MAIN PAPER

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/pst.1667

Published online in Wiley Online Library

How to test hypotheses if you must

Andrew P. Grieve*

Drug development is not the only industrial-scientific enterprise subject to government regulations. In some fields of ecology and environmental sciences, the application of statistical methods is also regulated by ordinance. Over the past 20 years, ecologists and environmental scientists have argued against an unthinking application of null hypothesis significance tests. More recently, Canadian ecologists have suggested a new approach to significance testing, taking account of the costs of both type I and type II errors. In this paper, we investigate the implications of this for testing in drug development and demonstrate that its adoption leads directly to the likelihood principle and Bayesian approaches.

Keywords: hypothesis tests; null hypothesis significance tests; type I error; type II error; power; planning of experiments; sample sizing; Neyman–Pearson lemma; likelihood principle; sampling frame; Lindley's paradox; Bayesian test

1. INTRODUCTION

Four recent papers by Mudge and colleagues [1–4] are likely to be of interest to practising scientists as they address a question that statisticians are often asked: how do I choose the type I error? They join other papers in the ecology and marine environmental literature that challenge the accepted dogma of significance and hypothesis testing [5–9].

Significance tests, in their modern form, have been around since the early 1920s. It is therefore perhaps surprising that there are still associated with them issues that are opaque to researchers. Part of the reason for this is that today's standard practice in using statistical procedures is a hybrid procedure composed of elements of two of the major competing statistical schools, those associated with RA Fisher on the one hand and Jerzy Neyman and Egon Pearson on the other hand.

What is meant by standard practice? In the experimental sciences, the standard practice of statistical design and analysis can be characterised as follows. Before an experiment is to be conducted, the experimenter chooses the following:

- the probability of committing a type I error, α, typically 0.05 or 0.01, or the significance level;
- an appropriate null hypothesis of no treatment effect;
- an alternative hypothesis representing a treatment effect magnitude that is of interest or one that is expected to be achieved – this is the critical effect size referred to by Mudge *et al*. [1]; and
- a sample size to give the probability of committing a type II error at a prescribed level, β, typically 0.1 or 0.2.

After the experiment is completed, the experimenter does the following:

- estimates the treatment effect;
- calculates the *p*-value and, if it is smaller than the significance level, declares statistical significance; and
- determines a confidence interval for the treatment effect.

This hybrid approach [10,11] combines Fisher's null hypothesis, significance test and *p*-value with Neyman and Pearson's alternative hypothesis, type I and type II errors, power and Neyman's confidence intervals.

One issue that is sometimes less clear to practitioners is that there is a relationship between type I and type II error rates, and this has a consequence for decision-making. In simple terms, if the probability of type I error is increased, in which event we require a less stringent decision criterion to declare a positive outcome, we reduce the probability of a type II error or increase the power. In contrast, requiring a more stringent decision criterion reduces both the chances of declaring a false positive and of declaring a true positive. Clearly, the probabilities of type I and type II errors are inversely related. The following question then arises: how do we choose appropriate values for α and β because by changing one, we influence the other.

Neyman and Pearson's original solution held α as fixed and chose the decision criterion, or critical region, in order to minimise the probability of type II error and hence maximise the power [12]. However, as they themselves noted, these

> …Two sources of error can rarely be eliminated completely; in some cases it will be more important to avoid the first, in others the second. We are reminded of the old problem considered by LAPLACE of the number of votes in a court of judges that should be needed to convict a prisoner. Is it more serious to convict an innocent man or to acquit a guilty? That will depend upon the consequences of the error; is the punishment death or fine; what is the danger to the community of released criminals; what are the current ethical views on punishment? From the point of view of mathematical theory all that we can do is to show how the risk of the errors may be controlled and minimised. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator.

ICON Adaptive Trials Innovation Centre, Icon Plc, Marlow, Buckinghamshire, UK

**Correspondence to: Andrew P. Grieve, ICON Adaptive Trials Innovation Centre, Icon Plc, 2 Globeside, Globeside Business Park, Marlow, Buckinghamshire, SL7 1HZ, UK. E-mail: Andrew.Grieve@ICONPLC.com*

The Neyman–Pearson solution was to fix α and then to choose the decision criterion to minimise ß, but of course, this is not the only approach. More modern authors have taken up this theme. For example, Oakes [13] opined that

> The extent to which scientific caution need be exercised and the importance of discovery of an effect (alternatively the cost of making type 1 and type 2 errors) will vary from situation to situation. This would imply that conventional significance levels should be abandoned and that with any particular piece of research α should be set with regard to the costs in hand,

and more recently, Senn [14] noted:

> The Neyman–Pearson lemma does not justify that minimising the type II error rate whilst holding the type I error rate at the same predetermined level on any given occasion is a reasonable rule of behaviour.

It is interesting to note that whilst many statisticians working in the drug development industry might expect the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) guideline on statistical principles for drug trials to be strict about the setting of type I error rate, in fact, it provides support for a more relaxed attitude to the choice of the type I error.

> Conventionally the probability of type I error is set at 5% or less or as dictated by any adjustments made necessary for multiplicity considerations; the precise choice may be influenced by the prior plausibility of the hypothesis under test and the desired impact of the results. [15]

The approach proposed by Mudge *et al*. [1] is aimed precisely at the proposition stated by Oakes. Their approach, instead of fixing the type I error rate and minimising the type II error rate, consists of choosing an optimal value for the type I error rate by minimising the weighted sum of the error rates in which the weights are related to the costs of each type of error. It is important to note that in introducing their approach, it would appear that they are implicitly assuming that the sample size for the experiment has been chosen by some process that is independent of both the error rates. However, Mudge *et al*. suggest that sample size can also be determined by minimising the weighted sum of error rates [2], an idea we will look at later.

In this paper, I pull together a number of threads against the background of the Mudge *et al*. proposal [1]. For the purpose of developing and presenting the basic idea, we will assume that there is a single primary outcome and that the variance is known and will restrict attention to simple null hypotheses and one-sided tests. This does not weaken the argument, nor does it mean that the basic principle is not more generally applicable, and we will indicate how its use can be broadened to complex hypotheses, cases in which the variance is unknown and so on.

In Section 2, we address the following questions: what is a null hypothesis significance test (NHST) and what are the main issues in its use? In Section 3, we explore whether the probabilities of type I and type II errors are relevant to NHSTs. Section 4 reviews the accepted approach to choosing the sample size for a study. Section 5 looks at the basic proposal introduced by Mudge *et al*. [1] and develops an analytic solution for the optimal α. Section 6 investigates the implication of the optimal choice of the type I and type II error rates by determining the ratio of weights, which leads to standard type I and type II error rates. Section 7 looks at how the minimised weighted sum of errors can be used to design studies as suggested by Mudge *et al*. [2]. Section 8 looks at a modified form of the Neyman–Pearson lemma and shows how it leads to the likelihood principle. Section 9 considers the implication of the likelihood principle for clinical trials and gives two examples in which sampling frames can give rise to contentious issues. In Section 10, a connection between the rejection region based on minimising the weighted sum of errors and a Bayesian test of a simple null hypothesis is established, and implications for Lindley's paradox are considered [16,17]. In the final discussion section, the argument is made that the results presented cast doubt on the dogma of controlling the type I error at all costs.

2. WHAT IS A NULL HYPOTHESIS SIGNIFICANCE TEST?

Although there were earlier examples of significance tests, their systematic introduction was due to RA Fisher. The principle underlying significance tests is the search for a sensible test statistic, say T, whose distribution could be completely specified if the appropriate so-called null hypothesis were true but which would also be sensitive to departures from the null hypothesis. Once data have been collected and the corresponding value of the statistic, t, has been calculated, the next step is to derive the probability $P(T \geqslant t|H_0)$, corresponding to the chance that the statistic would be as, or more, extreme as the observed value if the null hypothesis were true. If this probability is small, then one may conclude either that an unlikely event has occurred by chance or that the null hypothesis is false. This probability is now called a p-value, which is a term coined by Deming [18] and is one of the most widely used – and abused – statistical concepts.

Over the years, there has been considerable debate about the desirability of using such NHSTs. Whole books have been written in their favour [19], as well as against them [20,21], with considerable emphasis on their use and abuse. There are a number of themes that recur in these debates. Here are perhaps the five most important.

First, it is often the case that the null hypothesis is almost certainly untrue, and we might ask ourselves if it is worth testing such an unlikely null hypothesis. In drug development, sponsors will have evidence for the efficacy and safety of their new drug before entering a phase III development; indeed, this phase of drug development is generally referred to as the confirmatory phase, and therefore, it is unlikely that the exact null hypothesis is true, which is not to say that the new drug necessarily delivers an effect of clinical importance.

Second, estimation of the underlying parameter is often more informative than testing a hypothesis concerning the parameter. From the late 1980s onwards, there was increasing clamour in the medical literature to oust hypothesis testing from its preeminent position and to replace it with confidence intervals. The campaign was endorsed by the International Committee of Medical Journal Editors [22] and culminated in a *British Medical Journal*-backed publication of a confidence interval cookbook [23,24].

Third, if we choose a large enough sample, we can almost certainly declare differences of no biological relevance to be impressively statistically significant. It is important to remember that statistical significance is not equivalent to biological relevance. This will be further discussed later.

Fourth, failure to reject a null hypothesis does not in itself mean that the hypothesis is necessarily true, as we can arrange for non-significance by choosing an inadequate sample size. This is a case of the aphorism 'absence of evidence does not mean evidence of absence' [25] and is related to the so-called rule of three [26]. This says that if you test a new drug on n patients and see no adverse events, you are not justified in claiming that the drug has no risk because the upper limit of the 95% confidence interval for the true rate of adverse events is approximately $3/n\%$. For example, zero adverse events from testing 20 patients gives an upper 95% confidence interval of 15%, so we may be unable to exclude rates of risk that are of clinical importance.

Finally, the p-value is not the probability that the null hypothesis is true, although scientists often believe it to be. This belief suggests that scientists would be happier with a Bayesian approach to hypothesis tests, which does look to determine the probability of hypotheses. This is no less true of confidence intervals. Scientists would like to regard a confidence interval as a fixed interval within which the true parameter lies with a predetermined probability of success, again a Bayesian interpretation. As an illustration of this, I would offer the following passage:

> …the proper interpretation of confidence interval requires that we consider a large number of hypothetical random samples (each of the same size). Then '95% confidence' means that approximately 95% of the 95% confidence intervals from those random samples would include the unknown true value, and about 5% would not. Because the true fraction in the population is unknown, it is impossible to tell if the 95% confidence interval of 28% to 55% that was obtained from the observed sample data actually included the true fraction. Strictly speaking, we cannot even tell how likely the 95% confidence interval of 28% to 55% is to include that unknown fraction. Nevertheless, the usual interpretation is that we are 95% confident that the unknown true value is between 28% and 55%. [27]

I have previously described this passage as 'Magnificent. But surely not logical' [28].

3. ARE TYPE I AND TYPE II ERRORS RELEVANT IN NULL HYPOTHESIS SIGNIFICANCE TESTS?

In the book *Design of Experiments*, Fisher considers issues related to the sample size of experiments. He notes that by

> …increasing the size of the experiment [either by enlargement or repetition], we can render it more sensitive, meaning by this that it will allow of the detection…of a quantitatively smaller departure from the null hypothesis. Since in every case the experiment is capable of disproving, but never of proving this hypothesis, we may say that the value of the experiment is increased whenever it permits the null hypothesis to be more readily disproved. [29]

Whilst this is not an explicit acknowledgement of the importance of power and Fisher never acknowledged such an importance, nonetheless, the implicit implication of this statement is support for more powerful tests. Stephen Senn has pointed out to me that in correspondence with Chester Bliss [30], Fisher argued that the choice between alternative significance tests should be based on 'no more than the experience that one test of significance gave more frequently significant results than another', that is on empirical power (also [14]). As we have seen, power depends upon the choice of the probability of a type I error, and Fisher, despite many authors having associated him with the use of fixed significance levels, in his later writings was clearly against this idea.

> No scientific worker has a fixed level of significance from year to year, and in all circumstances, he rejects hypothesis; he rather gives his mind to each particular case in the light of his evidence of ideas. [31]

Neyman argued that without type II errors 'no purely probabilistic theory of tests is possible' [32], and Fisher might have accepted this as far as it related to acceptance sampling but not to scientific hypothesis testing [33].

It is of course possible to take a pragmatic view of the number of experimental units required for a particular experiment and ask the following question: how many experimental units does the budget allow me to test? Given such a resource, what effect size am I able to detect with a given power? I have previously termed this as resource sizing, and it is deeply unsatisfying on a number of levels.

First, resource sizing is generally synonymous with underpowering, an issue that has been recognised for at least half a century [34,35] and still exists [36]. It might be thought that the cash-rich pharmaceutical industry would be immune to resource sizing, but that is not so. In any given year, a pharmaceutical company has a fixed budget for research and development and a portfolio of projects on which that money can be spent. A consequence is that at any particular time, the budget for an individual sponsor is finite despite an understandable desire to run as many development programmes as possible. Resources therefore are constrained. In contrast, many biotechnology companies will in general have only one or two development programmes but will be resource constrained by the need to restrict the 'burn' of money or the need to raise additional capital from investors. The last 15 years has seen a high failure rate in phase III clinical trials. Kola and Landis report an average failure rate in phase III trials in drug development of 45% and higher for specific therapeutic areas. For example, the failure rate is as high as 60% in oncology programmes [37]. Whilst there has been a recent slight improvement [38], such high failure rates may be partly due to underpowering studies by resource sizing.

Second are ethical issues. There are two types of ethics that are typically associated with human medical research – individual and collective ethics [39,40]. Individual ethics recognises the primacy of the individual and is aimed at doing what is best for the subjects in the current trial. In contrast, collective ethics is aimed at doing what is best for all future patients who will benefit from the results of the current trial. Unsurprisingly, there is a tension between these two principles, which is recognised in the Declaration of Helsinki. The declaration comes down on the side of the individual – 'Concern for the interests of the subject must always prevail over the interest of science and society' [41]. Recent ethical discussions on clinical trial designs have largely concentrated on adaptive designs and individual ethics. One perceived advantage of some adaptive designs is their ability to allocate a disproportionate percentage of patients to the best treatment, or dose. This will provide a differential advantage for some, although not all, patients but may have the disadvantage of slowing recruitment to trials in chronic diseases because informing patients that they are more likely to receive the 'best treatment' later in the trial may result in them withholding consent until they judge that their chance of receiving the 'best treatment' has risen sufficiently. This

is termed accrual bias. Designing underpowered trials by resource sizing has collective ethical consequences because an underpowered negative trial may deter independent researchers from studying the same mechanism of action and potentially deprive future patients of an efficacious drug, or drugs. Another question is whether it is ethical to plan a trial that, because the trial has been designed based on resource sizing, is unlikely to have the nominal protocol power. In such circumstances, should patients be randomised? An argument has been made against proscribing underpowered studies because it would prevent independent investigators, without access to substantial funding, from carrying out research and would limit the availability of important information to be combined in future meta-analyses [42]. There are, of course, counterarguments [36].

Third, from a regulatory perspective, it is inappropriate to utilise resource sizing. The ICH guideline on statistical principles in clinical trials is clear that there needs to be justification for all the inputs into the sample size calculations:

- the means and variances;
- response, or event, rates; and
- the clinically meaningful difference.

What is the basis for the choice that is made? For phase III clinical trials, it is expected that the assumptions underlying a design will come either from the literature or from earlier trials in the development programme [15]. The basis for the choice of the treatment effect to be detected may be based on 'a judgement concerning the minimal effect which has clinical relevance in the management of patients' or on 'a judgement concerning the anticipated effect of the new treatment', neither of which is related to resource sizing.

4. THE SCIENTIFIC APPROACH TO SAMPLE SIZING

If resource sizing is inappropriate, how should studies be sample sized? The 'scientific' planning of experiments as it has evolved is conceptually very simple. For illustrative purposes, we will assume that our interest lies in designing a single-arm study; that the main variable of interest is normally distributed with known variance, σ^2; that the effect size we are interested in detecting is δ_0, variously referred to as the clinically relevant difference [43] or minimally clinically important difference (MCID) [44]; that the one-sided significance level (probability of a type I error or false-positive rate) is α_0; and that the probability of a type II error (false-positive rate) is β_0.

With these assumptions, the sample size, n, to test the null hypothesis $H_o : \delta = 0$ against the alternative hypothesis $H_A : \delta = \delta_0$ is given by the formula

$$n = \frac{\sigma^2 \left(Z_{1-\alpha_0} + Z_{1-\beta_0}\right)^2}{\delta_0^2} \tag{1}$$

obtained from the requirement in Figure 1 that the coloured regions have the specified magnitudes. Although this approach is not appropriate in all circumstances, the central limit theorem allows it to be used in many cases after a suitable transformation.

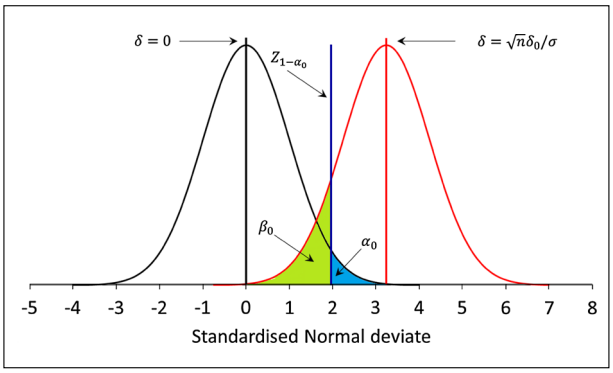

Figure 1. Determination of sample size.

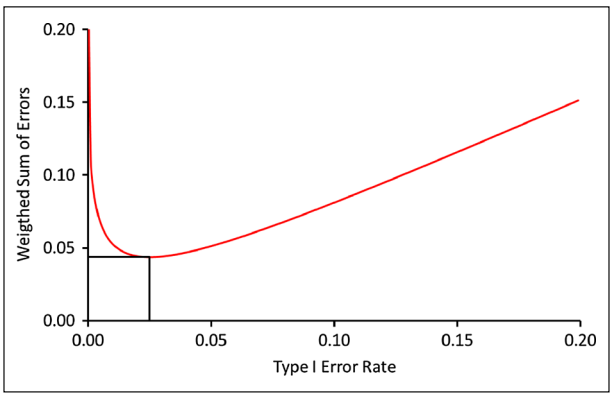

Figure 2. The weighted sum of type I and type II errors as a function of the type I error $\alpha(\omega = 3)$.

5. DETERMINING THE OPTIMAL α

For a given n, α and δ_0, we can determine the probability of a type II error to test the null hypothesis $H_o : \delta = 0$ against the alternative hypothesis $H_A : \delta = \delta_0$ by inverting Equation (1) to give

$$\beta = 1 - \Phi\left(\theta + Z_\alpha\right) \tag{2}$$

where $\theta = \sqrt{n}\delta_0/\sigma$. For a given ω – either the 'relative prior probabilities of the null and alternate hypotheses being true' [1] or the relative costs of the errors – the weighted sum of the probabilities of type I and type II errors is given by

$$\Psi = \frac{\omega\alpha + 1 - \Phi\left(\theta + Z_\alpha\right)}{\omega + 1} \tag{3}$$

which, for a fixed ω, is a function of α alone. The functional relationship between Ψ and α is illustrated in Figure 2 for the case $\omega = 3$, that is, a case in which the cost of a type I error is three times that of a type II error. It is clear that a minimum value exists.

The value of α, which minimises Ψ, is obtained by solving the equation

$$\frac{d\Psi}{d\alpha} = 0 = \omega - \frac{\phi\left(\theta + Z_\alpha\right)}{\phi\left(Z_\alpha\right)} \tag{4}$$

where

$$\phi(x) = \sqrt{\frac{1}{2\pi}} e^{-\frac{x^2}{2}}$$

The solution to (4) is

$$\alpha = \Phi\left(-\frac{\ln(\omega)}{\theta} - \frac{\theta}{2}\right) \tag{5}$$

Some immediate consequences of this solution are apparent. First, the corresponding value of the probability of the type II error is

$$\beta = 1 - \Phi\left(-\frac{\ln(\omega)}{\theta} + \frac{\theta}{2}\right)$$

and the minimum weighted sum is

$$\Psi = \frac{\omega\Phi\left(-\frac{\ln(\omega)}{\theta} - \frac{\theta}{2}\right) + \Phi\left(\frac{\ln(\omega)}{\theta} - \frac{\theta}{2}\right)}{\omega + 1} \tag{6}$$

Second, if $\omega = 1$, the equal-costs case, then the minimum sum occurs when $\alpha = \beta$.

In the unknown variance case, we can replace (2) by

$$\beta = 1 - F_{\upsilon}(\theta, t_{\upsilon,\alpha})$$

where $F_{\upsilon}(\theta, t)$ is the CDF of a noncentral t-distribution with υ degrees of freedom and noncentrality parameter θ and $t_{\upsilon,\alpha}$ is the one-sided $1 - \alpha$ critical value of the t-distribution with υ degrees of freedom. The corresponding weighted sum of errors is

$$\Psi = \frac{\omega\alpha + 1 - F_{\upsilon}(\theta, t_{\upsilon,\alpha})}{\omega + 1}$$

from which the minimum α is the solution to

$$\frac{d\Psi}{d\alpha} = 0 = \omega - \frac{f_{\upsilon}(\theta, t_{\upsilon,\alpha})}{f_{\upsilon}(0, t_{\upsilon,\alpha})} \tag{7}$$

where $f_{\upsilon}(\theta, t)$ is the density function of a noncentral t-distribution with υ degrees of freedom and $f_{\upsilon}(0, t)$ the noncentral t-distribution with υ degrees of freedom. There is no analytic solution to this equation, but numerical solutions are available either using a method such as *regula falsi* or using a search method [1].

The result (7) is applicable to a general class of problems. Suppose that the power of a test S, with critical value s_{α}, depends on a noncentral distribution function $H(\phi, s)$ characterised by a single noncentrality parameter φ. The weighted sum of errors is

$$\Psi = \frac{\omega\alpha + 1 - H(\phi, s_{\alpha})}{\omega + 1}$$

from which the minimum α is the solution to

$$\frac{d\Psi}{d\alpha} = 0 = \omega - \frac{h(\theta, s_{\alpha})}{h(0, s_{\alpha})}$$

where $h(\theta, s)$ is the noncentral density associated with s.

6. WEIGHTS LEADING TO STANDARD TYPE I AND TYPE II ERROR RATES

Suppose the sample size has been chosen on the basis of Equation (1), so that the experiment has been designed based on a fixed α_0 and, in order to achieve a power of $1 - \beta_0$, a sample size of n per group is required. Then given a value of ω, we can determine an optimal α from (4). A question of interest is which value of ω makes α_0 the solution of (3)? Answering this question allows us to understand what typical values of α_0 and β_0, for example

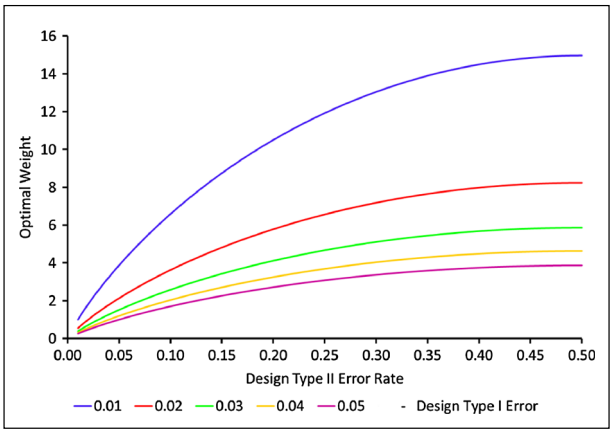

Figure 3. Optimal weights to give standard type I and type II errors.

0.05 for α_0 and 0.1 and 0.2 for β_0, imply in terms of the relative importance of type I and type II errors, that is, in terms of ω.

From (1), we have that

$$\frac{\sqrt{n}\delta_0}{\sigma} = \theta = Z_{1-\alpha_0} + Z_{1-\beta_0} = -Z_{\alpha_0} + Z_{1-\beta_0}$$

and substituting this into (3) gives

$$\omega = \frac{\phi\left(Z_{1-\beta_0}\right)}{\phi\left(Z_{\alpha_0}\right)}$$

Note that the solution is independent of δ_0. Furthermore, substitution of (5) into (2) shows that under these assumptions, the corresponding β will be precisely β_0. Figure 3 displays the value of ω that is required to give a range of values for α_0 and β_0 as the solution to Equation (4).

Often, we are interested in a one-sided type I error rate of 0.025 and a type II error rate of 0.1, a typical pair of values for phase III clinical trials run to achieve marketing authorisation of a new drug. These values give rise to an optimal weight of just over 3, implying that the cost of a type I error is three times that of a type II error. Figure 2 illustrates that this is indeed the optimal solution for this configuration of values. On the other hand, if the type II error rate is decreased to 0.05, then the relative cost changes, with the optimal weight being 1.76, in turn implying that the null hypothesis is *a priori* 75% more likely than a type II error.

In the unknown variance case, there is no analytic solution, and we need to resort to numerical approaches. Nonetheless, there are some general observations that can be made. First, the optimal ω is no longer independent of δ_0. Second, the optimal value of ω is larger than in the known variance case. Third, as n becomes large, corresponding to small values of δ_0, the optimal value of ω converges to the known variance case, which is no more than recognising that as n becomes large, we can replace the t-test by a z-test by using s^2, the sample estimate of σ^2, in place of σ^2 Figure 4 illustrates this latter point by showing the optimal weights for a t-density as a function of the standardised treatment effect δ_0/σ as well as the normal-density weight, which is independent of δ_0/σ.

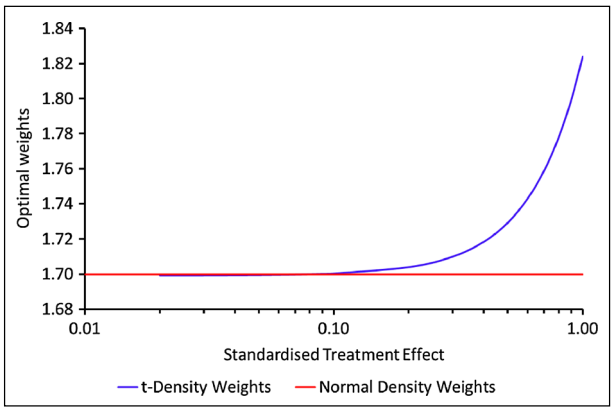

Figure 4. Optimal *t*-density weights as a function of the standardised treatment effect.

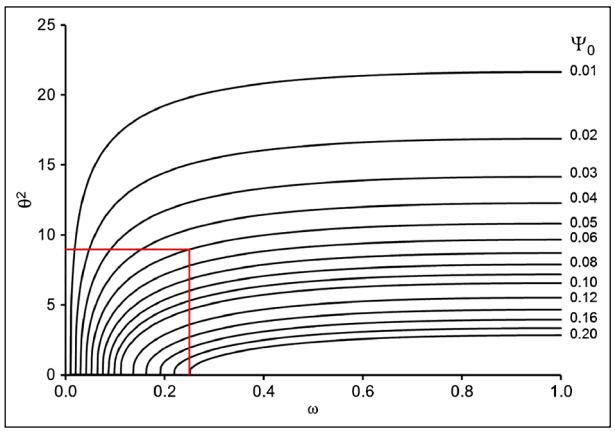

Figure 5. Sample size factor to control the weighted (ω or ω^{-1}) sum of errors to be $\leqslant \Psi_0$.

7. SAMPLE SIZING BASED ON MINIMISING THE SUM OF ERRORS

Mudge *et al.* made the suggestion that the sample size of a study could be determined from the requirement that the weighted sum of errors be minimised [2]. To see how this might be achieved, note that (6) is a function of ω and $\theta = \sqrt{n}\delta_0/\sigma$ and so can be written as $\Psi(\omega\theta)$. A couple of properties of $\Psi(\omega, \theta)$ are useful for determining the sample size. First, for a given value of ω, $\Psi(\omega, \theta) = \Psi(\omega^{-1}, \theta)$, and this simplifies the calculations because we need only solve for values of ω between 0 and 1, which then automatically provide solutions between 1 and infinity. Second, as $\theta \to 0$, $\Psi(\omega, \theta) \to \omega/(\omega + 1)$, which implies that there are values of $\Psi(\omega, \theta)$ that are unobtainable for a given ω.

Suppose that we wish to control the minimum weighted sum of errors at a maximum of Ψ_0 and that for a given ω, a solution is available in terms of $\theta = \sqrt{n}\delta_0/\sigma$. The appropriate sample size can then be determined as $n = \theta^2\sigma^2/\delta_0^2$, which has the same structural form as (1). There is no analytic solution for $\theta = \sqrt{n}\delta_0/\sigma$ in terms of ω and Ψ_0, but a numerical solution is trivially found using the method of *regula falsi*.

Figure 5 provides values of θ^2 as a function of ω and Ψ_0, thereby enabling the sample size to be determined. For example, suppose we wish to control the sum of errors to be no more than 0.05 and that the cost of a type I error is four times more costly than a type II error. Remembering that $\Psi(\omega, \theta) = \Psi(\omega^{-1}, \theta)$, we can read off the value of θ^2 as 9 so that for a standardised effect size of $\delta_0/\sigma = 0.5$, the required sample size is $n = 36$. By the traditional sample sizing approach, this corresponds to a one-sided $\alpha_0 = 0.025$ and $\beta_0 = 0.15$.

8. MINIMISING THE SUM OF ERRORS AND THE NEYMAN–PEARSON LEMMA

We have already noted that the original Neyman–Pearson lemma was developed for the case in which for a fixed type I error rate, α, a critical region is chosen so that the power, $1 - \beta$, is maximised. The critical region can be determined using a Lagrange multiplier. Now, suppose that instead of the usual conditions of the Neyman–Pearson lemma, we wish to choose a critical region to minimise the weighted average of the type I and type II error rates in which the weights are the costs of the errors ω_0 and ω_1. If $R(x)$ denotes the critical region based on data x, then the problem is to determine $R(x)$ to minimise

$$\begin{aligned}\Psi &= \omega_0 \text{Prob(type I error)} + \omega_1 \text{Prob(type II error)} \\ &= \omega_1 - \int_{R(x)} \left[\omega_1 p\left(x|H_1\right) - \omega_0 p\left(x|H_0\right)\right] dx\end{aligned}$$

Following the Lagrange multiplier proof of the original Neyman–Pearson lemma, the preceding expression can be minimised by choosing $R(x) = \{x : \omega_1 p\left(x|H_1\right) > \omega_0 p\left(x|H_0\right)\}$, which maximises the integral, corresponding to the region in which the likelihood ratio, λ, satisfies

$$\lambda = \frac{p(x|H_1)}{p(x|H_0)} > \frac{\omega_0}{\omega_1} = \omega \tag{8}$$

For our simple case, under the null hypothesis, the likelihood is given by

$$\begin{aligned}l\left(X; H_0\right) &\propto \left(\sigma^2\right)^{-\frac{n}{2}} \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}\left(x_i - \mu_0\right)^2\right\} \\ &= \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left((n-1)s^2 + n\left(\bar{x} - \mu_0\right)^2\right)\right\}\end{aligned}$$

and under the alternative hypothesis, it has the form

$$\begin{aligned}l\left(X; H_A\right) &\propto \left(\sigma^2\right)^{-\frac{n}{2}} \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}\left(x_i - \mu_0 - \delta_0\right)^2\right\} \\ &= \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left((n-1)s^2 + n\left(\bar{x} - \mu_0 - \delta_0\right)^2\right)\right\}\end{aligned}$$

from which the likelihood ratio is

$$\frac{l\left(X; H_A\right)}{l\left(X; H_0\right)} = \lambda = \exp\left\{-\frac{n}{2\sigma^2}\left[-2\left(\bar{x} - \mu_0\right)\delta_0 + \delta_0^2\right]\right\}$$

Requiring that $\lambda > \omega$ implies

$$\exp\left\{-\frac{n}{2\sigma^2}\left[-2\left(\bar{x} - \mu_0\right)\delta_0 + \delta_0^2\right]\right\} > \omega$$

which in turn implies

$$\frac{\sqrt{n}\left(\bar{x} - \mu_0\right)}{\sigma} > \sqrt{n}\frac{\delta_0}{2\sigma} + \frac{\sigma}{\sqrt{n}\delta_0}\ln(\omega) \tag{9}$$

and this is the optimal solution given by (5). Stephen Senn has pointed out that this approach introduces the type II error as a crucial element of inference. Whilst this is not immediately apparent from (9), it is implicit in that the criterion is a function of the standardised treatment effect – $\sqrt{n}\delta_0/\sigma$ – that the study has been planned to detect. This is entirely reasonable if one is interested in discriminating between $\mu = \mu_0$ and $\mu = \mu_0 + \delta_0$.

9. THE LIKELIHOOD PRINCIPLE AND SAMPLING FRAMES

What is the practical implication of this last result? In reality, this is a restatement of the likelihood principle, or it is another example of Senn's view 'that likelihood is really the more fundamental concept' [14]. In simple terms, the likelihood principle says that how the data are arrived at is irrelevant to the inferences that are to be drawn.

To illustrate this point, we modify an example given by Lindley and Philips [45]. Suppose a single-arm, open-label clinical trial is run with a primary endpoint that is binary, success or failure. We consider four scenarios:

Scenario 1. It is planned that in a fixed-sample study, 12 (n) patients are to be treated. Of the 12 patients treated, nine (r) respond successfully. Under the null hypothesis that the success is 50%, the p-value is calculated from

$$\sum_{i=9}^{12} \binom{12}{i} 0.5^{12} = 0.073$$

Scenario 2. It is planned that patients will be treated until nine (r) are treated successfully. The study is run, and the ninth success occurs when 12 (n) patients have been treated. Under the null hypothesis that the success is 50%, the p-value is calculated from the cumulative negative-binomial distribution function

$$\sum_{k=9}^{12} \binom{k-1}{9-1} 0.5^{k} = 0.033$$

Immediately, we can recognise the impact of the stopping rule on p-values because scenario 1 is significant, but scenario 2 is not.

Scenario 3. It is planned that patients will be recruited for a fixed period, say 2 weeks. At the end of the period, 12 patients have been recruited, of which nine are successfully treated. How can we determine a p-value? We could assume, for example, that the recruitment rate was such that the number of patients recruited in 2 weeks had a Poisson distribution with mean 10 and then look for more extreme cases. Some thought needs to be given as to how to define more extreme. For example, are both 8 successes out of 10 and 13 successes out of 15 more successful? If this definition is acceptable, then the p-value is 0.079. This p-value, however, depends on the Poisson mean assumption; if this were to be changed, then so would the p-value. If the Poisson mean is 5, the p-value increases to 0.180, whilst if it is 20, the p-value is reduced to 0.018.

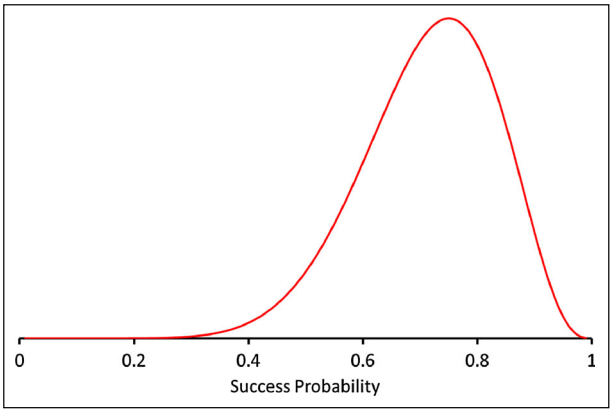

Figure 6. Likelihood function for the success probability of a single-arm experiment ($n = 12, r = 9$).

Scenario 4. It is planned to recruit 50 patients, but funding runs out when 12 patients have been recruited, of which nine are successfully treated. How should the p-value be calculated? We could condition on the sample size achieved, as an ancillary statistic, but is that appropriate? We cannot repeat the experiment, either in reality or as a thought experiment, and therefore, it is difficult to see how we can embed it in a sampling frame that would allow us to calculate a p-value. That was not the case in scenario 3 because we could contemplate the variable sample size and allow for it in the calculation of the p-value.

The calculation of the p-value was simple for scenarios 1 and 2, more complicated for scenario 3 and perhaps impossible for scenario 4. Despite these difficulties, the likelihood function for the unknown success proportion, displayed in Figure 6, is the same for each scenario:

$$\pi^9(1-\pi)^3$$

The likelihood function can be used to estimate π and to make inferences through a likelihood interval. Alternatively, if we were willing to entertain a prior distribution for π, Bayes' theorem can be used to combine the two sources of information to allow posterior inferences to be drawn.

The issue of the sampling frame can be a difficult problem for traditional statistical approaches, and these are not just purely academic or theoretical issues but arise in practical, medical research problems. Here are two illustrations.

The use of dynamic allocation techniques to keep treatment groups balanced with respect to a medium to large number of prognostic, stratification, factors has been a topic of continuing concern to regulatory authorities. The European Medicines Agency's 2003 guidance on baseline covariate adjustment remarked that these methods remained controversial and 'applicants are strongly advised to avoid such methods' [46]. The updated version of this guidance in 2013 is less dogmatic, requiring sponsors to consider carefully issues of bias and type I error control and suggesting that 'the use of re-randomization methods in the analysis should be considered' [47]. Whilst the perception has been that the Food and Drug Administration (FDA) has been more willing to accept the dynamic allocation, they have in the past required sponsors to justify calculating the p-value by ignoring the allocation process. Ebbutt *et al*. report a request by the FDA for a re-randomization analysis of a study that used

minimisation, a particular type of dynamic allocation, to show whether the *p*-value calculated by standard procedures ignoring the allocation method was appropriate. They report that the results were different if the randomisation test took account of the order in which patients entered the study [48].

A second example concerns a study that used a randomised play-the-winner (RPW) design introduced by Wei and Durham, a so-called urn model [49]. At the beginning of an RPW trial, an urn contains ε balls, each of two colours (blue and red), representing two treatments. When a patient is to be treated, a ball is chosen at random (with replacement) from the urn. When the patient's outcome is known, the urn content is updated as follows.

If the patient was allocated to treatment t and responded positively, φ balls of colour t are added to the urn; otherwise, γ of colour s (the complement of t) are added. In time, the urn will contain a higher proportion of coloured balls associated with the more successful treatment. This is called an RPW($\varepsilon, \varphi, \gamma$) design.

Bartlett *et al*. considered the treatment of neonates with severe respiratory failure for whom the expected outcome is death [50]. They proposed comparing extracorporeal membrane oxygenation (ECMO) with a traditional ventilator. Phase I trials had suggested a greater than 50% survival rate on ECMO compared with a less than 20% survival rate using an optimal conventional ventilator. An RPW design was chosen because of the speedy outcome, and the anticipated difference in response would imply a small sample size and also would make it ethically difficult to propose equal randomisation. The course of the trial is illustrated in Figure 7. The urn initially contained one blue and one red ball, and the trial proceeded as follows:

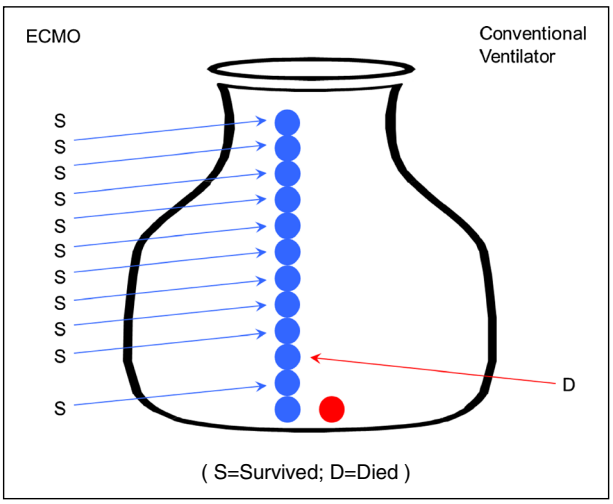

Figure 7. Randomised play-the-winner design comparing extracorporeal membrane oxygenation with a conventional ventilator for the treatment of neonatal respiratory failure [50].

(1) A patient was randomised to ECMO and survived. A blue ball was added to the urn.
(2) A patient was randomised to ventilator and died. A blue ball was added to the urn.
(3) A patient was randomised to ECMO and survived. A blue ball was added to the urn etc.

After 11 patients had been allocated to ECMO, and all survived, and one patient had been allocated to the conventional ventilator and died, the study was stopped by the independent safety committee, and statistical significance was declared.

This was an (in)famous study because of the extreme nature of the resulting imbalance between ECMO and the traditional ventilator groups. Could this have been avoided? Almost certainly. The problem lies in the choice of the starting configuration – namely one ball of each colour. In these circumstances, the randomisation imbalance in favour of either treatment is already 2:1 after the first patient's result is known, in this case in favour of ECMO. By the time the fifth patient was to be randomised, the allocation imbalance was 5:1 in favour of ECMO. It would have been preferable to start with three, four or five balls of each colour, or to conduct a permuted block of 10 patients before adaptive allocation began, thereby preventing a severe imbalance occurring too early in the course of the study. One of the issues with this study was the lack of acceptance of the results in the wider scientific community. As Dragalin has argued, adaptive designs, defined as multistage study designs that use accumulating data to decide on how to modify aspects of the study, need to do so without undermining the validity and integrity of the trial, and this includes 'credibility, interpretability, and persuasiveness of the study results to a broader scientific community' [51,52]. In the case of the ECMO study, one of the reasons that the results were not immediately accepted by the medical community was that there was no unanimity amongst statisticians as to the evidential value of the results. To illustrate, Figure 8 shows a range of *p*-values that have been proposed by statisticians [53–55]. These *p*-values are based on a range of assumptions with the following examples: (i) fixing both marginal totals, which leads to Fisher's exact test; (ii) an analysis that ignores the design and assumes complete randomisation; and (iii) an analysis conditioning on the observed sequence of responses. Some statisticians have argued that because the design was sequential without knowledge of the stopping rule, the sample space remains undefined and, as such, no significance test can be carried out and, therefore, no *p*-value calculated.

The associate editor suggested an alternative scenario in which a completely sequential experiment is run with an analysis after every patient. The study is planned to continue until the response rate is at least 75%, and the question 'how does the likelihood principle deal with the bias of this estimator' is raised. The likelihood principle has to do with observed data and not with data

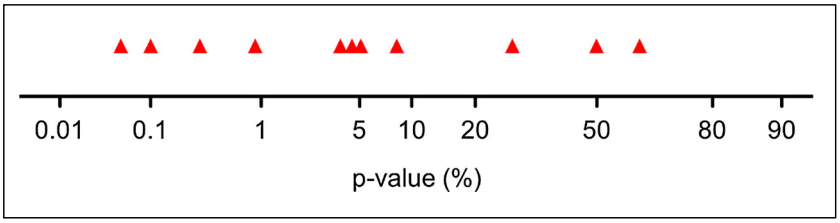

Figure 8. Proposed *p*-values associated with the results of the extracorporeal membrane oxygenation study.

that might have occurred but did not. There is clearly an issue here for a frequentist statistician, and I would probably agree with Senn, who in a different context has remarked that this 'is a weakness of the classical notion of unbiasedness but accept that given the conventional definition of unbiasedness there may be a problem' [56]. Of course, the fact that the likelihood principle holds that stopping rules are irrelevant to inference is not accepted by many statisticians, preeminent amongst them being Armitage: 'I feel that if a man deliberately stopped an investigation when he had departed sufficiently far from his particular hypothesis, then "Thou shalt be misled if thou dost not know that." If so, prior probability methods seem to appear in a less attractive light than frequency methods, where one can take into account the method of sampling' [57].

10. BAYESIAN CONSIDERATIONS

Pericchi and Pereira have pointed to a connection between the rejection region based on minimising the weighted sum of type I and type II errors and a Bayesian 'test of the null hypothesis' [58]. Suppose that π_0 and π_1 are the prior probabilities of the null hypothesis ($H_0 : \mu = \mu_0$) and the alternative hypothesis ($H_1 : \mu = \mu_0 + \delta_0$), respectively, then a simple application of Bayes' theorem produces the posterior probability of the null hypothesis in the form

$$P(H_0|x) = \frac{\pi_0 p(x|H_0)}{\pi_0 p(x|H_0) + \pi_1 p(x|H_1)}$$

A Bayesian 'test of the null hypothesis' in which rejection occurs if $P(H_0|x) < 0.5$ implies

$$\begin{aligned} P(H_0|x) = \frac{\pi_0 p(x|H_0)}{\pi_0 p(x|H_0) + \pi_1 p!(x|H_1)} &< \frac{1}{2} \\ \Rightarrow \pi_0 p(x|H_0) &< \pi_1 p(x|H_1) \\ \Rightarrow \frac{p(x|H_1)}{p(x|H_0)} &> \frac{\pi_0}{\pi_1} \end{aligned}$$

corresponding to the modified Neyman–Pearson criterion (8).

Pericchi and Pereira make two further contributions. First, they argue that the Lindley paradox is not primarily a distinction between Bayesian and non-Bayesian approaches but arises because of the traditional approach of fixing the type I error and maximising power. The paradox can be resolved if the weighted error approach is adopted. To see this, we can compare the traditional rejection rule

$$\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} > Z_{1-\alpha} \tag{10}$$

with the rejection rule based on weighted errors given by (9). In the case of the former, as the sample size increases, 'there is a positive probability of a false rejection' [58] because n can always be chosen to meet the rejection criterion (10). In contrast, as the sample size increases, so does the right-hand side of (9) with the consequence that the rejection criterion becomes ever more stringent so that in the limit, the type I error converges to zero as does the type II error.

Second, they generalize the result that for a test of a simple hypothesis versus simple hypothesis the optimal criteria to minimize the weighted sum of errors is the likelihood ratio with threshold given by the ratio of weights, to a general hypothesis:

$$H_0 : \mu \in \mathrm{M}_0 \text{ versus } H_1 : \mu \in \mathrm{M}_1$$

The corresponding solution replaces the likelihood ratio as the optimal criteria with the Bayes factor, with the threshold remaining the ratio of weights.

11. DISCUSSION

The distinction between statistical significance and scientific, biological or clinical relevance has long been recognised and remains a topic of interest to scientists and statisticians alike [59–63]. Whilst the former can to some degree be guaranteed by an appropriate choice of a large sample size, the latter is not in the control of the scientist or pharmaceutical sponsor except insofar that they can influence key opinion leaders within the relevant scientific community. Of course, the cost of 'guaranteeing' statistical significance will be high, and we have already argued that many sponsors would wish to increase the MCID in order to reduce the sample size – resource sample sizing. Part of the problem is that the traditional approach to statistical inference fixes the required level of evidence, in whatever context, independent of the sample size.

The example given by Peter Freeman indicates a serious consequence of this approach [64]. In the example, a constant p-value of 0.041 could be achieved for treatment estimates ranging from 0.25 to 0.000722 with a 100 000-fold increase in sample size. Partially on the basis of this example, Peter Freeman argues that we should

> Allow for sample size when interpreting any p-value. A p-value of 0.05 from a small sample can be quite strong evidence, but one of 0.05 from a large sample is always very weak evidence and in extreme cases provides evidence in favour of the hypothesis. Always require a p-value of 0.001 or less in large samples (n >200, say) before declaring anything significant. [64]

This quote supports the discussion in the previous section of the decision criterion (9). Whilst it is in general true that increasing the sample size should require a more stringent criterion for rejecting the null hypothesis, the absolute amount of evidence depends on the ratio of costs. The claim by Freeman that we need to require an extremely small p-value for sample sizes greater than 200 is based on likelihood ratio threshold of 4; in other words, he is effectively assuming that type I errors are four times more costly than type II errors. For other assumptions, a different level of evidence will be required but will still increase with sample size.

Freeman suggests that the argument is not understood by statisticians because the standard practice in group sequential designs is to begin with a small nominal significance level at the first interim and then to increase the nominal significance level as the study progresses through the subsequent interims to the final analysis, whereas the idea that the evidence criterion should become more stringent with increasing sample size suggests the converse.

The 2014 Ebola virus outbreak has highlighted the need to take into account the consequences of type I and type II errors. The treatment of infected international medical staff with the unapproved and untested-in-humans experimental drug ZMapp is no doubt related to the high mortality rates associated with the virus.

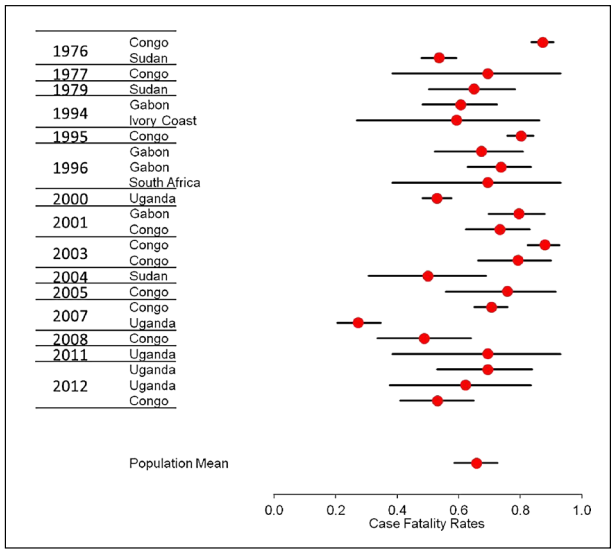

Figure 9. Chronology of Ebola virus outbreaks and case fatality rates.

Figure 9 shows a Bayesian hierarchical meta-analysis of case fatality rate data taken from the WHO (www.who.int/mediacentre/factsheets/fs103/en/). If and when ZMapp is subjected to testing in a randomised trial, what should be the level of evidence that is required to establish its efficacy? The approach outlined in this paper provides a rationale for reducing the level of evidence required in recognising the particular importance of not committing a type II error.

An alternative to increasing the type I error would be to formally incorporate the historical data provided in Figure 9 into the analysis of a future study. The use of historical data in this way is not new, and their use in both preclinical [65] and clinical contexts [66] goes back nearly 40 years. Pharmaceutical statisticians have been at the forefront of recent developments in Bayesian approaches to the use of historical information, particularly control information. Examples of these developments are the power prior [67], although there are issues with its original formulation [68–70], predictive priors [71] and commensurate priors [72]. Viele *et al*. provide a recent, useful review [73]. This work is exclusively Bayesian, but there are some traditional, frequentist, analogues. Tarone, for example, constructs a test for trend in proportions based on a Cochran–Armitage statistic with an adjustment to the concurrent control data depending upon the historical information [74]. Of course, the combination proposal could also be applied to such tests.

The general result (6) does not depend on the particular example, and it is not new. Lindley and Savage in joint work pointed out that in determining a particular pair of type I and type II error rates, the only form that indifference curves can have is parallel straight lines in which the slope is the negative of the ratio of the priors associated with the null and alternative hypotheses [75–77]. Cornfield considered minimising a linear function of the two errors, in which the slope parameter 'measures the undesirability, or the cost, of an error of the first kind relative to one of the second kind' and stated that it is easy to show that the rejection region for such a case must consist of all likelihood ratio values that exceed the fixed slope [78]. DeGroot argued that it is better to minimise a weighted sum of type I and type II error than to specify a value of type I error and then minimise type II error [79]. More recently Bernardo and Smith have noted the result, pointing out that 'minimising a linear combination of the two types of error is the only coherent way of making a choice' of α and β 'in the sense that no other procedure is equivalent to minimising an expected loss' [80]. Finally, Spiegelhalter *et al*. refer to Cornfield's paper, in particular the consequence of the results for sequential analyses [81]. They quote Cornfield:

> …it is clear that the entire basis for sequential analysis depends upon nothing more profound than a preference for minimizing β for given α rather than minimizing their linear combination. Rarely has so mighty a structure and one so surprising to scientific common sense, rested on so frail a distinction and so delicate a preference. [78]

ACKNOWLEDGEMENTS

I am grateful to Luis Pericchi for instructive discussions on the topic of this paper and for providing relevant unpublished papers. I am also indebted to the associate editor and two referees for helpful comments on an earlier draft.

REFERENCES

[1] Mudge JF, Baker LF, Edge CB, Houlahan JE. Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS ONE* 2012; **7**:e32734.
[2] Mudge JF, Barrett TJ, Munkittrick KR, Houlahan JE. Negative consequences of using $=$ 0.05 for environmental monitoring decisions: a case study from a decade of Canada's environmental effects monitoring program. *Environmental Science and Technology* 2012; **46**:9249–9255.
[3] Mudge JF, Edge CB, Baker LF, Houlahan JE. If all of your friends used $\alpha = 0.05$, would you do it too? *Integrated Environmental Assessment and Management* 2012; **8**:563–564.
[4] Baker LF, Mudge JF. Making statistical significance more significant. *Significance* 2012; **9**(3):29–30.
[5] Peterman RM. Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Sciences* 1990; **47**:2–15.
[6] Peterman RM, M'Gonigle M. Statistical power analysis and the precautionary principle. *Marine Pollution Bulletin* 1992; **24**:231–234.
[7] Mapstone BD. Scalable decision rules for environmental impact studies: effect size, type I, and type II errors. *Ecological Applications* 1995; **5**:401–410.
[8] Field SA, Tyre AJ, Jonze N, Rhodes JR, Possingham HP. Minimizing the cost of environmental management decisions by optimizing statistical thresholds. *Ecology Letters* 2014; **7**:669–675.
[9] Field SA, O'Connor PJ, Tyre AJ, Possingham HP. Making monitoring meaningful. *Australian Ecology* 2007; **32**:485–491.
[10] Gigerenzer G. *Adaptive Thinking: Rationality in the Real World*. Oxford University Press: New York, 2000.
[11] Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L. *The Empire of Chance*. Cambridge University Press: Cambridge, UK, 1989.
[12] Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London Series A* 1933; **231**:289–337.
[13] Oakes M. *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. John Wiley & Sons: Chichester, UK, 1986.
[14] Senn S. You may believe you are a Bayesian but you are probably wrong. *Rationality Markets and Morals* 2011; **2**:48–66.
[15] International Conference of Harmonisation. E9: statistical principles for clinical trials, 1996. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf (Accessed July 2014).
[16] Lindley DV. A statistical paradox. *Biometrika* 1957; **44**:187–192.
[17] Bartlett MS. A comment on D.V. Lindley's statistical paradox. *Biometrika* 1957; **44**:533–534.
[18] Deming WE. *Statistical Adjustment of Data*. John Wiley and Sons: New York, 1943.

[19] Chow SL. *Statistical Significance: Rationale, Validity and Utility*. Sage Publications: London, 1996.
[20] Morrison DE, Henkel RE (eds.) *The Significance Test Controversy*. Aldine: Chicago, 1970.
[21] McCloskey DN, Ziliak ST. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press: Ann Arbor, 2008.
[22] International Committee of Medical journal. Uniform requirements for manuscripts submitted to medical journals. *Annals of Internal Medicine* 1998; **108**:258–265 (*British Medical Journal* 1998; 296:101-105).
[23] Gardner MG, Altman DG. *Statistics with Confidence*. BMJ Publishing Group: London, 1989.
[24] Altman DG, Machin D, Bryant TN, Gardner MG. *Statistics with Confidence* (2nd edn). British Medical Journal Publications: London, 2000.
[25] Altman DG, Bland JM. Absence of evidence is not evidence of absence. *British Medical Journal* 1995; **311**:485.
[26] Hanley JA, Lippmann-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. *Journal of the American Medical Association* 1983; **249**:1743–1745.
[27] Braitman LE. Confidence intervals extract clinically useful information from data. *Annals of Internal Medicine* 1988; **108**:296–8.
[28] Grieve AP. Letter to the Editor. *Royal Statistical Society News and Notes* 1992; **18**(7):3–4.
[29] Fisher RA. *The Design of Experiments* (8th edn). Oliver and Boyd: Edinburgh, 1966.
[30] Fisher RA. *Statistical Methods and Scientific Inferences*. Hafner: New York, 1956.
[31] Bennett JH. *Statistical Inference and Analysis Selected Correspondence of R.A. Fisher*. Oxford University Press: Oxford, 1990.
[32] Neyman J. A note on an article by Sir Ronald Fisher. *Journal of the Royal Statistical Society Series B* 1956; **18**:288–294.
[33] Fisher RA. Statistical methods and scientific inference. *Journal of the Royal Statistical Society Series B* 1955; **17**:69–78.
[34] Cohen J. The statistical power of abnormal social psychological research: a review. *Journal of Abnormal and Social Psychology* 1962; **65**:145–153.
[35] Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial – survey of 71 negative trials. *New England Journal of Medicine* 1978; **299**:690–694.
[36] Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. *Journal of the American Medical Association* 2002; **288**:358–362.
[37] Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nature Drug Discovery* 2004; **3**:711–715.
[38] Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. *Nature Biotechnology* 2014; **32**:40–51.
[39] Palmer CR. Ethics, data-dependent designs, and the strategy of clinical trials: time to start learning-as-we-go? *Statistical Methods in Medical Research* 2002; **11**:381–402.
[40] Palmer CR, Rosenberger WF. Ethics and practice: alternative designs for phase III randomized clinical trials. *Controlled Clinical Trials* 1999; **20**:172–186.
[41] World Medical Association. 52nd Assembly, *Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects*. World Medical Association: Edinburgh, 2000.
[42] Edwards SJL, Lilford RJ, Braunholtz D, Jackson J. Why 'underpowered' trials are not necessarily unethical. *The Lancet* 1997; **350**:804–807.
[43] Lachin JM. Sample size determinations for r × c comparative trial. *Biometrics* 1977; **33**:315–324.
[44] Chuang-Stein C, Kirby S, Hirsch I, Atkinson G. The role of the minimum clinically important difference and its impact on designing a trial. *Pharmaceutical Statistics* 2011; **10**:250–256.
[45] Lindley DV, Phillips LD. Inference for a Bernoulli process (a Bayesian view). *The American Statistician* 1976; **30**:112–129.
[46] EMEA, 2003. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003639.pdf (Accessed July 2014).
[47] EMA, 2013. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/06/WC500144946.pdf (Accessed July 2014).
[48] Ebbutt A, Kay R, McNamara J, Engler J. The analysis of trials using a minimisation algorithm. *Statisticians in the Pharmaceutical Industry Annual Conference Report*, London: PSI, 1997; pp. 12–15.
[49] Wei LJ, Durham S. The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association* 1978; **73**:838–43.
[50] Bartlett RH, Roloff DW, Cornell RG, Andrews AF, Dillon PW, Zwischenberger JB. Extracorporeal circulation in neonatal respiratory failure: a prospective randomised trial. *Paediatrics* 1985; **76**:479–487.
[51] Dragalin V. Adaptive designs: terminology and classification. *Drug Information Journal* 2006; **40**:425–435.
[52] Dragalin V. Sequential methods in multi-arm clinical trials. *Sequential Analysis* 2010; **29**:444–462.
[53] Begg CB. On inferences from Wei's biased coin design for clinical trials (with discussion). *Biometrika* 1990; **77**:67–484.
[54] Wei LJ. Exact two-sample permutation tests based on the randomized play-the-winner rule. *Biometrika* 1988; **75**:603–606.
[55] Ware JH. Investigating therapies of potentially great benefit: ECMO. *Statistical Science* 1989; **4**:298–306.
[56] Senn S. A note regarding meta-analysis of sequential trials with stopping for efficacy. *Pharmaceutical Statistics* 2014; DOI: 10.1002/pst.1639.
[57] Armitage P. *Contribution to the discussion in The Foundations of Statistical Inference* Savage L (ed.) Methuen: London, 1962.
[58] Pericchi LR, Pereira CAB. Changing the paradigm of fixed significance levels. Testing hypothesis by minimizing sum of errors type I and type II. *Brazilian Journal of Probability and Statistics* 2014; http://imstat.org/bjps/papers/BJPS257.pdf.
[59] Bhardwaj SS, Camacho F, Derrow A, Fleischer AB, Feldman SR. Statistical significance and clinical relevance: the importance of power in clinical trials in dermatology. *Archives of Dermatology* 2004; **140**:1520–1523.
[60] Martínez-Abraín A. Statistical significance and biological relevance: a call for a more cautious interpretation of results in ecology. *Acta Oecologica* 2008; **34**:9-11.
[61] Panagiotakos DB. The value of *p*-value in biomedical research. *The Open Cardiovascular Medicine Journal* 2008; **2**:97–99.
[62] Tajer CD. Therapeutic trials, statistical significance, and clinical relevance. *Revista Argentina Cardiologica* 2010; **78**:385–390.
[63] Kaul S, Diamond GA. Trial and error: how to avoid commonly encountered limitations of published clinical trials. *Journal of the American College of Cardiology* 2010; **55**:415–427.
[64] Freeman PR. The role of *P*-values in analysing trial results. *Statistics in Medicine* 1993; **12**:1443–1452.
[65] Dempster AP, Selwyn MR, Weeks BJ. Combining historical and randomized controls for assessing trends in proportions. *Journal of the American Statistical Association* 1983; **78**:221–227.
[66] Pocock SJ. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases* 1976; **29**: 175–188.
[67] Ibrahim JG, Chen MH. Power prior distributions for regression models. *Statistical Science* 2000; **15**:46–60.
[68] Duan YY, Smith EP. Evaluating water quality using power priors to incorporate historical information. *Environmetrics* 2006; **17**: 95–106.
[69] Duan YY, Smith EP, Ye KY. Using power priors to improve the binomial test of water quality. *Journal of Agricultural, Biological and Environmental Statistics* 2006; **11**:151–168.
[70] Neuenschwander B, Branson M, Spiegelhalter D. A note on the power prior. *Statistics in Medicine* 2009; **28**:3562–3566.
[71] Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter D. Summarizing historical information on controls in clinical trials. *Clinical Trials* 2010; **7**:5–18.
[72] Hobbs BP, Carlin BP, Sargent DJ. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis* 2012; **7**:1–36.
[73] Viele K, Berry S, Neuenschwaner B, Amzal B, Chen F, Enas N, Hobbs B, Ibrahim JG, Kinnersley N, Lindborg S, Micallef S, Roychoudhury S, Thompson L. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics* 2014; **13**:41–54.
[74] Tarone RE. The use of historical control information in testing for a trend in proportions. *Biometrics* 1982; **38**:215–220.
[75] Savage LJ. The foundations of statistics reconsidered. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Neyman J (ed.), Vol. 1. University of California Press: Berkeley; 1961; pp. 575–586.

[76] Savage LJ. Bayesian Statistics. In *Recent Developments in Information and Decision Processes*, Machol RE, Gray PE (eds). Macmillan: New York; 1962; pp. 161–194.
[77] Lindley DV. *Bayesian Statistics: A Review*. SIAM: Philadelphia, 1972.
[78] Cornfield J. Sequential trials, sequential analysis and the likelihood principle. *The American Statistician* 1966; **20**:19–23.
[79] DeGroot MH. *Probability and Statistics* (2nd Edn). Addison-Wesley: New York, 1975.
[80] Bernardo J, Smith AFM. *Bayesian Theory*. John Wiley & Sons: Chichester, UK, 1994.
[81] Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials & Health-Care Evaluation*. John Wiley & Sons: Chichester, UK, 2003.