STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:3189–3209
Published online 19 May 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3603

Measurement in clinical trials: A neglected issue for statisticians?

Stephen Senn[1,*,†] and Steven Julious[2]

[1]*Department of Statistics, University of Glasgow, Glasgow G12 9LL, U.K.*
[2]*University of Sheffield, Sheffield, U.K.*

SUMMARY

Biostatisticians have frequently uncritically accepted the measurements provided by their medical colleagues engaged in clinical research. Such measures often involve considerable loss of information. Particularly, unfortunate is the widespread use of the so-called 'responder analysis', which may involve not only a loss of information through dichotomization, but also extravagant and unjustified causal inference regarding individual treatment effects at the patient level, and, increasingly, the use of the so-called number needed to treat scale of measurement. Other problems involve inefficient use of baseline measurements, the use of covariates measured after the start of treatment, the interpretation of titrations and composite response measures. Many of these bad practices are becoming enshrined in the regulatory guidance to the pharmaceutical industry. We consider the losses involved in inappropriate measures and suggest that statisticians should pay more attention to this aspect of their work. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: dichotomies; responder analysis; baselines; number needed to treat; titration; ordered categorical data

1. INTRODUCTION

Medical statisticians continue to be very busy with clinical trials, not just in planning and analysing them, but in developing yet better methods to do so. A quarter of a century ago, before the launch of *Statistics in Medicine*, there was no specialist journal for medical statisticians, the nearest to such a thing being, perhaps, *Biometrics*. Now we not only have *Statistics in Medicine* (since 1982), but also have the *Journal of Biopharmaceutical Statistics* (1991), *Statistical Methods in Medical Research* (1992), *Biostatistics* (2000), *Pharmaceutical Statistics* (2002) and *Statistics in Biopharmaceutical Research* (2008). Many papers in these journals are concerned at least partially

*Correspondence to: Stephen Senn, Department of Statistics, University of Glasgow, Glasgow G12 9LL, U.K.
†E-mail: stephen@stats.gla.ac.uk

Contract/grant sponsor: Engineering and Physical Research Council

Received 24 September 2008
Accepted 19 March 2009

with methodological issues to do with clinical trials. On the other hand, our impression is that measurement in clinical trials is a relatively neglected topic. It seems to be a general principle that the physician decides what is clinically relevant and measures it and the statistician analyses the measures without questioning their relevance.

There are some noticeable exceptions such as in quality of life or in areas where comparatively new technologies are coming to the fore such as in imaging or surrogacy [1] with [2] discussing how to assess new endpoints objectively. In endpoint validation statisticians have made, and continue to make, important contributions but it could be debated whether these contributions have been proactive or reactive.

An important exception among statisticians has been David Hand. He pointed out some years ago that measurement is not necessarily a simple matter [3]. He has now published an important monograph on the subject [4]. Yet many other statisticians, even when being innovative, ingenious and insightful as regards the analyses that they bring to bear on particular problems have been uncritical as regards the measures they deal with (we give an example in Section 2.13 below).

In this paper we seek to open up the debate on measurement. Our objective is to encourage the medical statisticians to be more critical as regards this and to encourage them when undertaking collaborative research to contribute actively to the issue of measurement.

2. SOME EXAMPLES OF UNFORTUNATE PRACTICES IN MEASUREMENT IN CLINICAL TRIALS

The list here is by no means exhaustive, but it suffices to show that there are some problems with the current practice.

2.1. Use of baselines to construct change scores

For many indications the main outcome variable, Y, is something that can be measured also at baseline, X. A common practice is to use as the main outcome variable difference from baseline, $D=Y-X$, where this is sometimes referred to as a *change score*. A generally more efficient approach, however, is to use analysis of covariance (ANCOVA) [5].

As is well known, a naïve t-test-type analysis just on the change score will have increased variability compared with the raw outcome alone if the correlation coefficient between the baseline and outcome is less than 0.5 [6]. This point is dealt with in more detail in Section 2.3. Clinical relevance is often given as a justification for using change from baseline, but this can easily be challenged. If a series of randomized clinical trials were carried out and each one analysed in one of three ways using raw outcomes alone; change score or ANCOVA, then from trial-to-trial the relative position of these three measures would differ. However, these differences would be purely random and the three meta-analyses of all the trials we could then carry out would be simply measuring the same thing [7].

It should be noted that provided the baseline X is used as a covariate it makes no difference whether D or Y is used as the outcome [8]: a fact widely recognized and acknowledged in regulatory guidelines [9]. Indeed using D but with X as a covariate is an approach commonly applied by many medical statisticians to reassure their medical colleagues while continuing to apply their preferred analysis.

DOI: 10.1002/sim

2.2. *Percentage change from baseline*

Percentage change from baseline is a popular choice of measure and is calculated as

$$100\frac{Y-X}{X}=100\left(\frac{Y}{X}-1\right) \tag{1}$$

The working part of this is simply

$$\frac{Y}{X}$$

This measure compounds the error of inefficient use of baseline with unfortunate scale of measurement. Even if X and Y are normally distributed, their ratio will not be and it will only be approximately normal if the means are large compared with the standard deviations.

Ratios are not good candidates for parametric analysis. However, when data are anticipated to take a log-normal form then ratios would be the focus of interest. For example, for pharmacokinetic assessments for bio-equivalence data are usually assumed to be log-normal [10, 11]. The observed data in this instance are usually log transformed for analysis. A log transformation reduces the problem to the same as that in Section 2.1 [7]. Of course, it can be useful to back-transform point estimates and confidence intervals to aid interpretation.

2.3. *Crude corrections in general*

The above are special cases of the more general phenomenon of crude correction of outcomes using covariates. Such crudely corrected outcomes have several drawbacks, in particular if the measures are subsequently used in regression through their strong assumption that the intercept is at zero [12].

A notorious example is given by the electrocardiogram (ECG) illustrated in Figure 1. It is common to identify various key points on the ECG trace. The interval between one heart beat and the next is conventionally measured as the difference between subsequent points R on the trace. Thus, the reciprocal of the RR interval is the heart rate. The QT interval is the time between the start of the Q wave to the end of the R wave (as highlighted in Figure 1). If the QT interval (conventionally measured in milliseconds) increases, this can be an indicator of problems with the heart. However, if the increase is only because the RR interval has increased, then such QT prolongation is usually unimportant. Hence, it seems appropriate to correct the QT interval using

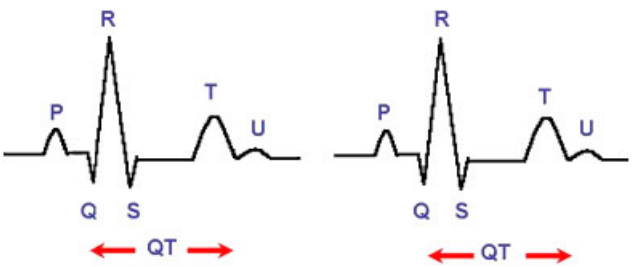

Figure 1. An illustration of QT and R intervals.

DOI: 10.1002/sim

the RR interval in some way. Such corrected QT values are labelled as QT_c and two approaches are commonly employed: Bazett's correction [13] and Fridiricia's correction [14] given below

$$QT_c = \frac{QT}{\sqrt{RR}} \text{ Bazzet's}$$
$$QT_c = \frac{QT}{\sqrt[3]{RR}} \text{ Fridericia's} \tag{2}$$

Note that the RR interval is usually close to 1 s, so that the usual impact of these formulae is to leave the general magnitude of the QT interval apparently largely unchanged. Of course, it will affect the ranking of a set of values and if it did not do this there would be hardly any point. This *apparent* lack of change of scale gives the illusion that the correction has not changed the units of QT. In fact this is not the case—unless there is an implied regression coefficient that just happens to be 1 and whose units are not mentioned in the correction. Otherwise, since both RR and QT are measured in terms of time, then dimensional analysis shows that the units of Bazzett's are in fact the square root of time and the units of Fridericia's the square of the cube root of time. This is bizarre, to say the least. In fact, this is one area at least where there is a growing realization that something is wrong.

In a review of the correction formulae commonly employed, Malik compared 31 such measures on a data set of 1402 patients on a variety of treatments [15]. Of these 31 approaches applied to the 318 patients on beta-blockers, three indicated shortening of the QT interval and 28 indicated prolongation, many of them significantly so. However, when using regression analysis, Malik found that there was no significant effect of beta-blockers on the QT interval. This regression approach is surely the way in which the problem ought to be handled. In our opinion what is really necessary to study is the effect of treatment on the joint distribution of RR and QT. In view of the fact that RR can be commonly influenced without too much harm, and in any case in view of the fact that the basic length of the cardiac cycle is statistically more fundamental than the subsections that make it up, an appropriate factorization of the joint distribution is in terms of the marginal distribution of RR and the conditional distribution of QT given RR—analogous to the use of ante-dependence models in repeated measures [16, 17]. Thus, we can study the effect of drugs on RR and then via some form of regression analysis the effect on QT is not predicted by the effect on RR.

Note however that it may be necessary to study other parameters than just means and conditional means to capture the full effect of drugs.

The ICH guideline on QT_c [18] recognizes the issues with different correction factors. It also discusses dichotomization. As well as suggesting that we should note the proportion of QT_c values greater 500 ms, it suggests cutoffs of about 450, 480 and 450 as well as investigating changes from baseline of >30 and >60 ms saying 'Multiple analyses using different limits are a reasonable approach'. The cutoffs are also consistent with CPMP [19]. We will discuss the issues with cutoffs later in the paper.

2.4. Correction for post-randomization covariates

There is in any case a danger in correcting a measure such as QT by a measure such as RR that is itself affected by treatment. The danger is that we remove some of the treatment effect. This is why a conditional measure such as QT_c, however the correction is done, cannot be safely interpreted unless the effect of the treatment on the correcting factor itself is studied first. Unfortunately, there are a number of examples in medicine where this principle is not observed and an outcome

DOI: 10.1002/sim

measure is corrected by another and then naively nominated as the main outcome variable. The correcting variable being forgotten. Here are two common examples we know of.

The first concerns provocation tests in asthma, whereby broncho-constriction is provoked by exposure to exercise, allergens, cold air, methacholine or histamine as the case may be. It is common to calculate the percentage drop from the value just before provocation. This is then either directly used as an outcome measure or used as a guide to titrate the provocation—the dose of which is then used as the outcome. What is being calculated is a measurement of the form

$$100\left(\frac{Y_1 - Y_2}{Y_1}\right) \tag{3}$$

This is superficially similar to the percentage change from the baseline measurements criticized previously and inherits, of course, all of its deficiencies but has further more serious problems [20, 21]. The difference is that unlike X, Y_1 is measured after treatment, albeit before provocation. The mistake that has been made here is not to rethink a standard medical test. The provocation test as conventionally run is an adequate means of distinguishing *patients* according to the severity of asthma in the absence of treatment; it is not, under the standard protocol, unless the measurement procedure is changed, an adequate means of comparing *treatments*.

The second example concerns response measures in Herpes Zoster. Richard Kay has drawn attention to the fact that a common measure is the time to cessation of pain after some intermediate stage such as say time to disappearance of rash [22]. Such a measure has the unfortunate property that, other things being equal, the longer it takes for the rash to disappear, the better the treatment will appear.

The fact is however that any variable measured after randomization is potentially a response to treatment and thus any correction using post-randomization covariates should be done with caution.

2.5. The effect of titrations on other measures

The effect of titrations is to destroy the meaning of other measurements [7, 20, 21]. Unfortunately, this fact is not always appreciated. Thus, examples can be found where patients are assessed in terms of the time exercising on a treadmill until the onset of the symptoms of angina but are also measured as regards other outcomes. Such other outcomes cannot meaningfully be used to compare treatments. As regards the symptoms of angina, the trial proceeds to this foregone conclusion and it is only the amount of exercise it takes to reach this conclusion that has meaning for the purpose of comparing treatments.

Similar problems arise if titrations are repeated. This is commonly done in trials in asthma. For example, Higham *et al.* [23] describe a five period cross-over trial comparing two doses of salbutamol; two doses of salmeterol and placebo using the dose of methacholine required to induce a 20 per cent drop in FEV_1 ($PD_{20}FEV_1$) as an outcome measure. $PD_{20}FEV_1$ was measured before the administration of each treatment but also 30 min and 120 min after. But it is quite plausible, and indeed was observed to be the case, that the amount of methacholine required to produce a 20 per cent drop in FEV_1 will be greater 30 min after the administration of a more effective treatment than it will be when given the same time after less effective treatment. In what sense, therefore can the comparison of the treatments at 120 min be meaningful? After all, nobody would agree to a procedure in which patients, having been randomized to the particular treatment that they were to receive, were then given different doses of methacholine according to which treatment group they

DOI: 10.1002/sim

were in [7, 20–22]. The fact that this has occurred anyway as part of the measurement process was overlooked by these investigators and is regularly overlooked in this field.

2.6. Ordinal data treated as categorical

Despite the fact that it is now more than 25 years since Peter McCullagh's important read paper on modelling ordered categorical data [24], many investigators still treat ordered categories as if they were unordered rather than using, say, the proportional odds model. Consider a scale with four outcomes used in a clinical trial with two arms. If these categories are analysed as if they were unordered using, for example, a two-by- four contingency table or a log-linear model, the effect of treatment is judged using three degrees of freedom. However, the resulting test has low power for plausible treatment effects in order to cover some very implausible ones. Consider the case where the 'response' is one of 'poor', 'moderate', 'good' or 'excellent' and the effect of treatment is generally to move patients from the lower categories to the higher one. There are $4!=24$ possible permutations of the columns of the two-by-two table, each of which produces the same value of the chi-square statistic. Of course one of these is the one that corresponds to a complete reversal of the pattern and hence is relevant to the extent that one is open to the possibility that the treatment is worse than the control, and therefore wishes to calculate a two-tailed test. The other patterns are both implausible *a priori* and not indicative of a useful treatment and it seems foolish to power a test to detect them at the cost of power for detecting both more plausible and useful patterns.

Clearly, the dependence of the proportional odds model on the assumption of proportionality can be overstressed. Suppose that two different statisticians would cut the same three-point scale at different cut points. It is hard to see how anybody who could accept either dichotomy could object to the compromise answer produced by the proportional odds model.

2.7. Cutoffs and equivalence and non-inferiority trials with ordinal data

It is worth noting that the concentration for binary cutoffs here and throughout the paper is in the context of superiority trials. For non-inferiority and equivalence trials the issues are more complicated. The reason for this is that although the data, as collected, are frequently ordinal in form, in many ways for non-inferiority and equivalence trials this is the wrong scale on which to base inference. The rationale for this is due to the objective of the trial.

For a superiority trial the objective is to assess whether two populations differ. This assessment is conventionally done through a P-value. When analysing ordinal data, there are a number of ways of determining this P-value. For the proportional odds model to be appropriate, the odds-ratio for each cumulative 2×2 should be equal across all k categories, i.e. $\mathrm{OR}_1=\mathrm{OR}_2=\mathrm{OR}_3=\cdots=\mathrm{OR}_k$. In practice the individual observed ORs will deviate randomly from the overall OR. However, given random deviation only, the overall estimate, and inference, will be valid.

For non-inferiority, and equivalence, trials we wish to determine whether two populations do not differ. This assessment is primarily done through a confidence interval where, for a non-inferiority trial, we wish to determine whether the lower bound is greater than some pre-specified non-inferiority margin [10, 25, 26]. This is operationally the same as doing a one-sided test. However, it is the determination and interpretation of this non-inferiority margin, which is the issue here.

A non-inferiority margin is usually expressed in terms of an irrelevant difference. This is an effect below which, if we can prove the difference between investigative treatment and control is no worse than, we can demonstrate non-inferiority. For a non-inferiority trial both the investigative and control treatments are usually active such that the objective is to show that one treatment is

DOI: 10.1002/sim

no worse than the other. In superiority trials clinically relevant differences are discussed. Here, the effect is the difference of interest between an investigative treatment and control (usually placebo) above which we can say that the difference is a genuine one.

For superiority trials Julious *et al.* [27] have discussed how pre-specified cutoffs could be used to determine a treatment effect for designing a superiority trial. Extending these arguments you can use the cutoffs to determine non-inferiority limits for ordinal data. This is when the crux of the problem is encountered. This is because for a non-inferiority trial the clinically irrelevant difference has a part in the inference, whereas the analogous clinically relevant difference only affects the planning of a superiority trial. It is unlikely that the clinically relevant difference will be constant on the log-odds scale. For example, decision analytic considerations might suggest a constant margin on the probability scale and in practice varying margins are used. These will not translate into constant differences on the log-odds scale. For example, suppose we wished to design a trial where two health-related quality of life questionnaires, the hospital anxiety and depression scale (HADS) [28] and the Rotterdam symptom checklist (RSCL) [29] would be considered as the primary endpoints. For HADS a score of 0–7 would indicate that a patient was assessed as 'normal', whereas for the RSCL a score of 0–10 would indicate that a patient was a 'non-case'. For both scales there would be no point demonstrating non-inferiority with an overall assessment of the odds-ratio if it cannot be proven at the clinically meaningful cutoffs.

To resolve such a problem, obviously we could do some form of a step-down procedure. First test the overall odds-ratio and if non-inferior test the odds-ratio at a cutoff for non-inferiority. However, such an approach would be driven by the least efficient comparison, i.e. the one on the dichotomous cutoff.

Note what we are saying here for non-inferiority and equivalence trials does not hold for responses which could be treated as continuous which also sometimes have binary cutoffs applied, as we will discuss later.

2.8. Grouped predictors

As explained in the introduction, on the whole we are concerned here with defects in measuring *outcomes*. However, a very common fault is also to take a continuous predictor and divide into a number of groups. Quintile groups, or fifths, as defined by the four quintiles are very popular in epidemiology, for example. The misleading argument that is often given is that we are unwilling to make the linearity assumption that using the predictor in an ANCOVA requires. However, this is clearly a red-herring since remaining with a continuous predictor would not commit us to using a simple straight-line model. Other alternatives exist such as using higher-order polynomials, fractional polynomials [30] or splines. If the stratification is in quintile groups, superior uses of the four degrees of freedom surely exist.

To highlight some of our points, take the example where patients are divided into two groups according to whether their age is greater than or equal to 65 or not. The following two estimation procedures are equivalent:

1. Estimate the predicted value for each age group separately as the average value of the response Y for that age group.
2. Replace each patient's real age by the average age for the group to which that patient belongs. Call this *group mean age*. Regress the real response on group mean age and estimate the predicted response.

DOI: 10.1002/sim

This shows that such division into two groups can also be described as a straight-line fit, albeit a very crude one. In almost all circumstances we would expect the use of the original ages to be superior. The exception might be if a genuine dramatic change point were expected, as say with a trial containing pre- and post- menopausal women [31, 32].

Good reviews of various approaches for creating suitable prognostic models are given by Harrell *et al.* [33] and Royston *et al.* [34] and also by Frank Harrell in his book [35]. For a critical examination of the habit of dichotomizing predictors see Royston *et al.* [34].

2.9. The losses in creating dichotomies

In Section 2.10 we will discuss issues with dichotomization of ordinal responses. Here we discuss the effect dichotomization has on data of a plausibly normal form. As is well known, the Pitman efficiency of the sign test compared with the t-test is [36]

$$\frac{2}{\pi} \approx 64 \text{ per cent}$$

However, this is an optimistic analogy to the losses attendant in practice on dichotomizing when running parallel group trials, because it would only be appropriate if a median split were employed. However, nearly all dichotomous measures are defined in advance of the results being seen and many on the basis of extreme values seen in patient populations. This reduces the efficiency considerably beyond that of the sign test compared with the t-test.

Suppose we have a continuous outcome approximately normally distributed and suppose that the control group mean is μ and without further loss of generality that the variance is 1. Under the null hypothesis, the variance must be the same in the treatment group. Suppose that the effect of treatment is to bring about a small perturbation, $\Delta(\mu)$, in the value of μ and that we have the standard allocation of equal numbers of patients, n, to each arm of the trial. The number of patients required to run the trial to a satisfactory precision will then be proportional to

$$1/[\Delta(\mu)]^2 \tag{4}$$

Now suppose that we have dichotomized at some value k and proceed to use $\theta = \Phi(k-\mu)$, where $\Phi()$ is the cumulative density function of the standard normal, as the basis for characterizing 'response' in the control group. The control group proportion is estimated with a variance proportional to

$$\theta(1-\theta) = \Phi(k-\mu)[1-\Phi(k-\mu)]$$

On the other hand, the perturbation $\Delta(\theta)$ of the value of θ in the control group will be given approximately by

$$\frac{\partial \Phi(k-\mu)}{\partial \mu}\Delta(\mu) = -\phi(k-\mu)\Delta(\mu)$$

where $\phi()$ is the probability density of the standard normal. The number of patients required to run the trial to satisfactory precision is now proportional to

$$\frac{\Phi(k-\mu)[1-\Phi(k-\mu)]}{[\phi(k-\mu)]^2[\Delta(\mu)]^2} \tag{5}$$

DOI: 10.1002/sim

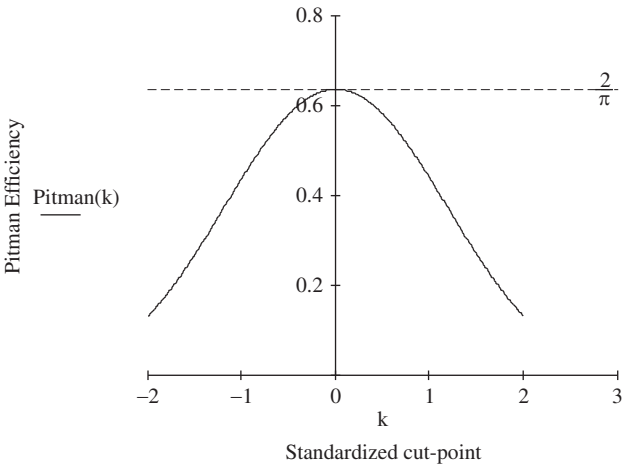

Figure 2. Pitman asymptotic relative efficiency as a function of the standardized cut-point.

Noting that without loss of generality we can set $\mu=0$, then the ratio of the term previously found to this one is

$$\frac{[\phi(k)]^2}{\Phi(k)[1-\Phi(k)]} \tag{6}$$

Of course, when $k=0$, $\theta=0$, $\Phi(k)=1-\Phi(k)=0.5$, $\phi(k)=1/\sqrt{2\pi}$ and we have the value of the Pitman efficiency of $2/\pi$ previously noted. However, this is the maximum value attained and Figure 2 below shows a plot of relative efficiency as a function of k.

Note that an approach that is sometimes used is to classify patients using the dichotomy normal/abnormal, where abnormality is often defined in terms of a number of standard deviations from the mean. The value of two for k is popular. The relative efficiency where this is the case is 13 per cent.

Dichotomies can be used when designing trials as discussed earlier to quantify a treatment effect [27]. In addition even when kept on a continuous or ordinal scaled you can still interpret the data in terms of a response that you would estimate for a binary outcome such as an odds-ratio. The issue is when they are used for the analysis also [24, 37].

2.10. Continuous scales, ordered categories and dichotomies

Of course, the greatest losses are when continuous scales are dichotomized. Ordered categorical scales are also frequently dichotomized and continuous scales are sometimes turned into ordered categories. Both these involve losses of information [38, 39]. Indeed a common practice is to use cut-points on an ordinal scale to categorize patients into groups for design and analysis [27, 40]. This, of course, either results in an inflation of the sample sizes to maintain equivalent precision or to a loss of power.

Assuming that the overall mean response for each category is the same for different numbers of categories, Table I can be constructed, which shows the sample size inflation required to maintain

DOI: 10.1002/sim

Table I. Correction factor for the sample size to be used when the number of categories is less than or equal to 5.

Number of categories	Mean proportions anticipated	Correction factor
2	$\bar{p}_1=\bar{p}_2=0.5$	1.333
3	$\bar{p}_1=\bar{p}_2=\bar{p}_3=0.333$	1.125
4	$\bar{p}_1=\bar{p}_2=\bar{p}_3=\bar{p}_4=0.25$	1.067
5	$\bar{p}_1=\bar{p}_2=\bar{p}_3=\bar{p}_4=\bar{p}_5=0.2$	1.042

equivalent power to a continuous scale [38]. Cochran and Hopkins produced a similar table but written in terms of loss of power [36].

From Table I we can see that for the optimum mean responses we would anticipate a 33 per cent increase compared with a continuous response as opposed to just 5 per cent for 5 categories [38]. Thus, what these results show is that dichotomizing could lead to a serious inflation of the sample size.

Such increases in the sample size are not trivial. If you had a trial where each patient cost £20000 every increase of 100 in the sample size would add £2000000 to the trial's cost.

It should be noted that though this increase in the sample size could be considered on the optimistic side. We are comparing here the increase in sample size compared with the situation where a t-test would be used in the primary analysis. The situation would be even worse if an ANCOVA might be an option as will now be discussed.

2.11. The further losses in using change scores to create dichotomies

The point about losses due to dichotomization will now be further expanded to include the issue of the use of the baselines. Provided that change scores are not dichotomized, it is always possible to recover the information in the baselines in an ANCOVA. This, as has already been noted, is then equivalent to analysing the raw outcomes in an ANCOVA [5, 7, 8]. However, if dichotomies are used the full recovery is not possible. In fact, in practice when dichotomies are used, the baseline is *not* fitted as a covariate. There is thus a double loss involved.

Suppose that the baseline and outcome have approximately the same variance, σ^2, and correlation, ρ. Then the variance of the change score estimator $D=Y-X$ is

$$2(1-\rho)\sigma^2 \tag{7}$$

and this, as is well known, is greater than the variance of the raw outcome Y if $\rho<\frac{1}{2}$ [5, 7]. A further problem with the score, however, is that it is not independent of the baseline value, having a covariance $\sigma^2(1-\rho)$. In fact, the ANCOVA adjusted score can be constructed from a general estimator of the form $Y-\beta X$ by finding either the value of β that minimizes the variance of the score, which variance in general must have the form

$$\sigma^2(\beta^2-2\beta\rho+1)=\sigma^2[(\beta-\rho)^2+1-\rho^2] \tag{8}$$

or the value that makes the covariance with X,

$$\sigma^2(\rho-\beta) \tag{9}$$

DOI: 10.1002/sim

zero. By inspection, for either of these two approaches, the value is obviously $\beta=\rho$ and by substituting this value in (9) the residual variance is then

$$\sigma^2(1-\rho^2) \tag{10}$$

Thus, the relative efficiency of the change score compared with ANCOVA is the ratio of (10) to (7)

$$\frac{1-\rho^2}{2(1-p)}=\frac{1+\rho}{2} \tag{11}$$

For example, if the correlation between the baseline and outcome is about 0.7, then the relative efficiency is 85 per cent. If the cut-point has been chosen at $k=1$, the further loss in dichotomization is an efficiency penalty of 44 per cent and the combined effect is to produce an efficiency of 37 per cent.

This overstates slightly the advantage of ANCOVA, since the analysis so far has assumed that nuisance parameters are known. In practice, they are not and will have to be estimated and also in practice the covariate values will not be balanced between treatment groups. There is thus some loss or efficiency due to non-orthogonality compared with using the change score [7]. However, the expected loss is equivalent to one patient and hence not particularly important [41, 42].

2.12. Post hoc dichotomization

Of course, a further problem arises if the cut-point for defining response is not predefined [43]. A little exercise one of us (SAJ) does when teaching is to ask students to simulate data from two uniform distributions and then at every possible cutoff on the continuous scale to dichotomize the data and do a chi-squared test. We tell them to choose the cutoff for presentation to be the one with the smallest P-value with a sample size of 100 in each of two groups and data sampled from the random uniform distribution of 0–100. This approach would falsely gives a positive result on an average 50 per cent of the time.

2.13. Inappropriate definitions of response

The sort of dichotomy discussed in Section 2.9 is based on a single cut-point and a continuous measure with no further side conditions. In fact, far more inappropriate definitions of response can be found. Consider, for example, this definition from [44], of which the most important word is the first.

> 'Arbitrarily, response criteria for antihypertensive therapy include the percentage of patients with a normalization of blood pressure (reduction SBP <140 mmHg and DBP <90 mmHg) and/or reduction of SBP $\geqslant 20$ mmHg and/or DBP $\geqslant 10$ mmHg. Results obtained should be discussed in terms of statistical significance and in relation to their clinical relevance'.

The operation jointly of systolic blood pressure (SBP) and diastolic blood pressure (DBP) to define the response here is slightly obscure, but it is clear that it makes it much easier for a patient to respond if just hypertensive than not. Consider the following definition in [45].

> 'In accordance with the guidelines for the clinical evaluation of antihypertensive agents, patients with a blood pressure level greater than or equal to $\frac{160}{95}$ mmHg after receiving a placebo for 4 weeks were enrolled in the study. The test drug was then administered for

DOI: 10.1002/sim

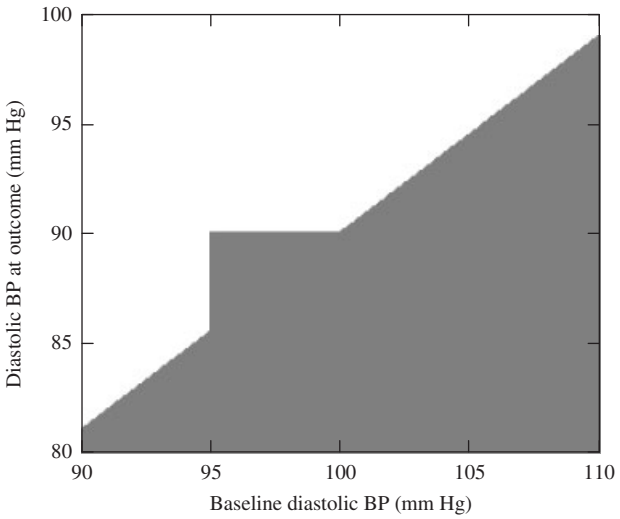

Figure 3. Response region as defined as a function of diastolic blood pressure at baseline and outcome from Goetghebeur *et al.* [46].

12 weeks. A decrease in blood pressure by $\frac{20}{10}$ mmHg from the pretreatment level or a blood pressure level below $\frac{149}{89}$ mmHg was considered to indicate a response'.

Again, the fact that response is defined in terms of a joint requirement for DBP and SBP makes a discussion of the behaviour of this rule rather complicated; however, it suffices to show that the requirement in terms of DBP alone is extremely bizarre to illustrate that this definition is most unsatisfactory. Consider, therefore, the simpler definition found in Goetghebeur *et al.* [46]

'The treatment is called successful if either the patient has gone down from a baseline diastolic blood pressure of $\geqslant$95 mmHg$\leqslant$90 mmHg or has achieved a 10 per cent reduction in blood pressure from baseline'.

Figure 3 shows the response region plotted in terms of DBP at baseline and DBP at outcome for such a trial. A 'responder' is any patient whose baseline and outcome DBP place him or her either below and to the right of the solid line or to the right of the vertical and below the horizontal dotted lines.

The indicator functions for the response is

$$I(X,Y)=(Y<0.9X)\cup[(X\geqslant 95)\cap(Y\leqslant 90)] \tag{12}$$

and if X, Y are jointly distributed as a normal distribution with parameters μ_X, μ_Y, σ_X^2, σ_Y^2, ρ, then the conditional distribution of Y given X, $f(Y|X=x)$ is normal with mean

$$\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$$

DOI: 10.1002/sim

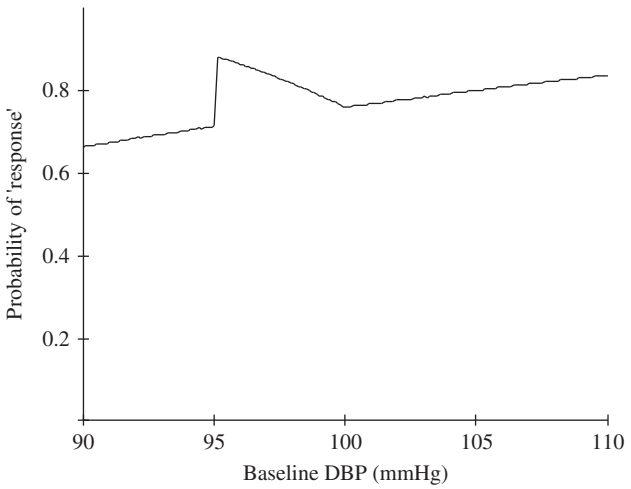

Figure 4. Probability of 'response' as a function of baseline DBP.

and variance $\sigma_Y^2(1-\rho^2)$. The probability of response, R, as a function of X can then be obtained as

$$E[I(X,Y)|X=x]=\int_{-\infty}^{\infty} f(Y|X=x)I(x,Y)\,\mathrm{d}y \tag{13}$$

Figure 4 shows the probability of a patient 'responding', as given by (13) as a function of baseline blood pressure if the values of parameters of the normal distribution are

$$\sigma_X=\sigma_y=10\,\mathrm{mmHg},\quad \mu_X=100\,\mathrm{mmHg},\quad \mu_Y=90\,\mathrm{mmHg},\quad \rho=0.7$$

An example of the issues in defining a response through some arbitrary cut-point is highlighted through the simple worked example with data taken from the DASH study [47]. The analysis presented is all our own and was not the way the study was originally analysed.

The part of the study was a cross-over trial to assess the effect of three levels of intake of sodium (low, intermediate and high) on blood pressure. For this analysis we arbitrarily defined a responder as being someone who had a 10 per cent reduction in their blood pressure. Figure 5 plots the percentage of baseline at outcome against baseline for subjects on the low sodium diet (subjects anticipated to have the greatest response). The subjects with stars are those who are defined as responders at the low intake. The filled circles are subjects who were defined as responders on the other intakes but not on the lowest intake.

This figure demonstrates the arbitrary nature of cut-points, as most of the subjects who were defined as responders on other intakes but not on the lowest are still mainly clustered around the cutoffs.

3. INDIVIDUAL RESPONSE TO TREATMENT

In this section we concentrate particularly on issues surrounding attempts to identify what statisticians might refer to as patient by treatment interaction. A very good discussion of the difficulties

DOI: 10.1002/sim

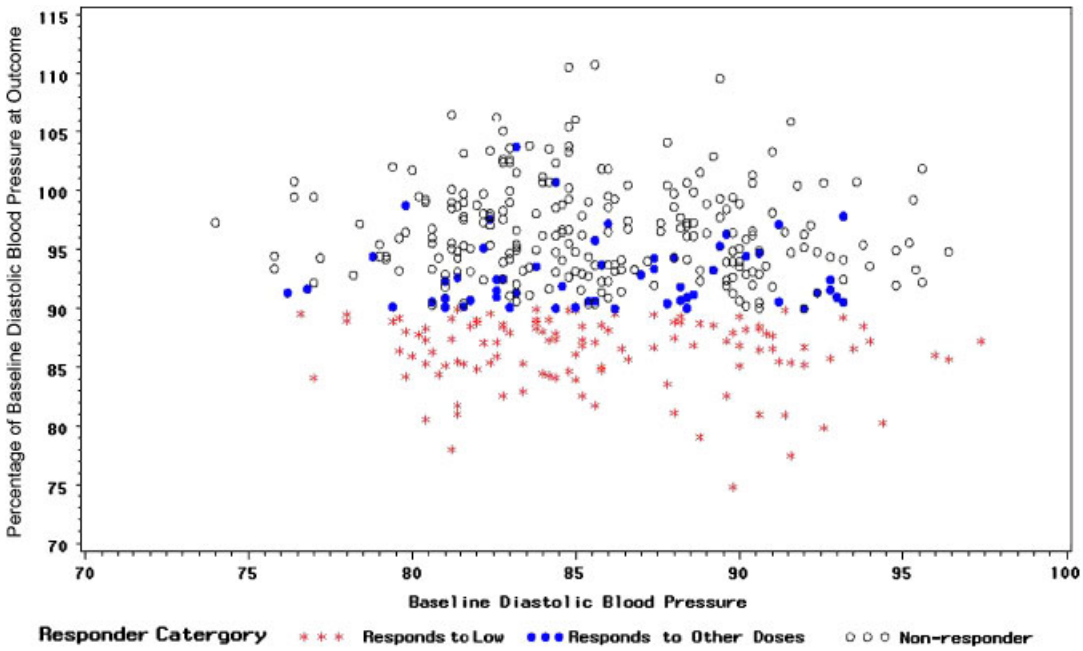

Figure 5. Percentage of baseline diastolic blood at outcome pressure against baseline diastolic blood pressure (mmHg) for the lowest intake of sodium.

in doing this was given in a read paper on quality of life several years ago [48]. Unfortunately, too many medical researchers and even some statisticians seem to be unaware of the problem of identifiability outlined in that paper.

3.1. Fallacies of responder analysis

A common fallacy concerns the so-called responder analysis, a thing that is almost *de rigueur* now in regulatory submissions. For example, Kieser *et al.* [49], who were inspired to write a paper describing a realistic approach to sample size determination in view of the common requirements to analyse responders in clinical trials, referred to such requirements in European guidelines on obesity, diabetes mellitus, Parkinson's disease, depression, schizophrenia and Alzheimer's disease [50–55]. In fact, there are many other indications with such requirements. Frequently associated with such definitions are extravagant causal interpretations of 'response' which we will now discuss in context with cross-over trials.

3.1.1. Issues with responder analyses in context with cross-over trials. A suggestion of Guyatt *et al.* [56] to define a responder is that you calculate the proportion of patients who respond better under one treatment than another in order to calculate the number needed to treat (NNT), a way of reporting binary outcomes that will be picked up in Section 3.2. They propose that in a cross-over trial you might calculate the difference between the observed outcome under treatment and the observed outcome under control and compare this with some clinically relevant difference to see whether the patient has had a superior response. This ignores the fact that because of within-patient

DOI: 10.1002/sim

variability it is possible for a treatment to have identical long-term effects on a group of patients but apparently yield quite different results from patient to patient in the short term [57–60].

Consider the case where in fact we carry out two cross-over trials using exactly the same patients for the second trial that we used for the first. A consequence would be that we have managed to compare two treatments twice for each patient. We can then attempt a classification of the bivariate distribution of 'responses', whereby such 'responses' are determined using the method of Guyatt *et al.* [56]. Suppose we have 24 patients and the results are as given in Table II below. This would, indeed, then support the sort of conclusion that Guyatt *et al.* propose that their method justifies.

This is not, however, what their method produces, since it is based on a single cross over. The situation would then be as given in Table III, which shows only one of the margins from Table II and not the joint distribution over two cross-over trials.

To assert that the position would be as it is in Table II if only the further data to produce it had been collected requires an untestable (since the data have not been collected) and unreasonable assumption since the joint distribution that would be produced by studying these patients could equally well be that given in Table IV, which is the condition that corresponds to complete independence [60]. Thus for this pattern, it makes no sense to talk about 'responders' and

Table II. Joint pattern of responses in two cross-over trials using the same patients when perfect correlation obtains.

		Second cross over		
		Responders	Non-responders	Total
First cross over	Responders	24	0	24
	Non-responders	0	8	8
	Total	24	8	32

Table III. Pattern of responses in a single cross over.

		Total
First cross over	Responders	24
	Non-responders	8
	Total	32

Table IV. Joint pattern of responses in two cross-over trials using the same patients when independence obtains.

		Second cross over		
		Responders	Non-responders	Total
First cross over	Responders	18	6	24
	Non-responders	6	2	8
	Total	24	8	32

DOI: 10.1002/sim

'non-responders'. In fact it would be better if this terminology, which carries with it a considerable causal freight it cannot support, were abandoned.

3.2. Number needed to treat (NNT)

The NNT is defined as the reciprocal of the risk reduction in treating patients with one treatment rather than another [61]. It is thus the reciprocal of the risk difference and was originally proposed by Laupacis *et al.* [62]. We have no objection to the NNT as a *concept* when it is applied to genuine dichotomies as opposed to dichotomized measures. It is not, however, even in the case of genuine binary outcomes a useful way to summarize the results of a trial [63–66], since the population in a clinical trial cannot be regarded as a random sample of any target population and will almost certainly differ in terms of background risk from such a population. This is important, as to interpret an NNT you must also know what the risks are in each group or at least the baseline risk [67].

The trouble with the NNT is that it is highly unlikely to be additive, thus even if we wish to calculate an NNT for a particular group of patients in order to determine the value of treatment for them, it would be best to use a measure that is more likely to be additive, such as the log-odds-ratio, as a starting point for a prediction [61]. Whatever your view, and we personally find Hutton and Grieve's structures against this measure as commonly used compelling [63, 64], it is totally unacceptable to create dichotomies purely in order to be able to calculate NNTs.

3.3. Composite response measures

Perhaps because of an understandable concern about multiplicity, there is a tendency for trialists to combine measures to produce a single composite endpoint. This is very commonly the case in survival analysis, where the patient is at risk from a number of possible events and it may be decided to measure the time until the first of these occurs. For example, in the CAPRICORN study [68], which compared carvedilol to placebo in patients with a confirmed myocardial infarction and left ventricular dysfunction, 'the primary endpoint was all-cause mortality or hospital admission for cardiovascular problems'. There was no significant difference ($P=0.30$) between the groups as regards this measure and the estimated hazard ratio was 0.92 with a 95 per cent confidence interval of 0.8–1.07. On the other hand the more serious endpoint of all-cause mortality was significantly lower in the carvedilol group ($P=0.03$) hazard ratio 0.77, confidence interval 0.60–0.98. The more serious endpoint was originally the only primary endpoint with composite endpoint added as a co-primary when lower than expected mortality was observed from a blinded analysis.

Despite the rather modest evidence of effect on a secondary endpoint, the investigators chose to interpret this as indicating a beneficial treatment effect, a conclusion, however, that other studies supported. For example, in the similar COPERNICUS study, the primary endpoint was all-cause mortality, which was significantly reduced for carvedilol compared with placebo ($P=0.0014$ adjusted for interim analysis) [69] as was the combined risk of death or hospitalization for a cardiovascular reason ($P=0.00002$) [68].

Of course, if a sequential analysis is to be run and some stopping rule is to be agreed upon then since ultimately a stop–go decision is required this implies some sort of trade-off between endpoints possibly involving a composite of the sort used. However, as regards actually evaluating the evidence, such endpoints are far from ideal. If there is no universal agreement as to what should be the primary endpoint (for CAPRICORN and COPERNICUS different choices were made), then it is not clear as to why an arbitrary choice should be useful for others. As more and more evidence

DOI: 10.1002/sim

are obtained for a treatment what is really needed is evidence as to what it does, not evidence as to how well it does what we should like it to do. Consider, for example, the now controversial VIGOR study of rofecoxib compared with naproxen in rheumatoid arthritis [70]. This found a highly significant benefit for rofecoxib compared with naproxen as regards gastrointestinal events (relative risk for rofecoxib:naproxen, 0.5; 95 per cent confidence interval, 0.3–0.6; $P<0.001$) but reported an even more significant benefit for naproxen as regards myocardial infarction (relative risk for rofecoxib: naproxen, 5; 95 per cent confidence interval, 1.4–10).

What a patient and his or her physician need to make an informed choice is information on both of these together with estimates of background absolute risk in order to translate the finding of the trial into action at the individual level. Of course, here there is little danger of forming a composite from gastrointestinal and cardiac events but a similar point applies perhaps *mutatis mutandis* to endpoints that are commonly combined.

For a discussion of approaches to individual decision making see [71–73].

4. DISCUSSION

We have already referred to the regrettable tendency of statisticians to regard measurement as being uniquely the province of the physician. The physician, of course, carries out the measurements but that does not mean that the statistician does not have the right to criticize such procedures. Unfortunately, it is not hard to find examples where statisticians have failed to sufficiently question what is being provided. Consider, for example, the paper we used to provide our example of 'response' measures in hypertension [46]. The paper is a sophisticated contribution to the literature on missing observations but could the actual measure not have been questioned?

Even experienced medical statisticians are apt to make these claims. For example, Lewis in an article entitled, 'In Defence of the Dichotomy' [74] writes

> '...it is not lack of serious consideration that leads to dichotomizing data and the use of responder analysis..... It is a determined attempt to understand if the effect of a drug, shown to be statistically significant on a poorly understood scale of measurement, has any clinical significance'.

But what makes a measure such as that considered for response in hypertension clinically relevant? Can it be right, for this would surely happen, that simply by increasing the precision with which we measure blood pressure or basing it say on the average of three determinations, either of which would reduce the relevant values of σ_X and σ_Y, we should change the probability of response, even though nothing will actually have changed in the way the treatment affects the patients [7]?

Lewis also writes, 'Patients *do* differ in their responses to treatments, whether or not our usual statistical models take account of this fact. This is part of the reason for the current major interest in pharmacogenetics'.

But this is a statement of faith, not fact. For most diseases we simply have not run the sort of repeated period cross-over trials that would let us identify variations in response. Dichotomizing a continuous measure does nothing to address this issue. Even when dichotomized we cannot distinguish between the cases represented by Tables II and IV and in any case there are grounds to believe that some of the expectations for genetic variation in treatment response are not well founded [58]. Also, in our opinion, the desire of doctors to indulge in patient-specific prescribing

DOI: 10.1002/sim

is not as great as often supposed since at the moment there is the technology already to do so and yet this option is usually passed up. This is because one factor which is very well known to be predictive of efficacy is weight (or some other measure of a person's size). When undertaking a pharmacokinetic analysis across doses or any between group comparison of pharmacokinetics weight would be standardly fitted as covariate. This is as you would expect as the amount of drug in the body for a given dose is related to the size of the actual body thus for example, other things being equal, Arnold Schwarzenegger would be expected to have lower drug exposures than Kate Moss for a given dose. In turn the amount of exposure to a given drug is related to the efficacy of that drug. Yet despite this, and with the exception of intravenous administration, weight or body size is not usually allowed for when individually prescribing as the use of a basic pair of scales is seen as making the business of treating an individual patient far too complex.

The commonest argument any statistician encounters while trying to reform the business of measuring effects is that such measurements cannot be changed because it has always been done this way, or thought in this way. All such claims must be rejected as unscientific. Precedence is a justification for lawyers not scientists and it is logic not precedence that has to determine the way we measure. Of course, experience is also relevant, but if the experiment of not dichotomizing is not tried then we shall indeed be stuck with what is common practice whether or not such practice is good.

5. CONCLUSION

The losses involved in choosing inappropriate measures are not negligible. For example, to take the case of dichotomized continuous outcomes, it is no exaggeration to claim that if all such dichotomies in clinical trials were abandoned tomorrow we would not only see an immediate gain in efficiency in carrying out clinical trials but *pace* claims to the contrary, an improvement in interpretability [75]. There are many other practices we have drawn attention to, however, which are even more misleading. In particular correction for post-baseline covariates, although a less common habit that of dichotomizing, has an even greater potential to mislead. It is time that statisticians took the matter of measurement seriously. We believe that there is considerable opportunity for improving the quality of pharmaceutical trials.

ACKNOWLEDGEMENTS

We would like to thank the US National Institute of Health (NIH) for providing us with the data from the DASH study under their data sharing policy. Stephen Senn's work was partially supported by the Engineering and Physical Research Council's Simplicity, Complexity and Modelling (SCAM) project.

REFERENCES

1. FDA. Critical path opportunities report, 2006. Available from: http://www.fda.gov/oc/initiatives/criticalpath/reports/opp_report.pdf (Last accessed 3 August 2008).
2. Julious SA, Zariffa N. ABC of pharmaceutical trial design. *Pharmaceutical Statistics* 2006; **1**:45–54.
3. Hand DJ. Statistics and the theory of measurement. *Journal of the Royal Statistical Society Series A—Statistics in Society* 1996; **159**:445–473.
4. Hand D. *Measurement*: *Theory and Practice*. Arnold: London, 2004.
5. Hills M, Armitage P. The two-period cross-over clinical trial. *British Journal of Clinical Pharmacology* 1996; **8**:7–20.

DOI: 10.1002/sim

6. Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Statistics in Medicine* 1992; **11**:1685–1704.
7. Senn SJ. *Statistical Issues in Drug Development*. Wiley: Chichester, 1997.
8. Laird N. Further comparative analyses of pre-test post-test research designs. *The American Statistician* 1983; **37**:329–330.
9. CPMP. Points to consider for guidance on adjustment for baseline covariates. *CPMP/EWP/2863/99*, 2003.
10. Julious SA. Tutorial in biostatistics—sample sizes for clinical trials with normal data. *Statistics in Medicine* 2004; **23**:1921–1986.
11. Julious SA, Debarnot CA. Why are pharmacokinetic data summarized by arithmetic means? *Journal of Biopharmaceutical Statistics* 2000; **10**:55–71.
12. Kronmal RA. Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society Series A—Statistics in Society* 1993; **156**:379–392.
13. Bazett HC. An analysis of the time-relations of electrocardiograms. *Heart*: *A Journal for the Study of the Circulation* 1920; **7**:353–370.
14. Fridiricia LS. Die Systolendauer im elektrokardiogramm bei normalen menschen und bei herzkranken. *Acta Medica Scandinavica* 1920; **53**:469–486.
15. Malik M. The imprecision in heart rate correction may lead to artificial observations of drug induced QT interval changes. *Pacing and Clinical Electrophysiology* 2002; **25**:209–216.
16. Gabriel KR. The model of ante-dependence for data of biological growth. *Bulletin de l'Institut International de Statistique* (*Paris*) 1961; **39**:253–264.
17. Kenward MG. A method for comparing profiles of repeated measurements. *Applied Statistics—Journal of the Royal Statistical Society Series C* 1987; **36**:296–308.
18. ICH E14. The clinical evaluation of QT/QT_c interval prolongation and proarrhythmic potential for non-antiarryhthmic drugs. *ICH E14*, 2005.
19. CPMP. Points to consider: the assessment of the potential QT interval prolongation by non-cardiovascular medicinal products. *CPMP/986/96*, 1997.
20. Senn SJ. The use of baselines in clinical trials of bronchodilators. *Statistics in Medicine* 1989; **8**:1339–1350.
21. Senn SJ. Statistical issues in short term trials in asthma. *Drug Information Journal* 1993; **27**:779–791.
22. Kay R. Some fundamental statistical concepts in clinical-trials and their application in herpes-zoster. *Antiviral Chemistry and Chemotherapy* 1995; **6**:28–33.
23. Higham MA, Sharara AM, Wilson P, Jenkins RJ, Glendenning GA, Ind PW. Dose equivalence and bronchoprotective effects of salmeterol and salbutamol in asthma. *Thorax* 1997; **52**:975–980.
24. McCullagh P. Regression models for ordinal data. *Journal of the Royal Statistical Society B* 1980; **42**:109–142.
25. CHMP. Guideline on the choice of the non-inferiority margin. *EMEA/CPMP/EWP/2158/99*, 2005.
26. CPMP. Points to consider on switching between superiority and non-inferiority. *CPMP/EWP/482/99*, 2000.
27. Julious SA, George S, Machin D, Stephens RJ. Sample sizes for randomized trials measuring quality of life in cancer patients. *Quality of Life Research* 1997; **6**:109–117.
28. Snaith RP, Zigmond AS. The hospital anxiety and depression scale. *British Medical Journal* (*Clinical Research Edition*) 1986; **292**:344.
29. de Haes JC, van Knippenberg FC, Neijt JP. Measuring psychological and physical distress in cancer patients: structure and application of the Rotterdam symptom checklist. *British Journal of Cancer* 1990; **62**:1034–1038.
30. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates—parsimonious parametric modeling. *Applied Statistics—Journal of the Royal Statistical Society Series C* 1994; **43**:429–467.
31. Julious SA. Inference and estimation in a change point regression problem. *Journal of the Royal Statistical Society Series D—The Statistician* 2001; **50**:51–61.
32. Lees B, Molleson T, Arnett TR, Stevenson JC. Differences in proximal femur bone density over two centuries. *Lancet* 1993; **341**:673–675.
33. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**:361–387.
34. Royston P, Sauerbrei W, Altman DG. Modeling the effects of continuous risk factors. *Journal of Clinical Epidemiology* 2000; **53**:219–220.
35. Harrell FE. *Regression Modeling Strategies*. Springer: New York, 2001.
36. Cochran WG, Hopkins CE. Some classification problems with multivariate qualitative data. *Biometrics* 1961; **171**:10–32.
37. Moser BK, Coombs LP. Odds ratios for a continuous outcome variable without dichotomizing. *Statistics in Medicine* 2008; **23**(12):1843–1860.

DOI: 10.1002/sim

38. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in 2 group comparisons. *British Medical Journal* 1995; **311**:1145–1148.
39. Whitehead J. Sample size calculations for ordered categorical-data. *Statistics in Medicine* 1993; **12**:2257–2271.
40. Julious SA, George S, Campbell MJ. Sample sizes for studies using the short form 36(SF-36). *Journal of Epidemiology and Community Health* 1995; **49**:642–644.
41. Atkinson AC. Optimum biased-coin designs for sequential treatment allocation with covariate information. *Statistics in Medicine* 1999; **18**:1741–1752.
42. Burman C-F. *On Sequential Treatment Allocations in Clinical Trials*. Department of Mathematics, Chalmers University of Technology, 1996.
43. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using optimal cut-points in the evaluation of prognostic factors. *Journal of the National Cancer Institute* 1994; **86**:829–835.
44. CHMP. Note for guidance on clinical investigation of medicinal products in the treatment of hypertension. *CPMP/EWP/3020/3*, 2004.
45. Kuramoto K. Double-blind studies of calcium antagonists in the treatment of hypertension in Japan. *Journal of Cardiovascular Pharmacology* 1989; **13**(Suppl. 1):S29–S35.
46. Goetghebeur E, Molenberghs G, Katz J. Estimating the causal effect of compliance on binary outcome in randomized controlled trials. *Statistics in Medicine* 1998; **17**:341–355.
47. Sacks FM, Svetkey LP, Vollmer WM, Appel LJ, Bray GA, Harsha D, Obarzanek E, Conlin PR, Miller III ER, Simons-Morton DG, Karanja N, Lin PH. Effects on blood pressure of reduced dietary sodium and the dietary approaches to stop hypertension (DASH) diet. DASH—sodium collaborative research group. *The New England Journal of Medicine* 2001; **344**:3–10.
48. Cox DR, Fitzpatrick R, Fletcher AE, Gore SM, Spiegelhalter DJ, Jones DR. Quality-of-life assessment—can we keep it simple. *Journal of the Royal Statistical Society Series A—Statistics in Society* 1992; **155**:353–393.
49. Kieser M, Röhmel J, Friede T. Power and sample size determination when assessing the clinical relevance of trial results by 'responder analyses'. *Statistics in Medicine* 2004; **23**:3287–3305.
50. CPMP. Note for guidance on clinical investigation of drugs used in weight control. *CPMP/EWP/281/96 Rev.1*, 2006.
51. CPMP. Note for guidance on medicinal products in the treatment of Alzheimer's disease. *CPMP/EWP/553/95*, 1997.
52. CPMP. Note for guidance on clinical investigation of medicinal products in the treatment of schizophrenia. *CPMP/EWP/559/95*, 1998.
53. CPMP. Note for guidance on clinical investigation of medicinal products in the treatment of Parkinson's disease. *CPMP/EWP/563/95*, 1998.
54. CPMP. Note for guidance on clinical investigation of medicinal products in the treatment of depression. *CPMP/EWP/518/97, Rev 1*, 2002.
55. CPMP. Note for guidance on clinical investigation of medicinal products in the treatment of diabetes mellitus. *CPMP/EWP/1080/00*, 2002.
56. Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. *British Medical Journal* 1998; **316**:690–693.
57. Senn SJ. Applying results of randomised trials to patients. N of 1 trials are needed. *British Medical Journal* 1998; **317**:537–538.
58. Senn SJ. Individual therapy: new dawn or false dawn. *Drug Information Journal* 2001; **35**:1479–1494.
59. Senn SJ. Author 's reply to Walter and Guyatt. *Drug Information Journal* 2003; **37**:7–10.
60. Senn SJ. Individual response to treatment: is it a valid assumption? *British Medical Journal* 2004; **329**:966–968.
61. Cook RJ, Sackett DL. The number needed to treat—a clinically useful measure of treatment effect. *British Medical Journal* 1995; **310**:452–454.
62. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine* 1988; **318**:1728–1733.
63. 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.
64. Hutton JL. Numbers needed to treat: properties and problems (with comments). *Journal of the Royal Statistical Society A* 2000; **163**:403–419.
65. Senn SJ. Odds ratios revisited. *Evidence-Based Medicine* 1998; **3**:71.
66. Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses—sometimes informative, usually misleading. *British Medical Journal* 1999; **318**:1548–1551.

DOI: 10.1002/sim

67. Julious SA. Issues with number needed to treat. *Statistics in Medicine* 2005; **24**:3233–3235.
68. Dargie HJ. Effect of carvedilol on outcome after myocardial infarction in patients with left-ventricular dysfunction: the CAPRICORN randomised trial. *Lancet* 2001; **357**:1385–1390.
69. Packer M, Coats AJ, Fowler MB, Katus HA, Krum H, Mohacsi P, Rouleau JL, Tendera M, Castaigne A, Roecker EB, Schultz MK, DeMets DL. Effect of carvedilol on survival in severe chronic heart failure. *The New England Journal of Medicine* 2001; **344**:1651–1658.
70. Bombardier C, Laine L, Reicin A, Shapiro D, Burgos-Vargas R, Davis B, Day R, Ferraz MB, Hawkey CJ, Hochberg MC, Kvien TK, Schnitzer TJ. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. *New England Journal of Medicine* 2000; **343**:1520–1528.
71. Ashby D, Baron DN. Harmonizing multiple-choice question marks with essay marks. *Medical Education* 1986; **20**:321–323.
72. Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. *British Medical Journal* 1995; **311**:1356–1359.
73. Hilden J, Habbema JDF. The marriage of clinical-trials and clinical decision. *Statistics in Medicine* 1990; **9**:1243–1257.
74. Lewis JA. In defence of the dichotomy. *Pharmaceutical Statistics* 2004; **3**:77–79.
75. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine* 2006; **25**:127–141.

DOI: 10.1002/sim