PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. 2003; **2**: 239–240 (DOI:10.1002/pst.090)

VIEWPOINT

Disappointing dichotomies

Stephen Senn*,†
Department of Statistics, University of Glasgow

A regrettably common means of judging the effect of treatments is by 'responder analysis': at the end of the trial every patient is classified according to a binary variable with 'responded' and 'did not respond' as the alternatives [1]. Sometimes the classification is natural. For example, in a trial in anti-infectives it may be possible to separate patients into 'cured' and 'not cured' categories, and this may seem to be the only meaningful thing to do. Or, in a trial in lung cancer, it may seem natural to use some standard time of follow-up and classify the patients as either alive or dead at the end. However, even in such cases a simple analysis using a chi-square test on a fourfold table, or, where covariate information is employed, its more sophisticated cousin, logistic regression, may be unwise. To take the case of lung cancer, in the long run we are all dead, and if the follow-up time has been chosen unwisely we may fail to find important genuine differences between treatment groups. A better approach may be to use the time of death as the outcome variable and hence survival analysis rather than logistic regression.

However, such dichotomies are often far from natural and instead arbitrarily constructed from continuous (or nearly continuous) measurements. An approach common in many areas, hypertension and depression to name but two, is to compare the result at outcome to that at baseline and then classify patients as responders or not depending on whether some arbitrary threshold of difference has been reached [2, pp. 118–119]. There are many reasons to why this is undesirable. The first is that it is inefficient and leads to at the very least a 40% increase in the sample size required [3]. (If the cut-point has been chosen unwisely, matters can be *much* worse.) The second is that it is vulnerable to the sort of trends that clinical trials are designed to control. Consider, for example, a trial for hypertension in which patients have been selected for inclusion on the basis of a single baseline measurement, compared to one in which some form of repeated measurement over a period has been used. Other things being equal, in the former case there is likely to be a greater regression to the mean effect. Hence the response rate in both control and intervention groups is likely to be higher. The expected difference between the groups will not necessarily be the same, either on the probability scale or on the log-odds scale. Third, such an approach makes an inefficient use of the baseline measurements. Being based on change scores, a double inefficiency is introduced. The dichotomy is more inefficient than the change score, as already discussed, and the change score is already more inefficient than analysis of covariance [2, pp. 106–108]. Finally, it encourages naïve and inappropriate judgements of causality. It is inherently liable to be extravagantly interpreted as showing whether a given patient did or did not gain a benefit from treatment. However, every patient in a clinical trial could show the same true degree of benefit to treatment but through measurement error some would have a difference that exceeded the response threshold and some would not. We could then make erroneous judgements about the proportion showing response [4].

Unfortunately, such inefficient measures are becoming enshrined in regulatory practice. For

*Correspondence to: Stephen Senn, Department of Statistics, University of Glasgow, 15 University Gardens, Glasgow G12 8QQ.
†E-mail: stephen@stats.gla.ac.uk

example, the CPMP guideline on multiplicity more or less takes it for granted that such measures will be used and, indeed, are desirable [5]. The current fashionable obsession with numbers need to treat is unfortunately fuelling the demand for dichotomies [6]. There are now a number of guidelines for specific therapeutic areas that require responder analyses. They are often defended in terms of 'clinical relevance', but in my opinion this phrase is simply a mantra that is chanted to justify bad habits. Take the case of blood pressure. Trialists seem to have forgotten that this is itself a surrogate measure (admittedly a very familiar one) for strokes, kidney damage, eye damage, heart disease and so forth. If clinically relevant measures are needed, then these therapeutic outcomes are relevant. Diastolic blood pressure as a continuous outcome will predict these sequelae of hypertension better than if dichotomized.

The net consequence of dichotomizing continuous data is that trials are much bigger than they need be, that our inferences are poorer and that we are wasting both resources and lives. Thinking about appropriate measurement is an important part of any science. Physicists take it very seriously. Trialists should do the same.

REFERENCES

1. Altman DG. Statistics in medical journals: some recent trends. *Statistics in Medicine* 2000; **19**: 3275–3289.
2. Senn SJ. *Statistical issues in drug development*. Wiley: Chichester, 1997.
3. Whitehead J. Sample size calculations for ordered categorical data. *Statistics in Medicine* 1993; **12**:2257–2271.
4. Senn SJ. Author's reply to Walter and Guyatt. *Drug Information Journal* 2003; **37**:7–10.
5. Committee for Proprietary Medicinal Products. Points to consider on multiplicity issues in clinical trials. 2002; 1–11.
6. Grieve AP. The number needed to treat: a useful clinical measure or a case of the Emperor's new clothes? *Pharmaceutical Statistics* 2003; **2**: 87–102.