BMJ 2015;351:h5496 doi: 10.1136/bmj.h5496 (Published 20 November 2015)

ANALYSIS

Clinical importance cannot be ruled out using mean difference alone

Christopher Cates and **Charlotta Karner** argue that information about individual patient responses should be included in the clinical assessment of treatments

Christopher Cates *senior clinical research fellow*[1], Charlotta Karner *health technology assessment analyst lead*[2]

[1]Population Health Research Institute, St George's, University of London, Cranmer Terrace, London SW17 0RE, UK; [2]BMJ Technology Assessment Group

Clinicians and patients who face treatment choices want to know whether the results of randomised trials are clinically important, as well as statistically significant. The result is considered statistically significant if the probability that it is a chance finding is less than one in 20 (P value is less than 0.05). But the P value does not measure the size of the difference between treatment and comparator and therefore is not sufficient to assess clinical importance. The next question is: does the treatment make enough difference in outcome to be worth using?

The minimum important difference (MID) is the least change in a measurement that is judged to matter to the person being treated and is best assessed by patients themselves. For outcomes such as quality of life, which are measured as scores on a continuous scale, the MID is the minimum change in score required for the patient to decide that their treatment has been effective.[1] The MID can be used to evaluate the clinical importance of treatments, but there are different ways to do this.[2 3]

In this article we consider how to assess whether statistically significant results are also clinically important. We show that when the mean difference is statistically significant, clinical importance can be better judged using a combination of the size of the mean difference and the spread of individual responses to treatment.

More than just the mean difference

NICE guideline groups (among others), currently use the GRADE (grading of recommendations assessment, development, and evaluation) approach to evaluate continuous outcomes in clinical trials. The GRADE approach uses the estimate of the mean difference and its 95% confidence interval (CI) to assess whether there is a statistically significant difference between the treatment and its comparator.[4]

In large trials or meta-analyses the 95% CI can be very narrow, so a small mean difference may be statistically significant, but is it also clinically important? To answer this question the mean difference is compared with the MID. When the 95% CI of a mean difference does not reach the MID threshold, guideline writers using GRADE may conclude that the treatment is not clinically worthwhile, even if the result is statistically significant.

The 95% CI around a mean difference is where we are 95% sure that the average treatment effect in the population is located—it does not describe the range of results that we would expect for individual participants. The 95% CI of the mean difference is 3.92 times as wide as the standard error of the mean. By contrast, individual patient results are spread more widely, with 95% of patients having results within an interval that is 3.92 times the standard deviation. So, although the 95% CI of the mean difference may be less than the MID, this does not mean that 95% of individual patients fail to achieve the MID threshold.

We need more information about the distribution of individual responses to determine the clinical relevance of the treatment. One way is to look at how many people on treatment and on placebo had a response at least as great as the MID. Such individuals have been described as "responders," and this approach as a "responder analysis."[2 3]

Worked example

We compared the two approaches (95% CI of a pooled mean difference and responder analysis) using data obtained for a Cochrane systematic review of tiotropium versus placebo in randomised controlled trials of patients with chronic obstructive pulmonary disease (COPD).[5]

In 11 672 participants from nine trials in the review quality of life was measured with the St George's Respiratory Questionnaire. This scale runs from zero to 100, with higher scores representing a lower quality of life. The MID is a reduction of four units from baseline, which has previously been

Correspondence to: C Cates ccates@sgul.ac.uk

shown to be clinically important on the basis that it was the average change in score from the patients who judged that their treatment had been "slightly effective."[6]

Mean difference

The trials in this review found an average reduction of 2.89 units on the St George's scale for tiotropium compared with placebo. The large numbers of participants in the trials led to a narrow 95% CI around this estimate of the mean. The authors said, "Compared to placebo, tiotropium treatment significantly improved the mean quality of life (mean difference of −2.89; 95% CI −3.35 to −2.44)." The forest plot shows that the 95% CI of the estimated mean from all the trials (shown as the width of the diamond) is clearly statistically significant (P<0.00001), but it does not reach the MID threshold of a four unit reduction (shown as a dotted vertical line) (fig 1⇓).

It is tempting to conclude that the treatment has no clinically significant effect.[6 7] To assess whether this is correct, we looked at the number of responders—those who showed a reduction of at least four units—on tiotropium and on placebo.

Responder analysis

The results from each patient are rarely included in trial reports,[8] but one of the authors (CK) was able to obtain unpublished information from the trial sponsors about the number of responders in each arm.

The risk ratio of responders on tiotropium compared with placebo was 1.25 (95% CI 1.20 to 1.31) (fig 2⇓), which means that the chance of being a responder is between 20% and 31% higher in patients given tiotropium than in those given placebo. Interpreting the clinical importance of this treatment effect using ratios alone is not possible; we also need to know how many people responded to tiotropium or placebo to establish the absolute difference that the treatment makes.

Overall, 1988 of 5108 patients given placebo (39%) were responders. These are shown as 39 green faces in a Cates plot (fig 3⇓). We applied the pooled risk ratio (and its 95% CI) to this average risk on placebo to demonstrate the expected benefit of the treatment. The 10 yellow faces in the Cates plot show that for every 100 people with COPD given tiotropium for an average of a year, 10 more people (95% CI 8 to 12) would respond. The 51 red faces represent people who would not respond on either placebo or tiotropium.

The responder analysis shows that around one in 10 more participants had a noticeable benefit of treatment with tiotropium than with placebo. The number needed to treat for one additional patient on tiotropium to achieve the MID is 11 (95% CI 9 to 13). This can be used to assess the clinical relevance of the treatment, even if the mean difference does not reach the MID.

In another example, responder analysis was reported alongside the mean difference for a trial of high frequency oscillation in neonates born before 29 weeks of gestation.[9] Despite a seemingly small mean difference in lung function, responder analysis showed a clinically important difference—high frequency oscillation was associated with a lower proportion of children whose lung function results were below the 10th centile in later life (37%) than conventional ventilation (47%).[9]

We are not proposing that responder analyses should be used instead of the mean difference, as there would be considerable loss of statistical power.[3] Rather, we think that clinical importance should not be assessed using the 95% CI of the mean difference alone, but in combination with responder analysis or other methods of presenting individual responses, such as dot plots.[8]

Individual patient responses

In the tiotropium trials each participant was randomised to receive either active treatment or placebo, meaning that we cannot directly compare the response to each intervention in the same person. The mean difference between tiotropium and placebo in individual patients can only be calculated in studies in which patients are given the active treatment and placebo in random order in succession; such studies have been called "n of 1" studies.[10] This would allow us to gauge whether tiotropium was consistently better or worse than placebo for individual patients with COPD, as Senn has previously pointed out.[2 11 12]

So, even though we can see that there are more responders in the population given tiotropium, we cannot tell which patients would benefit more on tiotropium than they would have done on placebo.

Implications for practice

Guideline writers cannot fully evaluate the clinical importance of treatments without looking at both the mean difference and individual responses of participants in trials. Statistical significance should be calculated for continuous outcomes using the mean difference, but clinical importance needs further information about individual responses. Population benefit of a treatment cannot be ruled out in guidelines or systematic reviews just because the 95% CI of the mean difference fails to reach the MID.

Implications for research

Clinical trials should specify in their protocol that they will report the distribution of results in individual participants as well as the mean difference. Researchers should publish plots of individual results and responder analyses in clinical trial reports. Then guideline writers, systematic reviewers, and clinicians could use this information, as well as the mean difference, to assess the clinical importance of treatment effects measured as continuous outcomes in randomised trials.

We thank Stephen Senn, Janet Peacock, and the peer reviewers for their helpful suggestions and improvements.

Contributors and sources: CC is funded by the NIHR for his work as the coordinating editor of the Cochrane Airways group; he is also responsible for the statistical accuracy of reviews published by this group. As a former GP he has a longstanding interest in translating evidence into practice. CK is a systematic reviewer who was previously a research assistant with the Cochrane Airways group and is first author of the systematic review discussed in this paper. CC had the idea for the paper; both authors wrote it together. CC is guarantor.

Provenance and peer review: Not commissioned; externally peer reviewed.

Competing interests: We have read and understood BMJ policy on declaration of interests and declare that we have no competing interests. All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Key messages

- Mean difference is the most powerful way to measure statistical significance of continuous outcomes
- But mean difference alone is not suitable for assessing clinical importance of treatments
- Responder analyses are also needed to interpret the clinical importance of treatment effects
- Researchers should be encouraged to publish both mean differences and details of individual responses from their clinical trials

1 McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *JAMA* 2014;312:1342-3.
2 Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. *BMJ* 1998;316:690-3.
3 Snapinn SM, Jiang Q. Responder analyses and the assessment of a clinically relevant treatment effect. *Trials* 2007;8:31.
4 Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence—imprecision. *J Clin Epidemiol* 2011;64:1283-93.
5 Karner C, Chong J, Poole P. Tiotropium versus placebo for chronic obstructive pulmonary disease. *Cochrane Database Syst Rev* 2012;7:CD009285.
6 Jones PW. Interpreting thresholds for a clinically significant change in health status in asthma and COPD. *Eur Respir J* 2002;19:398-404.
7 Jones PW, Beeh KM, Chapman KR, Decramer M, Mahler DA, Wedzicha JA. Minimal clinically important differences in pharmacological trials. *Am J Respir Crit Care Med* 2014;189:250-5.
8 Schriger DL, Savage DF, Altman DG. Presentation of continuous outcomes in randomised trials: an observational study. *BMJ* 2012;345:e8486.
9 Zivanovic S, Peacock J, Alcazar-Paris M, et al. Late outcomes of a randomized trial of high-frequency oscillation in neonates. *N Engl J Med* 2014;370:1121-30.
10 Guyatt G, Sackett D, Taylor DW, Ghong J, Roberts R, Pugsley S. Determining optimal therapy—randomized trials in individual patients. *N Engl J Med* 1986;314:889-92.
11 Senn S. Individual response to treatment: is it a valid assumption? *BMJ* 2004;329:966-8.
12 Senn S. Applying results of randomised trials to patients N of 1 trials are needed. *BMJ* 1998;317:537.

Accepted: 05 October 2015

Cite this as: *BMJ* 2015;351:h5496

Figures

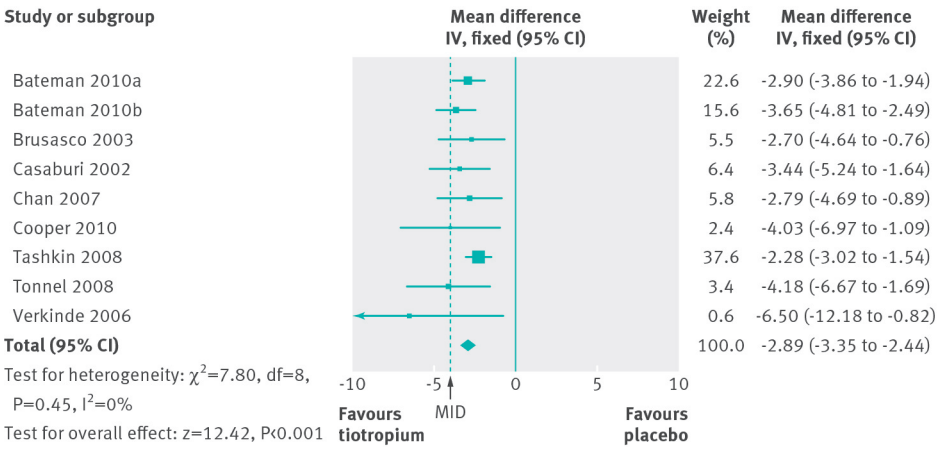

Fig 1 Forest plot of quality of life (total score on St George's Respiratory Questionnaire) for tiotropium versus placebo.[5] SE=standard error of the mean, IV=inverse variance. The dotted line indicates the four unit threshold for minimum important difference (MID).

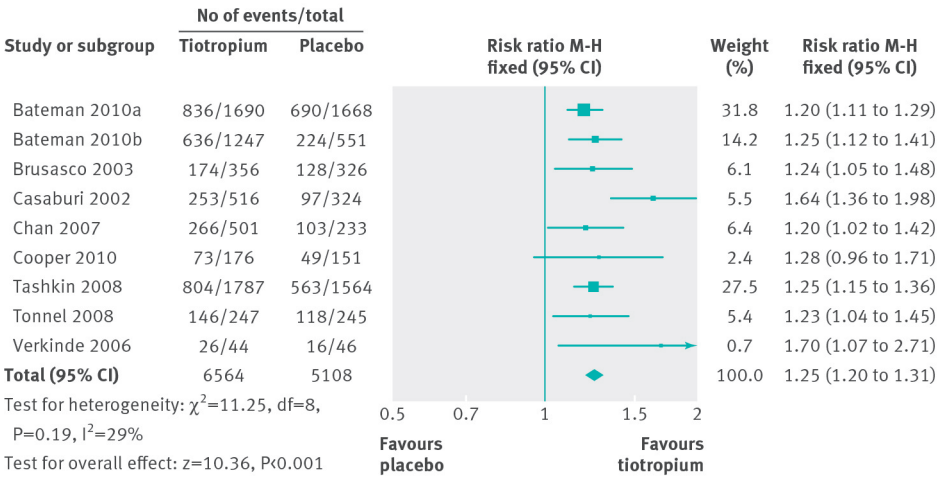

Fig 2 Forest plot of the number of people who improved by at least four units in quality of life (total score on St George's Respiratory Questionnaire) for tiotropium versus placebo.[5] M-H=Mantel-Haenszel.

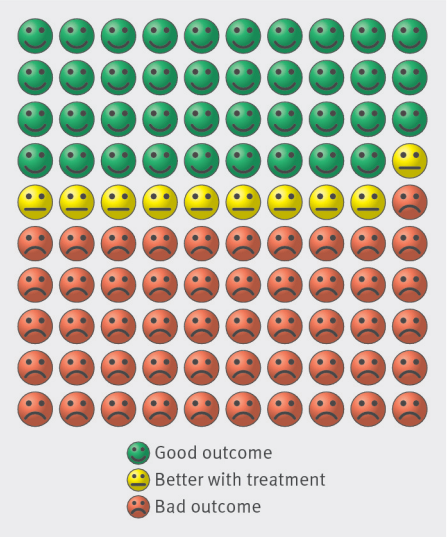

Fig 3 Cates plot of responders measured by a four unit improvement in St George's Respiratory Questionnaire on tiotropium and placebo (created using Visual Rx at www.nntonline.net)