PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. 2005; **4**: 187–201
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/pst.175

Assurance in clinical trial design

MAIN PAPER

Anthony O'Hagan[1,*,†], John W. Stevens[2] and Michael J. Campbell[3]
[1] *Centre for Bayesian Statistics in Health Economics, University of Sheffield, UK*
[2] *AstraZeneca R&D Charnwood, Loughborough, UK*
[3] *School of Health and Related Research, University of Sheffield, UK*

Conventional clinical trial design involves considerations of power, and sample size is typically chosen to achieve a desired power conditional on a specified treatment effect. In practice, there is considerable uncertainty about what the true underlying treatment effect may be, and so power does not give a good indication of the probability that the trial will demonstrate a positive outcome.

Assurance is the unconditional probability that the trial will yield a 'positive outcome'. A positive outcome usually means a statistically significant result, according to some standard frequentist significance test. The assurance is then the prior expectation of the power, averaged over the prior distribution for the unknown true treatment effect.

We argue that assurance is an important measure of the practical utility of a proposed trial, and indeed that it will often be appropriate to choose the size of the sample (and perhaps other aspects of the design) to achieve a desired assurance, rather than to achieve a desired power conditional on an assumed treatment effect. We extend the theory of assurance to two-sided testing and equivalence trials. We also show that assurance is straightforward to compute in some simple problems of normal, binary and gamma distributed data, and that the method is not restricted to simple conjugate prior distributions for parameters. Several illustrations are given. ***Copyright © 2005 John Wiley & Sons, Ltd.***

Keywords: ***assurance; Bayesian analysis; Bayesian clinical trial simulation; binary data; design of experiments; expected power; power; preposterior analysis; prior distribution; sample size***

1. INTRODUCTION

Randomized controlled trials are often conducted with the primary objective of comparing the efficacy of two treatments, and the specific objective may be to demonstrate the *superiority* of one treatment over the other, the *non-inferiority* of one treatment with respect to the other, or the *equivalence* of the two treatments. A clinical trial can be thought of as having three distinct phases, the design phase, the conduct phase during which the data are collected, and the analysis phase. During the analysis phase, we make statistical inferences by calculating point estimates, *p*-values

*Correspondence to: Anthony O'Hagan, Centre for Bayesian Statistics in Health Economics, University of Sheffield, Hicks Building, Sheffield S3 7RH, UK.
† E-mail: a.ohagan@sheffield.ac.uk

Pharmaceutical STATISTICS

(posterior probabilities) and confidence (credible) intervals depending on whether we are adopting a frequentist or Bayesian perspective.

At the design phase the trial depends on many things, including its objectives. A conventional (frequentist) design depends on the null and alternative hypotheses, the significance level of the test (the Type I error) and the power of the test (1 minus the Type II error). As usual, the Type I error is the probability of rejecting the null hypothesis, H_0, when it is actually true, and the Type II error is the probability of not rejecting the null hypothesis when an alternative hypothesis, H_1, is true.

1.1. Power in clinical trial design

An important component of conventional (frequentist) design is the concept of power. In general, once the structure of the design, the hypothesis to be tested and the significance level required for the test (usually set at 5%) have been decided upon, the sample size is determined by considering the power of the test. Specifically, the sample size is chosen to be just large enough to achieve a desired power at some specified value of the treatment effect (and any other required parameters such as the population variance(s)). Power, the probability that we will reject the null hypothesis if the true treatment effect equals the requisite value [1, 2], is conventionally set at 80% or 90%, and it is often recommended that this value is made to be as large as possible, particularly when trials are difficult or impossible to repeat.

The specification 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, where this is larger' [3]. However, there is no guarantee that the true underlying treatment effect will be equal to (or greater than) the assumed value. Therefore, power does not quantify the probability that the trial will actually end in the null hypothesis being rejected. Yet, from the perspective of the trial sponsor, this probability is extremely important when deciding whether to conduct the trial at all or what sample size to choose for the trial.

We begin by assuming that we wish to design a clinical trial to test a null hypothesis against some alternative hypothesis. Let the event R denote rejection of the null hypothesis. The conventional (frequentist) approach to determining sample size proceeds by specifying a desired power, $\pi^* = 1 - \beta$, and assumed parameter values, $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, corresponding to the specified treatment effect (and any other required parameters), and the sample size is chosen to solve

$$\pi(\boldsymbol{\theta}^*) = \pi^*, \tag{1}$$

where $\pi(\cdot)$ denotes the power function

$$\pi(\boldsymbol{\theta}) = P(R \mid \boldsymbol{\theta}).$$

This is the *conditional* probability of R, conditional on the unknown true value of the parameter(s), for example the unknown but true treatment effect.

It is immediately clear that the precise specification of the treatment effect is critical in determining the ability of the trial to reject the null hypothesis. As discussed above, the choice of a 'clinically relevant' treatment difference is not unambiguous. Other unknown parameters such as sampling variances must be fixed at prior estimated values. Thus, in practice, the choice of $\boldsymbol{\theta}^*$ inevitably has an element of arbitrariness. In recognition of this, the sensitivity of a sample size calculation is often evaluated by varying the assumptions over some range. However, evidence that is available from published data, previous trials or expert opinion will indicate the relative plausibility of these different assumed parameter values and it is unlikely that these will be equally plausible. A synthesis of the available evidence for the treatment effect (and any other required parameters) should be regarded as an important contribution to the design of a clinical trial and this evidence should be incorporated into sample size determination.

1.2. Assurance in clinical trial design

As an alternative to power, Brown *et al.* [4] advocated calculating probabilities such as $P(R \mid E_2)$, the probability of rejecting the null hypothesis *conditional* on the event, E_2, that treatment 2 is genuinely superior. However, like power this is a conditional probability because it makes the assumption that treatment 2 is superior, and gives an over-optimistic estimate of the probability of actually rejecting the null hypothesis.

O'Hagan and Stevens [5] instead specified a desired 'assurance' of an outcome, γ^*, as the *unconditional* probability that a trial will lead to a specific outcome and chose the sample size to satisfy

$$\gamma = \gamma^*,$$

where $\gamma = P(R)$. The term 'assurance' and the symbol γ are used to emphasize that it is an *unconditional* probability and to distinguish it from power, $1 - \beta$. In fact, the assurance of the event R is the *expected power*

$$P(R) = E(\pi(\boldsymbol{\theta})), \tag{2}$$

where the expectation is with respect to the (prior) probability distribution of $\boldsymbol{\theta}$. The use of assurance in this way avoids the need to condition on a fixed treatment effect, $\boldsymbol{\theta}^*$, at the design stage, rather it quantifies the ability of the trial to achieve a desired outcome *based on the available evidence*.

The concept of assurance, although not necessarily this terminology, has been used in work dating back to the 1980s. The idea of computing the Bayesian prior assurance of success in a trial whose data will be analysed in conventional frequentist ways was described by Spiegelhalter and Freedman [6] and is called the 'hybrid classical–Bayesian' approach by Spiegelhalter *et al.* [7, Section 6.5]. It has also been called 'expected power' in early work, or 'average power' [8].

The concept of power has proved to be extremely valuable in the development of methodology for designing clinical trials, and its familiarity will no doubt ensure its continued use by many. Nevertheless, we believe that assurance is a more relevant measure, particularly from the perspective of the trial sponsor, and that this will be increasingly recognized as practitioners become familiar with its interpretation and usage. In practice, at least until assurance is more widely used, trial designers may find the consideration of both measures useful.

1.2.1. Parameter uncertainty

The use of assurance at the design stage entails a Bayesian perspective because it requires the specification of a prior distribution for the unknown population parameters. The prior distribution expresses prior *uncertainty* about the fixed but unknown true value of $\boldsymbol{\theta}$ corresponding to the specified treatment effect (and any other required parameters), and describes the relative plausibility of different parameter values. The prior probability that, say, a treatment effect δ exceeds some value δ^* represents a degree of confidence that this critical value will be exceeded.

Adopting a Bayesian perspective at the design stage is completely reasonable because the design of a trial is the sponsor's decision and its outcome is the sponsor's risk, so the design should take into account all information available to the sponsor to increase the chance that the trial will achieve a satisfactory outcome. In general, we will assume that at the analysis stage the analysis will be conducted based on frequentist significance tests or confidence intervals without incorporating the prior information with the trial data, except in Section 6, where we indicate how assurance can also be used in conjunction with a Bayesian analysis of the trial data. The prior distribution is used simply to determine the chance of obtaining relevant trial outcomes.

If the prior information for the unknown treatment effect is sufficiently strong, then the required sample size may be zero because the prior information alone is sufficient to produce a positive outcome without the need for any more sample data. It is clearly undesirable and unrealistic to express prior information this strong, particularly when the intention is to generate data

Pharmaceutical STATISTICS

for submission to a drug regulatory authority. On the other hand, if the prior information is weak then even infinite sample sizes may not be sufficient to achieve a desirable assurance of a positive outcome. In fact, as will be seen in several examples later in the paper, the prior distribution provides an upper bound on the achievable assurance of a positive outcome, unlike power where it is possible to generate any desirable value with suitable specification of the assumptions.

The specification of the prior distribution may be little more than an approximate judgement, in which case the derived assurances will also be approximate but may nevertheless be of real value. Where assurances are important, perhaps because the clinical trial is large enough for there to be a financial risk if it has a low probability of achieving a successful outcome, or because they contribute to strategic decisions regarding further development, it will be appropriate to spend substantial effort on specifying the prior distribution. An experienced facilitator might work with a number of experts to elicit the best available clinical and scientific judgements. However, it is beyond the scope of this paper to deal further with the process of specifying prior distributions; the interested reader is referred to the wide-ranging review of Garthwaite *et al*. [9].

1.2.2. Applications of assurance

Assurance can be applied to any meaningful outcome of a trial, and need not have any direct connection to power. For example, significance levels in superiority trials are usually two-sided, so that the null hypothesis of no treatment difference is rejected if the data suggest sufficiently strongly either (a) that treatment 1 is superior or (b) that treatment 2 is superior. To the trial sponsor, these may not be equivalent outcomes. If we let R_i denote the outcome of rejecting the null hypothesis with data suggesting that treatment i is superior, then we might separately be interested in the assurances $\gamma_1 = P(R_1)$ and $\gamma_2 = P(R_2)$. In this case, the overall assurance $\gamma = P(R) = \gamma_1 + \gamma_2$ of rejecting the null hypothesis is the expected power, but the individual $P(R_i)$ have no direct relation to the usual definition of power for such a trial.

Power is conventionally determined at an assumed value of a single outcome such as the primary efficacy variable. However, the outcome may be defined using more complex decision rules including safety and efficacy measurements.

The outcome may also be defined to involve the true parameter value(s) such as the outcome that the null hypothesis is rejected *and* the null hypothesis is actually false. For example, a trial sponsor may wish to see the experimental treatment demonstrated to be superior, since that may lead to regulatory approval, but the benefit to the company will be seriously compromised if it later turns out that the drug is not actually superior to current treatments. In this case, if E_2 denotes the event that treatment 2 is genuinely superior to treatment 1, then interest is in the assurance $P(R_2, E_2)$.

Assurances figure naturally in more formal decision-theoretic optimal design. For instance, the financial returns from developing a drug may depend on many different outcomes such as the event of getting a statistically significant result in favour of the experimental treatment in a Phase 3 clinical trial, getting a positive expected net benefit in a cost-effectiveness analysis, or having a sufficiently large treatment effect to obtain market penetration against existing competitors.

Analysis using assurance to date has been confined to the outcome of a positive result in one-sided testing, where data are normally distributed with known variance, and where a conjugate normal prior distribution is assumed for the unknown treatment effect. In this paper, we emphasize the generality and value of computing assurances for all kinds of trial outcomes. Specifically, we explore extensions to two-sided testing and to non-inferiority and equivalence trials; to outcomes dependent on combinations of the data and the true values of parameters; to normally distributed data with unknown variance; to binary data; and to non-conjugate prior distributions. Assurance is a key input to decision-making during drug development and provides a reality check on other methods of trial design. Whether

they are directly related to power calculations or not, the key feature of assurances is that they are unconditional probabilities, and so do not condition on the true value of $\boldsymbol{\theta}$. As such, assurances of relevant trial outcomes have great value in the planning of trials.

Section 2 of this paper presents explicit formulae for assurances in the case of normally distributed data with known variance and normal prior distribution. The assumption of a conjugate prior distribution is quite restrictive, and as soon as we move from the simple framework of Section 2, whether in terms of the data structure or the prior distribution, we cannot obtain explicit formulae for assurances. Nevertheless, computation is straightforward and efficient using Bayesian clinical trial simulation as described in Section 3. Sections 4 and 5 present the calculation of assurance in simple randomized, non-sequential clinical trials with normally distributed (with unknown variances) and binary data. In Section 6 we discuss the practical use of assurance and review other Bayesian approaches to sample size calculations.

2. THE SIMPLE NORMAL CASE

2.1. Data-based outcomes

Suppose that a randomized controlled trial is to be conducted with patients randomized to one of two treatments, with n_i patients allocated to treatment i $(i = 1, 2)$, and suppose that the jth patient receiving treatment i will yield a continuous response x_{ij} that we can assume is normally distributed. Thus, the model is that

$$x_{ij} \sim N(\mu_i, \sigma_i^2), \text{ independent.}$$

We address first the case where the population variances σ_i^2 are assumed known, so that the only unknown parameters are the population means, μ_1 and μ_2. The sufficient statistics are the sample means $\bar{x}_1$ and $\bar{x}_2$, having sampling distributions $\bar{x}_i \sim N(\mu_i, \sigma_i^2/n_i)$ conditional on the parameters. All the usual significance tests are based on the difference $\bar{x}_2 - \bar{x}_1$, whose sampling distribution is $N(\delta, \tau^2)$, where

$$\delta = \mu_2 - \mu_1, \qquad \tau = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2},$$

and τ is known. This is the conditional distribution of the same mean difference which is used in computing power, but for assurance we need the unconditional distribution.

Now suppose that the prior distribution for the effect difference δ has the conjugate normal form $\delta \sim N(m, v)$. Then the unconditional distribution is

$$\bar{x}_2 - \bar{x}_1 \sim N(m, \tau^2 + v).$$

From this we can compute the assurance of rejecting the null hypothesis in any of the standard significance tests (or of obtaining any other outcome defined in terms of $\bar{x}_2 - \bar{x}_1$).

2.1.1. *One-sided superiority trial*

A one-sided $100\alpha\%$ significance test of the null hypothesis that $\delta = 0$ (or $\delta \leqslant 0$) against the alternative that $\delta > 0$ will reject the null hypothesis if $\bar{x}_2 - \bar{x}_1 > \tau Z_\alpha$, where Z_α is the upper $100\alpha\%$ significance point of the standard normal distribution. The assurance of this outcome is

$$\gamma = P(\bar{x}_2 - \bar{x}_1 > \tau Z_\alpha) = \Phi\left(\frac{-\tau Z_\alpha + m}{\sqrt{\tau^2 + v}}\right), \tag{3}$$

where Φ denotes the standard normal distribution function (and, in particular, $\Phi(-Z_\alpha) = \alpha$).

2.1.2. *Two-sided superiority trial*

The $100\alpha\%$ two-sided test of the null hypothesis $\delta = 0$ against the two-sided alternative $\delta \neq 0$ rejects the null hypothesis if $|\bar{x}_2 - \bar{x}_1| > \tau Z_{\alpha/2}$. As in the one-sided case, the assurance that the null hypothesis is rejected with data favouring treatment 2 is

$$\gamma_2 = P(\bar{x}_2 - \bar{x}_1 > \tau Z_{\alpha/2}) = \Phi\left(\frac{-\tau Z_{\alpha/2} + m}{\sqrt{\tau^2 + v}}\right), \tag{4}$$

Pharmaceutical STATISTICS

while the assurance of a rejection with data favouring treatment 1 is

$$\gamma_1 = P(\bar{x}_2 - \bar{x}_1 < -\tau Z_{\alpha/2}) = \Phi\left(\frac{-\tau Z_{\alpha/2} - m}{\sqrt{\tau^2 + v}}\right). \quad (5)$$

The overall assurance of rejecting the null hypothesis is therefore $\gamma = \gamma_2 + \gamma_1$.

2.1.3. Non-inferiority trial

For non-inferiority testing, we set up the null hypothesis that $\delta \leqslant -\delta^*$, where δ^* is a clinically meaningful treatment difference, and test against the alternative that $\delta > -\delta^*$. The null hypothesis is rejected if $\bar{x}_2 - \bar{x}_1 > -\delta^* + \tau Z_\alpha$, with assurance

$$\gamma = P(\bar{x}_2 - \bar{x}_1 > -\delta^* + \tau Z_\alpha) = \Phi\left(\frac{-\tau Z_\alpha + m + \delta^*}{\sqrt{\tau^2 + v}}\right). \quad (6)$$

2.1.4. Equivalence trial

The conventional approach to testing equivalence [10] is to confirm the hypothesis of equivalence at the $100\alpha\%$ level if a $100(1-\alpha)\%$ two-sided confidence interval lies entirely within the region of equivalence. Letting the equivalence region be $[-\delta^*, \delta^*]$, this occurs if $-\delta^* + \tau Z_{\alpha/2} \leqslant \bar{x}_2 - \bar{x}_1 \leqslant \delta^* - \tau Z_{\alpha/2}$. The assurance of this is

$$\gamma = P(-\delta^* + \tau Z_{\alpha/2} \leqslant \bar{x}_2 - \bar{x}_1 \leqslant \delta^* - \tau Z_{\alpha/2}) = \Phi\left(\frac{\delta^* - \tau Z_{\alpha/2} - m}{\sqrt{\tau^2 + v}}\right) - \Phi\left(\frac{-\delta^* + \tau Z_{\alpha/2} - m}{\sqrt{\tau^2 + v}}\right). \quad (7)$$

2.2. Outcomes involving the true effect difference

As discussed previously, the outcome may also be defined to involve the true parameter value such as the outcome that the null hypothesis is rejected in favour of treatment 2 *and* that treatment 2 is indeed superior, $P(\bar{x}_2 - \bar{x}_1 > \tau Z_\alpha,\ \delta > 0)$. This assurance is for an event that involves the true effect difference ($\delta > 0$) as well as the test outcome ($\bar{x}_2 - \bar{x}_1 > \tau Z_\alpha$), and so the calculation requires the (unconditional) joint distribution of both $\bar{x}_2 - \bar{x}_1$ and δ. Their marginal distributions are $N(m, \tau^2 + v)$ and $N(m, v)$ respectively, and their joint distribution is easily found to be bivariate normal with these margins and covariance $\mathrm{cov}(\bar{x}_2 - \bar{x}_1, \delta) = v$.

Assurances can then be derived explicitly in terms of the bivariate standard normal distribution function

$$\Phi_2(a, b, r) = \int_{-\infty}^{a}\int_{-\infty}^{b} \frac{1}{2\pi\sqrt{1-r^2}} \exp - \left\{\frac{1}{2(1-r^2)}(x^2 + y^2 - 2rxy)\right\} \mathrm{d}x\, \mathrm{d}y.$$

For instance, for a $100\alpha\%$ one-sided superiority trial the assurance that the null hypothesis is rejected in favour of treatment 2 *and* the true value of δ is positive (and, in particular, the null hypothesis is false) is

$$P(\bar{x}_2 - \bar{x}_1 > \tau Z_\alpha,\ \delta > 0) = \Phi_2\left(\frac{-\tau Z_\alpha + m}{\sqrt{\tau^2 + v}}, \frac{m}{\sqrt{v}}, \sqrt{\frac{v}{\tau^2 + v}}\right). \quad (8)$$

2.3. Example 1

We suppose that a Phase 2 superiority trial is to be conducted to assess the effect of a new drug in reducing C-reactive protein (CRP) in patients with rheumatoid arthritis. CRP is a marker for disease severity, so this trial is intended to indicate the potential for the new drug in delivering clinically meaningful benefits. The outcome variable is a patient's reduction in CRP after 4 weeks relative to baseline, and the principal analysis will be a two-sided test of superiority at the 5% significance level.

The population variance of CRP reduction is assumed to be known and equal in the two patient groups, with values $\sigma_1^2 = \sigma_2^2 = 0.0625$. We suppose that the analyst specifies that the test is required to have 80% power (i.e. $\pi^* = 0.8$) to detect a

treatment effect of $\delta^* = 0.2$, leading to a proposed trial size of $n_1 = n_2 = 25$ patients (see [2]).

For the calculation of assurance, we suppose that the elicitation of prior information about the unknown treatment effect δ from a relevant expert gives the mean, $E(\delta) = 0.2$, and variance, $\text{var}(\delta) = 0.06$. If we assume a normal prior distribution for δ, we can compute assurances from (4) and (5) with $m = 0.2$, $v = 0.06$ and $\tau^2 = 0.125/n$, assuming that the trial is required to have equal sample sizes $n_1 = n_2 = n$ in each treatment group. With $n = 25$, we find $\gamma_2 = 0.595$. Thus, even though the prior expectation of δ equals the value, 0.2, at which the trial was found to have 80% power, there is only 60% assurance of a positive significant result. The explanation is clearly seen in Figure 1, which shows the power function (strictly, the power function for a 2.5% one-sided test) and the prior density of δ. The assurance is the expectation of the power function (solid line) with respect to the prior (dotted). There is a large prior probability of δ taking a value for which the power is very much lower than 0.8. In particular, there is a 0.207 prior probability that the new drug is less effective than placebo.

In fact, even with infinite sample sizes, where we will learn the true value of the treatment effect δ, the assurance cannot exceed $\Phi(0.2/\sqrt{0.06}) = 0.793$, that is, the prior probability that the new drug is indeed superior. At $n = 25$, the assurance $\gamma_2 = 0.595$ is 75% of this maximum assurance, while at $n = 100$ the assurance reaches $\gamma_2 = 0.701$, or 88% of the maximum.

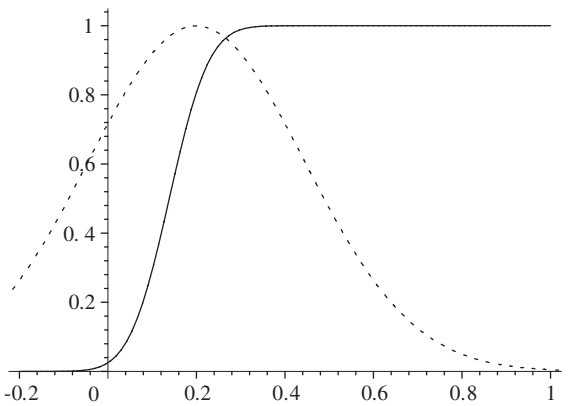

Figure 1. Power (solid line) and prior density of δ (dotted line), Example 1. For clarity, the prior density has been scaled to have value 1 at the mode.

2.4. Example 2

In this example, we suppose that the prior mean and variance of δ in Example 1 were obtained from a slightly more detailed elicitation of prior belief. Suppose that the expert actually gave a prior probability of 0.5 that $\delta = 0$, that is, that there really is no difference between the treatments. Alternatively, with prior probability 0.5 the expert specified that $\delta \sim N(0.4, 0.04)$. This leads to the same prior mean and variance as given in Example 1, but with a distinctly non-normal prior distribution.

It is simple to show that the assurance is actually

$$\gamma_2 = 0.5\left(0.025 + \Phi\left(\frac{-1.96\sqrt{0.125/n} + 0.4}{\sqrt{0.04 + 0.125/n}}\right)\right),$$

since with probability 0.5 the null hypothesis of no treatment effect is true, in which event there is just a 0.025 chance of getting a positive significant result. At $n = 25$, we now have $\gamma_2 = 0.458$. This seems like a small assurance, but the prior probability that $\delta > 0$ is now only 0.488. An assurance of 0.458 may in practice be an adequate basis for conducting a trial, as we will discuss in Section 6.

At $n = 100$, $\gamma_2 = 0.487$. However, remember that this includes at least a probability of 0.0125 that the drug is ineffective but we get a positive significant result by pure chance. The analogue of (8) is $P(\bar{x}_2 - \bar{x}_1 > \tau Z_{\alpha/2},\ \delta > 0) = 0.473$. Increasing the sample size from 25 to 100 now increases assurance by a very small amount, and would only be worth doing if the potential value of the drug is very large.

Notice in these two examples how the assurance is affected by the form of the prior distribution. The first example makes an inappropriate assumption of a normal distribution for δ, whereas the expert's beliefs are more accurately represented by the distinctly non-normal prior distribution in

Pharmaceutical STATISTICS

Example 2. As a result, the value of assurance obtained in Example 1 is unrealistically high.

3. COMPUTATION BY BAYESIAN CLINICAL TRIAL SIMULATION

In Section 2 we showed that for a sufficiently simple problem with a suitable prior distribution it is possible to derive mathematical expressions for assurances, although this would be the exception in practice. In general, such closed-form expressions will not be available and assurances will need to be calculated numerically. The most useful general-purpose technique is based on simulation.

Simulation is already widely used to compute power and related properties of complex clinical trials. The technique, often known as *clinical trial simulation*, entails drawing a large number of random values for the trial *data*. To compute the power of a test, the test is performed on each simulated set of data and the proportion of times that the null hypothesis is rejected estimates the power. It is important to note that in this simulation the parameter vector $\boldsymbol{\theta}$ is fixed at the requisite values $\boldsymbol{\theta}^*$ at which the power π^* is desired. When the method is used for sample size determination, for instance, the simulation is repeated for different sample sizes until a sample size is obtained that yields the desired power.

To compute assurances, we must extend this method to incorporate sampling from the prior distribution of $\boldsymbol{\theta}$, and we call the resulting technique *Bayesian clinical trial simulation* (BCTS). The process simply involves sampling a new value of $\boldsymbol{\theta}$ each time before sampling the data. To compute the assurances of outcomes $A_1, A_2, \ldots, A_k$, the general algorithm is as follows:

1. Define counters I for iteration and $T_1, T_2, \ldots, T_k$ for the assurances, and set all counters to 0. Set the required number, N, of iterations.
2. Sample $\boldsymbol{\theta}$ from the prior distribution.
3. Sample the sufficient statistics using the model and the sampled value of $\boldsymbol{\theta}$.
4. For $j = 1, 2, \ldots, k$, increment T_j if the outcome A_j has occurred.
5. Increment I. If $I < N$, go to step 2.
6. For $j = 1, 2, \ldots, k$, estimate assurance $\gamma_j = P(A_j)$ by T_j/N.

Notice that in step 3 it is not necessary to sample all of the trial data. It is enough to sample the values of sufficient statistics (and strictly only those that are required for the test). If we were to use this method for the simple model of Section 2, after sampling δ from its prior distribution, we could sample the two sufficient statistics $\bar{x}_1$ and $\bar{x}_2$. However, to compute the relevant test statistic we only actually need to sample $\bar{x}_2 - \bar{x}_1$ from its $N(\delta, \tau^2)$ distribution.

Notice that assurances for any number of outcomes of interest may be computed from a single set of BCTS simulations of $\boldsymbol{\theta}$ and the sample sufficient statistics.

4. NORMAL DATA WITH UNKNOWN VARIANCES

4.1. Assurance calculations

Now suppose that we have the normal sampling framework of Section 2, but the variances are not known. We first make the common assumption that $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Then the standard two-sided test rejects the null hypothesis if $|t| > t_{\alpha/2}$, where

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\hat{\sigma}\sqrt{n_1^{-1} + n_2^{-1}}},$$

$\hat{\sigma}^2 = (n_1 + n_2 - 2)^{-1} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ and $t_{\alpha/2}$ is the two-sided significance point of the Student t distribution with $n_1 + n_2 - 2$ degrees of freedom.

The sampling distribution of t is a non-central t distribution (which only becomes an ordinary Student t distribution if $\delta = 0$). The power function requires evaluation of the distribution function of the non-central t distribution. Since the non-centrality parameter depends on both the treatment difference δ and the sampling variance σ^2, both must be specified in the usual frequentist procedure to determine power, and thence to choose a sample size. Although approaches exist that deal explicitly with this framework (e.g. [11, 12]), in practice the sample is invariably large

enough for $t_{\alpha/2}$ to be indistinguishable from $Z_{\alpha/2}$. As a result, the required sample size will be almost exactly the same as in the case of known variance. However, the variance is in reality unknown, and to obtain the sample size it is necessary to provide an assumed value for σ^2. Recognizing also that the necessary sample size may depend very strongly on this assumed value, it is usual to do some sensitivity analyses exploring alternative σ^2 values. There is then no clear mechanism for deciding which of the resulting range of sample sizes should be chosen, and even less is there any clear idea of how this translates into assurance of success for the trial sponsor.

Computation of assurance is possible by numerical integration over the space of δ and σ^2 if a subroutine for computing the distribution function of the non-central t distribution is available. However, BCTS is very simple and effective in this case. The appropriate version of the general algorithm in Section 3 for the case of computing γ_1 is given below and WinBUGS code to implement it is provided in the Appendix. It is trivial to alter the algorithm to compute assurances of other outcomes.

1. Set counters $I = P = 0$. Set required number, N, of iterations.
2. Sample δ and σ^2 from their joint prior distribution.
3. Sample $\bar{x}_2 - \bar{x}_1 \sim N(\delta, (n_1^{-1} + n_2^{-1})\sigma^2)$ and $(n_1 + n_2 - 2)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n_1+n_2-2}$. Compute t.
4. Increment P if $t > t_{\alpha/2}$.
5. Increment I. If $I < N$, go to step 2.
6. Estimate γ_1 by dividing P by N.

It is simple to modify this for one-sided testing; the critical value $t_{\alpha/2}$ becomes t_α, the counter P is used for assurance and S for strong assurance. Modification for equivalence testing is also straightforward.

If we do not assume equality of variances, there is no longer a simple exact significance test available. The usual method is to apply an approximate t test. For instance, Satterthwaite [13] defines the test statistic

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{(\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2)}}, \tag{9}$$

$\hat{\sigma}_i^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$, and approximates its distribution under the null hypothesis by a t distribution with degrees of freedom chosen to match moments. The Bayesian clinical trial simulation algorithm is again readily modified. At step 2, we sample from the joint prior distribution of λ, σ_1^2 and σ_2^2. At step 3, we sample $\bar{x}_2 - \bar{x}_1 \sim N(\lambda, n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2)$, $(n_1 - 1)\hat{\sigma}_1^2 \sim \chi^2_{n_1-1}$ and $(n_2 - 1)\hat{\sigma}_2^2 \sim \chi^2_{n_2-1}$ before computing t from (9) and comparing it with the appropriate t significance point.

4.2. Example 3

Continuing with Example 2, we suppose that the population variances are unknown but equal in the two treatment groups, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and incorporate elicited prior information about σ^2. We will assume that although the expert gave an estimate of 0.0625 for σ^2, there was evidence to suggest that it could be as large as 0.16 or as small as 0.0225. We represent this information using a lognormal prior distribution for σ^2 such that $\ln \sigma^2 \sim N(-2.77, 0.7)$. This gives 0.0625 as the median of the distribution, although the mean is 0.089, reflecting the skewness of the distribution.

The usual approach to sample size setting would obtain $n = 9$ for $\sigma^2 = 0.0225$, but $n = 63$ for $\sigma^2 = 0.16$. In response to this, the trial designer must compromise, and might choose a more conservative sample size than 25, such as $n = 40$.

The WinBUGS program in the Appendix implements the BCTS in this case, and at $n = 25$ we find that $\gamma_2 = 0.44$, while at $n = 40$ we have $\gamma_2 = 0.46$ and at $n = 100$ we have $\gamma_2 = 0.48$. These results are very similar to those obtained for Example 2, the assurance being only slightly reduced by the uncertainty concerning σ^2. From the perspective of the developer of the supposed new drug, the extra assurance obtained by increasing sample size from 25 to 40 per treatment group is unlikely to justify the increased size (and longer recruitment time). The frequentist concern about sensitivity of the sample size to different values for the unknown population variance is seen to be largely irrelevant in this example, which suggests that a risk-averse approach of choosing

Pharmaceutical STATISTICS

(say) the upper limit of a 95% confidence interval of a variance estimate as the value to use in a sample size calculation leads to unnecessarily large studies.

5. BINARY DATA

Now consider a clinical trial in which the primary outcome measure is binary, so that the data comprise r_i successes observed out of n_i patients for treatment i ($i = 1, 2$). Letting θ_i denote the underlying population success rate for patients receiving treatment i, the null hypothesis for one-sided or two-sided tests is that $\theta_1 = \theta_2$. A standard test is based on approximate normality of the sample proportions $p_i = r_i/n_i$, although Fisher's exact test is sometimes applied [14, Chapter 10]. For instance, a simple two-sided test rejects the null hypothesis if $|Z| > Z_{\alpha/2}$, where

$$Z = \frac{p_2 - p_1}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \tag{10}$$

is approximately $N(0, 1)$ if the null hypothesis is true [15]. Notice that the denominator in (10) makes a further approximation by replacing each θ_i by its estimate p_i. A common further approximation is to replace $p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$ in the denominator by $\bar{p}(1-\bar{p})(n_1^{-1} + n_2^{-1})$, where $\bar{p} = (r_1 + r_2)/(n_1 + n_2)$. To derive the power function for the test based on (10) it is necessary to assume that the normal approximation is exact, and thereby to ignore sampling variation in the denominator. In particular,

$$P(R_1 \mid \theta_1, \theta_2) \approx \Phi\left(-Z_{\alpha/2} + \frac{\theta_2 - \theta_1}{\sqrt{\theta_1(1-\theta_1)/n_1 + \theta_2(1-\theta_2)/n_2}}\right), \tag{11}$$

where R_1, as before, is the event of rejecting the null hypothesis with a positive sample treatment difference. Fleiss *et al.* [15] provide a simple approximation to the sample size; see also [16, Chapter 3].

Note that $P(R_1 \mid \theta_1, \theta_2)$ depends on both parameters θ_1 and θ_2, not just the difference. In the usual terminology for binary outcomes, it depends on the incidence θ_1 in the control group as well as the treatment effect. As discussed in Section 4.1, this means that in order to determine a suitable sample size we need also to specify the control group incidence, and it is usual to examine sensitivity to this nuisance parameter. A lack of clarity thereby arises over what power is actually achieved by the final choice of sample size. The assurance calculations naturally handle prior uncertainty in both the treatment effect and the control group incidence.

Even with the approximations used to derive (11), it is not possible to obtain an analytical formula for $\gamma_1 = P(R_1)$. It is necessary to use computational methods. By numerical integration we can find the expectation of (11) with respect to the joint prior distribution of θ_1 and θ_2. However, this will be approximate in the same way that (11) is approximate. BCTS, in contrast, computes the exact assurances for the test based on (10); see the WinBUGS code in the Appendix.

5.1. Example 4

We now suppose that the drug considered in Example 1 has demonstrated its ability to reduce CRP in the Phase 2 trial, and it is now planned to run a Phase 3 trial comparing it with methotrexate, which is standard first-line therapy for rheumatoid arthritis. The outcome measure is ACR20 after 6 months of treatment, which is a widely-used composite measure that roughly indicates a 20% improvement in symptoms. Thus, θ_1 is the proportion of patients who achieve ACR20 with methotrexate and θ_2 is the proportion achieving ACR20 with the new drug. We are interested in the assurance of the trial demonstrating superiority to methotrexate.

Prior information concerning θ_1 is specified via $E(\theta_1) = 0.2$, $\mathrm{sd}(\theta_1) = 0.08$. The expected value is based on published results for methotrexate (e.g. [17]) and the standard deviation includes uncertainty about how closely the baseline disease characteristics of patients recruited to this trial will match those in earlier studies. We represent this prior information by a beta

distribution: $\theta_1 \sim \text{Be}(5, 20)$. There is more uncertainty about the efficacy of the new drug, but on the basis of its performance during development and the efficacy of related drugs the company's assessment is given by $E(\theta_2) = 0.4$, $\text{sd}(\theta_2) = 0.17$, which corresponds to a beta distribution $\text{Be}(3, 4.5)$. However, the development team recognizes that there is a chance that the drug will simply be ineffective, that despite its effect on CRP the drug will not in fact act effectively on rheumatoid arthritis. They therefore allow a 0.15 probability that θ_2 follows the $\text{Be}(2, 23)$ distribution, which effectively represents a placebo, and it follows the $\text{Be}(3, 4.5)$ distribution with probability 0.9. The prior distributions are assumed independent in this example.

The trial is planned with $n_1 = 200$ patients in the control (methotrexate) group and $n_2 = 400$ in the treatment group. The primary analysis is a two-sided significance test for superiority at the 5% level. Using (11), these sample sizes give approximately 80% power to detect an improvement from $\theta_1 = 0.2$ to $\theta_2 = 0.3$ (i.e. a 50% treatment effect).

The assurance γ of a significant and positive test result ($Z > 1.96$, where Z is given by (10)), was computed using the WinBUGS program given in the Appendix. We find $\gamma = 0.635$. Notice that, although the prior distributions are such that the new drug is expected to achieve an efficacy far higher than the $\theta_2 = 0.3$ assumed for the power calculation, the assurance is still lower than the power. This is because there is only a 0.74 prior probability that the new drug is genuinely more effective than methotrexate, and the assurance cannot exceed this value no matter how large the sample sizes may be.

6. DISCUSSION

6.1. Practical uses of assurance

The concept of assurance is quite different from the conventional notion of power, and more relevant to decision-making. Whereas a power calculation tells the trial sponsor the probability of successfully rejecting the null hypothesis if the true value of the unknown treatment effect δ is the specified δ^*, the assurance specifies the *unconditional* probability of a successful outcome. Its use is not limited to probabilities of specific outcomes of hypothesis tests but can extend to other relevant trial outcomes. For instance, we have suggested that $P(\bar{x}_2 - \bar{x}_1 > \tau Z_\alpha, \delta > 0)$, which is the assurance of obtaining a significant positive effect in a superiority trial *and* of the test drug being genuinely superior, may well be meaningful for a decision-maker.

The examples in this paper have demonstrated the potential for BCTS to compute assurances that are directly relevant to decision-making and trial design. The methodology extends easily to complex trials with realistic prior knowledge structures. An important further extension is to consider a sequence of trials, such as the Phase 2 and Phase 3 trials discussed in the examples. BCTS can be used prior to the Phase 2 trial to assess the assurance that the drug will pass Phase 2 *and* go on to show superiority to the active comparator methotrexate in Phase 3. This would mean building a model of the relationship between the surrogate outcome of CRP reduction and the meaningful clinical outcome ACR20, including probability distributions to represent uncertainty about the parameters of that relationship. These might be based on published data sources but could also involve the elicitation of expert judgement. It is likely that the development sequence would also include an intermediate trial to establish the dose for Phase 3, which would entail also modelling the dose–response relationship. Although this leads to complex modelling, the potential to advise the whole development strategy of a new drug can fully justify the exercise.

In such a case, assurances of many more aspects of the trial outcomes may be relevant. For instance, Table I presents assurances for a range of different outcomes in a BCTS study that was conducted to compare three different development strategies that had been proposed for a drug to treat stroke patients. The analysis covered a 'preliminary Phase 3' trial as well as a full Phase 3 study. Strategies S1 and S2 were to conduct these trials with two different doses of the new drug,

Pharmaceutical STATISTICS

Table I. Assurances for three different strategies.

Outcome assurances	S1	S2	S3
Failure at Phase 2b	0.000	0.000	0.145
Failure at preliminary Phase 3	0.056	0.068	0.023
Failure for futility at futility analysis	0.050	0.102	0.034
Failure for safety at interim analysis	0.049	0.021	0.044
Failure for futility at interim analysis	0.039	0.076	0.025
Success at interim analysis	0.253	0.267	0.226
Failure for efficacy at final analysis	0.063	0.102	0.052
Failure for safety at final analysis	0.128	0.068	0.118
Success at final analysis	0.362	0.296	0.333
Failure	*0.385*	*0.437*	*0.441*
Success	*0.615*	*0.563*	*0.559*

while for strategy S3 the dose was to be determined by first running a 'Phase 2b' trial. In all three strategies, the Phase 3 trial would incorporate a futility analysis and an interim analysis prior to the final analysis, allowing the possibility of stopping the trial early. Success was defined as not only obtaining a significant result in a test of superiority but also satisfying a safety condition.

We see from Table I that strategy S1 gives the highest overall assurance of success, but note that S2 gives a slightly higher chance of success at the interim analysis, and so may lead to a quicker and less costly adoption of the drug. Strategy S3 has the lowest chance of success, but allows for the option to terminate development early in a Phase 2b trial, thereby saving the cost of a large Phase 3 trial. Another possible benefit of S1 is that it seems to be more effective at identifying safety problems. Consideration of an extensive range of assurance probabilities is typically important for well-informed planning of trials.

For someone who is familiar with the conventional approach to sample size setting based on power, it is easy to misinterpret the values obtained in assurance calculations. Clinical trials are almost invariably designed to achieve power of 80% or more at the chosen clinically meaningful treatment effect δ^*, and it is tempting to think then that there is an 80% chance of a significant result. But of course this figure only applies *conditionally* on δ equalling δ^*. The assurance figure is often much lower, because there is an appreciable prior probability that the true treatment difference is less than δ^*. Pharmaceutical companies explore very large numbers of chemicals in the search for an effective drug, so the chance that any one of these will be successfully demonstrated to be superior to existing treatments is small. Even at an advanced stage of development, such as entering Phase 2, a new drug still has a relatively low chance of being successfully launched. The company continues with development because the benefits of success are large enough to balance the cost of continuing development, even with a large risk of failure. In this context, assurance values such as the 0.458 found in Example 2 are not small, and may indeed be large enough to justify continuing with the proposed trial. The assurance is a realistic assessment of the chance of success with the drug, and it is the power of 80% or more that is misleading and may tend to induce unfounded optimism.

We have seen in Examples 1 and 2 that it is important to specify prior knowledge about the unknown true treatment effect carefully. There is a substantial literature on the elicitation of expert knowledge, with particular emphasis on avoiding risks such as over-confidence of experts. The investment of time and resources on careful quantification of knowledge and uncertainty is repaid in terms of reliable specification of assurances for the aid of planners and decision-makers.

6.2. Other Bayesian approaches to trial design

There is a substantial literature on Bayesian approaches to the design of experiments generally, and to clinical trials in particular. Many important references are reviewed by Joseph *et al.* [18], Pezeshk [19] and Spiegelhalter *et al.* [7], or contained in the excellent collection edited by Smeeton and Adcock [20]. Some other recent references are Katsis and Toman [21], Karnon [22], Pham Gia and Turkkan [23], Dendukuri *et al.* [24], Mayo and Gajewski [25] and De Santis *et al.* [26]. Almost all of this work employs a fully Bayesian approach in which the data obtained in the trial will be analysed using Bayesian methods, or used as inputs to a formal Bayesian decision-theoretic analysis. The prior distribution that is used for this analysis is also employed in formulating the design. O'Hagan and Stevens [5] present an alternative approach in which different prior distributions are used to design the study and to analyse the data. The study design is made on the basis of the assurance (calculated from the design prior distribution) that the analysis (performed using the analysis prior distribution) will yield an appropriate outcome.

Special cases of the O'Hagan and Stevens formulation are the usual Bayesian approach, in which the same prior distribution is used for both, and the stance adopted in the present paper, where the analysis prior distribution is effectively non-informative, so that the analysis follows conventional frequentist statistics and the outcomes are typically phrased in terms of significance tests. The latter seems appropriate in the present context of medical research, where the use of frequentist methods to analyse clinical trials is almost mandatory. The regulatory framework is changing, with a greater willingness to allow fully Bayesian analyses, but it is unlikely that this will embrace a fully Bayesian analysis in which the prior beliefs of the trial sponsors are also accepted by regulators for analysis of the data.

Bayesian clinical trial simulation is a generic tool that can compute assurances for any trial result, whether that is based on a Bayesian analysis of the data, frequentist significance tests or a formal decision analysis such as a decision by a health care provider to reimburse use of the drug on cost-effectiveness grounds.

ACKNOWLEDGEMENTS

We thank an anonymous referee for several suggestions that have improved the presentation of this paper.

REFERENCES

1. Pocock SJ. *Clinical trials*: *a practical approach*. Wiley: Chichester, 1983.
2. Machin D, Campbell MJ, Fayers P, Pinol A. *Statistical tables for the design of clinical studies*, 2nd edn. Blackwell Scientific: Oxford, 1997.
3. International Conference on Harmonisation. *ICH Harmonised Tripartite Guideline E9*: *Statistical Principles for Clinical Trials*, 1998. Available http://www.ich.org.
4. Brown BW, Herson J, Atkinson EN, Rozell ME. Projection from previous studies – a Bayesian and frequentist compromise. *Controlled Clinical Trials* 1987; **8**:29–44.
5. O'Hagan A, Stevens JW. Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making* 2001; **21**:219–230.
6. Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial based on subjective clinical opinion. *Statistics in Medicine* 1986; **5**:1–13.
7. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian approaches to clinical trials and health-care evaluation*. Wiley: Chichester, 2004.
8. Gillett R. An average power criterion for sample size estimation. *The Statistician* 1994; **43**:389–394.
9. Garthwaite PH, Kadane JB, O'Hagan A. Statistical methods for eliciting prior distributions. *Journal of the American Statistical Association* 2005; **100**: 680–701.
10. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence. *British Medical Journal* 1996; **313**:36–39.
11. Guenther WC. Sample size formulas for normal theory *t* tests. *American Statistician* 1981; **35**: 243–244.
12. Schouten HJA. Sample size formula with a continuous outcome for unequal group sizes and unequal variances. *Statistics in Medicine* 1999; **18**:87–91.

13. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics* 1946; **2**:110–114.
14. Swinscow TDV, Campbell MJ. *Statistics at square one*, 10th edn. BMJ Publishing Group: London, 2002.
15. Fleiss JL, Tytun A, Ury HK. A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics* 1980; **36**: 343–346.
16. Fleiss JL. *Statistical methods for rates and proportions*, 2nd edn. Wiley: New York, 1981.
17. Kremer JM, Genovese MC, Cannon GW, *et al*. Concomitant leflunomide in patients with active rheumatoid arthritis despite stable doses of methotrexate: a randomized double blind placebo controlled trial. *Annals of Internal Medicine* 2002; **137**:726–733.
18. Joseph L, du Berger R, Bélisle P. Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine* 1997; **16**:769–781.
19. Pezeshk H. Bayesian techniques for sample size determination in clinical trials: a short review. *Statistical Methods in Medical Research* 2003; **12**:489–504.
20. Smeeton NC, Adcock CJ (eds). Sample size determination. *The Statistician* 1997; **46**:127–283.
21. Katsis A, Toman B. Bayesian sample size calculations for binomial proportions. *Journal of Statistical Planning and Inference* 1999; **81**:349–362.
22. Karnon J. Planning the efficient allocation of research funds: an adapted application of a non-parametric Bayesian value of information analysis. *Health Policy* 2002; **61**:329–347.
23. Pham Gia T, Turkkan N. Determination of exact sample sizes in the Bayesian estimation of the difference of two proportions. *The Statistician* 2003; **52**:131–150.
24. Dendukuri N, Rahme E, Bélisle P, Joseph L. Bayesian sample size determination for prevalence and diagnostic studies in the absence of a gold standard test. *Biometrics* 2004; **60**:388–397.
25. Mayo MS, Gajewski BJ. Bayesian sample size calculations in phase II clinical trials using informative conjugate priors. *Controlled Clinical Trials* 2004; **25**:157–167.
26. De Santis F, Petrone Pacifico M, Sambucini V. Optimal predictive sample size for case-control studies. *Applied Statistics* 2004; **53**:427–441.

APPENDIX. WINBUGS CODE

We present here WinBUGS code for Examples 3 and 4. The two WinBUGS files can also be downloaded from http://www.shef.ac.uk/~st1ao/pub.htnl

```
** WinBUGS code for Example 3
model{
   delta <- eff * deltaeff            # These lines set out the
   eff ~dbern(0.5)                    # prior as in Example 3.
   deltaeff ~ dnorm(0.4, 25)          # They should be modified
   sigsq <- exp(logsigsq)             # as appropriate for
   logsigsq ~ dnorm(-2.77, 1.42857)   # any real example.
   t <- xbardiff * sqrt(precis)
   precis <- n/(2*sampvar)
   xbardiff ~ dnorm(delta, precis)
   nminus1 <- n-1
   nminusoversig <- nminus1/sigsq
   sampvar ~ dgamma(nminus1, nminusoversig)
   ass <- step(t-tsig)  # The mean of ass is the assurance
   }
list(n=25, tsig=2.06)    # Choose one set of input values
list(n=33, tsig=2.035)
list(n=35, tsig=2.03)
```



```
list(n=40, tsig=2.021)
list(n=50, tsig=2.009)
list(n=100, tsig=1.984)

** WinBUGS code for Example 4
model{
   theta1 ~dbeta(5,20)
   theta2 <- (eff * theff)+((1-eff)*thneff) # These lines set out
   eff ~dbern(0.85)      # the prior as in Example 4.
   theff ~dbeta(3,4.5)   # They should be modified
   thneff ~dbeta(2, 23)  # as appropriate for any real example.
   r1 ~dbin(theta1, n1)
   r2 ~dbin(theta2, n2)
   p1 <- r1/n1
   p2 <- r2/n2
   se <- sqrt((p1*(1-p1)/n1)+(p2*(1-p2)/n2))
   z <- (p2-p1)/se
   ass <- step(z-zsig)  # The mean of ass is the assurance
   }
list(n1=200, n2=400, zsig=1.96) # Sample sizes and critical value
```