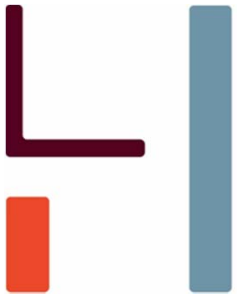
Minimally Important Differences
definitions, ambiguities and pitfalls

Stephen Senn



LUXEMBOURG
INSTITUTE
OF **HEALTH**
RESEARCH DEDICATED TO LIFE

Acknowledgements

Many thanks for the invitation

This work is partly supported by the European Union's 7th Framework Programme for research, technological development and demonstration under grant agreement no. 602552. "IDEAL"

It seems I could stop the talk here

Minimal important difference
"The smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management"

Jaeschke et al, 1989

Warnings

- What I know about quality of life could be written on the back of an envelope
- Although I know a lot more about clinical measures, I dislike dichotomies
- Many of you will find much to hate in this talk
- The rest of you may fall asleep

Outline

- Differences for planning
- Differences for interpreting treatment effects?
- Individual effects
- Conclusions?

Differences for planning

Clinically relevant differences?

The National Institute for Health Research view (2014)

- Talked about target differences
- Considered two approaches
  - A difference considered to be important
  - A realistic difference
- I will cover four, two of which are similar to the two here
- However, first some statistical basics

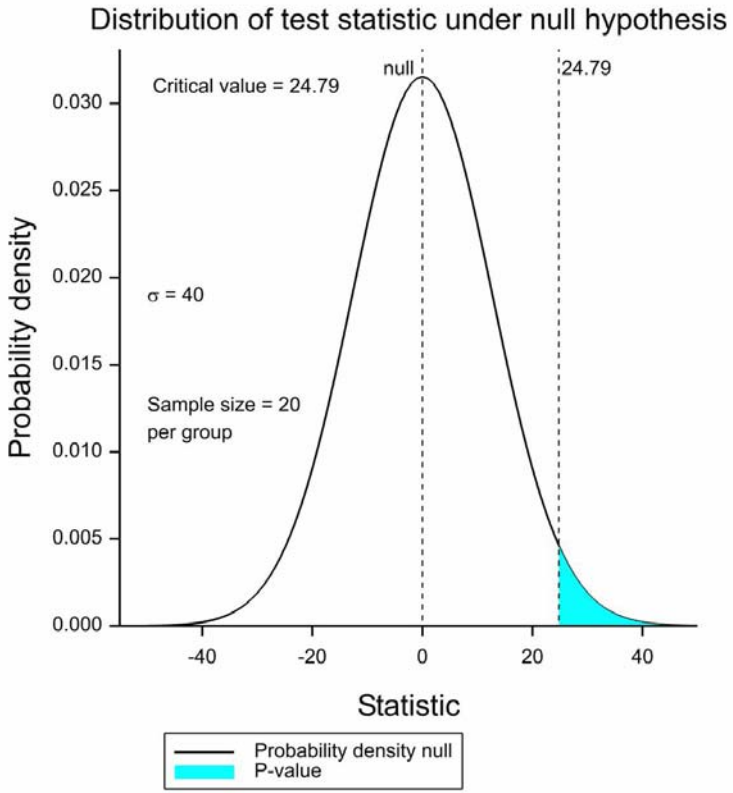
HEALTH TECHNOLOGY ASSESSMENT

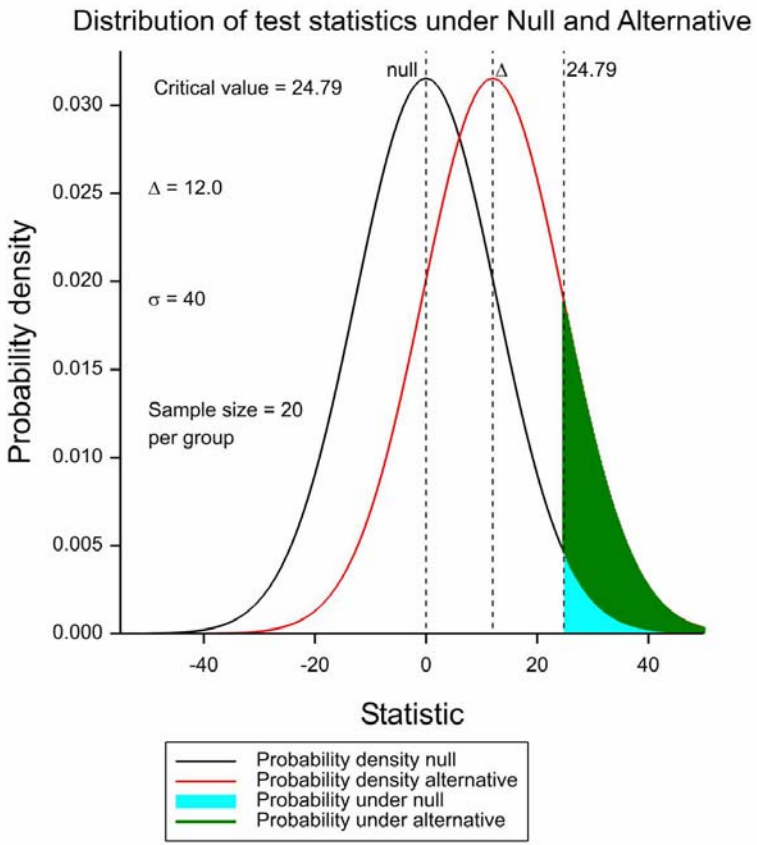
VOLUME 18 ISSUE 28 MAY 2014
ISSN 1366-5278

Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review

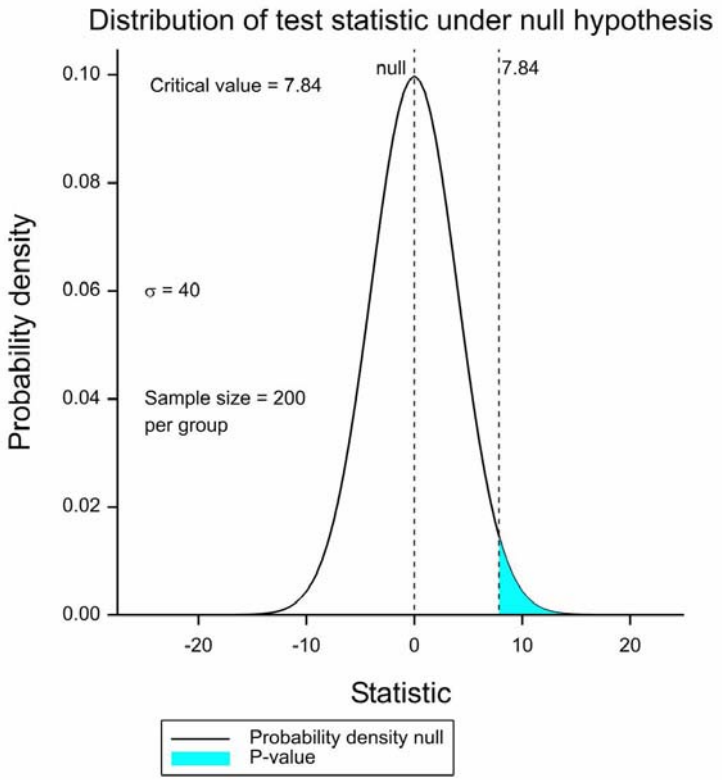
Jonathan A Cook, Jennifer Hislop, Temitope E Adewuyi, Kirsten Harrild, Douglas G Altman, Craig R Ramsay, Cynthia Fraser, Brian Buckley, Peter Fayers, Ian Harvey, Andrew H Briggs, John D Norrie, Dean Fergusson, Ian Ford and Luke D Vale

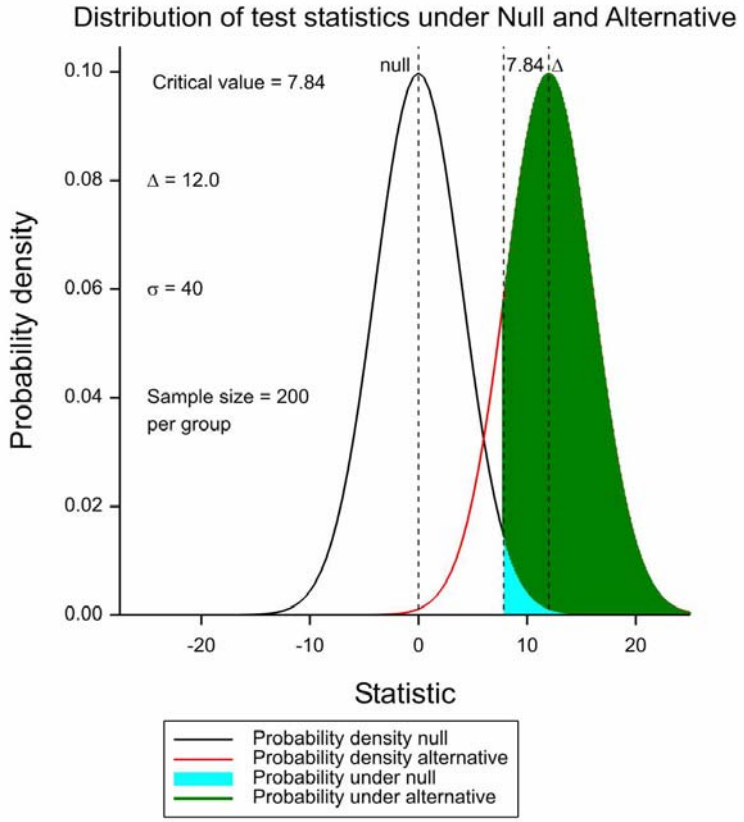
Hypothesis testing basics





Increasing the power





Delta force

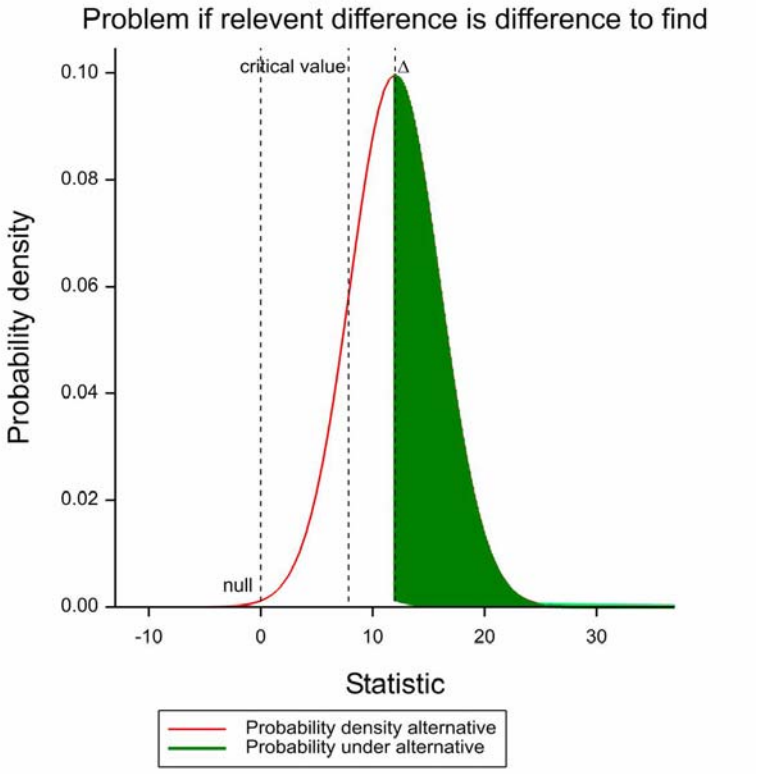
What is delta?

- The difference we would like to observe?
- The difference we would like to ‘prove’ obtains ?
- The difference we believe obtains
- The difference you would not like to miss?

The difference you would like to observe

This view is hopeless

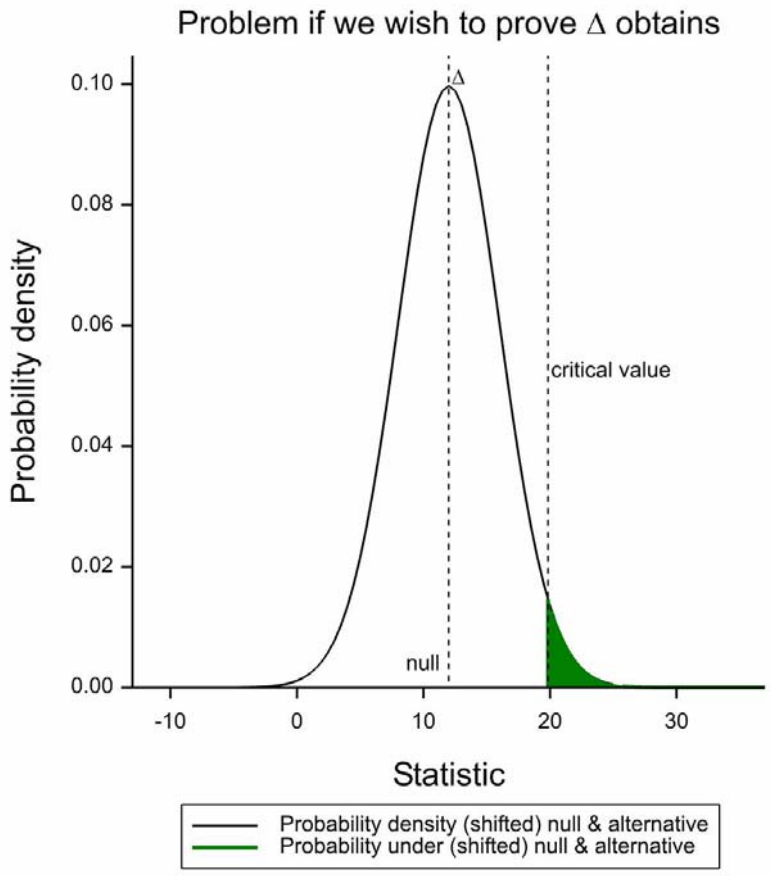
if Δ is the value we would like to observe and if the treatment does, indeed, have a value of Δ then we have only half a chance, not (say) an 80% chance, that the trial will deliver to us a value as big as this.



The difference we would like to 'prove' obtains ?

This view is even more hopeless

It requires that the lower confidence interval should be greater than Δ. This requires using Δ as a (shifted) null value and trying to reject *this*. If this is what is needed, the power calculation is completely irrelevant.



The difference we believe obtains

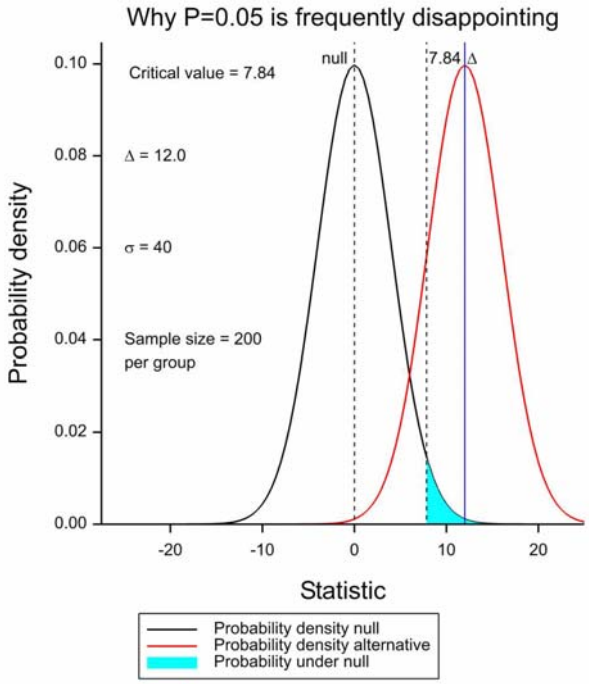
- This is very problematic
- It views the sample size as being a function of the treatment and not the disease
- It means that for drugs we think work less well we would use bigger trials
- This seems back to front
- If modified to a Bayesian probability distribution of effects it can be used to calculate assurance
- This has some use in deciding whether to run a trial

The difference you would not like to miss

- This is the interpretation I favour.
- The idea is that we control two (conditional) errors in the process.
  - The first is α, the probability of claiming that a treatment is effective when it is, in fact, no better than placebo.
  - The second is the error of failing to develop a (very) interesting treatment further.
- If a trial in drug development is not 'successful', there is a chance that the whole development programme will be cancelled.
- It is the conditional probability of cancelling an interesting project that we seek to control.

But be careful: P=0.05 is a disappointment

- To have a greater that 50% power for a significant result we must have that Δ > critical value
- But P=0.05 means the test statistic is just equal to the critical value
- Hence the result we see is less than the clinically relevant difference
  - 70% of Δ if you planned for 80% power



Differences for interpreting treatment effects?

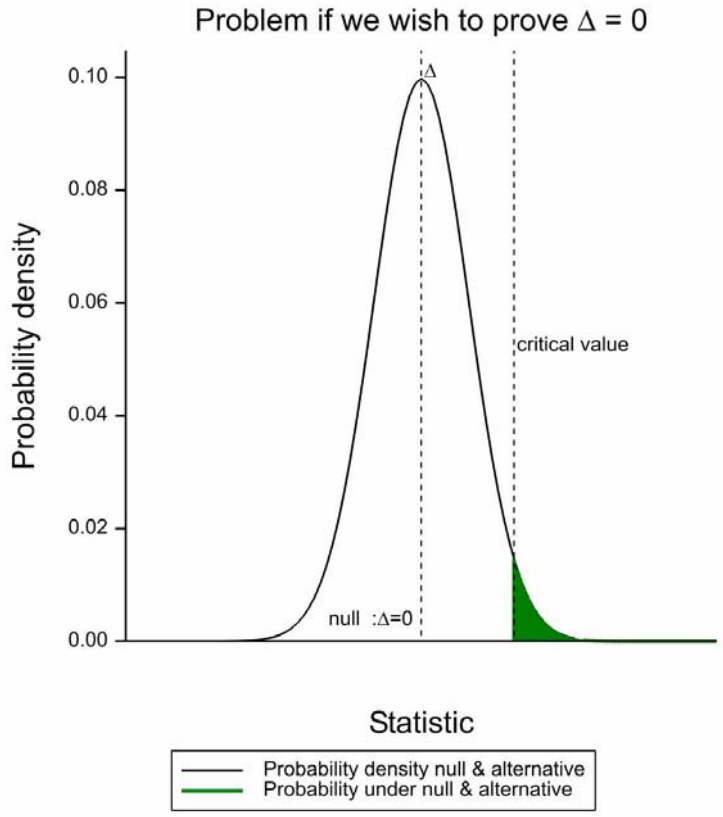
Minimally important differences?

A plan is not an inference

- You plan so as to have a reasonably low probability of missing an important effect
- In drug development, if you have a positive result, work goes on
- Furthermore, once the results are in, the plan is largely irrelevant
- You analyse the data you have
- Thus the clinically relevant difference has no direct effect on the inference

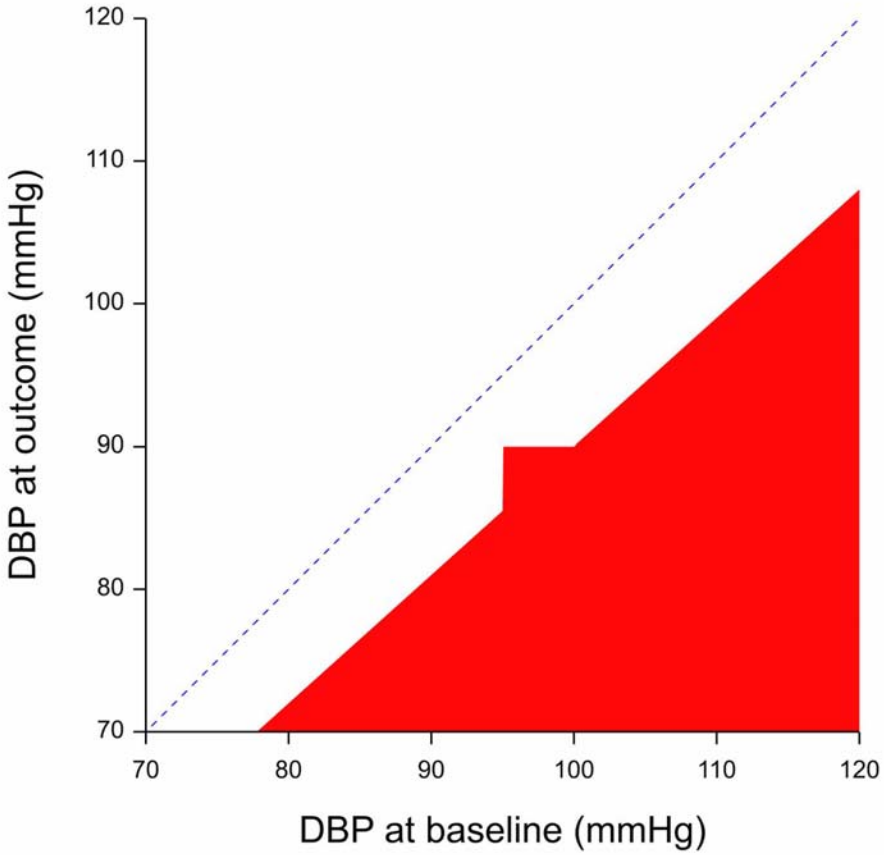
Clinically irrelevant difference

- Often used for so-called active controlled equivalence studies
  - Sponsor tries to show that the new treatment is not inferior to a standard by some agreed margin
  - Because if the new treatment really is similar to the existing one, the power of proving the difference is not 0 is just the type I error rate
- So we now look to prove that the new treatment cannot be inferior by more than an *irrelevant* amount

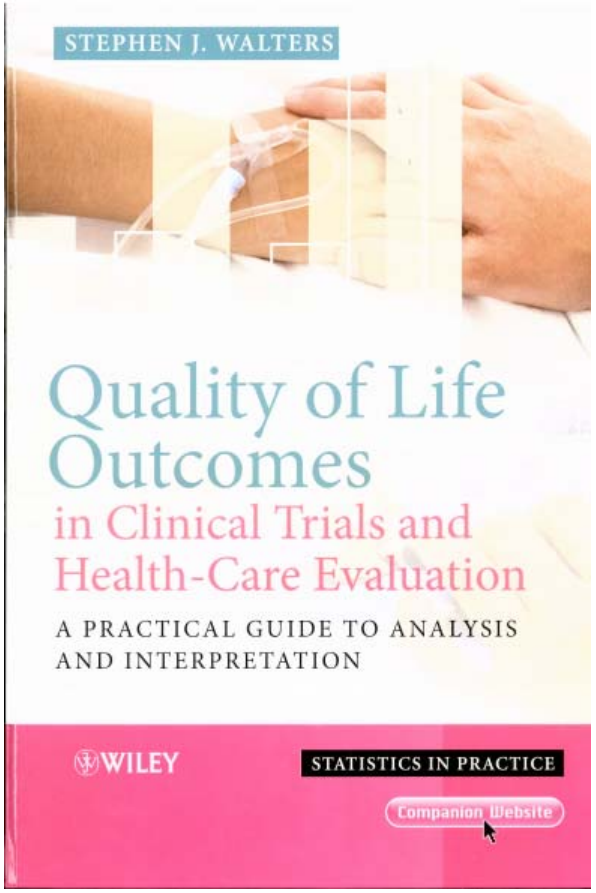


The problem

- Consider the example of hypertension
- A CPMP guideline from 1998 quotes 2mm Hg in diastolic blood pressure as *clinically irrelevant*
- A guideline from 2017 defines *response* as being normalisation (≥95mm to < 90mm) or a 10mm Hg drop
- So response is 5-10mm Hg and irrelevance is 2mm Hg



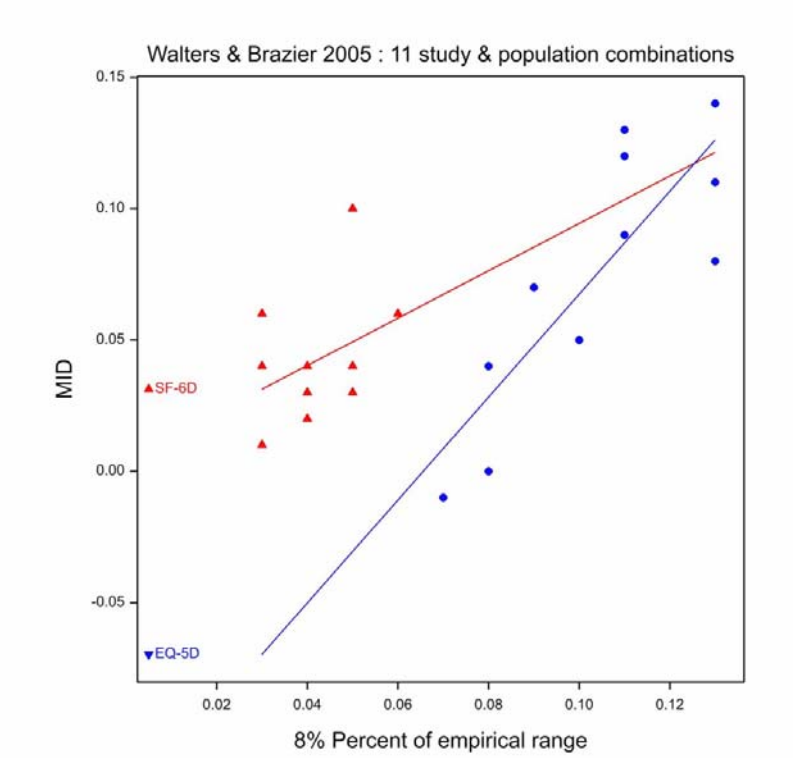
Establishing the minimal important difference



- Clinical or non-clinical anchor
- Mapping to other QoL scores
  - For example, single overall satisfaction question
- Distribution based approach
  - For example, $\frac{\sigma}{2}$
- Empirical rule
  - For example, 8% of theoretical or empirical range of scores

Walters & Brazier, 2005

- Took 11 study/ population/ follow-up combinations
  - Based on 8 studies
- SF-6D (0.29 to 1) and EQ-5D (-0.59 to 1) were available for each
- Calculated MID, SD/2, %of empirical range for both measures for patients who were defined as having had a meaningful response



Individual effects

Responder analysis?

Tiotropium v Placebo in Chronic Obstructive Pulmonary Disease

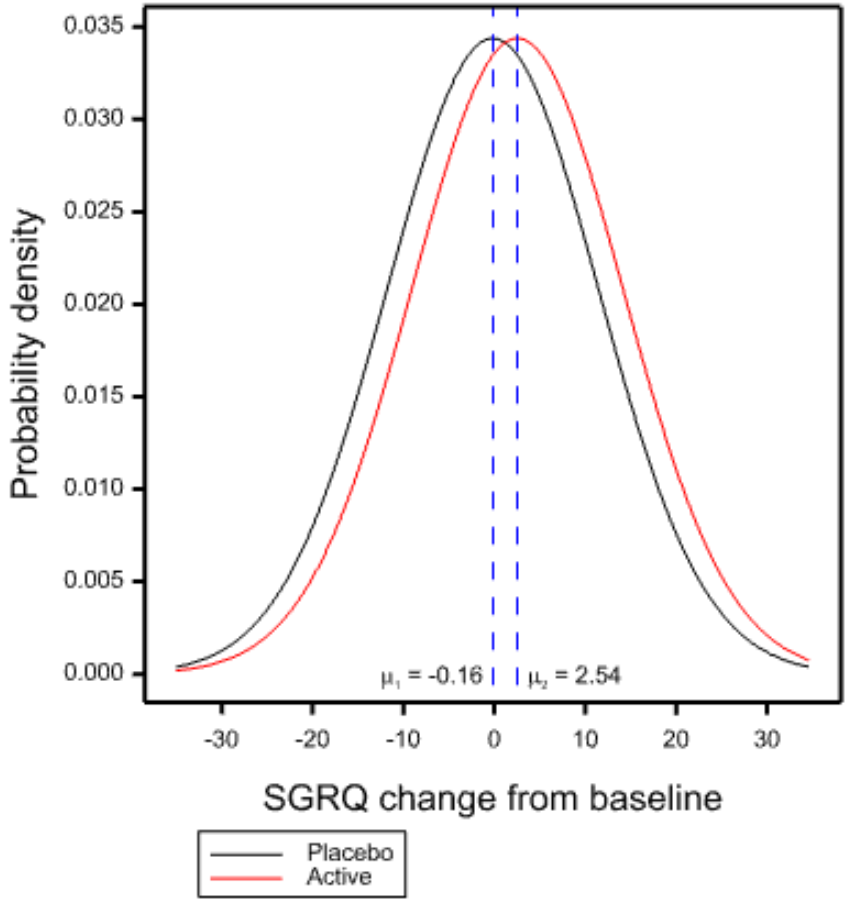
From the UPLIFT Study, *NEJM*, 2008

Significant differences in favor of tiotropium were observed at all time points for the mean absolute change in the SGRQ total score (ranging from 2.3 to 3.3 units, P<0.001), although the differences on average were below what is considered to have clinical significance (Fig. 2D). **The overall mean between-group difference in the SGRQ total score at any time point was 2.7** (95% confidence interval [CI], 2.0 to 3.3) in favor of tiotropium (P<0.001). **A higher proportion of patients in the tiotropium group than in the placebo group had an improvement of 4 units or more in the SGRQ total scores from baseline at** 1 year (49% vs. 41%), 2 years (48% vs. 39%), 3 years (46% vs. 37%), and **4 years (45% vs. 36%)** (P<0.001 for all comparisons).

(My emphasis)

The St George's Respiratory Questionnaire SGRQ

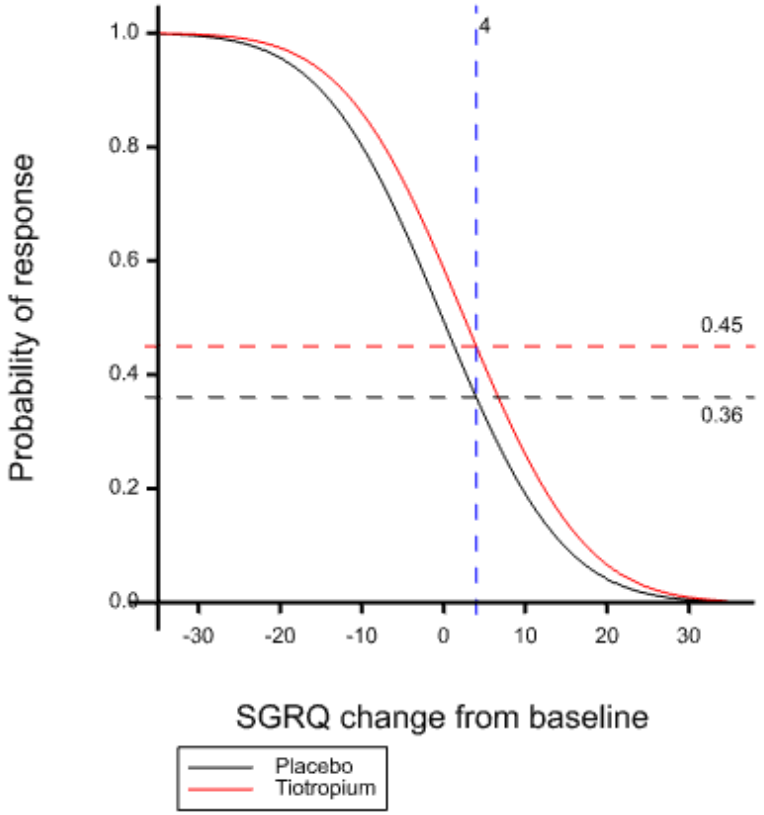
- Jones, Quirk, Baveystock, Littlejohn . A self-complete measure of health status for chronic airflow limitation, *American Review of Respiratory Disease*, **145**, (6), 1991
- 2466 citation by 2 March 2017
- 76 item questionnaire
  - Minimum score 0
  - Maximum score 100
  - Higher values worse
- Minimum important difference is generally taken to be 4 points



Imagined model

Two Normal distributions with the same spread but the Active treatment has a mean 2.7 higher.

If this applies, every patient under active can be matched to a corresponding patient under placebo who is 2.7 worse off



A cumulative plot corresponding to the previous diagram.

If 4 is the threshold, placebo response probability is 0.36, active response probability is 0.45.

In summary…this is rather silly

- If there is sufficient measurement error *even if the true improvement is identically 2.7*, some will show an 'improvement' of 4
- The conclusion that there is a higher proportion of *true* responders *by the standard of 4 points* under treatment than under placebo is quite unwarranted
- So what is the point of analysing 'responders'?

Who are the authors?

1. Tashkin, DP, Celli, B, Senn, S, Burkhart, D, Kesten, S, Menjoge, S, Decramer, M. A 4-Year Trial of Tiotropium in Chronic Obstructive Pulmonary Disease, *N Engl J Med* 2008.

Personal note. I am proud to have been involved in this important study and have nothing but respect for my collaborators. The fact that, despite the fact that two of us are statisticians, we have ended up publishing something like this shows how deeply ingrained the practice of responder analysis is in medical research. We must do something to change this.

Conclusions?

My personal advice

Conclusions

- Responder analysis is an unforgiveable sin
  - If used to create a primary variable for analysis it will increase your sample size by at least a half but usually much more
  - It is nearly always accompanied by quite unwarranted causal judgements
  - It has led to a lot of nonsense and hype re personalised medicine
- Present the results analysed using the original scale
- Let the reader and others use an MID if they want to interpret these results
- If you want to go beyond quoting mean effects you need
  - Repeated measures
  - Very smart statistics