12 Sivarajah N. *Nutritional status of the children in Jaffna*. Jaffna: University of Jaffna, 1993.
13 Natchinarkinian CS. Current pattern of health care and resource allocation. In: Arulanantham K, Rathneswaren S, Sreeharan N, eds. *Victims of war in Sri Lanka: a quest for health consensus*. London: Medical Institute of Tamils, 1994:15-24.
14 Nissan E. *Sri Lanka: a bitter harvest*. London: Minority Rights Group, 1990.
15 Amnesty International *Sri Lanka: wavering commitment to human rights*. London: AI, 1996.
16 Trawick M. Reasons for violence: a preliminary ethnographic account of the Tamil tigers. In: Gamage S, Watson IB, eds. *Conflict and community in contemporary Sri Lanka: "Pearl of the East" or the "Island of Tears"*? New Delhi: Sage, 1999: 139-163.
17 Unicef, SCF Coalition to stop the use of child soldiers. *The use of children as soldiers in the Asia-Pacific region*. London: Coalition to stop the use of child soldiers
18 University Teachers for Human Rights. *Children in the North-East War: 1985-1995*. Colombo: UTHR-J, 1995.
19 University Teachers for Human Rights. *The sun god's children and the big lie*. Colombo: UTHR-J, 2000.
20 Rosenbeck R. The Malignant post-Vietnam stress syndrome. *Am J Orthopsychiatry* 1985;55:2,319-32.
21 Amnesty International. *Children in South Asia*. London: AI, 1998.
22 Jareg E, McCallin M. *The rehabilitation of former child soldiers*. Geneva: International Catholic Child Bureau, 1993.
23 University Teachers for Human Rights-Jaffna. *From Manal Aru to Weli Oya*. Colombo: UTHR-J, 1994.
24 Machel G. *Impact of armed conflict on children*. New York: United Nations, 1996.

Peer review of statistics in medical research: the other problem

Peter Bacchetti

Peer review has long been criticised for failing to identify flaws in research. Here Peter Bacchetti argues that it is also guilty of the opposite: finding flaws that are not there

Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143-0560, USA
Peter Bacchetti *professor*
pbacchetti@epi.ucsf.edu

BMJ 2002;324:1271–3

The process of peer review before publication has long been criticised for failing to prevent the publication of statistics that are wrong, unclear, or suboptimal.[1 2] My concern here, however, is not with failing to find flaws, but with the complementary problem of finding flaws that are not really there.

My impression as a collaborating and consulting statistician is that spurious criticism of sound statistics is increasingly common, mainly from subject matter reviewers with limited statistical knowledge. Of the subject matter manuscript reviews I see that raise statistical issues, perhaps half include a mistaken criticism. In grant reviews unhelpful statistical comments seem to be a near certainty, mainly due to unrealistic expectations concerning sample size planning. While funding or publication of bad research is clearly undesirable, so is preventing the funding or publication of good research. Responding to misguided comments requires considerable time and effort, and poor reviews are demoralising—a subtler but possibly more serious cost.

This paper discusses the problem, its causes, and what might improve the situation. Although the main focus is on statistics, many of the causes and potential improvements apply to peer review generally.

Summary points

- Peer reviewers often make unfounded statistical criticisms, particularly in difficult areas such as sample size and multiple comparisons
- These spurious statistical comments waste time and sap morale
- Reasons include overvaluation of criticism for its own sake, inappropriate statistical dogmatism, time pressure, and lack of rewards for good peer reviewing
- Changes in the culture of peer review could improve things, particularly honouring good performance

The problem

Mistaken criticism is a general problem, but may be especially acute for statistics. The examples below illustrate this, including commonly abused areas (examples 1 and 2), non-constructiveness (1), quirkiness and unpredictability (3 and 4), and the potential difficulty of successful rebuttal (3 and 4).

Example 1: Grant review, US National Institutes of Health

"There is a flaw in the study design with regard to statistical preparation. The sample size appears small."

Because of uncertainties inherent in sample size planning, reviewers can always quibble with sample size justifications—and they usually do. The information needed to determine accurately the "right" sample size (a murky concept in itself) is often much more than available preliminary information. For example, even directly relevant preliminary information from 30 subjects provides an estimated variance with a threefold difference between the lower and upper ends of its 95% confidence interval, resulting in threefold uncertainty in a corresponding sample size calculation.[3] Often considerable uncertainty also exists about other relevant factors such as the size of the effect or association, an outcome's prevalence, confounding variables, adherence to study medication, and so on. Such uncertainties can be especially acute for highly innovative research.

Unfortunately, reviewers usually expect a "sample size calculation," with all the precision that "calculation" implies. This may be reasonable for studies based on extensive previous data but is unrealistic in many situations, particularly pilot or exploratory studies. In

this particular example the request for proposals specifically asked for pilot studies and prohibited phase III clinical trials, and the review provided no reasoning for the quoted criticism.

Example 2: Review for a leading bench science journal

"The statistical test used . . . is not appropriate for the multiple comparisons necessitated by this experimental design."

The authors were puzzled by this comment, because two groups differed substantially and a third was intermediate, all in keeping with their biological theory. They had not expected that gathering data on the intermediate condition would be interpreted as weakening their results. Because it is rarely acceptable to perform only a single statistical analysis in a study, this type of objection can usually be raised. Whether to adjust P values for multiple comparisons is controversial,[4 5] but reviews usually state the need for adjustment as accepted dogma. More importantly, I have rarely seen the issue raised in the classic situation where only one result of many has a small P value. Instead, some reviewers object routinely, even when most results have small P values and there is even a coherent pattern (for example, a treatment showing benefit by many different measures). In such situations, the results reinforce each other, rather than detracting from each other as required by the methods (usually Bonferroni adjustment) that reviewers often suggest.

Example 3: Review for a clinical specialty journal

"Figure 1 appears to be a ROC Curve at a 50% threshold. . . . it is not clear how well the system would have worked had other thresholds been chosen."

For those not familiar with receiver operating characteristic (ROC) curves, this is a self contradicting criticism because such curves display the tradeoffs from all possible cut offs of a prediction rule. The paper was an excellent first effort by a very junior lead author, but the deputy editor explicitly endorsed this and many other spurious and demoralising comments and rejected the revised paper despite our attempts diplomatically to rebut the errors. Another journal published essentially the same paper.

Example 4: Grant review for a disease specific foundation

"It is questionable whether a theoretical baseline value of zero should the use for statistical analysis of differences in the measurements of median matched difference. (sic)"

We found this nearly indecipherable even in the context of the entire review, but the criticism concerned a published study with a matched design and a corresponding statistical analysis (Wilcoxon signed-rank tests). Such methods boil down to testing whether within-pair differences are centred at zero, so the reviewer seemed to be objecting to this general strategy, an objection so spurious that it is almost impossible to rebut. How does one argue that no difference implies a difference of zero when a reviewer believes that empirical research is needed to verify or refute this? The study was not funded, even though the above comment was the only substantive criticism of the proposal. Essentially the same proposal was funded nine months later, after that reviewer had rotated off the committee.

Causes

Several factors may contribute to this problem, some common to all peer review. A pervasive factor is the desire to find something to criticise. Tannen recently documented the overvaluation of criticism and conflict both generally in Western popular culture and specifically in academia.[6] In addition, the notion that finding flaws is the key to high quality peer review is fairly explicit in some writings,[7–9] and developers of an instrument for rating review quality recently focused only on "completeness" and not on "whether the reviewer's judgment was correct."[10] A panel on peer review for the US National Institutes of Health acknowledged an overly critical climate, stating "Peer reviewers should eschew the common current tendency to find fault."[11] Finding flaws is certainly important, and scepticism and disputation are revered in scientific tradition. But when criticism is an end in itself rather than a tool for advancing knowledge, when finding flaws is imperative rather than the natural result of careful review with an open mind, then mistaken criticisms will arise.

The problem may be more acute in statistics because of two factors that are synergistic both with each other and with the need to criticise. The first is that reviewers see statistics as a rich area for finding mistakes. This perception is correct, because statistical errors are common. But areas such as sample size and multiple comparisons can be reflexively subjected to unfair and unhelpful criticism. In the case of clinical trials Meinert lists many other "universal" criticisms.[12] The second factor is many reviewers' poor understanding of statistics,[2] especially the belief that rules must be blindly followed. I am dismayed by how often my clients ask whether a particular approach would be "legal" or "against the rules" rather than "accurate" or "misleading." This misunderstanding of statistics as a body of seemingly arbitrary dogma leads many reviewers to perceive violations even when the research has not actually been harmed.

Finally, another pair of synergistic factors apply to peer review generally. The first is the frequent need to rush reviewing. This seems unlikely to improve, given

increasing emphasis on documented productivity and the accelerating pace of life generally.[13] The second, perhaps more important, is the lack of incentives. One recent editorial noted, "It is generally admitted that being a good referee does not lead to any tangible rewards with respect to career advancement."[14] Another noted that "the integrity of the scientific review process requires that the performance of reviewers be appropriately rewarded" and ended, "We do thank you."[15] This gratitude, while sincere, is emblematic of the inadequate rewards that reviewers can expect. The only likely concrete consequence of good reviewing is future requests for more reviews.

What might help?

Aside from widespread improvement in understanding of statistical methods (a worthy goal), care by reviewers and changes in peer review systems and culture may reduce mistaken statistical criticisms and improve peer review generally.

- As a reviewer, criticise statistical flaws only when you can explain how they specifically detract from the study. If you suspect a problem but cannot meet this criterion, recommend further review by an expert statistician.
- Raise criticisms in swampy areas such as sample size and multiple comparisons only when unavoidable. For sample size, this will usually mean that a proposal requires a clearly unrealistic scenario to achieve the stated goals. Keep in mind that a meaningful increase in sample size may be impossible, so the alternative is a sample size of zero—that is, not doing the study. Also remember that research in new areas must start somewhere, even when there are no preliminary data for sample size calculations. Concerns about sample size after a study is done can generally be refocused more directly on whether the authors have properly presented and interpreted the uncertainty in their results, particularly negative findings.

Changing the system

Any substantial improvement will probably require changes in the manuscript and grant review processes. Statistical reviewers are already used to some extent.[2] While much wider use may not be possible, obtaining a statistical review whenever subject reviewers raise statistical concerns might be workable. Research on double blinded and non-anonymous peer review has found little effect of these variations on who knows whose identities,[16–19] but mistaken criticism has not been directly addressed. Training or guidelines may help, particularly if they warn against stretching to find criticisms.

A change that would perhaps improve peer review even more would be to evaluate its quality and reward good performance. Because meaningful grading must reflect the substance of the reviews, including whether criticisms are correct and whether serious flaws have been overlooked, fellow reviewers of the same paper are perhaps best positioned to rate each other's performance. This would also promote reflection on one's own performance.

A simple form of reward would be to supplement long annual lists of all reviewers with much shorter honour rolls of those who have provided high quality reviews. Multiple honour rolls could address different aspects, such as helpfulness to editors, high ratings from fellow reviewers, or good marks from rejected authors on constructiveness. Paying attention to review quality might result in cultural changes. For example, top academic institutions may come to see failure to make at least one honour roll of a relevant journal as a serious weakness.

Conclusion

Peer review is a key part of the collective scientific process. Expecting it to work well on donated time, with little training and even less accountability or incentives, seems unrealistic. Changes in the systems and culture of peer review might improve things, notably less pressure to criticise, more training in reviewing skills, and less statistical dogmatism. The most promising change might be to better reward good performance.

I thank Professors Douglas G Altman and Steven N Goodman for helpful comments on an early draft of this paper.

Competing interests: None declared.

1 Altman DG. Statistical reviewing for medical journals. *Stat Med* 1998;17:2661-74.
2 Goodman SN, Altman DG, George SL. Statistical reviewing policies of medical journals: Caveat lector? *J Gen Intern Med* 1998;13:753-6.
3 Matthews JNS. Small clinical trials: are they all bad? *Stat Med* 1995;14:115-6.
4 Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990;1:43-6.
5 Goodman SN. Multiple comparisons, explained. *Am J Epidemiol* 1998;147:807-12.
6 Tannen D. *The argument culture: moving from debate to dialogue*. New York: Random House, 1998.
7 Callaham ML, Baxt WG, Waeckerle JF, Wears RL. Reliability of editors' subjective quality ratings of peer reviews of manuscripts. *JAMA* 1998;280:229-31.
8 Godlee F, Gale CR, Martyn CN. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports. *JAMA* 1998;280:237-40.
9 Kassirer JP, Campion EW. Peer review: crude and understudied, but indispensable. *JAMA* 1994;272:96-7.
10 Black N, van Rooyen S, Godlee F, Smith R, Evans S. What makes a good reviewer and a good review for a general medical journal? *JAMA* 1998;280:231-3.
11 Alberts BM, Ayala FJ, Botstein D, Frank E, Holmes EW, Lee RD, et al. *Recommendations for change at the NIH's Center for Scientific Review: phase 1 report*. www.csr.nih.gov/bioopp/select.htm (accessed 17 September 2001).
12 Meinert CL. *Clinical trials: design, conduct, and analysis*. New York: Oxford University Press, 1986:276-7.
13 Gleick J. *Faster. The acceleration of just about everything*. New York: Pantheon Books, 1999.
14 Pros and cons of open peer review. *Nat Neurosci* 1999;2:197-8.
15 Bloom FE. The importance of reviewers. *Science* 1999;283:789.
16 Justice AC, Cho MK, Winkler MA, Berlin JA, Rennie D. Does masking author identity improve peer review quality? A randomized controlled trial. *JAMA* 1998;280:240-2.
17 van Rooyen S, Godlee F, Evans S, Smith R, Black N. Effect of blinding and unmasking on the quality of peer review. *JAMA* 1998;280:234-7.
18 Goldbeck-Wood S. Evidence on peer review—scientific quality control or smokescreen? *BMJ* 1999;318:44-5.
19 van Rooyen S, Godlee F, Evans S, Black N, Smith R. Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *BMJ* 1999;318:23-7.

Endpiece

Only external things

He was a strong man in business and in politics, but these are external things: only a fool gives his soul to them.

Robertson Davies (1913-95), *What's bred in the bone*, London: Penguin, 1986