Title: Estimating risk from underpowered, but statistically significant, studies: Was APPROVe on TARGET?‡

Short Title: Estimating risk in underpowered studies

Authors:
Adam La Caze BPharm, BA(Hons), PhD. School of Pharmacy, The University of Queensland
Stephen Duffull, Dip Pharm, M Pharm, PhD, School of Pharmacy, University of Otago

SUMMARY

The importance of statistical power is widely recognised from a pre-trial perspective, and when interpreting results that are not statistically significant. Though it has been discussed, it is less well recognised is that poor power can lead to inflated estimates of the effect size when statistically significant results are observed. We call this bias ‘significant-result bias’. There are few methods available for estimating significant-result bias. We use trial simulations to quantify significant-result bias, and briefly apply this method to estimate possible significant-result bias in the rate of thrombotic events observed in the APPROVe trial. Statistically significant results, on outcomes for which there is empirical evidence of poor power, may provide inflated estimates of the size of effect. If independent evidence is available to judge the likely effect size of an underpowered statistical test, trial simulations can provide a method for quantifying this bias.

KEY WORDS: Bias; Clinical Trials; Rofecoxib; Power

‡ This is a preprint of La Caze A, Duffull, S. (2011). Estimating risk from underpowered, but statistically significant, studies: Was APPROVe on TARGET? Journal of Clinical Pharmacy and Therapeutics. Follow the link to the published paper.
WHAT IS KNOWN AND OBJECTIVE

The primary objective of most clinical trials is confirmatory: does the intervention under investigation bring about the expected change in the primary endpoint? Clinical trials are designed and analysed to meet this objective. Sufficient numbers are enrolled to ensure that the number of primary endpoints observed is large enough to reliably discriminate between the treatment under investigation and the control. But trials report more than just the primary hypothesis test, and the statistical tests conducted on these endpoints are often far less reliable. Despite the potential for unreliability secondary endpoints or subgroup analyses are often relevant to clinicians. Take, for instance, safety-related endpoints. Safety is important to therapeutic decisions, and yet safety-related endpoints are infrequently the primary endpoint of clinical trials.

The statistical problems associated with interpreting the results of secondary endpoints and subgroup analyses are well recognised (1–3), as are the problems of an uncritical reliance on $P$ values (4–6). Post hoc data-derived analyses are particularly problematic (7). But even if the endpoints, subgroups and the planned analysis are specified a priori, secondary endpoints and subgroup analyses often suffer from poor power. While secondary endpoints and subgroup analyses are not necessarily underpowered, they often will be. The size of the trial is decided on the basis of the estimated effect size for the primary endpoint; if the secondary endpoint is a safety endpoint the effect size will often be smaller, and the power of the subgroup analysis will be lower. There is little argument regarding the importance of power prior to the trial. It is similarly well recognised that non-statistically significant results from an underpowered test provide little evidence for the lack of an effect. But there is considerably less recognition of the
implications of poor pre-trial power for interpreting results that happen to be statistically significant.

A simulation study, in which the simulated trial was sufficiently powered to test the endpoint in the overall sample, showed that 7–26% of subgroup analyses provided a statistically significant result when there was no overall effect from treatment, and that only one of the two subgroups gave a statistically significant result in 41–66% of cases when there was overall effect from treatment (2). Charles E. Land explained 100-fold discrepancies in estimates of cancer risk on the basis of underpowered studies (8). Low estimates of effect size arise when studies are not large enough to detect small effects (i.e. the statistical test fails to reject a false null hypothesis). High estimates can arise from underpowered tests because those trials that provide statistically significant results do so because the effect from this instance of the trial is both large enough to be observed and fall into the rejection region required by the statistical test. Smaller effect sizes, that are closer to the true effect size, are more likely to occur but these would not yield a statistically significant result and hence would be ignored as an underpowered negative result. The average of all statistically significant results, for an underpowered test, will therefore overestimate the effect size (as discussed by Land (8)).

The problem of statistically significant results from underpowered endpoints is a form of publication bias. Publication bias results from the over reliance, over publication, or over reporting of statistically significant results from small studies. Asymmetrical funnel plots have been used to diagnose instances of publication bias (3, 9, 10). Funnel plots graph effect size against sample size. Asymmetrical funnel plots show a significant reduction in effect size as the sample size, and power, of the included studies increases. Funnel plots are only viable once considerable data are available. While funnel plots help to diagnose instances of publication
bias once they have occurred, they do not help to assess the role of publication bias as data accumulates.

A statistically significant result observed from a truly underpowered test will over-estimate the effect size. Of course, given power is rarely considered for secondary endpoints or subgroup analyses, it is very difficult for the clinician to appropriately interpret a statistically significant result from a secondary endpoint or subgroup analysis. Assume, for instance, that there is independent evidence that suggests the statistical test is underpowered. Should the clinician assume the observed result is overestimated? Or should the observed data be taken to undermine the assessment of the test’s power? It is no surprise that clinicians are advised to refrain from placing weight on the findings of secondary outcomes or subgroup analyses (1).

There is a standard reply to this problem. Rather than rely on underpowered secondary endpoints or subgroup analyses, clinicians are advised to treat these findings as ‘hypothesis generating’ and await the combination of a number of trials into meta-analyses, or a large-scale trial on the issue in question. While this is good advice, this reply fails to assist decision-makers while further data is collected, and it fails to recognise that the results of meta-analyses on secondary endpoints can be equivocal (11, 12).

It is natural for clinicians to interrogate data on secondary endpoints and subgroups to answer important clinical questions. The dangers of interpreting statistically significant findings from secondary endpoints or subgroup analysis in the same way that statistically significant findings are interpreted for the primary endpoint emphasise the importance of ensuring this problem is well recognised. We use trial simulations to quantify the problem of interpreting underpowered
endpoints or subgroup analyses and present a brief case study using data on COX-2 inhibitors from the APPROVe and TARGET trial.

COMMENT

The risk of misinterpreting underpowered, but statistically significant studies, is quantified in a simple simulation. A hypothetical trial is simulated in which the true difference between two treatments is a small positive number. The distribution of true differences was assumed to be Normal. One thousand trial outcomes were simulated each with 5, 50, 150, 200, 400, 500, 800 and 1000 patients. For those trial outcomes that showed a statistically significant result (P-value < 0.05 using a one-tailed t-test) the “significant-result bias” was calculated. We define “significant-result bias” as the difference between the observed mean difference and the true mean difference divided by the true mean difference for only those results that were statistically significant. Power was estimated empirically as the proportion of statistically significant results. Figure 1 shows the significant-result bias of statistically significant trial outcomes plotted against the power of the study. Note here that significant-result bias is not actual study bias.

Figure 1 shows that studies that have low power (high beta-error), and provide statistically significant results, confer substantial significant-result bias. A power as low as 10% produces an average size of effect in excess of 5 times greater than the true size of effect. In contrast, when power approaches 80% the significant result bias shrinks to negligible levels. This result is neither surprising nor controversial when presented in general terms. This relatively simple method, however, provides a way of quantifying possible significant-result bias in published trials.
The APPROVe study was powered to determine whether rofecoxib would reduce the number of adenomatous polyps in patients with a history of colonic adenomas (15). In a secondary analysis it was found that the rofecoxib cohort was at an increased risk of thrombotic events. The incidence per year in the rofecoxib group was 1.5% (n=1287) and in the placebo group was 0.77% (n=1299). This provided a statistically significant excess in risk of 0.72 events per 100 patient years ($P$-value=0.008, Relative Risk 1.9).

APPROVe was not explicitly powered to test whether rofecoxib increases the risk of thrombotic events. To calculate the power of APPROVe to test this endpoint the true excess risk of thrombotic events in patients taking rofecoxib needs to be known. The true value is obviously unknown and an approximation to the true value is highly contentious. In three clinical studies, the reported excess in risk of thrombotic events in patients taking rofecoxib ranges from -0.21 to 0.97 events per 100 patient years (16–18). For the purposes of our analysis we use the excess in risk observed in the very large TARGET trial (19) to judge the power APPROVe had to test the risk of thrombotic events in patients taking rofecoxib.

TARGET compared the incidence of cardiovascular events with lumiracoxib versus ibuprofen or naproxen. The observed difference in incidence of thrombotic events between the lumiracoxib group and comparator was 0.11 events per 100 patient years, the $P$-value was not statistically significant. Lumiracoxib is approximately 3 fold more selective for COX-2 than rofecoxib (20). Using TARGET assumes that the risk of thrombotic events is a class effect of COX-2 inhibitors.
We simulated 10,000 clinical studies with n=2568 patients each with the same fraction of patients receiving rofecoxib as in the APPROVe study. We set our alpha level to be 0.008, which corresponds to the \( P \)-value found in the APPROVe study. The expected true incidence of rofecoxib-induced cardiovascular events was set to that of lumiracoxib from the TARGET study (0.65%). From these 10,000 simulations, those trials that provided a significant difference based on a chi-squared test with \( P \)-value < 0.008 were kept and those that did not were discarded. The size of the effect was tabulated for only those trials that showed significant results (shown in Figure 2). In this figure the actual effect size seen in the APPROVe study is shown as a vertical line, which falls well within the likely predictions based on the incidence observed in TARGET. If this is the only source of difference then it appears that the APPROVe and TARGET studies support the same conclusion.

WHAT IS NEW AND CONCLUSION

When available, pre-experimental empirical evidence on the primary endpoint is usually incorporated into the set up of the statistical test. Pre-experimental evidence plays a role in appropriate study design and is used to ensure the statistical test conducted on the primary endpoint is adequately powered. Secondary analyses do not incorporate pre-experimental evidence in this way and hence may be underpowered. We have shown that inflation of the estimate of the effect size occurs in relation to the inverse of the power of the statistical test. This makes observed effect sizes difficult to interpret when statistical tests for secondary endpoints are likely to be underpowered. Assessing whether the statistical test on the secondary endpoint is indeed unpowered is challenging. Where plausible independent evidence (that is evidence from other well conducted trials) exists, the technique described above can help to quantify possible significant-result bias.
It will almost always be difficult to judge which statistical tests are indeed underpowered. Similarly, assigning a ‘pre-trial power’ to a statistical test in which the data are already available will almost always be contentious. (Any assessment of pre-trial power must be made on independent evidence and *not* on data observed in the trial under consideration). Despite these difficulties the possibility of significant-result bias should be considered, especially for unexpected or rare events such as some side effects. Providing the assumptions underpinning the assessment of significant-result bias are explicit, we see no problem in the provisional nature of the estimate.

There are important (and controversial) assumptions required to use TARGET to estimate the power of APPROVe. First, APPROVe and TARGET used different comparators. Whereas APPROVe was placebo controlled, TARGET compares a COX-2 inhibitor to traditional NSAIDs. The safety of traditional NSAIDs in terms of thrombotic events is currently inconclusive (21). Second, despite not being explainable in pharmacological terms, the possibility of differences between the thrombotic risks of rofecoxib and lumiracoxib is recognised. Informed debate on the likely pre-trial power of certain statistical tests, and the possibility of significant-result bias is likely to aid (rather than hinder) decision-making. Quantifying significant-result bias provides an alternative to waiting for a meta-analysis of similar trials (which, in any case may provide inconclusive and contentious results), or, in the case of rofecoxib, waiting for a large randomised trial that has cardiovascular events as its primary outcome (a trial that is unlikely to be funded, or given ethical approval).

More, larger, and better-designed studies, as well as meta-analyses, will, in the long run, “wash out” any significant-result bias from early studies. However, the considerable delay in awaiting
further data present considerable practical difficulties for those who need to make decisions regarding clinically important pharmacological effects from clinical trials. Power should be explicitly considered when interpreting statistically significant results from clinical trials. Simulations, like those presented, provide an avenue for estimating the possibility of significant-result bias.
References


