E13. If you want to publish, get your statistics right

Donald A. Berry

Chairman, Department of Biostatistics, and Frank T. McGraw Memorial Chair of Cancer Research, University of Texas, M.D. Anderson Cancer Center, Houston, USA

The title of this paper is somewhat misleading. Although literally correct, it might be read as suggesting that publication is an end in itself. It is not. Your goal should be publications that

- Draw appropriate conclusions
- Have clinical or scientific impact
- Are generalisable and verifiable by others

To achieve this goal, appropriate statistical thinking and statistical methodology are essential. On the other hand, getting work published may be possible with superficially correct statistics even though it meets none of the above criteria.

Subset analysis

I can best illustrate the distinction using an example. Suppose you assemble a database consisting of all adjuvant breast cancer patients treated at your home institution in the past 10 years. These were treated by different physicians who had different ideas about using chemotherapy. As a result, some patients received anthracycline-based chemotherapy (A) while others did not (~A), even though they might have had the same clinical and demographic characteristics. You are interested in addressing the benefits of A on disease-free survival (DFS). But you recognise that assignment to A or ~A was not randomised. So in your statistical analysis you adjust for important prognostic characteristics (stage, tumour size, grade, number of positive lymph nodes, oestrogen-receptor (ER) status, etc.) – that is, you use a multivariate model. You carry out a proportional hazards model for DFS and find that treatment (A vs ~A) is not statistically significant.

But you persist. You consider patients with ER-negative tumours separately in a multivariate analysis and find that A is statistically superior to ~A in the sense that the P-value for the treatment effect is <0.05. To be sure that you have done the analysis correctly, you provide the data for those patients having ER-negative tumours to a statistician, who verifies your conclusion. So you publish.

What's wrong with that? Everything! You have published a subset analysis without admitting it. You have committed the first sin of empirical science by telling only part of the story. When someone tries to confirm your results, either in the laboratory or in the clinic, they will likely fail. You succeeded in publishing because you (probably unwittingly) duped your statistician and the publication process. Subset analyses are inevitable and important, but you did this one wrong.

The above scenario occurs too frequently, although perhaps not in such a stark fashion. Researchers frequently ask statisticians to analyse data that have been "cleaned". Perhaps they have removed duplicates. Or averaged some observations. Or removed outlying observations. Or restricted to experimental units of most interest – perhaps based on a perusal of the data. Sometimes the most important statistical analyses have been carried out before the statistician sees the data. In the worst cases, the results as presented are useless for inference and detrimental to science.

Experimental design

Research involves experimentation. Statistics deals with designing experiments and subsequently analysing the results. Design and analysis are paired. Good statistical principles must be applied at the design stage, and at the design stage the eventual statistical analyses should be prospectively defined.

Laboratory scientists make several types of mistakes in setting up their experiments: (1) they do not control for ancillary factors that might be confounded with treatment effect, (2) they do not randomise experimental units to treatment, (3) they do not blind technicians as to treatment assigned, and (4) they selectively redo parts of the experiment because the results were not consistent with their expectations (and then they report only the revised results). Another type of mistake laboratory scientists make is not adequately describing the set-up and conduct of their experiment in manuscripts submitted for publication. An investigator who wants to confirm their results should be able to duplicate the experimental conditions.

Clinical researchers also make mistakes at the design stage, including these: (1) not specifying the primary endpoint and method of analysis, (2) not specifying timing of interim analyses and consequent decisions based on

1359-6349/$ – see front matter

interim results, and (3) not appropriately evaluating the design's operating characteristics (especially the false-positive rate and statistical power). Clinical research is more difficult than laboratory research in the sense that experimental units (patients) are very dear. A ubiquitous problem in clinical research design is the inability to meet target sample size. Implications of this circumstance on the inferences to be drawn can and should be addressed in the design. And steps should be taken to salvage an experiment that seems unlikely to meet its objectives. These might include a plan for synthesising the experiment's results with those from other experiments.

Statistical analysis

Laboratory experimentation usually allows for measuring many quantities. Extreme examples are cDNA and mRNA microarrays and proteomic spectra. Tens of thousands of quantities may be measured. But immunohistochemistry measurements of a half dozen markers per tumour are subject to the same inferential problems. Namely, with many possibilities, some observations will look "interesting" or even "tantalising". For example, it is a trivial matter to discriminate between tumours that recur early and those that recur late on the basis of the expression of a sufficient number of genes, even if those genes are irrelevant to the disease. The critical issue is confirming or validating the results. This can be done on a separate set of tumours or, using appropriate statistical techniques within a single experiment.

Clinical research has similar problems of "multiplicities", including subset analyses, as discussed above.

As indicated earlier, statistical analyses are laid out at the design stage. These should be followed when publishing results. However, not every experiment goes exactly as planned. Analyses not prospectively specified may seem reasonable and even necessary after the fact. These can be carried out and presented in publications, but with appropriate caveats and cautions. The guiding principle is to tell the whole story: what was planned, what happened to change the plan, and what analysis strategies were tried in addition to those actually presented. If the results are sufficiently dramatic then they may carry the day. More typically, they should be advertised and interpreted as being hypothesis-generating.

A common type of clinical end-point in breast cancer research is time to an event (death, recurrence, progression). These may be analysed via multivariate methods such as proportional hazards models, or displayed in Kaplan-Meier curves by treatment with an associated univariate test such as the logrank. The former is preferred statistically, but the latter may be designated the primary analysis to satisfy a regulatory authority. In any case, it behoves the investigator to assess treatment effects by considering hazards over time. For example, does the treatment modify the hazards early, but not in the long-term? Several examples are provided in Refs [1,2].

An additional reason for considering hazards over time is that patients will have different durations of follow-up. Therefore, early hazards will be more precisely known than will later hazards. And some estimated hazards are subject to change with additional follow-up while others (those for the earliest time periods) are not. It behoves the investigator to make clear which are subject to change, and by how much they may change. Regarding this last point, predictive distributions [1] are important tools for conveying the uncertainty present in view of missing or censored data.

References

[1] Berry DA (2003). Statistical Innovations in Cancer Research. In *Cancer Medicine e. 6*. Ch 33, pp 465–478. London: BC Decker. (Ed: Holland J, Frei T et al.].

[2] Berry DA, Smith T, Buzdar A (2004). Breast Cancer. In *Textbook of Clinical Trials*. New York: John Wiley & Sons (Ed: Machin D, Day S.) (To appear).