BMJ

BMJ 2014;348:g2215 doi: 10.1136/bmj.g2215 (Published 31 March 2014)

RESEARCH METHODS & REPORTING

Trap of trends to statistical significance: likelihood of near significant P value becoming more significant with extra data

When faced with a P value that has failed to reach some specific threshold (generally P<0.05), authors of scientific articles may imply a "trend towards statistical significance" or otherwise suggest that the failure to achieve statistical significance was due to insufficient data. This paper presents a quantitative analysis to show that such descriptions give a misleading impression and undermine the principle of accurate reporting.

John Wood *principal research associate*[1], Nick Freemantle *professor of clinical epidemiology and biostatistics*[1], Michael King *professor of primary care psychiatry*[2], Irwin Nazareth *professor of primary care and population science*[1]

[1]Research Department of Primary Care and Population Health and PRIMENT Clinical Trials Unit, University College London, London NW3 2PF, UK; [2]Division of Psychiatry and PRIMENT Clinical Trials Unit, University College London, London W1W 7EJ, UK

Introduction

P values that fail to reach the conventional significance level of P≤0.05 are regularly reported as if they were moving in that direction. Phrases such as "almost/approaching statistical significance" or, most tellingly, a "trend towards" statistical significance continue to find their way into papers in journals with high impact factors.[1] In this article, we examine the mathematical basis for this assumption and assess the extent to which a near significant P value may predict movement towards a future significant P value through the addition of extra data. We also explore the likelihood that extra data would actually result in a significant outcome and, lastly, the confidence one might have that a repeat experiment would independently give statistically significant results.

What does P value represent?

The clearest context in which to consider the correct interpretation of a P value is within a randomised trial. Fisher described how the "simple precaution of randomisation will suffice to guarantee the validity of the test of significance."[2] Random allocation of participants to groups ensures that only the play of chance or a real effect of treatment can explain any difference seen in outcome between the groups. A P value tells us how far chance alone can explain the observed difference and acts as a "snapshot" measure of the strength of evidence at the end of the trial.

Calculating extent to which "near significant" P value predicts subsequent significant one

As evidence accumulates, conclusions become more firmly based. A P value could easily be imagined as following a similar course, with "near significant" values seen as a failure to detect a real effect due to insufficient sample size (or, in technical language, as type II errors caused by insufficient power). Referring to a "trend towards significance" expresses the view that had the experiment recruited more people, the P value would have become more significant. This apparently orderly picture is contradicted in situations in which P values are monitored, say, for safety purposes in randomised trials. There we find that P values can fluctuate markedly between inspections, showing qualitatively both the large random component of a P value and that a new test conducted with additional data will not necessarily provide statistically stronger results.

In the model that follows, we use confidence intervals for the true size of a treatment effect together with our understanding of the nature of a P value to analyse the extent to which results might become more or less significant with the addition of extra data. Thereby, we can determine how justifiable it is to describe a near significant result in terms of a genuine trend.

Full technical details of our statistical approach are given in the appendix, but the outline is as follows. Consider a number of

Correspondence to: J Wood john.x.wood@ucl.ac.uk

Extra material supplied by the author (see http://www.bmj.com/content/348/bmj.g2215?tab=related#webextra)

participants allocated equally to (say) active treatment or placebo in an experiment designed to estimate the true size of the treatment effect on a continuous outcome. The resulting estimate will be subject to the play of chance, the influence of which reduces as the number of participants increases. Under standard assumptions, a confidence interval for the treatment effect, along with a test for significance, can be calculated. These two are directly related: knowing the one enables calculation of the other. Before the experiment starts, the P value is a random variable, the distribution of which depends on the (unknown) true treatment effect, the variability in outcome between trial participants receiving the same treatment (that is, the "error variance"), and the sample size.

Suppose the experiment is now carried out, providing a particular P value (P_1) which is, as usual, two sided (that is, pertaining to the probability of a difference between the groups without specifying which group is superior).[3] The extent to which this predicts a subsequent more significant P value can then be examined by imagining the collection of more observations. These will provide a further estimate of the treatment effect, which, combined with our original estimate, leads to an update of the P value (P_2). Like the original P value, the change from P_1 to P_2 depends on the true treatment effect and is also subject to the play of chance. However, our first data have already provided an estimate of the treatment effect, together with confidence intervals indicating the extent to which it may be higher or lower than this. Using a complete set of confidence intervals (that is, for all levels of confidence, not just the conventional 95%) to construct a "confidence distribution" for the true treatment effect,[4] and conjoining this with the sampling variance of the new estimate of the effect (in the same way as one would combine two probability distributions), enables a rational calculation of the confidence that P_2 will be less than P_1 (or vice versa). For this approach to be valid, we have to base the combination on some prior assumption about the size of the true treatment effect. However, although we need to specify a range (which can be quite wide) within which the true effect must lie, we assume no prior preference for any particular value or values within those limits.

Naturally, our confidence in predicting P_2 will depend on our current estimate of the true treatment effect, as well as its precision. It will also depend on how much new data we plan to add: the less we add, the more the direction of change is sensitive to chance fluctuation and vice versa. Working this through algebraically (justification and details are given in the appendix), we can estimate the percentage of the time that a P value will become less significant (and thus run counter trend). It turns out that this depends only on the observed P value and the relative amount of extra data, so adding 1000 pairs of participants to an original trial including 10 000 pairs will provide the same expectation of a less significant result as adding 10 pairs to an original trial of 100 pairs furnishing the same P value.

Table 1⇓ gives results for various combinations of P values (P_1) and amount of extra data envisaged. Although the chance of the test becoming less significant with the addition of more data is always less than 50%, it is in many circumstances substantial. For example, if our two sided P_1 from the original data is 0.08 (the sort of marginal value for which "trends" are often implied), we should expect that increasing the sample size by 10% will lead to results becoming less significant (P_2>0.08) some 39% of the time. If we added 20% extra data, the situation improves only marginally, as we can then expect P_2>0.08 around 35% of the time. Doubling the size of the study has more effect, when we should expect P_2>0.08 about 23% of the time. For comparison, if P_1=0.05, much the same chance (slightly smaller at 21%) exists that P_2>0.05—that is, of the result becoming non-significant—when the study size is doubled. This underlines the similarity of the situation on either side of the (artificial) P=0.05 dividing line. Even if we add 10 times the original sample, we should expect P_2>0.08 some 10% of the time given P_1=0.08, and P_2>0.05 just under 8% of the time given P_1=0.05. The likelihood that the P value becomes less significant is small only when we are already reasonably confident that the treatment is different from placebo (when $P_1 \leq 0.01$) and are considering the likely influence of a substantial amount of new data. However, it is the more marginal P values—such as P=0.08—that are of most practical interest. For these, the above figures show the inappropriateness of regarding them as being almost there on a journey towards statistical significance. Similarly, the results for a P value of 0.05 should militate against the conclusion that simply achieving this level of statistical significance means we are home and dry.

Furthermore, we may be interested in the likelihood that P_2 actually meets some specific level (rather than simply becoming less or more significant). Table 2⇓ gives results for the probability that P_2 does not reach significance at the conventional two sided P≤0.05, for various combinations of P_1 and amount of extra data envisaged, and these can be compared with those in table 1⇓. Thus for a P value of 0.08 and adding 20% more data, we should expect the updated results to remain non-significant at the P<0.05 level more than half the time.

Finally, we might consider the likelihood of obtaining a significant result from an exactly similar experiment, analysed completely separately from the original. For instance, it turns out that if the first experiment is just significant at the level of P=0.05, the probability that the results from this second experiment are significant at the same conventional two sided P=0.05 is 50%. Table 3⇓ shows corresponding results for other values of P_1, and it is interesting that, even when the first experiment is as convincing as P_1=0.001, a one in six chance still exists of failure to get a conventionally significant result in a new, replicate, trial. This is in contrast to the situation in which the results of the two trials are combined, when the chance of extra data "overturning" an original P value of 0.001 is very slim (see first column of table 2⇓).

Discussion

We have shown that a P value is by no means assured to become smaller even with the addition of quite a substantial proportion of extra data, a finding that undermines any claim of a trend towards statistical significance. This expression might be best considered a form of special pleading whereby authors, however unwittingly, are claiming something that their study has not achieved. The unpredictability of change in P values with additional data also illustrates the risk of concluding too much from a finding of modest statistical significance, such as two sided P<0.05 but >0.01. In that case, simply increasing the data by 20% could reasonably be expected to lead to less significant results about 30% of the time.

Instead of reporting trends, some authors imply that their results got close to or bordered on significance. Quite elaborate forms of words may be used, such as "teetering on the brink of significance."[5] They imply that, with a nudge (extra data), significance would plausibly have been achieved. However, the results in table 2⇓ show that achieving significance with extra data is often not likely and is always less likely than simple movement in the right direction (seen by comparison with table 1⇓, for P_1>0.05).

Table 3⇓ shows that repeating the original experiment exactly and analysing the new results by themselves means that the chance of achieving conventional significance second time round, when the original experiment did not, can be quite high—for instance, around one in three for an original P value of 0.15. However, table 3⇓ also shows that replicating a single significant result is by no means guaranteed: the chance is only 50-50 that experiment 2 will give a result significant at $P \leq 0.05$ when experiment 1 achieved P=0.05. The question of replicating results that are significant only at moderate levels (such as P=0.05) is analysed in detail from a different perspective by Johnson (2013),[6] who argues strongly that evidence thresholds should be made more stringent than this.

Overall, we would like to see greater recognition that individual P values in the region of 0.05 represent quite modest degrees of evidence, whichever side of the divide they lie on. We are not advocating that near significant P values or confidence intervals (to which exactly the same arguments apply) should be automatically brushed aside: some of these will undoubtedly have the status of "interesting hints."[7] Rather, our objection is to describing them as trends towards statistical significance or using any one of a number of phrases that carry a similar implication. This is not an academic argument about words: such terms are potentially misleading for the (quantitative) reasons given above. This point could usefully be addressed in CONSORT and other relevant reporting guidelines.[8]

Contributors: JW had the original idea for the paper and constructed the mathematical model. NF wrote the first draft. All authors contributed to the development of the idea with respect to important intellectual content, revised the manuscript, and approved the final version.

Funding: None of the authors received external funding for this work, which was undertaken as part of their normal academic duties.

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

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

1 Karunajeewa HA, Mueller I, Senn M, Lin E, Law I, Gomorrai PS, et al. A trial of combination antimalarial therapies in children from Papua New Guinea. *N Engl J Med* 2008;359:2545-57.
2 Fisher RA. The design of experiments. 8th ed. Oliver and Boyd, 1966:21.
3 Sedgewick P. One sided and two sided hypothesis tests. *BMJ* 2010;340:c2458.
4 Xie MG, Singh K. Confidence distribution, the frequentist distribution estimator of a parameter: a review. *Int Stat Rev* 2013;81:3-39.
5 Dorrance AM. Are macrophages the foot soldiers in the war waged by aldosterone against the heart? *Hypertension* 2009;54:451-3.
6 Johnson VE. Revised standards for statistical evidence. *Proc Natl Acad Sci U S A* 2013;110:19313-7.
7 Tukey JW. The philosophy of multiple comparisons. *Stat Sci* 1991;6:100-16.
8 CONSORT. The CONSORT statement. 2010. www.consort-statement.org/consort-statement/overview0/.

Accepted: 10 March 2014

Cite this as: *BMJ* 2014;348:g2215

Summary points

Describing near significant P values as "trends towards significance"' (or similar) is not just inappropriate but actively misleading, as such P values would be quite likely to become less significant if extra data were collected

Descriptions such as "on the brink of significance" are similarly misleading, as the chance that extra data would change a near significant P value into a significant one is even less than the chance of simply moving towards significance

Replicating significant results is a challenge; the chance that a repeat experiment analysed independently would give a non-significant result can be quite high, even if evidence from the original experiment looks strong

P values in the region of 0.05 represent quite modest degrees of evidence, whichever side of the divide they lie on

Tables

Table 1| Per cent of times P value would be expected to get less significant had extra data been collected, given current P value and amount of extra data

	Current (two tailed) P value						
Amount of extra data as % of current	**0.001**	**0.01**	**0.05**	**0.06**	**0.08**	**0.10**	**0.15**
1000	0.8	3.0	7.6	8.4	10.0	11.4	14.6
100	8.6	14.3	20.8	21.8	23.4	24.8	27.5
50	14.8	20.6	26.7	27.5	28.9	30.1	32.4
20	24.1	29.1	33.8	34.4	35.4	36.3	37.9
10	30.6	34.5	38.1	38.6	39.3	40.0	41.2
1	43.5	44.9	46.1	46.3	46.5	46.7	47.1
0.01	49.3	49.5	49.6	49.6	49.7	49.7	49.7

Table 2| **Per cent of times to expect non-significant result (two tailed test; α=0.05) on addition of extra data, given current P value and amount of extra data**

	Current (two tailed) P value						
Amount of extra data as % of current	**0.001**	**0.01**	**0.05**	**0.06**	**0.08**	**0.10**	**0.15**
1000	0.2	1.9	7.6	8.8	11.2	13.5	18.7
100	0.4	4.6	20.8	24.2	30.3	35.7	47.0
50	0.2	4.6	26.7	31.4	39.7	46.9	61.0
20	0.0	2.7	33.8	41.1	53.8	63.8	80.4
10	0.0	1.0	38.1	48.4	65.2	77.1	92.3

Table 3| **Per cent of times to expect non-significant result (two tailed test; α=0.05) of repeat experiment of same size, analysed independently**

Current (two tailed) P value	Per cent of times
0.001	17.3
0.01	33.2
0.05	50.0
0.06	52.2
0.08	55.9
0.10	58.8
0.15	64.4