Statistical errors in medical research – a chronic disease?

James Young

Statistician, Statistical Advisor for the Swiss Medical Weekly, Basel, Switzerland

It is over ten years now since Altman's "cri du coeur" – "The scandal of poor medical research" [1]. Since then editors of medical journals have made a concerted effort to improve the quality of medical research, with initiatives such as the CONSORT and other statements and uniform requirements for manuscripts submitted to medical journals (see www.icmje.org). But it would be fair to say that these initiatives have been slow to take effect [2, 3]. In this edition of the Swiss Medical Weekly, Strasak and colleagues review common statistical errors in medical research [4].

The authors make many excellent points. For example: studies lacking a pre-specified hypothesis should be clearly labelled as exploratory; conventional statistical inference requires randomisation or random sampling; and "where appropriate" does not constitute an adequate description of statistical methods. I would like to comment on three issues raised somewhat indirectly in their article: the role of the statistician in medical research, and the need for both interval estimates and multivariate methods.

The authors rightly stress the need to involve a statistician in study design. Recruiting a statistician is not an admission of inadequacy. Statistics, like medicine, has expanded to become a broad and complicated discipline. "Nowadays the catalogue of statistical methods is so very extensive that a working scientist is somewhat less than overjoyed at the prospect of having to learn yet another procedure" [5] – and that was in 1958. In one of a pair of valuable articles on how to carry out a randomised trial [6, 7] Sackett wrote: "Time taken to master [biostatistical] nuances is at the expense of maintaining clinical competence, a social life, a positive self-image and a sense of humour." Faced with these prospects, who would not want a statistician on their research team?

The authors also discuss the role of the statistician in the peer review process. Altman notes that "Evaluation of the scientific quality of research papers often falls to statisticians" [1]. Here I think the editors of this journal live up to the authors' expectations. I review almost all manuscripts with any sort of numerical information and my requests to see revised manuscripts are always honoured. Researchers should appreciate that statistical reviewers have only two to three hours in which to become familiar with material they have been working on for months if not years. It's only to be expected that statisticians will make mistakes and researchers should not hesitate to (gently) point out these mistakes in their reply. "Most statisticians want to be helpful. Like arrogant medics, arrogant statisticians are a dying race " [8]. "Statistical refereeing is a form of fire fighting" [1] however, and at this late stage, it is often impossible to resurrect what could have been valuable research.

Analysis may be the most visible activity of the statistician in medical research. But good design is more important; one can always re-analyse good data. So it is best to work with a statistician to develop a comprehensive protocol. Then to a certain extent, collecting and analysing data becomes a matter of following these guidelines and other members of the research team can carry out these activities, typically with the help of a statistician in more of an advisory role.

In general medical researchers have been slow to appreciate the need for both interval estimates and multivariate methods when analysing data, and even slower to appreciate the benefits of taking a Bayesian approach to analysis. Contrary to what the authors have written, Bayesians have no conceptual difficulty in asserting their probability of the truth of a null hypothesis given the data (see [9]). The Bayesian approach allows one to consider the evidence from this particular study in the light of prior evidence from other sources. When the prior evidence for a hypothesis is strong, a positive study is more likely to be a "true positive". "The mistake is to confuse an increment in support from a positive study with cumulatively strong support for the hypothesis" [10]. Focusing on cumulative support for a hypothesis is the key to avoiding spurious results. Fortunately most medical researchers are Bayesian in their discussion of what their results mean, even if not in their approach to analysis.

Many medical researchers still have an unhealthy pre-occupation with p-values [11]. Arguably hypothesis tests have their place – for any variable that was the subject of a formal sample size calculation. Then what is meant by a relevant difference has been defined and attitudes to the risk of "false positives" and "false negatives" have been asserted. The study will be designed to detect this difference and a hypothesis test is in order. But what if there has been no sample size calculation; what then is the use of a hypothesis test? A small p-value could merely reflect a difference that is clinically irrelevant. Collect large enough samples and a statistically significant difference is almost certain. On the other hand, a large p-value could merely reflect small samples with little power to detect a clinically relevant difference. What is really needed is an interval estimate of the size of the difference [12].

I think multivariate methods of analysis should be considered the rule, not the exception. This is particularly true for the many observational studies submitted to this journal. Observational studies have no randomisation, and statistical inference then relies on judgements of exchangeability within strata defined by covariates [13, 14]. That is, you believe that within each of the strata conceptually created by covariate adjustment (see [15] p. 96), whether a patient is exposed or not is essentially a random event. Here an estimate of a difference is a model-based inference and appropriate covariate adjustment is needed so that others will be convinced that under your model exchangeability is a reasonable assumption. Obviously one can never hope to measure all covariates that might influence exposure, but a sensible choice of a reasonable number of covariates should be sufficient because many covariates will be correlated. Therefore observational studies must be large enough to support appropriate covariate adjustment (see [16]).

Covariate adjustment is often advantageous even with data from a randomised trial. Adjustment using a baseline measure of outcome will lead to better precision in the estimate of a difference when outcomes are normally distributed, and reduce bias in this estimate with binary and survival outcomes [17]. Hence t and chi square and log rank tests are to statistics what cupping, bloodletting and leaches are to medicine: of historical interest, on rare occasions still useful, but largely superseded by superior methods. Instead the medical researcher needs to be familiar with their multivariate replacements: linear, logistic and proportional hazards regression.

Multivariate methods come with their own set of common errors. Resist the temptation to categorise continuous predictor variables [18]. Do not use automatic covariate selection methods or pretesting [16, 19]; instead pre-specify covariates in the protocol based on clinical reasoning [20]. Any subgroup analysis should be by estimating the effect of interactions between covariates and exposure [21, 22]. Watch for “pseudo-replication” (also known as clustering): observations that are not independent because of group membership (patients from same family or assessed by the same clinician) or because of repeated measurements made on the same individual. More advanced methods are needed for data of this sort [23–27]. More advanced methods are also needed when a number of exposure variables could potentially affect the outcome and interest lies in which exposures are the most important [16, 28].

And while the authors suggest it is not necessary to read “whole textbooks on statistical methodology”, some textbooks are easy to read and well worth the effort. I particularly like textbooks by Senn, Harrell, Kleinbaum and Klein, and Kirkwood and Sterne [15, 29–32].

The authors provide useful checklists and a comprehensive guide to the literature on statistical errors in medical research. Here I have added a few favourite references of my own focusing on observational studies and multivariate methods. A draft STROBE statement, designed to improve the quality of observational studies, is also available [33]. Prevention is definitely easier than a cure with this disease, so work with a statistician on study design; standard therapy is a multivariate analysis; and avoid p-values like the plague they undoubtedly are.

Correspondence:
Jim Young
Basel Institute for Clinical Epidemiology
University Hospital Basel
Hebelstrasse 10
CH-4031 Basel
Switzerland
E-Mail: jyoung@uhbs.ch

References

1 Altman DG. The scandal of poor medical research. BMJ. 1994;308:283–4.
2 Altman DG. Poor-quality medical research: what can journals do? JAMA. 2002;287:2765–7.
3 Senn S. Maintaining the integrity of the scientific record. Scientific standards observed by medical journals can still be improved. BMJ. 2002;324:169.
4 Strasak AM, Zaman Q, Ulmer H. Statistical errors in medical research – a review of common pitfalls. Swiss Med Wkly. 2006.
5 Bross IDJ. How to use ridit analysis. Biometrics. 1958;14:18–38.
6 Sackett DL. Why randomized controlled trials fail but needn’t: 1. Failure to gain “coal-face” commitment and to use the uncertainty principle. CMAJ. 2000;162:1311–4.
7 Sackett DL. Why randomized controlled trials fail but needn’t: 2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!). CMAJ. 2001;165:1226–37.
8 Sprent P. Statistics in medical research. Swiss Med Wkly. 2003;133:522–9.
9 Christensen R. Testing Fisher, Neyman, Pearson, and Bayes. Am Stat. 2005;59:121–6.
10 Savitz DA. Commentary: prior specification of hypotheses: cause or just a correlate of informative studies? Int J Epidemiol. 2001;30:957–8.

11 Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. Psychol Sci. 2004;15:119–26.

12 Braitman LE. Confidence intervals assess both clinical significance and statistical significance. Ann Intern Med. 1991;114: 515–7.

13 Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol. 1986;15:413–9.

14 Greenland S. Randomization, statistics, and causal inference. Epidemiology. 1990;1:421–9.

15 Senn S. Statistical issues in drug development. Chichester: Wiley; 1997.

16 Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004;66:411–21.

17 Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med. 2002; 21:2917–30.

18 Austin PC, Brunner LJ. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. Stat Med. 2004;23:1159–78.

19 Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996;49:907–16.

20 Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. Control Clin Trials. 2000;21: 330–42.

21 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.

22 Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol. 2004;57:229–36.

23 Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. Stat Med. 1998;17:1261–91.

24 Albert PS. Longitudinal data analysis (repeated measures) in clinical trials. Stat Med. 1999;18:1707–32.

25 Sullivan LM, Dukes KA, Losina E. Tutorial in biostatistics. An introduction to hierarchical linear modelling. Stat Med. 1999;18:855–88.

26 Senn S, Stevens L, Chaturvedi N. Repeated measures in clinical trials: simple strategies for analysis using summary measures. Stat Med. 2000;19:861–77.

27 Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. Stat Med. 2002;21:3291–315.

28 Greenland S. When should epidemiologic regressions use random coefficients? Biometrics. 2000;56:915–21.

29 Harrell FE. Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.

30 Kleinbaum DG, Klein M. Logistic regression: a self-learning text. 2nd ed. New York: Springer-Verlag; 2002.

31 Kleinbaum DG, Klein M. Survival analysis: a self-learning text. 2nd ed. New York: Springer-Verlag; 2005.

32 Kirkwood BR, Sterne JAC. Essential medical statistics. 2nd ed. Malden, Massachusetts: Blackwell; 2003.

33 von Elm E, Egger M. The scandal of poor epidemiological research. BMJ. 2004;329:868–9.

SMW

Established in 1871

Formerly: Schweizerische Medizinische Wochenschrift

Swiss Medical Weekly

Official journal of the Swiss Society of Infectious diseases, the Swiss Society of Internal Medicine and the Swiss Respiratory Society

The many reasons why you should choose SMW to publish your research