The ethics of sample size: The whole picture should be considered

Knottnerus, J.A.; Bouter, L.M.

published in
Journal of Clinical Epidemiology
2003

DOI (link to publisher)
10.1016/S0895-4356(02)00597-8

document version
Publisher's PDF, also known as Version of record

quence. If coumarin would be really inferior to aspirin ($RR_{true} > 1$), which is possible, then there is at most 5% chance to observe a false significant effect ($RR_{obs} < 1$ significantly) and to recommend the truly inferior treatment coumarin that in addition has the more burdensome regimen.

2. Notice the inconsistency

Again, *suppose that coumarin is really inferior* to aspirin, that is $RR_{true} > 1$. The two-sided significance level 0.05 implies a 2.5% chance to recommend the truly inferior treatment coumarin, that is $RR_{obs} < 1$ significantly. The one-sided significance level 0.05 implies a 5% chance to recommend the truly inferior treatment coumarin, and this greater chance to recommend a truly inferior treatment was chosen because of the added disadvantage of a more burdensome regimen. In my opinion, the chance to recommend a truly inferior treatment with a more burdensome regimen should be less than (or equal to) the chance to recommend a truly inferior treatment without a burdensome regimen, contrary to the advice by Knottnerus and Bouter [1]. A larger significance level may be chosen in an equivalence trial where the reference intervention is more burdensome. The next paragraph offers the well known solution.

3. Confidence intervals

Section 5.5 of an international (European Union, Japan, United States) guideline [3], on statistical principles for clinical trials, prefers the approach of setting the significance level for one-sided tests at half the conventional significance level for two-sided tests, because "this promotes consistency with the two-sided confidence intervals that are generally appropriate for estimating the possible size of the difference between two treatments." Moreover, a one-sided P-value of .04 will *not* convince [4] your colleagues if the 95% confidence interval for the true relative risk *includes* the null relative risk 1.0.

4. Interim analyses

In an interim analysis an O'Brien-Fleming 95% confidence interval for the true relative risk may be computed. The trial may be stopped when this interval excludes the null value $RR_{true} = 1$ and thus indicates statistical significance [5]. The trial may also be stopped when this interval, for example $0.92 \leqslant RR_{true} \leqslant 1.80$, excludes a relative risk that is clinically relevant. Notice that this example interval, from 0.92 to 1.80, would stop the trial only in case a one-sided test was chosen, not in case a two-sided test was chosen.

5. Unequal sample sizes

It is more efficient to have about the same number of events in each treatment group, instead of the same number of patients [6,7]. Suppose that investigators demand 90% power to detect the relative risk $RR_{true} = 2/3$. If $n_{aspirin} = (2/3)n_{coumarin}$ is chosen, then 1,833 patients should enter the study, 733 on aspirin and 1,100 on coumarin, and 220 events are expected in 1833×2.5 patient years. In case of equal group sizes, 1840 patients should enter the study and 230 events are expected. Warning: in case unequal sample sizes are planned for a survival analysis, wrong answers are obtained from a sample size formula for comparing proportions in a Pearson chi-squared test.

6. Conclusion

Knottnerus and Bouter [1] rightly explained that a one-sided test may be preferred in certain cases, but they should pay attention to the chance to recommend a truly inferior treatment. The significance level for a one-sided test should be half the appropriately chosen significance level for a two-sided test. In case interim analyses are planned, a one-sided view may stop a trial earlier than a two-sided view.

Hubert J. A. Schouten
Department of Methodology and Statistics, University of Maastricht, P.O. Box 616, NL 6200 MD Maastricht, The Netherlands
E-mail: hubert.schouten@stat.unimaas.nl

References

[1] Knottnerus JA, Bouter LM. The ethics of sample size: two-sided testing and one-sided thinking. J Clin Epidemiol 2001;54:109–10.
[2] Hellemons BSP, Langenberg M, Lodder J, Vermeer F, Schouten HJA, Lemmens Th, van Ree JW, Knottnerus JA. Primary prevention of arterial thromboembolism in non-rheumatic atrial fibrillation in primary care: randomised controlled trial comparing two intensities of coumarin with aspirin. BMJ 1999;319:958–64.
[3] ICH E9 Expert Working Group. Statistical principles for clinical trials. Statist Med 1999;18:1903–42.
[4] Pocock SJ. When to stop a clinical trial. BMJ 1992;305:235–40.
[5] Chow SC, Liu JP. Design and analysis of clinical trials: concepts and methodologies. New York: Wiley; 1998. p 413.
[6] Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. Stat Med 1982;1:121–9.
[7] Hsieh FY. Comparing sample size formulae for trials with unbalanced allocation using the logrank test. Stat Med 1992;11:1091–8.

PII: S0895-4356(02)00596-6

AUTHOR'S REPLY

The ethics of sample size: The whole picture should be considered

At variance with Schouten's interpretation, we aim to optimize, rather than "minimize" study size, to expose enough but not too many subjects to an experimental regimen to answer the key question at stake [1–3]. The first and foremost concern is to specify that question, with the preferred statistical approach being the consequence of it. As we formulated earlier [1]: a research hypothesis is no "random shooting" but expresses scientific uncertainty regarding a plausible, potentially clinically important effect. Before a trial one is not totally ignorant of what is to be tested. The problem at issue is generally

not as neutral as: "is the principal treatment better than the reference (A > B), or does it equal the reference (A = B), or is it worse than the reference (A < B)?" Often, the question is "whether the principal treatment is indeed better than the reference," considering that an advantage of the principal over the reference intervention is a reasonable but not sufficiently tested assumption. A proper research question is in many cases "one-sided" [4]. A further consideration is: what are the consequences of the possible outcomes (A > B, A = B, or A < B) in terms of study conclusions and recommendations for practice. In many instances there is no reason to distinguish between equivalence and inferiority of the more burdensome intervention. We maintain our view that one-sided testing and a corresponding sample size estimation is the preferred approach if (a) the scientific hypothesis to be tested is obviously one-sided, or if (b) only a clear advantage in effect of the principal over the reference intervention would have consequences for practice.

Starting from the principle of one-sided testing while maintaining the same power, this will imply a smaller study size than if two-sided testing is chosen, unless—as Schouten advocates—in the one-sided option half the significance level as used in the two-sided approach would be applied. Of course, a more stringent significance level will decrease the probability of a false significant result if the principle intervention is actually equivalent or inferior. Where to draw the line is a matter of judgement, rather than being evident and unequivocal. Why not use a one-sided significance level of 0.025 or even more strict, 0.01 or 0.001, instead of 0.05? This depends on the research topic, clinical implications, the acceptable amount of uncertainty, and ethical considerations, and has substantial consequences for the study budget. But this choice is fully independent of the essential difference between one- and two-sided testing. In the first option, there is one clear alternative for the null hypothesis (A > B). In the second, there are in fact two alternative hypotheses (A > B and A < B). Accordingly, we do not agree with Schouten's implicit suggestion that for both situations the same sample size should be used. Instead, testing two alternative hypotheses requires more evidence, that is, a larger sample size than testing one.

A further comment on Schouten's analysis is that, in line with the frequentists' view on hypothesis testing, he ignores the prior probability of the alternative hypothesis (e.g., A > B) being true given prior scientific work and reasoning. We favor the Bayesian approach, and believe that the posterior probability (after the study being done), that an observed statistically significant effect of A is, in fact, falsely significant, is lower as the prior probability of A being actually better is higher. Also, the higher the prior probability of the alternative hypothesis A > B being true, the more a one-sided hypothesis will be justified and preferable. In testing a well-prepared one-sided hypothesis with a one-sided significance level set at 0.05, the posterior probability of a false significant effect may, therefore, be even lower than in testing a "scientifically neutral" hypothesis at the .025 level.

Schouten prefers setting the significance level for one-sided tests at half the "conventional" significance level for two-sided tests, also "because this promotes consistency with the two-sided confidence intervals that are generally appropriate for estimating the possible size of the difference between two treatments," and states that an interval including RR = 1.0 would not be convincing to colleagues. In response, we emphasize that the 95% confidence interval is not a fixed thing, nor a natural phenomenon. In case of a one-sided hypothesis, one is interested in whether the interval limit directed towards the null value (RR = 1.0) would include 1.0 or not. For that purpose, in case of a significance level of 0.05, a one-sided 90% limit would suffice. Those who think that this is too unconventional of difficult to interpret may choose not to use confidence intervals in cases like this. Significance testing would be sufficient, because the fundamental issue in the discussed example is not a matter of estimation and quantification but of hypothesis testing. That is, deciding on whether the principal treatment to be tested is indeed better that the reference or not.

The fact that in case of a one-sided focus (hypothesis) a trial might be stopped earlier (that is, being conclusive in an earlier stage) compared with the two-sided approach is simply a consequence of the fundamental difference between both approaches. In fact, in the "one-sided situations" outlined in our earlier contributions, an earlier stop would be not only justified but even highly desirable. As we earlier have stipulated [2], the possibility to apply stopping rules is not the monopoly of the two-sided approach. Such rules can also be implemented according to a one-sided approach, both regarding primary end points and adverse effects [5–7].

Finally, in speaking about superior or inferior treatments, one must distinguish between studies focussing on primary clinical endpoints (which may include intentionally evaluated adverse effects, e.g., when looking at the incidence of both stroke and bleeding in a warfarin trial) and studies especially designed to detect or exclude infrequent and typically unsuspected adverse effects. The latter type of studies have specific requirements, regarding both methodology and sample size estimation [8].

J. André Knottnerus, Lex M. Bouter

References

[1] Knottnerus JA, Bouter LM. The ethics of sample size: two-sided testing and one-sided thinking. J Clin Epidemiol 2001;54:109–10.
[2] Knottnerus JA, Bouter LM. Hypothesis testing complexity in the name of ethics: response. J Clin Epidemiol 2002;55:210–11.
[3] Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002;288:358–62.
[4] Feinstein AR. Clinical epidemiology. The architecture of clinical research. Philadelphia: W.B. Saunders Company; 1985.
[5] Demets DL, Ware JH. Group sequential methods in clinical trials with a one-sided hypothesis. Biometrika 1980;67:651–60.
[6] Berry DA, Ho CH. One-sided stopping boundaries for clinical trials: decision-theoretic approach. Biometrics 1988;44:219–27.
[7] Kumar KV, Powell MR, Waligora JM. Early stopping of aerospace medical trials: application of sequential principles. J Clin Pharmacol 1994;34:596–98.
[8] Turbert-Bitter P, Manfredi R, Lellouch J, Begaud B. Sample size calculations for risk equivalence testing in pharmacoepidemiology. J Clin Epidemiol 2000;53:1268–74.

PII: S0895-4356(02)00597-8