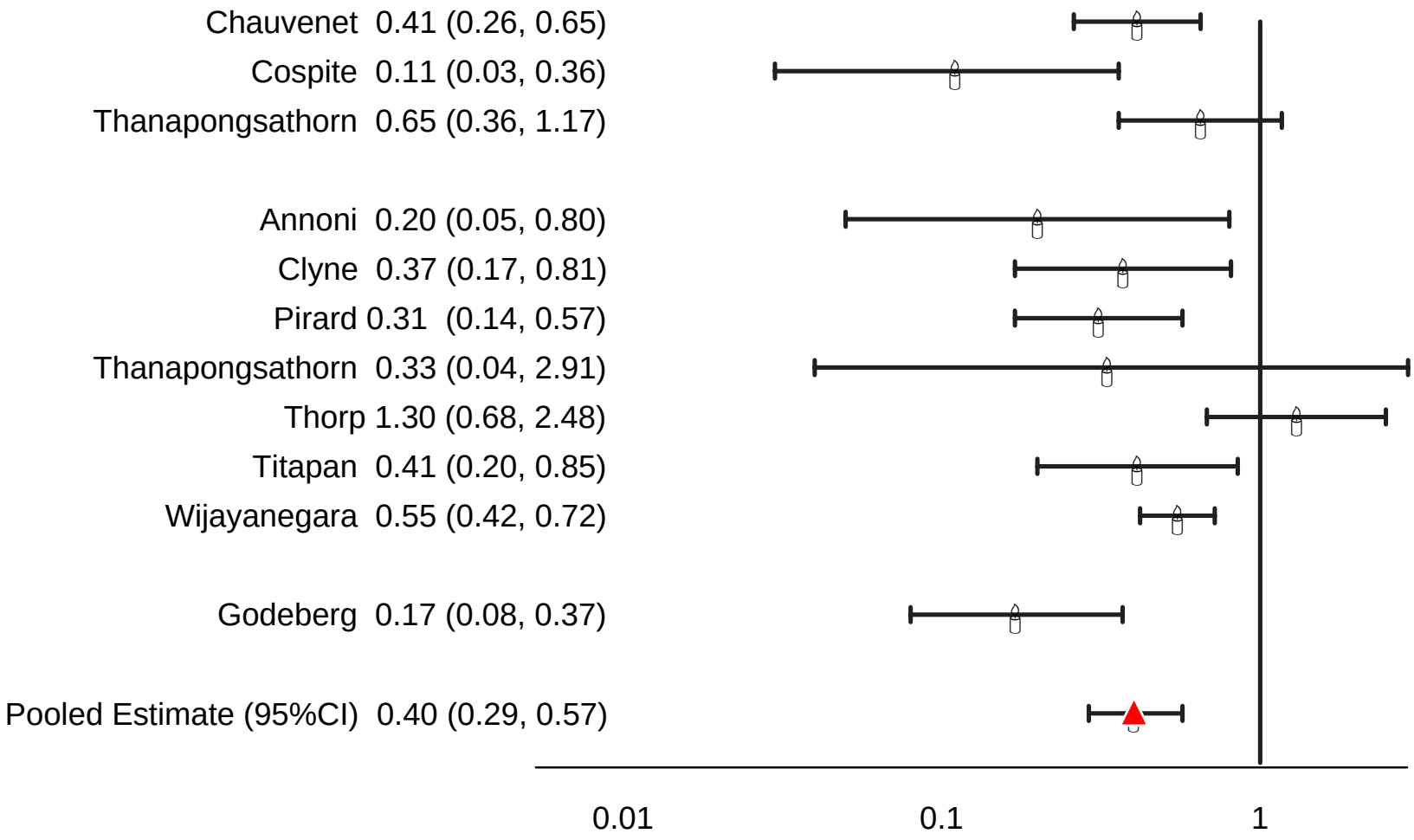
Systematic review of RCTs of thrombolysis in acute stroke

- use Rankin threshold 2 to 3
  - 2 minor handicap
  - 3 moderate handicap
  - proportion "dead or disabled"
- "death or dependency"
  - 55.2% in thrombolysis, 68.3% in control
  - 42% odds reduction
  - 13.1% absolute risk reduction
  - NNT 7 to 8

Flavanoids for Hemorrhoids

- venotonic agents
  - mechanism unclear, 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)

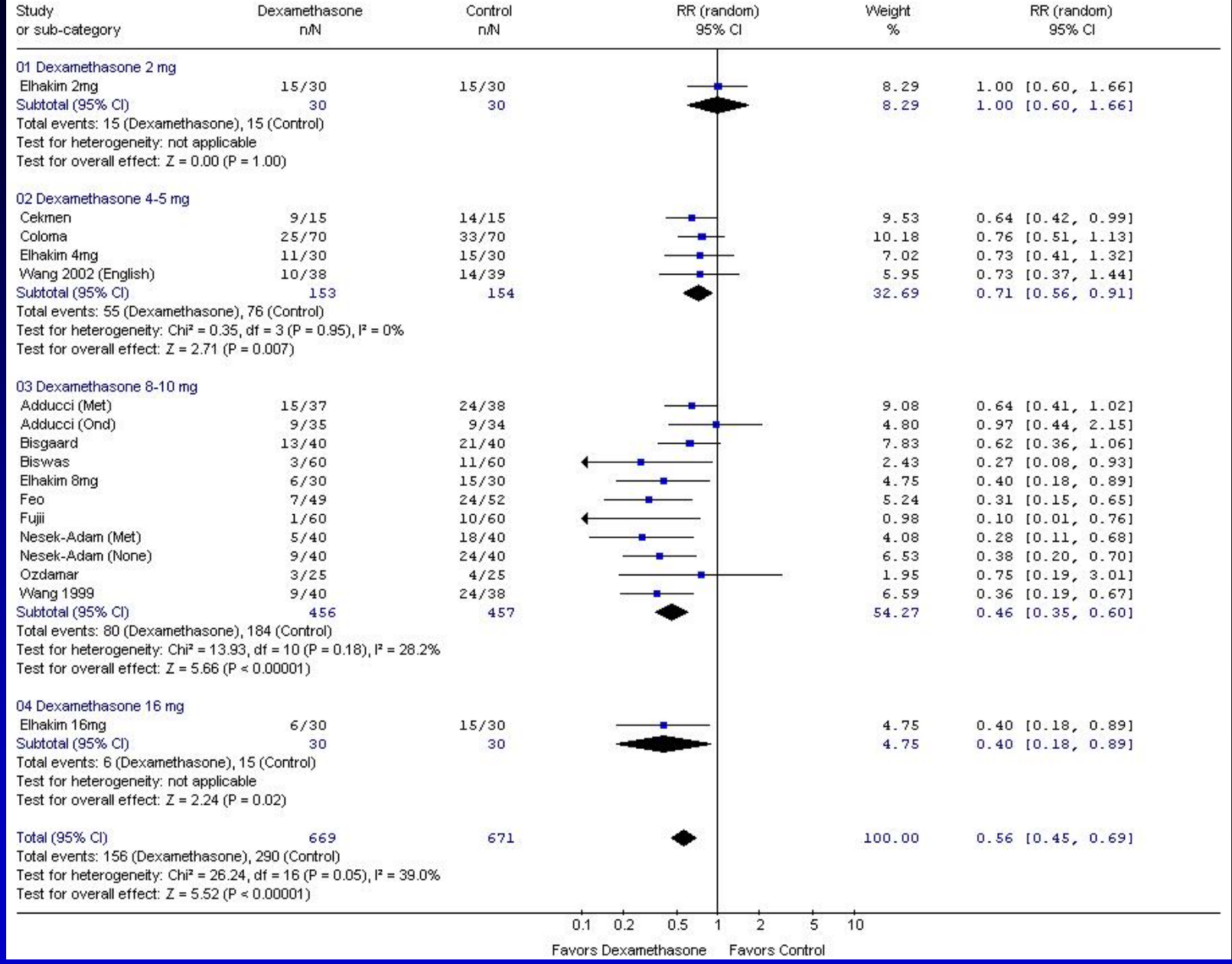


Systematic reviews, meta-analysis

- seldom have original data from individual studies to apply thresholds
- individual studies my use different PROs to measure same concepts

Steroids for laparoscopic Cholecystectomy

- systematic review
- nausea and vomiting
  - 16 RCTs
- pain
  - 5 RCTs

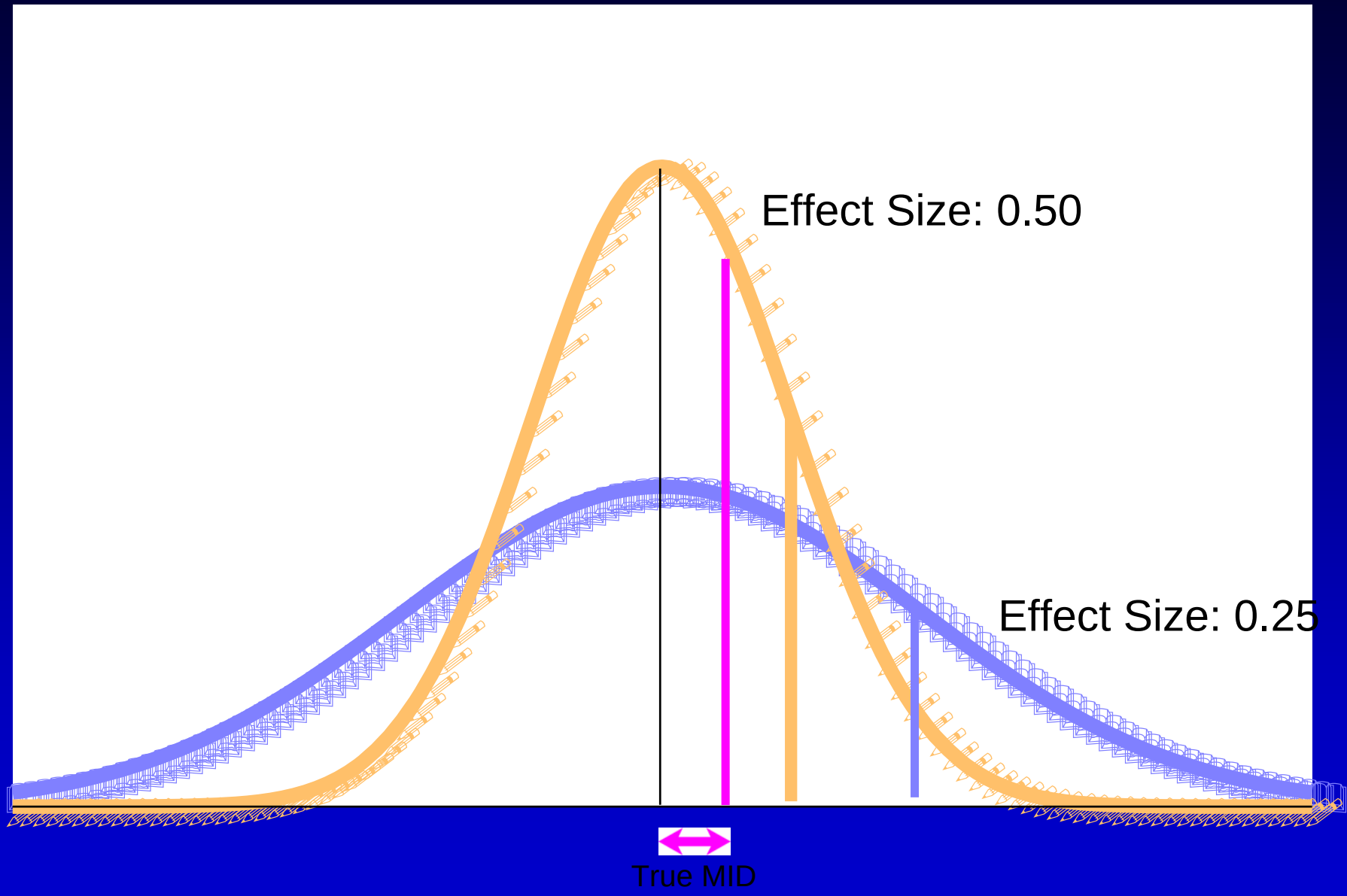


Pain – no dichotomies, multiple measures – what to do?

Methods of reporting	Pooled estimate and 95% CI	
	COPD meta-analysis	**Dexamethasone meta-analysis**
Category 1: Methods derived from standard deviation units		
i) SMD	SMD=0.73 (95% CI 0.49 to 0.96)	SMD=-0.79(95%CI -0.17 to -1.41)
ii) Conversion of SMD to dichotomies (to OR) (Suissa)	[1]OR=3.22 (95% CI 2.24 to 4.74)	[2]OR=0.21 (95% CI 0.04 to 0.76)
iii) Conversion of SMD to OR (Hasselbad/Hedges)	OR=3.74 (95% CI 2.42 to 5.68)	OR=0.23 (95% CI 0.08 to 0.74)
iv) Conversion of SMD to dichotomies (to NNT) (Suissa)	[1]NNT=3.6 (95% CI 2.7 to 5.4)	[2]NNT=6.7 (95% CI 5.3 to 22.7)
Category 2: Not derived from SMD, not relying on MID		
i) Conversion of all instruments to the most popular	[3]MD=0.76 (95% CI 0.64 to 0.88)	[4]MD=4 (95%CI 3.2 to 4.7)
iv) Ratio of means (ROM)	[5]Not applicable	ROM=0.87 (95% CI 0.78 to 0.98)
Category 3: Not derived from SMD, depends on knowledge of MID		
i) Minimally important difference (MID) units	[6]MD=1.75 (95% CI 1.37 to 2.13)	[7]Not applicable
ii) Conversion of MDs to dichotomies (to OR)	OR=3.52 (95% CI 2.60 to 4.76)	[7]Not applicable
iii) Conversion of MDs to dichotomies (to NNT)	NNT=4.3 (95% CI 3.34 to 6.0)	[7]Not applicable

Effect size

- divide each effect by standard deviation
- ultimate result in SD units
- between "effect size" or SMD
- within standardized response mean


Effect Size: 0.50
Effect Size: 0.25
True MID

Methods of reporting	Pooled estimate and 95% CI	
	COPD meta-analysis	Dexamethasone meta-analysis
Category 1: Methods derived from standard deviation units		
i) SMD	SMD=0.73 (95% CI 0.49 to 0.96)	SMD=-0.79(95%CI -0.17 to -1.41)

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

Making SD units interpretable

- convert back to natural units of most popular measure
- pooled effect 0.5 SD units
  - pooled SD of on 100 point scale 20
  - effect in natural units 10
- vulnerability
  - effect size distortion by heterogeneity
  - SD of studies using most popular varies

Avoiding heterogeneity problem

- convert all measures to units of most popular
  - most popular 0 – 100
  - alternative 0 to 7
- multiply all scores by 100/7
  - get weighted mean difference
  - alternative to standardized mean difference
- result in natural units

Methods of reporting	Pooled estimate and 95% CI	
	COPD meta-analysis	Dexamethasone meta-analysis
Category 2: Not derived from SMD, not relying on MID		
i) Conversion of all instruments to the most popular	[3]MD=0.76 (95% CI 0.64 to 0.88)	[4]MD=4 (95%CI 3.2 to 4.7)

- vulnerable to differences in instrument properties
  - assumes clinicians understand natural units (MID)
  - vulnerable to interpretation problem

Outcomes	Illustrative Comparative Risks (95% CI)		Relative Effect (95% CI)	Number of participants (studies)	Quality of the Evidence[1] (GRADE0	Comments
	Assumed risk with Placebo	Corresponding risk with Dexamethasone				
Post-operative pain (B) Measured on a scale from 0, no pain, to 100, worst pain imaginable.[4]	The mean post-operative pain scores with placebo ranged from 43 to 54	The mean pain score in the intervention groups was on average **15 (3 to 27) lower**		539 (5 studies)	**Low**[2,3]	Scores estimated based on an SMD of 0.79 (95% CI -1.41 to -0.17)[4]

1- Quality rated from 1 (very low quality) to 4 (high quality), 2- Evidence limited by heterogeneity between studies, 3- Evidence limited by imprecise data (small sample size or event rate), 4- A standard deviation of 0.5 represents a moderate difference between groups,

MID units

- Cochrane review of respiratory rehabilitation for COPD
- using 16 trials, we compared the existing method with the MID method
- trials employed two widely used disease-specific HRQL instruments
  - Chronic Respiratory Disease Questionnaire (CRQ)
  - St. Georges Respiratory Questionnaire (SGRQ)

Results

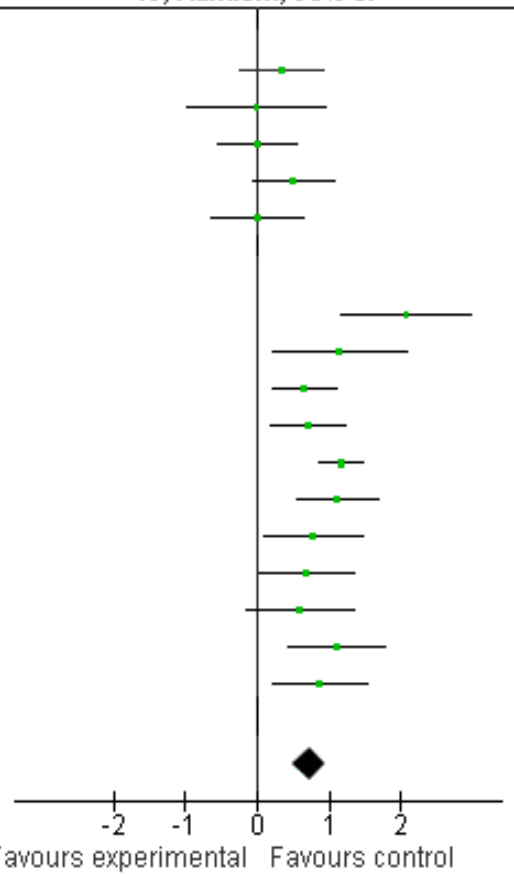
CRQ	**Mean Difference (95% CI)**
Dyspnea	1.06 (0.85, 1.26)
Emotional Function	0.76 (0.52, 1.00)
Fatigue	0.92 (0.71, 1.13)
Mastery	0.97 (0.74, 1.20)
Overall	0.94 (0.57, 1.32)
SGRQ	
Activities	4.78 (1.72, 7.83)
Impacts	6.27 (2.47, 10.08)
Symptoms	4.68 (0.25, 9.61)
Overall	6.11 (3.24, 8.98)

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: Tau² = 0.13; Chi² = 35.82, df = 15 (P = 0.002); I² = 58%

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

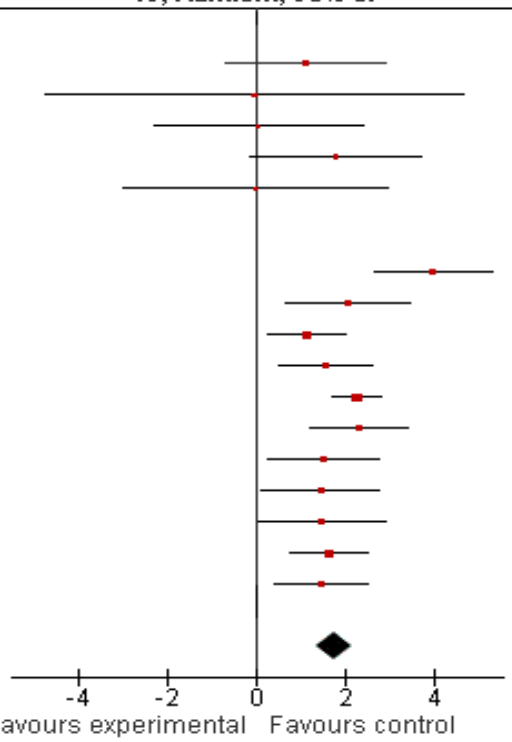


Results - MID Units

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)



MID Units

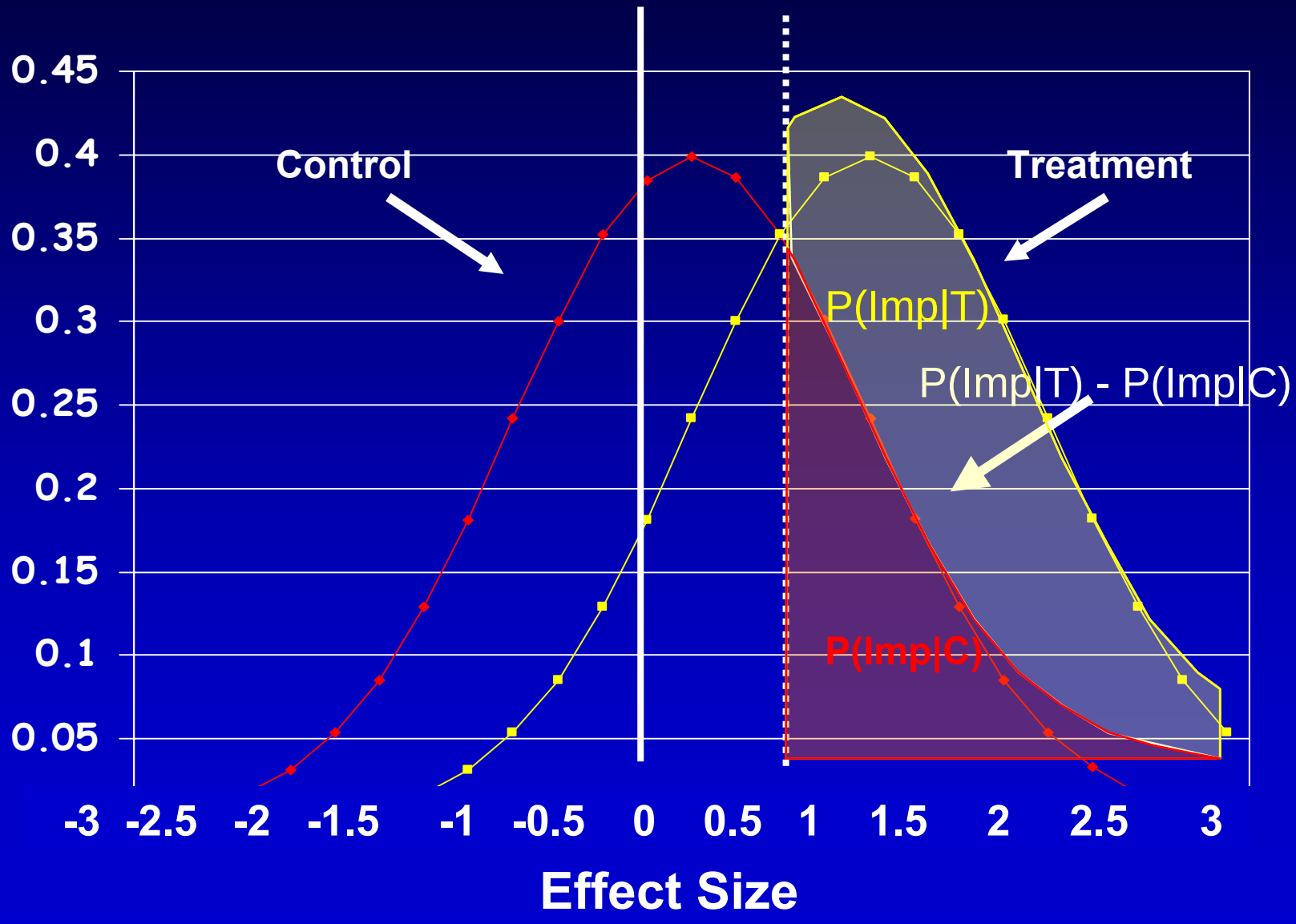
- suggests a large effect:
  - the pooled estimate twice the smallest difference patients perceive as important
- MID approach
  - prevents introducing inconsistency depending on the SD
  - intuitive interpretation
  - vulnerable to all-or-nothing misinterpretation

Applying the MID

- assume MID is 0.50 and patients mean improvement is 0.25
- 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

No dichotomies in primary studies

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



Approaches to dichotomizing

- Suissa: output risk in control and intervention group
- assumes normal distribution
  - but not equal variance
  - requires specification of control group risk
- specify control group risk and generate intervention group risk
- from event rates generate odds ratio

Approaches to dichotomizing

- Hasselbad and Hedges
- assume logistic distribution
  - doesn't require control event risk
  - assumes normality, equal variance
- Cox and Snell
- Kraemer – ROC, AUC
- all can generate OR, RD, NNT
- all vulnerable to heterogeneity, normality

Furukawa
Suissa approach to generate NNTs

Control group response rate	10	20	30	40	50	60	70	80	90
ES = 0.2	25	17	14	13	13	13	15	20	33
ES = 0.5	9	6	5	5	5	5	7	9	16
ES = 0.8	5	4	3	3	4	4	5	7	12
ES = 1.0	4	3	3	3	3	4	4	6	11

Methods of reporting	Pooled estimate and 95% CI	
	COPD meta-analysis	Dexamethasone meta-analysis
Category 1: Methods derived from standard deviation units		
i) SMD	SMD=0.73 (95% CI 0.49 to 0.96)	SMD=-0.79(95%CI -0.17 to -1.41)
ii) Conversion of SMD to dichotomies (to OR) (Suissa)	[1]OR=3.22 (95% CI 2.24 to 4.74)	[2]OR=0.21 (95% CI 0.04 to 0.76)
iii) Conversion of SMD to OR (Hasselbad/Hedges)	OR=3.74 (95% CI 2.42 to 5.68)	OR=0.23 (95% CI 0.08 to 0.74)
iv) Conversion of SMD to dichotomies (to NNT) (Suissa)	[1]NNT=3.6 (95% CI 2.7 to 5.4)	[2]NNT=6.7 (95% CI 5.3 to 22.7)

Ratio of Means (RoM)

$$RoM = \frac{mean_{exp}}{mean_{control}}.$$

- Requires estimate of variance of this ratio - this can be estimated using the delta method:
  - $Var_{ln(RoM)} = \frac{var_{exp}}{(mean_{exp}^2)} + \frac{var_{control}}{(mean_{control}^2)}$

Avoiding heterogeneity problem: Ratio of means

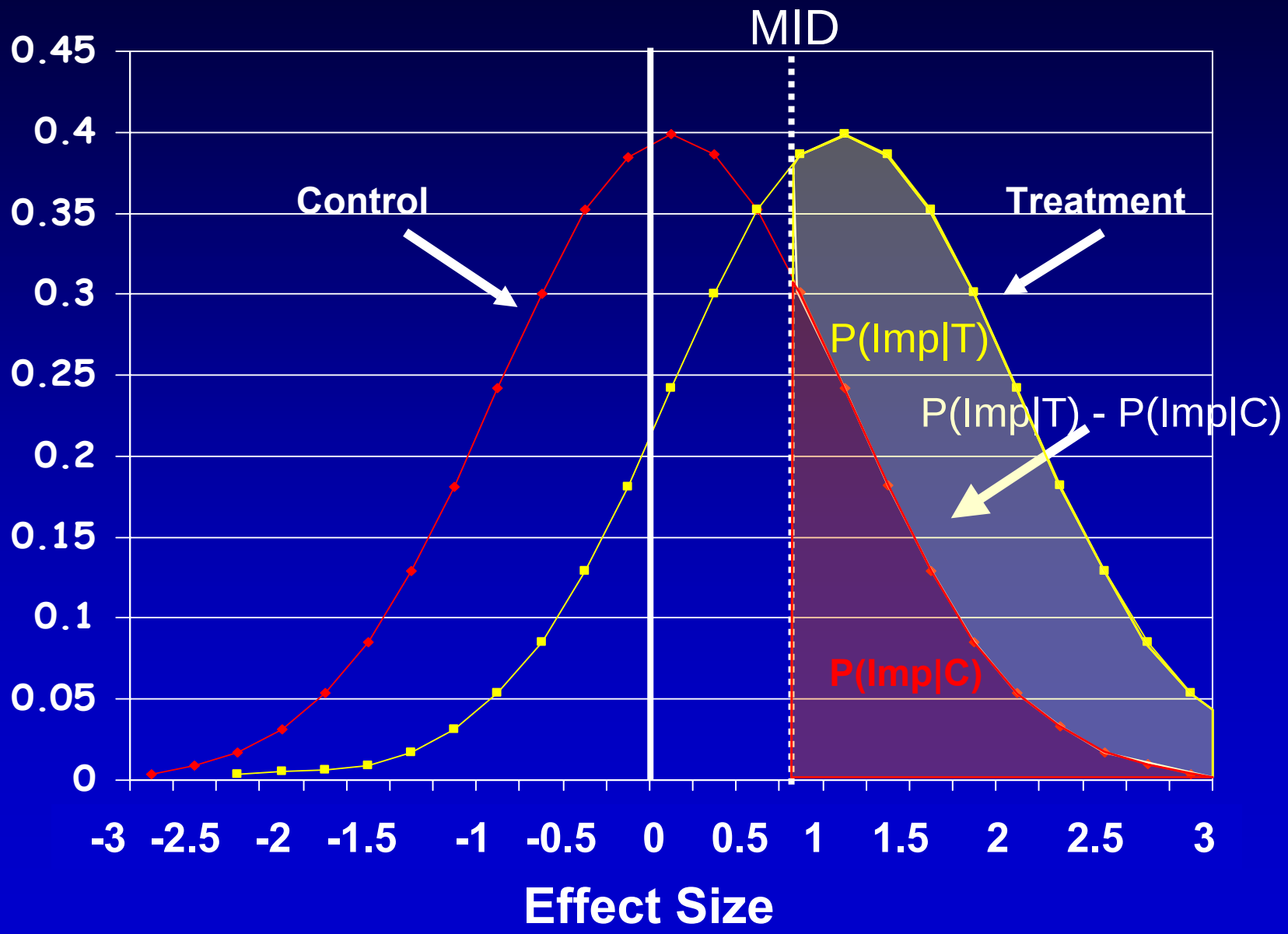
- analogous to relative risk
  - greater absolute difference with greater control risk
- requires natural zero
- cannot use if results reported as change

Methods of reporting	Pooled estimate and 95% CI	
	COPD meta-analysis	Dexamethasone meta-analysis
Category 2: Not derived from SMD, not relying on MID		
i) Conversion of all instruments to the most popular	[3]MD=0.76 (95% CI 0.64 to 0.88)	[4]MD=4 (95%CI 3.2 to 4.7)
iv) Ratio of means (ROM)	[5]Not applicable	ROM=0.87 (95% CI 0.78 to 0.98)

Avoiding heterogeneity problem

- back to MID
- effect in MID units
- dichotomy – risk difference

No dichotomies in primary studies



Methods of reporting	Pooled estimate and 95% CI	
	COPD meta-analysis	**Dexamethasone meta-analysis**
Category 3: Not derived from SMD, depends on knowledge of MID		
i) Minimally important difference (MID) units	[6]MD=1.75 (95% CI 1.37 to 2.13)	[7]Not applicable
ii) Conversion of MDs to dichotomies (to OR)	OR=3.52 (95% CI 2.60 to 4.76)	[7]Not applicable
iii) Conversion of MDs to dichotomies (to NNT)	NNT=4.3 (95% CI 3.34 to 6.0)	[7]Not applicable

Methods of reporting	Pooled estimate and 95% CI	
	COPD meta-analysis	**Dexamethasone meta-analysis**
Category 1: Methods derived from standard deviation units		
i) SMD	SMD=0.73 (95% CI 0.49 to 0.96)	SMD=-0.79(95%CI -0.17 to -1.41)
ii) Conversion of SMD to dichotomies (to OR) (Suissa)	[1]OR=3.22 (95% CI 2.24 to 4.74)	[2]OR=0.21 (95% CI 0.04 to 0.76)
iii) Conversion of SMD to OR (Hasselbad/Hedges)	OR=3.74 (95% CI 2.42 to 5.68)	OR=0.23 (95% CI 0.08 to 0.74)
iv) Conversion of SMD to dichotomies (to NNT) (Suissa)	[1]NNT=3.6 (95% CI 2.7 to 5.4)	[2]NNT=6.7 (95% CI 5.3 to 22.7)
Category 2: Not derived from SMD, not relying on MID		
i) Conversion of all instruments to the most popular	[3]MD=0.76 (95% CI 0.64 to 0.88)	[4]MD=4 (95%CI 3.2 to 4.7)
iv) Ratio of means (ROM)	[5]Not applicable	ROM=0.87 (95% CI 0.78 to 0.98)
Category 3: Not derived from SMD, depends on knowledge of MID		
i) Minimally important difference (MID) units	[6]MD=1.75 (95% CI 1.37 to 2.13)	[7]Not applicable
ii) Conversion of MDs to dichotomies (to OR)	OR=3.52 (95% CI 2.60 to 4.76)	[7]Not applicable
iii) Conversion of MDs to dichotomies (to NNT)	NNT=4.3 (95% CI 3.34 to 6.0)	[7]Not applicable

Conclusions re interpretability

- if possible use natural dichotomies
- many approaches rely on SD units
  - suffer from problem of heterogeneity
  - important limitation
- approaches not relying on SD units preferable
  - ideally know MID
  - can present in MID units and proportions
  - approaches complementary