Effect measure for quantitative endpoints: Statistical versus clinical significance, or "how large the scale is?"

Cristian Baicus, Simona Caraiola

Clinica de Medicina Interna, and Réseau d'Epidémiologie Clinique International Francophone (RECIF), Spitalul Colentina, Soseaua Stefan cel Mare 19–21, Sector 2, 21125 Bucharest, Romania

ABSTRACT

Whenever a study finds a statistical significance for the difference between treatment and placebo, we must always ask ourselves if the difference is clinically important, too. In order to do this, we need to know at least how large the scale is, and to compare the size of the scale with the size of the effect. Sometimes, the effect of placebo is greater than the intrinsic effect of the drug. The results of these studies are expressed as averages of effects on patients who respond to treatments and patients who do not, so in our daily practice we must distinguish these categories, treating only the first.

Frequently when a study finds a statistical significance for the difference between treatment and placebo, the scientific society of the respective specialty introduces this new treatment in its guidelines [1], leaving apart the intense promotion by the pharmaceutical industry. However, we must always ask ourselves if the differences detected in continuous variables-endpoints with the tested treatment is clinically important. In order to do this, we need to know at least how large is the scale and how clinically significant is the treatment effect.

To demonstrate the difference between statistical and clinical significance we will choose a few examples. The first is a study concerning the effect of the locally applied diclofenac in knee osteoarthritis, published in CMAJ [2] in 2004 and reviewed in Evidence Based Medicine with the title "Topical diclofenac improved pain and physical function with no systemic side effects in primary osteoarthritis of the knee" [3]. The benefits of topical diclofenac and placebo were: 3.9 points versus 2.5 (difference: 1.4) on the pain scale; 11.6 points versus 7.1 (difference: 4.5) on the physical function scale; and 1.5 points versus 0.6 (difference: 0.9) on the stiffness scale, so the locally applied diclofenac reduced more than placebo all the parameters from the WOMAC scale, and the differences were statistically significant. As a result, along with the title, the commentator of Evidence Based Medicine wrote: "The 4 week blinded RCT by Bookman et al. in patients with primary OA of the knee found that topical diclofenac was significantly better than both vehicle controlled and placebo solutions in reducing WOMAC pain, physical dysfunction, and stiffness". Now we can ask ourselves what "significantly better" means?

In order to assess the magnitude of the effect and the clinical significance, we must find out what the WOMAC scale means, and searching on the Internet (by Google search), we find that this scale assigns 50 points for pain, 170 points for physical function and 20 points for stiffness. Comparing the size of the scale with the size of the effect, we have the right to think: on a 50 points scale, is a 1.4 points reduction of the pain clinically important? The same, on a 170 points scale, is a 5.9 points improvement of the physical function clinically significant? Moreover, one can observe another phenomenon: the intrinsic effect of the active substance (the difference between the total effect and the placebo effect) is lesser than the placebo effect (1.4 points against 2.5 obtained with placebo for pain, 4.5 points against 7.1

obtained with placebo for the improvement of the physical function, 0.2 points against 0.6 obtained with placebo for the pain in walking). In this situation, a philosophical dilemma emerges: how much attention deserves a drug which intrinsic effect is lesser than the placebo effect? The answer depends on how clinically important is this effect in a view of potential side effects.

Another study [4] (published in Journal of Rheumatology and reviewed in bmjupdates — www.bmjupdates.com) concerned the effect of ketoprofen patch in tendinitis of recent onset. The effect was measured on a 100 mm visual analogue scale. After one week of treatment, the decrease in pain was of 25.8±24.5 mm (37%) in the placebo group, and 38.4±25.6 mm (56%) in the treatment group. The difference was statistically significant (p=0.0013). Looking at the results, we see that the treatment decreased the pain with almost 40 mm (which is, this time, important on a scale of 100 mm), while placebo decreased it lesser, and the difference between them – therefore the intrinsic effect of ketoprofen – is of 38.4–25.8=12.6 mm. How clinically important is this effect giving into consideration side effects associated with the therapy?

Another example, from respiratory diseases, is the TRISTAN study [5], which investigated the effect of the combination salmeterol/fluticasone in chronic obstructive pulmonary disease (COPD). One of the studied outcomes was the quality of life measured by the St George's respiratory questionnaire (SGRQ), a questionnaire validated for COPD. In the Summary of this study we read: "Combination treatment produced a clinically significant improvement in health status", and in the Results section: "Only the combination group showed a clinically significant improvement in health status questionnaire score by week 52. The raw mean changes in health status total score were −4.3 (SD 10.8) by week 8 and −4.5 (SD 12.9) at week 52. The change in SGRQ score in the combination group over 52 weeks at the end of the study was significantly greater than that in both the placebo and fluticasone groups" — the difference was of 2.2 points. All we have to do, now, is to see if this difference is clinically important. In order to do this, we must see, first, what the St George's questionnaire represents. On the site of the American Thoracic Society we find essential information (http://www.atsqol.org/sections/instruments/pt/pages/george.html—the search was performed on January 2008). First, the questionnaire has 100 points scale, sowe can ask ourselves, how important is an improvement of 2.2 points? We learned that a treatment that improves the score is considered slightly efficacious for a 4 units change, moderately efficacious for an 8 units change, and very efficacious for a 12 units change. Placebo improved the score with almost 3 points, so it was not far from being slightly efficacious, while the combination salmeterol/fluticasone, which brought a 4.5 point improvement, was certainly "slightly efficacious". Concerning the statistical significance of the difference between the combination and placebo, p=0.0003, a high statistical significance for a slight clinical importance. The interpretation is: these 2.2 points of difference between treatment and placebo are not the result of chance, but the result of treatment. Or, using the richer information provided by the confidence interval, we could say that the treatment with the combination of salmeterol/fluticasone improves the SGRQ with 1.1 to 3.3 points better than placebo. Therefore the difference between placebo and treatment is below being "slightly efficacious".

Of course, it is possible that even a small decrease in a continuous variable to result in an important clinical effect. To decide if this is the case, one has to establish thresholds of clinical importance. If the simple comparison of the magnitude of the scale with the treatment effect looks over-simplistic, there are some apparently more scientific methods to decide the threshold of clinical importance, and this threshold is called "the minimal important difference" (MID), defined as "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" [6].

For the MID determination, two approaches exist [7]: one purely statistical and another more clinical, anchor based. The first is distribution based, relying on expressing an effect in terms of the underlying distribution of the results (the Cohen's standardized effect size, computed as the mean change divided by the standard deviation, and Norman's 0.5×standard deviation rule of the thumb) [8]. The anchor based method assesses the relationship between the Health Related Quality of Life (HRQoL) measure and an independent anchor, like the change on a global scale [7]. Both methods are imperfect, so both would have to be used and confronted in the evaluation of a HRQoL questionnaire, and there is a vast body of literature evaluating these questionnaires (including the WOMAC scale index and SGRQ [9,10]) in different settings, on different patients with different diseases).

We must stress the fact that the results of these studies are expressed as means[1]: the mean improvement of the St George score is 2.2, which might be the average between two patients who did not improve at all, and one patient who improved by 6.6 points. In our daily practice, we must try identifying that very patient who would benefit from the treatment.

Acknowledgements
We thank Drs. Pierre Duhaut and Victor Novack for their insights on earlier versions of this manuscript.

References

[1] Baicus C, Chivu R. Drug politics — economic viewpoint of a practitioner. Rev Med Chir Soc Med Nat Iasi 2004;108:674–8 [French].
[2] Bookman AA, Williams KS, Shainhouse JZ. Effect of a topical diclofenac solution for relieving symptoms of primary osteoarthritis of the knee: a randomized controlled trial. CMAJ 2004;171:333–8.
[3] Topical Diclofenac Improved Pain and Physical Function with No Systemic Side Effects in Primary Osteoarthritis of the Knee. Ann Cranney, Siobhan O'Donnell (Commentators). EBM 2005; 10:81.
[4] Mazieres B, Rouanet S, Guillon Y, Scarsi C, Reiner V. Topical ketoprofen patch in the treatment of tendinitis: a randomized, double blind, placebo controlled study. J Rheumatol 2005;8:1563–70.
[5] Calverley P, Pauwels R, Vestbo J, Jones P, Pride N, Gulsvik A, et al. for the TRISTAN (TRial of Inhaled STeroids ANd long-acting β2 agonists) study group. Combined salmeterol and fluticasone in the treatment of chronic obstructive pulmonary disease: a randomised controlled trial. Lancet 2003;361:449–56.
[6] Jaeschke R, Singer J, Guyatt GH. Measurement of health status. ascertaining the minimal clinically important difference. Cont Clin Trials 1989;10:407–15.
[7] Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health Qual Life Outcomes 2003;11(1):4.

[1] Although always expressed as means with standard deviations, for the pedantic statistician this is wrong: the composed scores from the HRQoL questionnaires are ordinal, and not scale variables. Even in scale variables, the distribution can depart form normal (often skewed to the right), so expressing them as mean with standard deviation can lead to overestimation of the statistics for the comparison.

[8] Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health related quality of life: the remarkable universality of half a standard deviation. Med Care 2003;41:582–92.
[9] Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. Arthritis Rheum 2001;45:384–91.
[10] Jones PW. Interpreting thresholds for a clinically significant change in health status in asthma and COPD. Eur Respir J 2002;19:398–404.