# Dichotomania: An Obsessive Compulsive Disorder that is Badly Affecting the Quality of Analysis of Pharmaceutical Trials

Stephen Senn
*University of Glasgow, Department of Statistics*
15 University Gardens,
Glasgow, UK
*stephen@stats.gla.ac.uk*

## 1. Introduction

The main emphasis of pharmaceutical statistics continues to be the design and analysis of clinical trials. Other topics are relatively neglected(Grieve, 2002; Lendrem, 2003; Senn, 2003b). For example, most major pharmaceutical companies will have several dozen statisticians working in development but it will be usual to have only half a dozen, if that, assigned to research and only one or two to manufacturing. There are various reasons for this but the influence of the regulator must not be underestimated. By the 1970s the Food and Drug Administration of America (FDA) was regularly using in-house statisticians in evaluating clinical trials in regulatory submissions and by 1979 had already issued a, ' Statistical Documentation Guide for Protocol Development and NDA Submission'(O'Neill, 1991) . (Here NDA stands for *New Drug Application*.) This was a logical reflection of the fact that statisticians had been amongst the prime movers for developing clinical trial methodology. (One thinks particularly of Bradford Hill.) The FDA's approach helped to promote the employment of statisticians in drug development by the pharmaceutical industry and this in its turn eventually found its reflection in the European regulatory agencies.

The industry also employs a number of leading statisticians to carry out methodological research of relevance to its concerns. However, they too seem to have been mainly involved in developments of relevance to the design and analysis of clinical trials. I am not suggesting that this is not an important field: it is and will continue to remain so. However, there is no doubt in my mind that other areas of activity of the pharmaceutical industry suffer from relative statistical neglect(Grieve, 2002; Lendrem, 2003; Senn, 2003b). These areas include, but are not limited to, in-vitro and pre-clinical drug research, forecasting in connection with marketing, post-marketing surveillance and risk analysis and decision analysis applied to selection and management of the candidate drugs in the development portfolio.

There is no question that considerable improvements to pharmaceutical quality can be made with the help of the development and application of statistical methods in these non-clinical areas. Since the title of the International Statistical Institute (ISI) session in which this paper is delivered is 'Improving Drug Quality,' it might have been logical to cover some of these fields. However, even within the field of design and analysis of clinical trials there is one matter that has been very poorly served by statisticians and which would benefit greatly from statistical input, and that is what this paper will concentrate on.

The matter in question is that of measurement and the reasons for its comparative neglect can only be speculated on. Whatever the reason, there seems to be an unfortunate compartmentalism in clinical research, whether in drug development or elsewhere, whereby it is assumed that the physician will choose the measure and the statistician will analyse it. In this paper I shall explain why this is not good enough. In many diseases a far greater contribution can easily be made by improving measurement than by improving trial design. As David Hand has recently reminded us in his excellent and important book (Hand, 2004), measurement matters.

The plan of the paper is as follows. In the second section I give some general examples of unfortunate practices in measurement. In the third section I consider the particular losses attendant upon turning continuous measures into binary ones. In the final section I consider some general lessons and offer some general conclusions. On the whole I shall concentrate on issues affecting outcome measures, that is to say measures that could act as '$Y$' on the left-hand side of a regression equation, rather than predictor variables, that could act as '$\boldsymbol{X}$' on the right-hand side. However, I shall make some brief remarks regarding predictor variables also.

## 2. Some examples of unfortunate practices in measurement in clinical trials

The list here is by no means exhaustive but suffices to show that there are some problems with 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, $X$ . Examples include pain in rheumatism (measured perhaps on a visual analogue scale), forced expiratory volume in one second ($FEV_1$) in asthma and diastolic blood pressure in hypertension. 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)(Hills & Armitage, 1979). (Provided the baseline *X* is used as a covariate it makes no difference whether *D* or *Y* is used as the outcome(Laird, 1983)) As is well known, the change score will have increased variability compared to the raw outcome alone if the correlation coefficient between baseline and outcome is less than 0.5. (This point is dealt with in more detail in section 3.) Clinical relevance is sometimes used as a justification for this practice, but this is just nonsense. If a series of randomised clinical trials were carried out and each one analysed in one of three ways, using raw outcomes alone, change-score or analysis of covariance, 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(Senn, 1997). Clinical relevance cannot decide the toss between these measures, only statistical considerations such as precision are relevant and these favour ANCOVA.

### 2.2. Percentage change from baseline.

This is also a popular choice of measure and is calculated as

$$100\frac{Y-X}{X} = 100\left(\frac{Y}{X} - 1\right) \quad (2.1).$$

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. Ratios are not good candidates, unless further transformed, for parametric analysis; it is better to log-transform both baseline and outcome. This reduces the problem to the same as that in section 2.1, from which it follows that the approach to be used is ANCOVA with $\log(Y)$ as the outcome and $\log(X)$ as the covariate(Senn, 1997).

### 2.3. Crude corrections in general

The above are special case 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(Kronmal, 1993). A notorious example is given by the electrocardiogram (ECG). 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.) 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 the RR interval in some way. Such corrected QT values are labelled QTc and two approaches are commonly employed. Bazett's correction divides the QT interval by the square root of the RR interval and Fridericia's correction by the cube root of the interval. Thus the formulae are as given below

$$QT_C = \frac{QT}{\sqrt{RR}} \quad \text{Bazzet's}$$
$$QT_C = \frac{QT}{\sqrt[3]{RR}} \quad \text{Fridericia's} \qquad (2.2)$$

Note that the RR interval is usually close to 1 second 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 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 coefficient that just happens to be 1 and whose units are not mentioned in the correction, then dimensional analysis shows that the units of Bazett's are in the square root of time and the units of Fridericia's in 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 realisation that something is wrong. In a review of correction formulae commonly employed, Malik compared 31 such measures

on a data-set of 1402 patients on a variety of treatments (Malik, 2002). 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 that the problem ought to be handled. In my 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 factorisation 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(Gabriel, 1961; Kenward, 1987). Thus we can study the effect of drugs on RR and then via some form of regression analysis the effect on QT that 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.)

### 2.4. Correction for post-randomisation covariates

There is in any case a danger in correcting a measure such as QT for a measure such as RR that is itself affected by treatment. The danger is that one removes some of the treatment effect. This is why a conditional measure such as QTc, 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 measure is corrected by another and then naively nominated as the main outcome variable, the correcting variable being forgotten. Here are two common examples I know of.

The first concerns provocation tests in asthma, whereby bronchoconstriction 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 and 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

$$100\left(\frac{Y_1 - Y_2}{Y_1}\right) \quad (2.3).$$

This is superficially similar to the percentage change from baseline measurements criticised previously and inherits, of course, all of its deficiencies but has further more serious problems(Senn, 1989; Senn, 1993). 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 standardly run is an adequate means of distinguishing *patients* according to 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(Kay, 1995). 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.

### 2.5. The effect of titrations on other measures

The effect of titrations is to destroy the meaning of other measurements(Senn, 1989; Senn, 1993, 1997). Unfortunately, this fact is not always appreciated. Thus examples can be found where patients are assessed in terms of time exercising on a treadmill until the onset of 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., 1997) describe a five period cross-over trial comparing two doses of salbutamol and two doses of salmeterol and placebo using the dose of methacholine required to induce a 20% drop in $FEV_1$ ($PD20FEV_1$) as an outcome measure. $PD20FEV_1$ was measured before administration of each treatment but also 30 minutes and 120 minutes after. But it is quite plausible, and indeed was observed to be the case, that the amount of methacholine required to produce a 20% drop in $FEV_1$ will be greater 30 minutes after administration of a more effective

treatment than it will be the same time after less effective treatment. In what sense, therefore can the comparison of the treatments at 120 minutes be meaningful? After all, nobody would agree that a procedure in which, patients, having been randomised to the particular treatment they were to receive, were then given different doses of methacholine according to which treatment group they were in(Kay, 1995; Senn, 1989; Senn, 1993, 1997). 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 25 years since Peter McCullagh's important read paper on modelling ordered categorical data(McCullagh, 1980), 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.

### 2.7. Grouped predictors

As explained in the introduction, on the whole I am 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 as defined by the four quintiles are very popular in epidemiology, for example. The misleading argument that is often given is that one is unwilling to make the linearity assumption that using the predictor in an analysis of covariance requires. However, this is clearly a red-herring. 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.

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 one 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 and this was used to construct the groups. However, remaining with a continuous predictor would not commit one to using a simple straight line model. Other alternatives exist such as using higher order polynomials, fractional polynomials(Royston & Altman, 1994) or splines. For example, if the alternative is quintile groups, superior uses of the four degrees of freedom surely exist. Good review of various approaches for creating suitable prognostic models are given by Harrell et al(Harrell et al., 1996) and Royston et al(Royston et al., 2000) and also by Frank Harrell in his book(Harrell, 2001).

## 3. Creating dichotomies

In this section I concentrate particularly on issues surrounding the creation of dichotomies

### 3.1. The losses in creating dichotomies

As is well known, the Pitman efficiency of the sign test compared to the t-test is

$$\frac{2}{\pi} \approx 64\% .$$

(See, for example, (Garthwaite et al., 2002) pp197-198.) However, this is an optimistic analogy to the losses attendant in practice on dichotomising 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 to the t-test.

Suppose we have a continuous outcome approximately Normally distributed and suppose that the control group mean is $\mu$ 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 $\mu$ 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

$$1/\left[\Delta(\mu)\right]^2 \quad (3.1)$$

Now suppose that we have dichotomised 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 characterising 'response' in the control group. The control group proportion is estimated with a variance proportional to

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

On the other hand the perturbation $\Delta(\theta)$ of the value of $\theta$ 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

$$\frac{\Phi(k-\mu)\left[1-\Phi(k-\mu)\right]}{\left[\phi(k-\mu)\right]^2\left[\Delta(\mu)\right]^2} \quad (3.2)$$

Noting that without loss of generality we can set $\mu = 0$, then the ratio of the term previously found(3.1) to this one(3.2) is

$$\frac{\left[\phi(k)\right]^2}{\Phi(k)\left[1-\Phi(k)\right]} \quad (3.3)$$

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 1 below shows a plot of relative efficiency as a function of $k$.

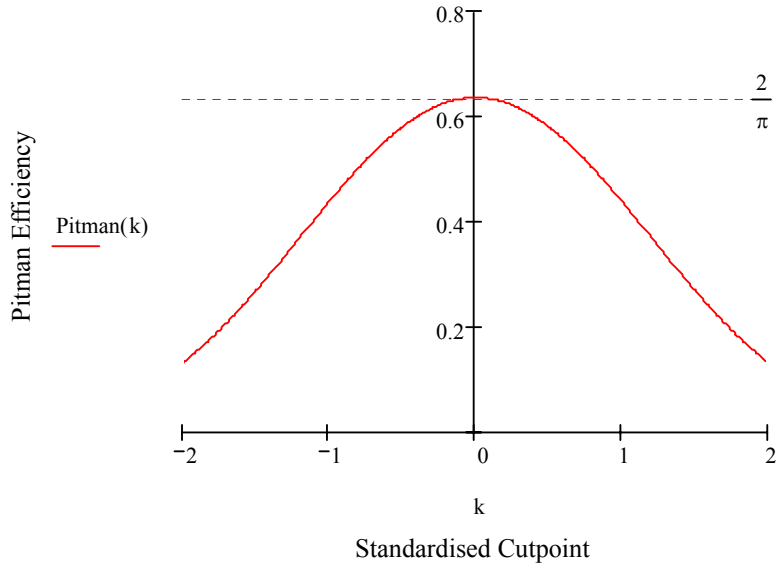


Figure 1. Pitman asymptotic relative efficiency as a function of the standardised cut-point.

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%.

### 3.2. The further losses in using change scores to create dichotomies

This has already been mentioned in section 2.1. The point is expanded here. Provided that change scores are not dichotomised, it is always possible to recover the information in the baselines in an analysis of covariance and this is, as has already been noted, is then equivalent to analysing the raw outcomes in an analysis of covariance(Hills & Armitage, 1979; Laird, 1983; Senn, 1997). 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 baseline and outcome have approximately the same variance, $\sigma^2$ and correlation $\rho$. Then the variance of the change score estimator $D = Y - X$ is

$$2(1-\rho)\sigma^2 \quad (3.4)$$

and this, as is well known, is greater than the variance of the raw outcome $Y$ if $\rho < 1/2$ (Hills & Armitage, 1979; Senn, 1997). 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 analysis of covariance adjusted score can be constructed from a general estimator of the form $Y - \beta X$ by finding either the value of $\beta$ that minimises the variance of the score, which variance in general must have the form

$$\sigma^2\left(\beta^2 - 2\beta\rho + 1\right) = \sigma^2\left[(\beta-\rho)^2 + 1 - \rho^2\right] \quad (3.5)$$

or the value that makes the covariance with $X$,

$$\sigma^2(\rho - \beta) \quad (3.6)$$

zero. By inspection, for either of these two approaches, the value is obviously $\beta = \rho$ and by substituting this value in (3.5) the residual variance is then

$$\sigma^2\left(1-\rho^2\right) \quad (3.7)$$

. Thus the relative efficiency of the change score compared to ANCOVA is the ratio of(3.7) to (3.4)

$$\frac{1-\rho^2}{2(1-p)}=\frac{1+\rho}{2} \quad (3.8).$$

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

This overstates slightly the advantage of analysis of covariance, 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 to using the change score(Senn, 1997). However, the expected loss is equivalent to one patient and hence not particularly important(Atkinson, 1999; Burman, 1996).

### 3.3. Foolish definitions of response

The sort of dichotomy discussed in section 3.1 is based on a single cut-point and a continuous measure with no further side-conditions. In fact, far more foolish definitions of response can be found. Consider, for example, this definition from the Committee of Proprietary Medicinal Products of the European Medicines Evaluation Agency, of which the most important word is the first:

> "Arbitrarily, response criteria for antihypertensive therapy include the percentage of patients with a normalisation of blood pressure (reduction SBP < 140 mmHg and DBP < 90 mmHg) and/or reduction of SBP ≥ 20 mmHg and/or DBP ≥ 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(Kuramoto, 1989)

> "In accordance with the guidelines for the clinical evaluation of antihypertensive agents, patients with a blood pressure level greater than or equal to 160/95 mm Hg after receiving a placebo for 4 weeks were enrolled in the study. The test drug was then administered for 12 weeks. A decrease in blood pressure by 20/10 mm Hg from the pretreatment level or a blood pressure level below 149/89 mm Hg was considered to indicate a response."

Again, the fact that response is defined in terms of a joint requirement for DBP and SBP makes 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 (Goetghebeur et al., 1998)

> "The treatment is called successful if either the patient has gone down from a baseline diastolic blood pressure of ≥ 95 mmHg ≤ 90mmHg or has achieved a 10 per cent reduction in blood pressure from baseline."

Figure 2 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.

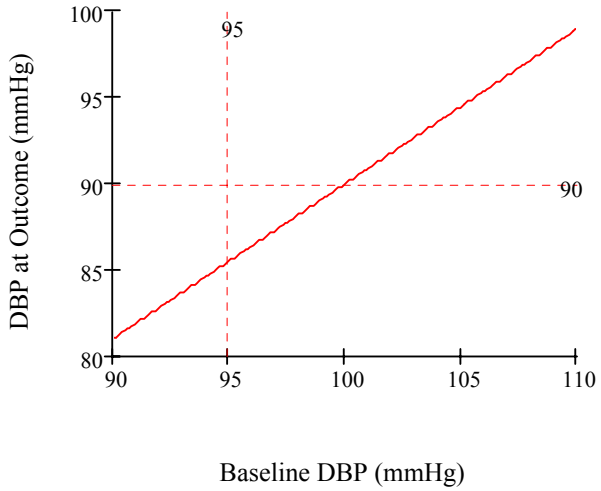


Figure 2 Response region as defined in (Goetghebeur, et al., 1998) as a function of diastolic blood pressure at baseline and outcome

The indicator functions for the response is

$$I(X,Y) = (Y < 0.9X) \cup \left[ (X \geq 95) \cap (Y \leq 90) \right] \quad (3.9)$$

and if $X, Y$ are jointly distributed as a Normal distribution with parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$, then the conditional distribution of $Y$ given $X$, $f(Y|X=x)$ is Normal with mean

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

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

$$E\left[ I(X,Y) \middle| X = x \right] = \int_{-\infty}^{\infty} f(Y|X=x) I(x,Y) dy \quad (3.10)$$

Figure 3 show the probability of a patient 'responding', as given by (3.10) as a function of baseline blood pressure if the values of parameters of the Normal distribution are

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

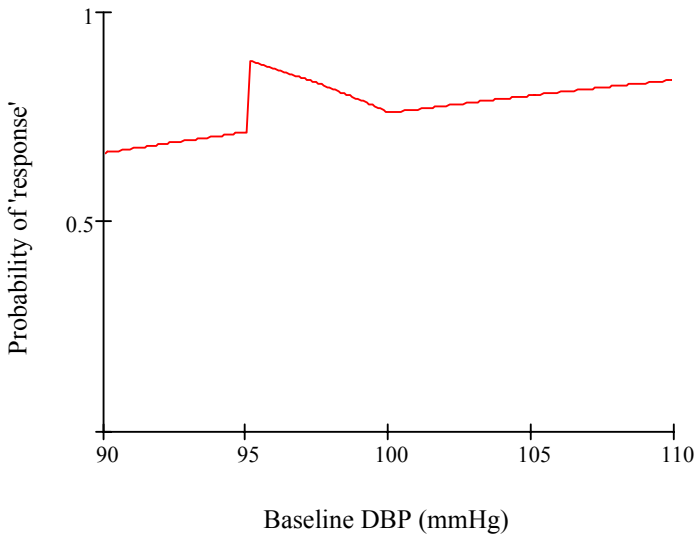


Figure 3. Probability of 'response' as a function of Baseline DBP

Clearly this is a ludicrous region and any scientists, whether physicians or statisticians, who imagine that such a measure can be safely interpreted are deluding themselves.

### 3.4. Fallacies of responder analysis

A similar delusion concerns so-called responder analysis, a thing that is almost *de rigueur* now in regulatory submissions. Allied to this are extravagant causal interpretations of 'response'. Consider, for instance, a suggestion of (Guyatt et al., 1998) that one calculate the proportion of patients who respond better under one treatment than another in order to calculate the number needed to treat, a way of reporting binary outcomes that will be picked up in section3.5. They propose that in a cross-over trial one might calculate the difference between the observed outcome under treatment and the observed outcome under control and compare this to some clinically relevant difference to see whether the patient has had a superior response. This ignores the fact that because of within-patient 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(Senn, 2004; Senn, 1998a; Senn, 2001, 2003a).

Consider the case where in fact we carry out two cross-over trials using exactly the same patients for the second trial we used for the first, so 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., 1998). Suppose we have 24 patients and the results are as given in Table 1 below. This would, indeed, then support the sort of conclusion that Guyatt et al propose that their method justifies.

Table 1 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 |

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 2, which shows only one of the margins from Table 1 and not the joint distribution over two cross-over trials.

Table 2 Pattern of responses in a single cross-over .

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

To assert that the position would be as it is in table 1 if only the further data to produce it had been collected requires an untestable (since they 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 3, which is the condition that corresponds to complete independence(Senn, 2004). Thus for this pattern, it makes no sense to talk about 'responders' and 'non-responders'. In fact it would be better if this terminology, which carries with it a considerable causal freight it cannot support, were abandoned.

Table 3 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 |

### 3.5. Number needed to treat

The number needed to treat (NNT) is the reciprocal of the risk reduction in treating patients with one treatment rather than another(Cook & Sackett, 1995). It is thus the reciprocal of the risk difference and was originally proposed by (Laupacis et al., 1988). I have no objection to the NNT as a *concept* when it is applied to genuine dichotomies as opposed to dichotomised measures. It is not, however, even in the case of genuine binary outcomes a useful way to summarise the results of a trial (Grieve, 2003; Hutton, 2000; Senn, 1998b; Smeeth et al., 1999), 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. The trouble with the NNT is that it is highly unlikely to be additive, thus even if one wishes to calculate a 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(Cook & Sackett, 1995). Whatever one's view, and I personally find Hutton and Grieve's strictures of this measure as commonly used compelling(Grieve, 2003; Hutton, 2000), it must surely be totally unacceptable to create dichotomies purely in order to be able to calculate NNTs.

## 4. Discussion

In this section I consider the obstacles in the way of proceeding to a more rational treatment of measures.

### 4.1. Division of responsibilities

I have already referred to the regrettable of tendency of statisticians to regard measurement as being 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 criticise such procedures. Unfortunately, it is not hard to find example where statisticians have simply failed to question what is being provided. Consider, for example, the paper we used to provide our third example of 'response' measures in hypertension. The paper is a sophisticated contribution to the literature on missing observations by extremely eminent and able medical statisticians but the measure is accepted without question. But anybody who is not appalled by this sort of measure has not thought about it seriously.

### 4.2. False claims about interpretability or relevance

Even experienced medical statisticians are apt to make these claims. For example, Lewis in an article entitled, 'In Defence of the Dichotomy' (Lewis, 2004)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."(pp77-78).

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 I measure blood pressure or basing it say on the average of three determinations, either of which would reduce the relevant values of $\sigma_X$ and $\sigma_Y$ , I should change the probability of response, even though nothing will actually have changed in the way the treatment affects the patients(Senn, 1997)(pp124-125)?

Elsewhere in the same article Lewis 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."(p78)

But this is again misleading. For most diseases we simply have not run the sort of repeated period cross-over trials that would let us identify variation in response. Dichotomising a continuous measure does nothing to address this issue. Even when dichotomized we cannot distinguish between the cases represented by Table 1 and Table 3 and in any case there are grounds to believe that some of the expectations for genetic variation in treatment response are not well founded(Senn, 2001). For example, it is a common experience that attempts to improve the power of trials by excluding non-responders are not particularly effective. It is true that we label such dichotomies 'response' but to interpret this *word* as implying something about differential effects of treatment is, in the words of Hobbes, to turn wise men's counters into the coins of fools.

### 4.3. Citing historical precedent

The commonest argument any statistician encounters trying to reform the business of measuring effects is that such measurement cannot be changed because one has always done it this way, or thought in this way. All such claims must be rejected as unscientific. 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 dichotomising 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 dichotomising are not negligible and the gains are illusory. It is no exaggeration to claim that if all dichotomised measures of the outcomes of 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. It is time that statisticians took the matter of measurement seriously. Using original continuous measures rather than their dichotomies would lead to a considerable improvement in the quality of pharmaceutical trials.

## REFERENCES


Atkinson, A. C. (1999). Optimum biased-coin designs for sequential treatment allocation with covariate information. *Statistics in Medicine* **18(**14**)**: 1741

Burman, C.-F. (1996). in *Department of Mathematics* (Gothenburg: Chalmers University of Technology)

Cook, R. J., & Sackett, D. L. (1995). The Number Needed to Treat - a Clinically Useful Measure of Treatment Effect. *British Medical Journal* **310(**6977**)**: 452

Gabriel, K. R. (1961). The model of ante-dependence for data of biological growth. *Bulletin Institut International Statistique (Paris)* **39**253

Garthwaite, P. H., et al. (2002). *Statistical Inference* (Second ed.: Oxford University Press)

Goetghebeur, E., et al. (1998). Estimating the causal effect of compliance on binary outcome in randomized controlled trials. *Stat Med* **17(**3**)**: 341

Grieve, A. P. (2002). Do statisticians count? A personal view. *Pharmaceutical Statistics* **1(**1**)**: 35

Grieve, A. P. (2003). The number needed to treat: a useful clinical measure or a case of the Emperor's new clothes? *Pharmaceutical Statistics* **2(2)**: 87

Guyatt, G. H., et al. (1998). Interpreting treatment effects in randomised trials [see comments]. *British Medical Journal* **316(**7132**)**: 690

Hand, D. (2004). *Measurement: Theory and Practice* (London: Arnold)

Harrell, F. E. (2001). *Regression Modeling Strategies* (New York: Springer)

Harrell, F. E., Jr., et al. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15(**4**)**: 361

Higham, M. A., et al. (1997). Dose equivalence and bronchoprotective effects of salmeterol and salbutamol in asthma. *Thorax* **52(**11**)**: 975

Hills, M., & Armitage, P. (1979). The two-period cross-over clinical trial. *British Journal of Clinical Pharmacology* **8**7

Hutton, J. L. (2000). Numbers needed to treat: properties and problems (with comments). *Journal of the Royal Statistical Society A* **163(**3**)**: 403

Kay, R. (1995). Some Fundamental Statistical Concepts in Clinical-Trials and Their Application in Herpes-Zoster. *Antiviral Chemistry & Chemotherapy* **6**28

Kenward, M. G. (1987). A Method for Comparing Profiles of Repeated Measurements. *Applied Statistics-Journal of the Royal Statistical Society Series C* **36(**3**)**: 296

Kronmal, R. A. (1993). Spurious Correlation and the Fallacy of the Ratio Standard Revisited. *Journal of the Royal Statistical Society Series a-Statistics in Society* **156**379

Kuramoto, K. (1989). Double-blind studies of calcium antagonists in the treatment of hypertension in Japan. *J Cardiovasc Pharmacol* **13 Suppl 1**S29

Laird, N. (1983). Further comparative analyses of pre-test post-test research designs. *The American Statistician* **37**329

Laupacis, A., et al. (1988). An Assessment of Clinically Useful Measures of the Consequences of Treatment. *New England Journal of Medicine* **318(**26**)**: 1728

Lendrem, D. (2003). Statistical support to non-clinical. *Pharmaceutical Statistics* **1(2)**: 71

Lewis, J. A. (2004). In defence of the dichotomy. *Pharmaceutical Statistics* **3(2)**: 77

Malik, M. (2002). The imprecision in heart rate correction may lead to artificial observations of drug induced QT interval changes. *Pacing Clin Electrophysiol* **25(2)**: 209

Mccullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society B* **42**109

O'neill, R. T. (1991), in *Biometri in der chemisch-pahramzuetischen Industrie*, ed. J. Vollmar (Stuttgart: Gustav Fischer), 31

Royston, P., & Altman, D. G. (1994). Regression Using Fractional Polynomials of Continuous Covariates - Parsimonious Parametric Modeling. *Applied Statistics-Journal of the Royal Statistical Society Series C* **43(**3**)**: 429

Royston, P., et al. (2000). Modeling the effects of continuous risk factors. *Journal of Clinical Epidemiology* **53(2)**: 219

Senn, S. (2004). Individual response to treatment: is it a valid assumption? *BMJ* **329(**7472**)**: 966

Senn, S. J. (1989). The use of baselines in clinical trials of bronchodilators. *Statistics in Medicine* **8(**11**)**: 1339

Senn, S. J. (1993). Statistical issues in short term trials in asthma. *Drug Information Journal* **27**779

Senn, S. J. (1997). *Statistical Issues in Drug Development* (Chichester: John Wiley)

Senn, S. J. (1998a). Applying results of randomised trials to patients. N of 1 trials are needed [letter; comment]. *British Medical Journal* **317(**7157**)**: 537

Senn, S. J. (1998b). Odds ratios revisited. *Evidence-Based Medicine* **3**71

Senn, S. J. (2001). Individual Therapy: New Dawn or False Dawn. *Drug Information Journal* **35(4)**: 1479

Senn, S. J. (2003a). Author ’s Reply to Walter and Guyatt. *Drug Information Journal* **37**7

Senn, S. J. (2003b). Lost opportunities for statistics. *Pharmaceutical Statistics* **2(1)**: 3

Smeeth, L., et al. (1999). Numbers needed to treat derived from meta-analyses--sometimes informative, usually misleading [see comments]. *British Medical Journal* **318(**7197**)**: 1548

**RÉSUMÉ**

Les analyses des essais cliniques emploient, trop souvent, une échelle binaire construite en sectionnant en deux parties, selon des critères plus ou moins arbitraires, une échelle originale de manière continue. Cela est même souvent proposé par les autorités d'enregistrement. Cette mauvaise habitude entraîne normalement une grande perte de précision avec la conséquence que le nombre de patients pour en tirer une conclusion utile doit être fortement augmenté. Je montre, avec des exemples, que ces pertes peuvent être très importantes et que les soi-disant avantages de cette pratique sont illusoires.