ELSEVIER

Journal of Clinical Epidemiology 62 (2009) 711−715

Journal of Clinical Epidemiology

The evidence provided by a single trial is less reliable than its statistical analysis suggests

George F. Borm[a,*], Oscar Lemmers[a], Jaap Fransen[b], Rogier Donders[a]

[a]*Department of Epidemiology, Biostatistics and HTA, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands*
[b]*Department of Rheumatology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands*

Accepted 22 September 2008

Abstract

Objective: To investigate whether a single trial can provide sufficiently robust evidence to warrant clinical implementation of its results. Trial-specific factors, such as subject selection, study design, and execution strategy, have an impact on the outcome of trials. In multiple trials, they may lead to heterogeneity that can be taken into account in the (random effects) meta-analysis. Single trials lack this method of estimating the impact of such factors, and this affects the credibility of the results.

Study Design and Setting: To indicate how much the precision of the results of a single trial might be overestimated, we calculated the ratio of the widths of the confidence intervals when heterogeneity was taken into account and when it was not.

Results: The ratios of the widths of the confidence intervals with and without between-study variability were 1.15, 1.41, and 2.00, when the heterogeneity I^2 values were 0.25, 0.50, and 0.75, respectively.

Conclusion: The results of a single trial should be interpreted with caution. When it is difficult to predict or determine how trial-specific factors influence the results, the best way to evaluate the performance of a treatment is to use multiple, possibly smaller, trials. © 2009 Elsevier Inc. All rights reserved.

Keywords: Clinical trial; Confidence interval; Type I error; Heterogeneity; Meta-analysis; P-value

1. Introduction

Even when multiple trials are designed to address the same question, their results and conclusions may differ. Possible explanations for these differences are chance variation or differences in the implementation of the trial, for example, the characteristics of the participants, the design and execution of the trial, the treatment administration or dosage, concomitant exposures, outcome assessment, and others. Other contributing factors might be the local health care system, the way the investigational site is organized, and even the statistical methods used in the analysis [1−3]. However, it is often impossible to identify exactly which factor(s) caused the differences between the results [4]. This (unexplained) heterogeneity means that study results must be interpreted with caution.

Besides the differences in treatment results between the trials, one can also find differences in treatment results between trials and clinical practice [5,6]. The latter differences are likely to be at least as large as those between the trials; hence, the (unexplained) heterogeneity of the trial results provides an indication of the discrepancy that may exist between the trial results and clinical practice. Consequently, the heterogeneity should be taken into account in the evaluation and interpretation of the results.

When only one trial has been carried out, there is no information available about possible heterogeneity; hence, it may not be prudent to generalize the results. Therefore, we investigated the possible consequences of this lack of information on the interpretation of the outcome of a single trial. Based on the results of our investigations, we discuss whether a single trial can provide sufficient evidence, or whether a series of (possibly smaller) trials is more appropriate.

The outline of the article is as follows: In Section 2, we make a quantitative assessment of the differences between an analysis that takes heterogeneity into account and one that does not. Both approaches are used in the meta-analysis of series of trials (random effects analysis and

* Corresponding author. Department of Epidemiology, Biostatistics and HTA, 133, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands. Tel.: +31-243617667; fax: +31-243613505.

E-mail address: g.borm@ebh.umcn.nl (G.F. Borm).

0895-4356/09/$ − see front matter © 2009 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2008.09.013

fixed effect meta-analysis), but when only one study is available, only the latter method can be used. We calculate how much the confidence intervals are narrower and the *P* values are smaller when the fixed effects approach is used, compared with the random effects method. Section 3 offers some examples, and in Section 4, we discuss whether a single trial can provide sufficient evidence of the value of a treatment to adopt it in clinical practice.

2. Confidence intervals, *P* values, and type I errors of single trials

When a trial compares the means of an efficacy variable between two groups, the observed difference between the means can be written as the sum of the ''true'' difference δ_{trial} and a random fluctuation e_{trial}:

$$d_{trial} = \delta_{trial} + e_{trial}. \quad (1)$$

Here, e_{trial} represents the within-trial variability. The mean of e_{trial} is 0 and its variance is σ^2_{trial}.

However, not all trials can be expected to have the same ''true'' efficacy, and the true trial difference δ_{trial} may vary around the global true difference δ:

$$\delta_{trial} = \delta + b_{trial}, \quad (2)$$

where b_{trial} has mean 0 and variance τ^2. Although the parameter τ^2 represents the heterogeneity, we measure the heterogeneity in the standard way by using I^2, the proportion of the total variance in the study result caused by heterogeneity [7]. In the Appendix, we explain the relationship between I^2 and τ^2 (available on the journal's website at www.jclinepi.com).

2.1. The trial difference vs. the global difference

The outcome d_{trial} is an estimate of the true efficacy δ_{trial} of the study. Using d_{trial} and (an estimate of) the variance σ^2_{trial}, confidence intervals for the study difference δ_{trial} can be calculated, or the null hypothesis $\delta_{trial} = 0$ can be tested.

To test the global null hypothesis $\delta = 0$, the heterogeneity must be known or at least estimated, in which case, the confidence intervals for the global treatment effect δ can also be calculated. It is possible to estimate δ, because the means of e_{trial} and b_{trial} are 0; hence, the outcome d_{trial} of the trial is not only an estimate of δ_{trial}, but also of δ. However, the confidence intervals for the global difference δ will be wider than the confidence intervals for the trial difference δ_{trial}, and the *P* value of a test of the global null hypothesis $\delta = 0$ will be larger than that of a test of $\delta_{trial} = 0$. To investigate the exact relationship between the confidence intervals and *P* values related to δ_{trial} and those related to δ, we first consider the case when a Z-test is used for the analysis, that is, when e_{trial} and b_{trial} each have a normal distribution with known variance.

2.2. Confidence intervals

The width $w_{trial,\alpha}$ of the 1 - α two-sided confidence interval for the trial difference δ_{trial} is the distance between the point estimate at the center and the lower (or upper) limit of the confidence interval. Similarly, $w_{global,\alpha}$ is the width of the confidence interval for the global difference δ. From Equation (A4) in the Appendix, it immediately follows that, for normally distributed e_{trial} and b_{trial} with known variances,

$$w_{global,\alpha} = w_{trial,\alpha}/\sqrt{1 - I^2}, \quad (3)$$

the ratio between the widths depends on the heterogeneity, but not on α.

Figure 1 shows the relation between the widths of the confidence intervals and heterogeneity. When the heterogeneity is low ($I^2 = 0.25$), the width of the confidence interval for δ is 115% of the width of the confidence interval for δ_{trial}. When the heterogeneity is moderate ($I^2 = 0.50$) and high ($I^2 = 0.75$), the width ratios are 141% and 200%, respectively.

2.3. P values

In the Appendix, we show that for a two-sided Z-test, the *P* value for the global null hypothesis $\delta = 0$ is

$$P_{global} = 2\left(1 - \Phi\left(\Phi^{-1}(1 - P_{trial}/2)\sqrt{1 - I^2}\right)\right), \quad (4)$$

where P_{trial} is the *P* value related to the trial null hypothesis $\delta_{trial} = 0$, and Φ is the standard normal cumulative distribution function.

When $I^2 = 0.25$ and P_{trial} takes the values of 0.05, 0.01, and 0.001, the P_{global} values are 0.09, 0.026, and 0.0044, respectively. When $I^2 = 0.5$, P_{global} is 0.17, 0.069, and 0.02, respectively. When $I^2 = 0.75$, the P_{global} values are 0.33, 0.20, and 0.10. Figure 2 illustrates the relationship between P_{global} and P_{trial} for each value of I^2.

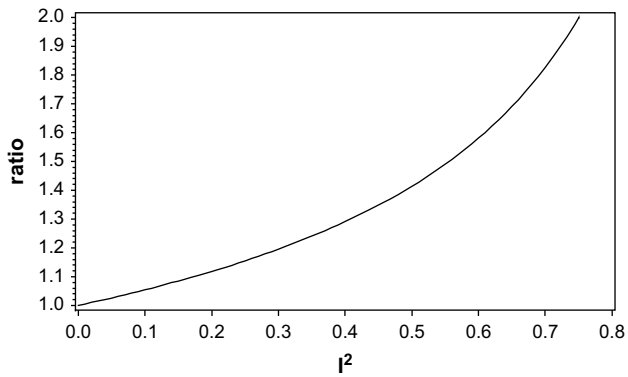

Fig. 1. Ratio between the width of the confidence interval for the global treatment effect δ and the width of the confidence interval for the trial treatment effect δ_{trial}.

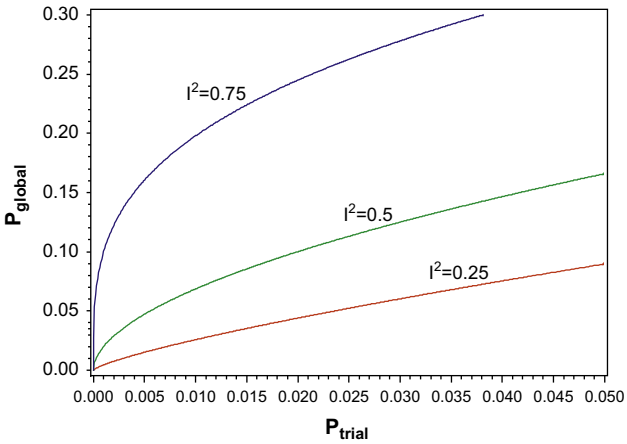

Fig. 2. P value of the two-sided test on the global null hypothesis $\delta = 0$ vs. the P value of the two-sided test on the trial null hypothesis $\delta_{\text{trial}} = 0$.

2.4. Type I error rates

Equation (4) immediately leads to the relationship

$$\alpha_{\text{global}} = 2\left(1 - \Phi\left(\Phi^{-1}(1 - \alpha_{\text{trial}}/2)\sqrt{1 - I^2}\right)\right). \quad (5)$$

between the two-sided error rates of the Z-tests on the global and trial null hypotheses.

Conversely,

$$\alpha_{\text{trial}} = 2\left(1 - \Phi\left(\Phi^{-1}(1 - \alpha_{\text{global}}/2)/\sqrt{1 - I^2}\right)\right). \quad (6)$$

Using this formula, we calculated the trial significance level that is required to keep the error level of the global test at a certain level. When $I^2 = 0.25$, 0.5, and 0.75, then the trial significance levels required to keep the global error rate at 0.01 are 0.0029, 0.00027, and 0.00000026, respectively. To keep the error rate at 0.05, the trial significance levels should be 0.024, 0.0056, and 0.000089, respectively.

2.5. General case

Although we assumed that the outcomes were normally distributed and that a Z-test was used, the results (3)–(6) will be approximately correct in several other situations:

1. Asymptotically, the Z-test and the t-test are equivalent. In practice, this means that Equations (3)–(6) are approximately correct for trials with a normally distributed outcome and 20 participants or more.
2. By the law of large numbers, the mean of the results of any experiment on large groups will approximately follow a normal distribution; hence Equations (3)–(6) will approximately be valid. The group size that is required depends on the distribution of the outcome variable, but 20 participants per group are usually sufficient. For trials with a dichotomous outcome, a normal approximation is considered appropriate when $nP > 5$ and $n(1 - P) > 5$, where n is the group size and P is the probability of one of the outcomes. A more conservative approach would be to set the boundary at 10 or 15, instead of 5.
3. Many analysis methods, even nonparametric ones, such as the Wilcoxon test, are based on a test statistic that approximately follows a normal or a t-distribution; hence, the formulas for the P values and error rates (4)-(6) will be approximately correct.
4. In the case of a lognormal distribution, the logarithm of the variable follows a normal distribution. Confidence intervals are calculated on a logarithmic scale and then transformed back. As a result, Equations (4)-(6) are valid, and we can use Equation (3) to recalculate the confidence interval on the log scale. The same approach can be used for the odds ratio and the hazard ratio. We illustrate this in the second example in the following section.

3. Examples

3.1. Disease-modifying antirheumatic drugs for rheumatoid arthritis

Choi et al. conducted a meta-analysis on the efficacy of combining disease-modifying antirheumatic drugs in patients with Rheumatoid Arthritis [8]. Although the trials were fairly heterogeneous in design and treatment modalities, they found a low to moderate heterogeneity of $I^2 = 0.33$; hence, it seems reasonable to assume that the heterogeneity in this therapeutic area will not exceed $I^2 = 0.33$. This information makes it possible to assess the robustness of the results of a new trial that evaluates a combination of disease-modifying anti-rheumatic drugs, because we can simply calculate the global confidence interval by using Equation (3), that is, by multiplying the width of the trial confidence interval by 1.2.

3.2. Electromagnetic fields as risk factors for leukemia

The heterogeneity of a series of trials that evaluate the possible impact of electromagnetic fields on the incidence of leukemia was 0.69 [7]. The primary outcome of the studies was the odds ratio of the incidences in the exposed and nonexposed group. Suppose that a new study that evaluates the impact of mobile telephone transmitters results in an odds ratio of 1.75, with 95% confidence interval 1.35-2.27. On the natural logarithm scale, this corresponds to a log odds ratio of 0.56, with confidence interval 0.30-0.82. As the log odds ratio approximately follows a normal distribution, we can adjust the confidence interval using Equation (3). When $I^2 = 0.69$, the adjustment factor is 1.8; hence, the global confidence interval is 0.09-1.03.

After back transformation, we obtain the adjusted confidence interval of 1.1-2.8 for the odds ratio.

A further refinement of this result is possible, because in the electromagnetic field trials series, the heterogeneity varied from 0.17 for poor-quality trials to 0.79 for high-quality studies. For $I^2 = 0.17$ and 0.79, the adjustment factors for the length of the confidence interval are 1.1 and 2.2, respectively. This results in the confidence intervals of 1-3.1 when the new study is of poor quality and 1.3-2.3 when it is of high quality.

4. Discussion

We discuss some major implications of our findings.

4.1. Robustness of the results of a single trial

When the results of only one trial are available, it may be wise to assess the possible impact of heterogeneity on its results, because heterogeneity is present in many trials and trial series. It may occur in 50% of trial series, and in 25% of the series, it may exceed $I^2 = 0.5$. In 12.5% of the series, it may even exceed $I^2 = 0.75$ [7]. The results described in Section 2 make it possible to assess the robustness of the trial and check whether the confidence interval of the main outcome of the trial is sufficiently narrow and far from 0 [9,10].

In general, the degree of heterogeneity is unknown and is likely to differ between therapeutic areas, types of interventions, and types of trials, although even the quality of the trial may play a role. In specific cases, it may be possible to obtain an estimate from meta-analyses of similar types of trials or interventions, but such an estimate and the resulting correction of the P values and confidence intervals will be at best only approximately correct.

When no (reliable) information about the heterogeneity is available, a P value of 0.01 may be a suitable criterion for statistical significance. It restricts the global error rate to 0.05 as long as the heterogeneity does not exceed 0.42, and this is probably the case in more than 70% of trials series [7]. A more conservative option would be $P = 0.001$, because it keeps the global error rate below 0.05 as long as the heterogeneity is less than 0.65 (approximately 80% of the trials series).

4.2. Large trials

It has been suggested that a single well-planned large trial is the best way to evaluate the efficacy of a treatment [11–13]. Such a trial should have a detailed protocol and be executed according to the highest standards. The power must be at least 90% and the inclusion criteria, treatments and endpoints should reflect clinical practice [5,6]. However, some authors expressed doubt about this suggestion and pointed out that, in some cases, the results of large trials have been challenged and refuted over the course of time [3,14,15]. The results in this paper offer theoretical and numerical support for their doubts.

It seems reasonable that in therapeutic areas that are known to produce variation in study results, or in which seemingly convincing results could not be confirmed by subsequent studies, it would be better not to draw conclusions from a single study, even if it is large [9,10]. A single trial may be sufficient only when the results of trials in a specific therapeutic area are fairly robust and variations in the treatment effects and patient populations are of minor importance [5,10,16]. However, even in this situation, a single trial may have its drawbacks, because the trial may be biased, not only by the way it is designed, but also by the way it is conducted. This bias may go unnoticed in one trial, but other subsequent studies may not all suffer from that same bias, so the global result may be more accurate.

4.3. Small trials and mega trials

An alternative to large trials is a series of smaller trials. Borm et al. have shown that a random effects analysis on two trials with 80% power, or on five trials with at least 30% power, keeps the error rate at the desired level [17]. Using the methods described in that study, we found that, for I^2 between 0 and 0.75, the error rates of random effects meta-analyses on series of two to 20 trials with 10% to 50% power were all between 0.045 and 0.055. Clearly, a series of small trials leads to correct P values.

Small trials could have other disadvantages. For example, they could lead to publication bias, because small trials with a nonsignificant result may not be published. However, when the trials have at least 30% power, the impact of publication bias and other possible disadvantages is likely to be small, and the results of a meta-analysis on the trials will be reliable [17].

Another option is a mega trial. A mega trial is a large multicenter trial that has sufficient power to address subgroup differences [3]. The outcome at each center can be seen as the result of a smaller (sub) study and the mega trial is similar to a meta-analysis on several smaller studies.

However, the heterogeneity may still be limited, because all centers of the mega trial use exactly the same protocol, follow the same procedures, share the same steering committee, and others. For the trial to produce realistic results that reflect (variations in) clinical practice, the trial procedures have to be sufficiently loose, and the protocol must grant great freedom to the investigators [6,14,16].

In addition, the heterogeneity between the centers of the mega trial should also be taken into account in the analysis. The results in Section 2 refer not only to the situation of a single trial vs. multiple trials, but also to the differences between a random effects analysis and a fixed effect analyses. This implies that a mega trial should be analyzed using random effects meta-analysis methods, as if it were a series of smaller trials, one in each center [18].

When the protocol of a mega trial is too restrictive, it may be difficult to translate the results into clinical practice. However, this may also be the case when multiple smaller trials are conducted that are exact clones of each other. Such ''sister trials,'' with exactly the same protocol, offer only weak evidence that results are replicable and it is preferable to have more variation in the design and the conduct of the trials [10,14,16].

5. Conclusion

The quantitative analysis described in this article offers an opportunity to assess the robustness of trial results. Even when a trial is large, its outcome should be interpreted with caution. When heterogeneity cannot be excluded, the best evidence of a treatment's clinical performance can be obtained by multiple, possibly smaller, trials that are independent and that differ in their protocols.

References

[1] Egger M, Smith GD, Altman DG, editors. Systematic reviews in health care. London: BMJ Publishing Group; 2001.
[2] Fletcher J. What is heterogeneity and is it important? BMJ 2007;334: 94−6.
[3] Shrier I, Platt RW, Steele RJ. Mega-trials vs. meta-analysis: precision vs. heterogeneity? Cont Clin Trials 2007;28:324−8.
[4] Chalmers TC, Berrier J, Sacks HS, Levin H, Reitman D, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline. II: replicate variability and comparison of studies that agree and disagree. Stat Med 1987;6:733−44.
[5] Flather M, Delahunty N, Collinson J. Generalizing results of randomized trials to clinical practice: reliability and cautions. Clin Trials 2006;3:508−12.
[6] Zwarenstein M, Oxman A. Why are so few randomized trials useful, and what can we do about it? J Clin Epidemiol 2006;59:1125−6.
[7] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557−60.
[8] Choy EHS, Smith C, Doré CJ, Scott DL. A meta-analysis of the efficacy and toxicity of combining disease-modifying anti-rheumatic drugs in rheumatoid arthritis based on patient withdrawal. Rheumatology 2005;44(11):1414−21.
[9] Points to consider on application with 1. Meta-analyses; 2. one pivotal study. London: EMEA; 2001.
[10] Guidance for industry. Providing evidence of effectiveness for human drug and biological products. Washington: FDA; 1998.
[11] Egger M, Ebrahim S, Smith GD. Where now for meta-analysis? Int J Epidemiol 2002;31:1−5.
[12] LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med 1997;337:536−42.
[13] LeLorier J, Gregoire G. Comparing results from meta-analyses vs large trials. JAMA 1998;280:518−9.
[14] Edwards SJL, Lilford RJ, Braunholz D, Jackson J. Why ''underpowered'' trials are not necessarily unethical. Lancet 1997;350:804−7.
[15] Ioannidis JP, Cappelleri JC, Lau J. Meta-analyses and large randomized, controlled trials. NEJM 1998;338:59−62.
[16] Matthews JNS. Small trials: are they all bad? Stat Med 1995;14: 115−26.
[17] Borm GF, den Heijer M, Zielhuis GA. Publication bias was not a good reason to discourage trials with low power. J Clin Epidemiol 2009;62:47−53.
[18] van Houwelingen JC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 2002;21:589−624.

Appendix

When the differences between the treatment groups in all studies have e a common variance σ^2, I^2 measures the percentage of total variance due to heterogeneity [1]:

$$I^2 = \frac{\tau^2}{\tau^2 + \sigma^2}, \tag{A.1}$$

where τ^2 is the between study variance.

In our article, we extend the definition of I^2 to a single trial and we define

$$I^2 = \frac{\tau^2}{\tau^2 + \sigma^2_{trial}}, \tag{A.2}$$

where σ^2_{trial} is the variance of difference between the treatment groups in the trial.

Then for the total variance σ^2_{global},

$$\sigma_{global} = \sqrt{\tau^2 + \sigma^2_{trial}}, \tag{A.3}$$

so (eq. A.2)

$$\sigma_{global} = \sigma_{trial}/\sqrt{1 - I^2}\,. \tag{A.4}$$

When the outcome of a trial is d_{trial}, the two-sided Z-test on the trial null-hypothesis δttial=0 has P value

$$P_{trial} = 2\left(1 - \Phi\left(\frac{|d_{trial}|}{\sigma_{trial}}\right)\right), \tag{A.5}$$

So

$$\frac{|d_{trial}|}{\sigma_{trial}} = \Phi^{-1}(1 - P_{trial}/2), \tag{A.6}$$

where Φ is the cumulative normal distribution function.

When the result of only one trial is available, the outcome d_{trial} of that trial must also be used to test the global null-hypothesis δ=0 and the P value for the global null-hypothesis δ=0 is (eq. A.3)

$$P_{global} = 2\left(1 - \Phi\left(\frac{|d_{trial}|}{\sqrt{\sigma^2_{trail} + \tau^2}}\right)\right)$$
$$= 2\left(1 - \Phi\left(\frac{|d_{trial}|}{\sigma_{trial}}\sqrt{\frac{\sigma^2_{trial}}{\sigma^2_{trial} + \tau^2}}\right)\right). \tag{A.7}$$

On the basis of eqs A.2 and A.6 it follows that

$$P_{global} = 2\left(1 - \Phi\left(\Phi^{-1}(1 - P_{trial}/2)\sqrt{1 - I^2}\right)\right). \tag{A.8}$$

References

1. Higgins JP, Thompson SG. *Quantifying heterogeneity in meta-analysis.* Stat med 2002; 21:1539-1558.