Trust, Bradford health authority, Bradford social services department, University of Bradford, York Health Economics Consortium at the University of York. The grant application signatories were JY, NS, KL, AF, and Angela Clegg. The trial steering group and project group members were JY, AF, NS, Susan Ince, Laura Hibbs, Angela Clegg, Joy Warburton, Jackie Hansford, Anne McAdam, Karen Mallinder, KL, JO'R, and JG. The study research team included Linda Dobrzanska, Helen Wright, Emma Tanner, Karen Mallinder, and JG.

Contributors: See bmj.com.

Funding: Health Foundation.

Competing interests: JY and JG have worked in the community hospital involved in this study.

Ethical approval: This study was approved by the research ethics committee of Bradford Hospitals NHS Foundation Trust.

1 Department of Health. *The NHS plan*. London: DoH, 2000.
2 Meads G. Rediscovering community hospitals. *Br J Gen Pract* 2001;51:91-2.
3 Seamark D, Moore B, Tucker H, Church J, Seamark C. Community hospitals for the new millennium. *Br J Gen Pract* 2001;51:125-7.
4 Department of Health. *Our health, our care, our say: a new direction for community services*. Jan 2006: Cm 6737. Ch 6 para 6.43.
5 Green J, Young J, Forster A, Mallinder K, Bogle S, Lowson K, et al. Effects of locality based community hospital care on independence in older people needing rehabilitation: randomised controlled trial. *BMJ* 2005;331:317-22.
6 *EuroQol EQ-5D user guide*. Version A (6/96). www.euroqol.org (accessed 5 July 2006).
7 Chartered Institute of Public Finance and Accountancy. *The health service financial database and comparative tool 2002*. London: CIPFA, 2002.
8 Department of Health. *Reference costs 2002*. Leeds: DoH, 2002.
9 Netten A, Curtis L. *Unit costs of health and social care 2002*. Canterbury, Kent: Personal Social Services Research Unit, University of Kent at Canterbury, 2002.
10 Briggs AH, Wonderling DE, Mooney CZ. Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. *Health Econ* 1997;6:327-40.
11 Briggs AH, Mooney CZ, Wonderling DE. Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. *Stat Med* 1999;18:3245-62.
12 Baker JE, Goldacre M, Muir-Gray JA. Community hospitals in Oxfordshire. *J Epidemiol Community Health* 1986;40:117-20.
13 Hine C, Wood VA, Taylor S, Charny M. Do community hospitals reduce the use of district general hospital inpatient beds? *J R Soc Med* 1996;89:681-7.
14 Cook PJ, Porter L. Community hospitals and district general hospital medical bed use by elderly people: a study of 342 general practitioner beds in Oxfordshire. *Age Ageing* 1998;27:357-61.

(Accepted 22 May 2006)

doi 10.1136/bmj.38887.558576.7C

Believability of relative risks and odds ratios in abstracts: cross sectional study

Peter C Gøtzsche

Abstract

Objective To compare the distribution of P values in abstracts of randomised controlled trials with that in observational studies, and to check P values between 0.04 and 0.06.

Design Cross sectional study of all 260 abstracts in PubMed of articles published in 2003 that contained "relative risk" or "odds ratio" and reported results from a randomised trial, and random samples of 130 abstracts from cohort studies and 130 from case-control studies. P values were noted or calculated if unreported.

Main outcome measures Prevalence of significant P values in abstracts and distribution of P values between 0.04 and 0.06.

Results The first result in the abstract was statistically significant in 70% of the trials, 84% of cohort studies, and 84% of case-control studies. Although many of these results were derived from subgroup or secondary analyses, or biased selection of results, they were presented without reservations in 98% of the trials. P values were more extreme in observational studies ($P < 0.001$) and in cohort studies than in case-control studies ($P = 0.04$). The distribution of P values around $P = 0.05$ was extremely skewed. Only five trials had $0.05 \le P < 0.06$, whereas 29 trials had $0.04 \le P < 0.05$. I could check the calculations for 27 of these trials. One of four non-significant results was significant. Four of the 23 significant results were wrong, five were doubtful, and four could be discussed. Nine cohort studies and eight case-control studies reported P values between 0.04 and 0.06, but in all 17 cases $P < 0.05$. Because the analyses had been adjusted for confounders, these results could not be checked.

Conclusions Significant results in abstracts are common but should generally be disbelieved.

Introduction

Abstracts of research articles are often the only part that is read, and only about half of all results initially presented in abstracts are ever published in full.[1] Abstracts must, therefore, reflect studies fairly and present the results without bias. This is not always the case. In a survey of 19 clinical trials that contained a mixture of significant and non-significant results, the odds were nine times higher for inclusion of significant results in the abstract.[2] Another survey found that bias in the conclusion or abstract of comparative trials of two non-steroidal anti-inflammatory drugs consistently favoured the new drug over the control drug in 81 trials and the control drug in only one.[3] And a survey of 73 recent observational studies found a preponderance of P values in abstracts between 0.01 and 0.05 that indicated biased reporting or biased analyses.[4]

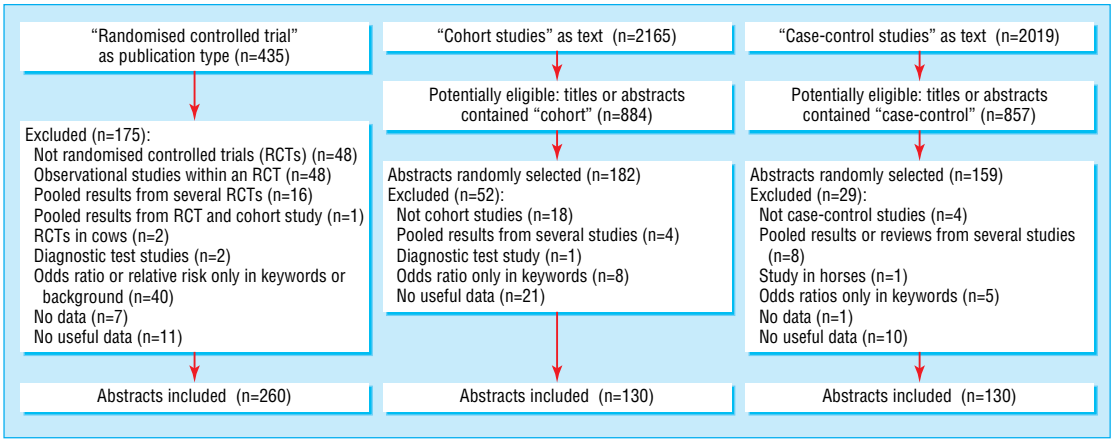
I explored in a large sample of research articles whether P values in recent abstracts are generally believable.

References w1-w19 and a table giving the recalculations for P values are on bmj.com

This article was posted on bmj.com on 19 July 2006: http://bmj.com/cgi/doi/10.1136/bmj.38895.410451.79

Nordic Cochrane Centre, H:S Rigshospitalet, DK-2100 Copenhagen Ø, Denmark
Peter C Gøtzsche *director*
pcg@cochrane.dk

BMJ 2006;333:231–4



Inclusion of abstracts

Methods

I compared the distribution of P values in abstracts of randomised controlled trials with that in observational studies. I also explored reasons for possible skewness, in particular for P values close to $P<0.05$, which is the conventional level of significance.

On 15 October 2004, I searched PubMed for all abstracts of articles published in 2003 that contained "relative risk" or "odds ratio" in any field. I found 7453 abstracts, 435 of which had the publication type "randomized controlled trial." After I excluded 175 irrelevant abstracts, mainly because they were not of randomised trials (figure), 260 trials that reported at least one binary outcome remained.

Of the 7453 abstracts, 2165 contained "cohort studies" and 2019 "case-control studies" as text words in any field. I randomised a subsample of these observational studies that, in addition, had either "cohort" (884) or "case-control" (857) in the title or abstract. I generated random numbers with Microsoft Excel and studied the abstracts in this order until I had 260 relevant ones, with half in each category. I excluded 62 and 29 ineligible abstracts, respectively, during this process (figure).

I took the first relative risk or odds ratio that was given and its P value. If the first result was a hazard ratio or a standardised mortality ratio, I accepted this. If a P value was not given, I calculated it from the confidence interval, when available, using the normal distribution after log transformation.[5] If the first result was not statistically significant, I noted whether the remainder of the abstract included any significant results.

To minimise errors, I downloaded the abstracts and copied the relevant text with the data into a spreadsheet, wrote the numbers in the appropriate columns, and checked the numbers against the copied text.

I compared the distributions of P values between trials and observational studies and between cohort studies and case-control studies, with the Mann-Whitney U test after categorisation.[4]

Finally, I checked whether P values between 0.04 and 0.06 were correct by comparison with the methods and results sections after retrieval of the full papers. I also double checked these data.

I used Stata (StataCorp, College Station, TX) for Fisher's exact test, Medstat (Wulff and Schlichting, Denmark) for the χ^2 test and the Mann-Whitney U test, and Review Manager (Nordic Cochrane Centre, Denmark) to calculate relative risks and odds ratios. When I could not reproduce the authors' P values, I contacted the authors for clarification, at least twice, in case of no reply.

Results

The first reported binary outcome in the abstract was the relative risk in 52% of the randomised trials, 35% of the cohort studies, and 4% of the case-control studies (table 1). This result was statistically significant ($P<0.05$) in 70% of the 260 trials, 84% of the 130 cohort studies, and 84% of the 130 case-control studies (table 2). P values were more extreme in observational studies than in trials ($P<<0.001$), and more extreme in cohort studies than in case-control studies ($P=0.04$). When I considered all results in the abstracts, 86%

Table 1 Measures of binary outcomes in 520 abstracts of research papers. Values are numbers (percentages)

Measure	Randomised trials (n=260)	Cohort studies (n=130)	Case-control studies (n=130)
Relative risk	135 (52)	46 (35)	5 (4)
Odds ratio	116 (45)	79 (61)	125 (96)
Hazard ratio	9 (3)	3 (2)	0
Standardised mortality ratio	0	2 (2)	0

Table 2 Distribution of P values in 520 abstracts of research papers. Values are numbers (percentages)

P interval	Randomised trials(n=260)	Cohort studies(n=130)	Case-control studies(n=130)
P<0.0001	24 (9)	38 (29)	20 (15)
0.0001≤P<0.001	25 (10)	16 (12)	14 (11)
0.001≤P<0.01	40 (15)	27 (21)	31 (24)
0.01≤P<0.02	20 (8)	7 (5)	16 (12)
0.02≤P<0.03	18 (7)	6 (5)	11 (8)
0.03≤P<0.04	25 (10)	7 (5)	11 (8)
0.04≤P<0.05	29 (11)	9 (7)	8 (6)
0.05≤P<0.10	16 (6)	2 (2)	2 (2)
0.10≤P<0.20	10 (4)	4 (3)	5 (4)
P≥0.20	53 (20)	14 (11)	12 (9)

(224/260), 93% (121/130), and 93% (120/130) gave significant results.

The distribution of P values in the interval 0.04 to 0.06 was extremely skewed. The number of P values in the interval $0.05 \le P < 0.06$ would be expected to be similar to the number in the interval $0.04 \le P < 0.05$, but I found five compared with 46, which is highly unlikely to occur ($P << 0.0001$) if researchers are unbiased when they analyse and report their data.

Only five trials had $0.05 \le P < 0.06$ whereas 29 trials had $0.04 \le P < 0.05$. (I included two abstracts where P was given as $P < 0.05$, which I assumed to be just below 0.05.[w1 w2]) I could check the calculations for four and 23 of these trials, respectively, and confirmed three of the four non-significant results. The fourth result was $P = 0.05$, which the authors interpreted as a significant finding; I got $P = 0.03$.[w3] Eight of the 23 significant results were correct; four were wrong,[w1 w4–w6] five were doubtful,[w7–w11] four could be discussed (see table A on bmj.com),[w2 w12–w14] and two were only significant if a χ^2 test without continuity correction was used (results not shown).[w15 w16]

The distribution of P values between 0.04 and 0.06 was even more extreme for the observational studies. Nine cohort studies and eight case-control studies gave P values in this interval, but in all 17 cases $P < 0.05$. Because the analyses had been adjusted for confounders, recalculation was not possible for any of these studies. One of the nine cohort studies and two of the eight case-control studies gave a confidence interval where one of the borders was one; in all three studies, this was interpreted as a positive finding,[w17–w19] although in one this seemed to be the only positive result out of six time periods the authors had reported.[w19]

Discussion

Significant results in abstracts should generally be disbelieved. I found a high prevalence of significant results in the abstracts of 260 randomised trials, 130 cohort studies, and 130 case control studies. I excluded abstracts that did not present useful data or any data at all for the first result, but this did not seem to have an effect. Of the 18 excluded trials (figure), 10 had significant results in the abstract for other outcomes, and four described positive findings; and all of the 32 excluded observational studies described significant or positive results in the abstract.

It was unexpected that so many abstracts of randomised trials presented significant results because a general prerequisite for trials is clinical equipoise—that is, the null hypothesis of no difference is often likely to be true. Furthermore, the power of most trials is low; the median sample size in group comparative trials that compared active treatments was only 71 in 1991.[6] Nevertheless, surveys have found significant differences in 71% of trial reports of hepatobiliary disease[7]; in 34% of trials of analgesics[8]; and in 38% of comparative trials of non-steroidal anti-inflammatory drugs, even though the median sample size per group was only 27.[3]

Ongoing research has shown that more than 200 statistical tests are sometimes specified in trial protocols.[9] If you compare a treatment with itself—that is, the null hypothesis of no difference is known to be true—the chance that one or more of 200 tests will be statistically significant at the 5% level is 99.996% ($=1-0.95^{200}$) if we assume the tests are independent. Thus, the investigators or sponsor can be fairly confident that "something interesting will turn up." Due allowance for multiple testing is rarely made, and it is generally not possible to discern reliably between primary and secondary outcomes. Recent studies that compared protocols with trial reports have shown selective publication of outcomes, depending on the obtained P values,[10–12] and that at least one primary outcome was changed, introduced, or omitted in 62% of the trials.[10]

The scope for bias is also large in observational studies. Many studies are underpowered and do not give any power calculations.[4] Furthermore, a survey found that 92% of articles adjusted for confounders and reported a median of seven confounders but most did not specify whether they were pre-declared.[4] Fourteen per cent of these articles reported more than 100 effect estimates, and subgroup analyses appeared in 57% of studies and were generally believed.[4]

Without randomisation, you would expect almost any comparison to become statistically significant if the sample size is large enough, since the compared groups would nearly always be different.[13] P values in observational research, therefore, can be particularly misleading and should not be interpreted as probabilities.[13] This fundamental problem is likely one of the reasons that the P values for cohort studies were the most extreme, as data from many big cohorts are published repeatedly.[4]

Because claimed cause-effect relations are so often false alarms, some experienced epidemiologists are not impressed by harms shown in observational studies, unless the risk is increased by at least three times.[14] This number should preferably be outside the confidence interval, since even an odds ratio of 20.5 fades, if the confidence interval goes from 2.2 to 114.0. Confidence intervals were available for the first result in 116 abstracts of the case-control studies, but only in six cases (5%) was the risk confidently increased by at least three times.

Although many of the significant results I identified in the abstracts were highly selective—for example, "The strongest mechanical risk factor," "The only

What is already known on this topic

Errors and bias in statistical analyses are common

A review of observational studies has found a preponderance of P values in abstracts between 0.01 and 0.05 that indicated biased reporting or biased analyses

What this study adds

A high proportion of abstracts of randomised trials and observational studies have significant results

Errors and bias in analysis and reporting are common

Significant P values in abstracts should generally be disbelieved

factor associated with," "The highest odds ratio"—few abstracts had any reservations about these data. I checked the 181 significant abstracts of randomised trials a second time but found only four reservations (2%), although subgroup or secondary analyses and adjustment for confounders in regression analyses were common, as shown by the frequent use of the odds ratio rather than relative risk (table 1). Accordingly, a trial survey found that most results of subgroup analyses found their way to the abstract or conclusion of the paper.[15]

To study bias during data analysis more closely, I focused on P values between 0.04 and 0.06, even though from a statistical perspective P values in this interval should be interpreted similarly, of course. Some of the significant results were wrong or doubtful. This agrees with a survey of drug trials, where it was usually not possible to check the calculations.[3] I found 10 trials in which significant results were erroneous and strongly suspected false positive results in another five, and in all cases the new drug was favoured over the active control drug.[3]

Significant results in abstracts should generally be disbelieved. The preponderance of significant results could be reduced if the following action was taken. Firstly, if we need a conventional significance level at all, which is doubtful,[16] it should be set at $P < 0.001$, as has been proposed for observational studies.[17] Secondly, analysis of data and writing of manuscripts should be done blind, hiding the nature of the interventions, exposures, or disease status, as applicable, until all authors have approved the two versions of the text.[18] And finally, journal editors should scrutinise abstracts more closely and demand that research protocols and raw data—both for randomised trials and for observational studies—be submitted with the manuscript.

I thank S H Arshad, B H Auestad, C Baker, D Bishai, P Callas, J T Connor, H Fjærtoft, T P George, K-T Khaw, L Korsholm, L C Mion, F-J Neumann, Y Sato, B-S Sheu, and P Talmud for clarifications on their studies.

Contributors: PCG is the sole contributor and is guarantor.

Funding: None.

Competing interests: None declared.

Ethical approval: Not needed.

1 Scherer RW, Langenberg P, von Elm E. Full publication of results initially presented in abstracts. *Cochrane Database Methodol Rev* 2005;2:MR000005.
2 Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials: a survey of three medical journals. *N Engl J Med* 1987;317:426-32.
3 Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal, antiinflammatory drugs in rheumatoid arthritis [amended in 1989;10:356]. *Controlled Clin Trials* 1989;10:31-56.
4 Pocock SJ, Collier TJ, Dandreo KJ, de Stavola BL, Goldman MB, Kalish LA, et al. Issues in the reporting of epidemiological studies: a survey of recent practice. *BMJ* 2004;329:883.
5 Bland JM, Altman DG. The use of transformation when comparing two means. *BMJ* 1996;312:1153.
6 Mulward S, Gøtzsche PC. Sample size of randomized double-blind trials 1976-1991. *Dan Med Bull* 1996;43:96-8.
7 Kjaergard LL, Gluud C. Citation bias of hepato-biliary randomized clinical trials. *J Clin Epidemiol* 2002;55:407-10.
8 Bland JM, Jones DR, Bennett S, Cook DG, Haines AP, Macfarlane AJ. Is the clinical trial evidence about new drugs statistically adequate? *Br J Clin Pharmacol* 1985;19:155-60.
9 Chan AW, Hróbjartsson A, Tendal B, Gøtzsche PC, Altman DG. Pre-specifying sample size calculations and statistical analyses in randomised trials: comparison of protocols to publications. Melbourne: *XIII Cochrane Colloquium*, 22-26 Oct, 2005:166.
10 Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;291:2457-65.
11 Chan AW, Krleza-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ* 2004;171:735-40.
12 Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ* 2005;330:753.
13 Greenland S. Randomization, statistics, and causal inference. *Epidemiology* 1990;1:421-9.
14 Taubes G. Epidemiology faces its limits. *Science* 1995;269:164-9.
15 Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;355:1064-9.
16 Sterne JAC, Smith GD. Sifting the evidence: what's wrong with significance tests? *BMJ* 2001;322:226-31.
17 Smith GD, Ebrahim S. Data dredging, bias, or confounding. *BMJ* 2002;325:1437-8.
18 Gøtzsche PC. Blinding during data analysis and writing of manuscripts. *Controlled Clin Trials* 1996;17:285-90.

(Accepted 1 April 2006)

doi 10.1136/bmj.38895.410451.79

A memorable patient

Diphtheria

In the 1940s we were still seeing occasional cases of diphtheria in our provincial hospital in New Zealand's North Island. Two I clearly remember died from overwhelming toxaemia; they had come from isolated country districts and presumably had not been immunised.

In April 1946, a 16 year old schoolgirl, also from an isolated country district, was admitted with severe respiratory obstruction due to diphtheria. She had widespread membrane formation in the mouth and throat, and, when I performed an urgent tracheotomy, we were alarmed to find a similar membrane lining the trachea. This feature of the case explained the utter frustration we experienced in the after care. We just could not obtain a satisfactory airway. In addition to the usual nursing care of a tracheotomy, I judiciously removed the outer tube periodically to endeavour to clear the trachea, but to no avail.

Three days after the operation, I was due to go off duty for the weekend. I discussed the care of my patient with the duty house surgeon and asked to be rung if there was any change in her condition.

Early on the Saturday evening while I was visiting friends the house surgeon rang and said rather sadly, "I think we are going to lose our patient; her breathing is much worse." I was back at the hospital within 10 minutes and proceeded to remove the tracheotomy tube. I discovered immediately what had happened; the diphtheritic membrane was lying, apparently free, in the trachea. With some trepidation, I proceeded to gently ease it out with a pair of Desjardins forceps. The membrane kept coming and did not seem to be attached at the lower end. We became quite excited and relieved to find we were holding a complete cast of the trachea and its commencing bifurcation.

We were so busy looking at this that we nearly forgot the patient. The staff nurse pointed out to us that the patient was trying to say something: she was mouthing, "I feel much better." From that moment she made a steady recovery and was able to go home some days later.

I can only leave the reader to imagine what I learnt from this case. Sixty years later I am still pleased and relieved that I went back to the hospital that Saturday evening.

Caleb Tucker *retired surgeon and hospital administrator, Wellington, New Zealand (candjtucker@xtreme.net.nz)*