Article:
Alhussain, Z. and Oakley, J. (2020) Assurance for clinical trial design with normally distributed outcomes: eliciting uncertainty about variances. Pharmaceutical Statistics, 19 (6). pp. 827-839. ISSN: 1539-1604

https://doi.org/10.1002/pst.2040

Received: 30 May 2019 | Revised: 14 February 2020 | Accepted: 12 May 2020

DOI: 10.1002/pst.2040

WILEY

MAIN PAPER

Assurance for clinical trial design with normally distributed outcomes: Eliciting uncertainty about variances

Ziyad A. Alhussain[1] | **Jeremy E. Oakley**[2]

[1]Mathematics Department, Faculty of Science in Zulfi, Majmaah University, Al Majma'ah, Saudi Arabia

[2]School of Mathematics and Statistics, The University of Sheffield, Sheffield, UK

Correspondence
Jeremy E. Oakley, School of Mathematics and Statistics, The University of Sheffield, The Hicks Building, Hounsfield Road, Sheffield, South Yorkshire S3 7RH, UK.
Email: j.oakley@sheffield.ac.uk

Funding information
Deanship of Scientific Research at Majmaah University, Grant/Award Number: R-1441-126

Summary

The assurance method is growing in popularity in clinical trial planning. The method involves eliciting a prior distribution for the treatment effect, and then calculating the probability that a proposed trial will produce a "successful" outcome. For normally distributed observations, uncertainty about the variance of the normal distribution also needs to be accounted for, but there is little guidance in the literature on how to elicit a distribution for a variance parameter. We present a simple elicitation method, and illustrate how the elicited distribution is incorporated within an assurance calculation. We also consider multi-stage trials, where a decision to proceed with a larger trial will follow from the outcome of a smaller trial; we illustrate the role of the elicited distribution in assessing the information provided by a proposed smaller trial. Free software is available for implementing our methods.

KEYWORDS

assurance, expert judgement, prior elicitation, variance elicitation

1 | INTRODUCTION

Assurance is a Bayesian alternative to a power calculation for choosing a sample size in a clinical trial. The aim of the assurance method is to provide a realistic assessment of the trial sponsor's probability of a "successful" trial. A prior distribution is elicited for the treatment effect, and the prior probability that the trial will be successful is calculated, for any success criteria that the trial sponsor wishes to consider (eg, that the observed treatment effect will be positive, and statistically significant at the appropriate size). This approach was first proposed by Spiegelhalter and Freedman,[1] and developed in O'Hagan and Stevens[2] and O'Hagan et al,[3] where the term "assurance" was used.

An extensive discussion of the benefits of the assurance method is given in Crisp et al[4] and Dallow et al.[5] These papers give an account of how assurance has been used on a large scale at GlaxoSmithKline, and describe several case studies. They describe how the *process* of deriving an assurance supports their decision making, in particular, how the elicitation of a prior distribution provides a formal assessment of the evidence and uncertainties regarding a treatment effect. They give examples where modifications are made to trial designs when it is been judged necessary to mitigate against risks and uncertainties identified at the elicitation stage; using the assurance method can result in more than just a modified sample size assessment.

Deriving an assurance requires a prior distribution for any relevant uncertain quantity. O'Hagan et al[3] consider the case of prior distributions for normally distributed and binomial data, and Ren and Oakley[6] presented elicitation

methodology for time-to-event data. Gasparini et al[7] also considered normally distributed data. Neither O'Hagan et al nor Gasparini et al discuss how one would elicit distributions for unknown population variances, and there is little guidance in the wider elicitation literature on how to elicit a distribution for a variance. We propose a method in this article, and show how to incorporate it within an assurance calculation.

In cases where there is greater prior uncertainty, trial sponsors may consider an adaptive design, or a multi-stage approach, where planning decisions about a latter stage trial may be informed by the results from an earlier stage. A case study involving the use of assurance in such a setting is given in Nixon et al.[8] For normally distributed data, it is again important to consider uncertainty in the variance, as the variance will affect how much information one gains for a given sample size. We illustrate how simulation can be used to investigate how a small study would reduce uncertainty about a treatment effect, to support the planning of a larger trial.

In the next section, we discuss the assurance method and note the role of variance distributions in assurance calculations. In Section 3, we review methods for eliciting a distribution for the treatment effect, and discuss elicitation of distributions for variances in Section 4. In Section 5, we show how the elicited distributions are incorporated within the assurance calculation, and we discuss the extension to multi-stage trial planning in Section 6. Free software is available to implement all our methods, and is described in the Appendix.

2 | ASSURANCE

In this section, we describe the assurance method, specifically in the context of a randomised controlled trial, where the observations in both the treatment and control arms are assumed to be normally distributed. We suppose that in the control arm, we have observations $X_1,...,X_{n_c} \overset{i.i.d.}{\sim} N(\mu_c,\sigma_c^2)$, and in the treatment arm, we have observations $Y_1,...,Y_{n_t} \overset{i.i.d.}{\sim} N(\mu_t,\sigma_t^2)$. We write $\mu_t = \mu_c + \delta$, so that we interpret δ as the treatment effect.

O'Hagan et al[3] consider the four cases of a one-sided superiority trial, a two-sided superiority trial, a non-inferiority trial and an equivalence trial. In this article, we will consider two-sided superiority trials only, but extension to the other cases would be straightforward (and would not change the methodology we are proposing here).

We suppose that the data will be analysed with a two-sample *t*-test:

$$T = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{S_X^2}{n_c} + \frac{S_Y^2}{n_t}}},$$

with T compared with the Student-t distribution with ν degrees of freedom computed using the Welch approximation.

For a power calculation, for a given n_t and n_c, we would fix values of δ, σ_t^2 and σ_c^2, and calculate the probability of observing data such that the null hypothesis is rejected, for a specified level of significance. We denote R to be the event of rejecting the null hypothesis, and write this probability as $Pr(R|\delta,\sigma_t^2,\sigma_c^2)$. (In practice, we might make some simplifications such as assuming $\sigma_t^2 = \sigma_c^2$, and that T can be compared with the standard normal distribution). In this calculation, we interpret δ as a "minimum clinically relevant" treatment effect: the smallest treatment effect that we would want our trial to be able to detect.

In the assurance method, we consider the unconditional probability of the *same* event R, but with a *different* interpretation of δ: we now interpret δ as the *true* treatment effect, elicit a prior distribution $\pi(\delta,\sigma_t^2,\sigma_c^2)$, and compute

$$Pr(R) = \int Pr(R|\delta,\sigma_t^2,\sigma_c^2)\pi(\delta,\sigma_t^2,\sigma_c^2)\,d\delta\,d\sigma_t^2\,d\sigma_c^2.$$

We emphasise that the event R is the same as that used in the power calculation: although we now have a prior distribution $\pi(\delta,\sigma_t^2,\sigma_c^2)$ we do *not* assume it will be used in the analysis of the trial data: we assume exactly the same (frequentist) analysis as that used in the power calculation. In general, we can think of a regulator or trial sponsor specifying an event R in which the trial outcome is "successful," and then we elicit a prior distribution *only* to assess the probability of achieving the "successful" outcome.

The computation of an assurance $Pr(R)$ is usually straightforward using Monte Carlo methods; the main effort required in any assurance calculation is in eliciting the prior distribution for all the uncertain parameters. Note that, although the assurance method typically does not involve any Bayesian analysis of clinical trial data, we can still make

use of elicitation methodology from any particular Bayesian analysis method that uses informative prior distributions. For example, Hampson et al[9] elicited a prior for use in the Bayesian analysis of binary outcome data; their elicitation method could be suitable if one wanted to compute an assurance for a trial with the same data type.

2.1 | The need to account for uncertainty in population variances

The main development in this article is in the elicitation and application of the prior for the variances σ_t^2 and σ_c^2. We will first comment on the importance of accounting for uncertainty in these parameters, relative to accounting for uncertainty in the treatment effect δ.

We consider three different risks faced by the trial sponsor, which we label as "primary," "secondary" and "tertiary," and discuss the role of eliciting uncertainty in understanding these risks. We define the primary risk as the risk that the treatment is either ineffective (or is not sufficiently effective for reimbursement), so that there is no prospect of any trial producing a successful outcome (ignoring the possibility of a Type I error). Clearly, only uncertainty about the treatment effect δ matters at this stage.

In the case that the treatment *is* effective, we define the secondary risk as the risk that the trial is unsuccessful in demonstrating effectiveness, and the tertiary risk as that the trial sample size is unnecessarily large; effectiveness would have likely been demonstrated with an appreciably smaller sample size. In assessing these two risks, one can argue that accounting for uncertainty in the standard deviations σ_t and σ_c is "equally important" as accounting for uncertainty in δ. (The comparison with standard deviation rather than variance is more appropriate, as σ_t, σ_c and δ are on the same scale).

We consider R to be the event of rejecting a null hypothesis of no treatment effect in a test of size α. First consider the case of equal variances $\sigma_t^2 = \sigma_c^2 = \sigma^2$ and equal sample sizes n per group. Using the same approximations one would use in a simple power calculation, an approximate expression for the probability of R conditional on all the parameters is

$$Pr\left(R|\delta,\sigma^2\right) = 1-\Phi\left(Z_{\alpha/2}-\frac{\delta}{\sigma}\times\sqrt{\frac{n}{2}}\right),$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution. We note that the two parameters δ and σ only influence $Pr(R|\delta, \sigma^2)$ through their ratio δ/σ. If we were to start from some baseline estimate of δ and σ, applying a percentage change upwards of one parameter would have the same effect on $Pr(R|\delta, \sigma^2)$ as applying the same percentage change downwards of the other parameter. The assurance $Pr(R|\delta, \sigma^2)$ is equally sensitive to changes in either parameter, and careful elicitation is equally important for both the treatment effect and the variance.

With unequal variances and sample sizes, the situation is a little more complex. We now have

$$Pr\left(R|\delta,\sigma_c^2,\sigma_t^2\right) = 1-\Phi\left(Z_{\alpha/2}-\delta/\sqrt{\frac{\sigma_c^t}{n_c}+\frac{\sigma_t^2}{n_t}}\right),$$

so that if only one standard deviation parameter (σ_t or σ_c) changes, the effect is not as great as the same relative change in δ. We will investigate sensitivity to changes in these parameters empirically in our examples.

3 | ELICITING A PRIOR DISTRIBUTION FOR THE TREATMENT EFFECT

We now consider how to elicit the prior distribution $\pi\left(\delta,\sigma_t^2,\sigma_c^2\right)$. For simplicity and ease of exposition, it is supposed that there is one female expert, and that the elicitation is conducted by a male facilitator. There are various general considerations when performing elicitation such as training of the experts, and how to manage (or combine opinions from) multiple experts. The focus of this article is solely on how to elicit judgements about a mean and variance, and we do not consider these other issues here. Guidance on these and other aspects of elicitation can be found elsewhere.[10-14] The practicalities of conducting elicitation in the assurance context are discussed in Dallow et al,[5] who adopt the SHELF approach to elicitation.[15]

We suppose that an expert would be equally willing to consider the treatment effect δ directly, or to consider the mean in the treatment group μ_t, given a hypothetical value for the mean in the control group μ_c (the expert might propose her own hypothetical value for μ_c, or an estimate may be available from previous trials with the same control arm). In the following discussion, we consider the former option.

General advice in elicitation methods is to ask experts about observable quantities, rather than parameters in statistical models.[16] Although not strictly observable, we think the mean of a normal distribution would be well-enough understood for an expert to make judgements about it directly. Hence, standard univariate elicitation methods[17] can be used to elicit a prior distribution for δ or $\mu_t|\mu_c$.

Such methods typically involve eliciting a small number of points from the expert's cumulative distribution function of δ: the expert judges $Pr(\delta \leq d_i) = p_i$, for $i = 1, ..., n$. We can specify $d_1, ..., d_n$ and ask the expert to provide $p_1, ..., p_n$, or vice-versa. For example, the expert can be asked to provide her quartiles, in which case p_1, p_2, p_3 are fixed at 0.25, 0.5 and 0.75, respectively, and the expert provides the corresponding values of d_1, d_2, d_3.

We then consider some parametric family of distributions, indexed by parameters θ, and choose θ to minimise

$$\sum_{i=1}^{n} (F(d_i;\theta) - p_i)^2,$$

where $F(.;\theta)$ is the cumulative distribution function from the chosen family with parameter θ. Both Gasprini et al[7] and O'Hagan et al[3] assume a normal distribution $\delta \sim N(m, v)$, so that we would have $\theta = (m, v)$. Since a full distribution has been chosen based on a small number of elicited probabilities, we would then feed back some additional quantiles or probabilities from this distribution to the expert, to check that the distribution is an acceptable representation of the expert's beliefs. We illustrate this in Figure 1, where we suppose that expert has provided her quartiles, and a normal distribution is fitted to her judgements. This approach can be implemented in R[18] using the package `SHELF`,[19] and is incorporated in our software.

3.1 | Mixture distributions

O'Hagan et al[3] and Dallow et al[5] also consider a mixture distribution where a non-zero probability is given to the event $\delta = 0$: the event that the treatment has no effect, and a conditional distribution is elicited for δ given that $\delta \neq 0$: the treatment has some effect. Mixture distributions can be specified in our software: the user provides $Pr(\delta = 0)$, and then judgements from the conditional distribution, following the approach described in the previous section.

To see the possible benefits of the mixture approach, consider the following example. Suppose an expert judges a 70% chance that the new treatment will have some (beneficial) effect, and also thinks that a treatment effect of 0.5 (on some appropriate scale) is "most likely." If we suppose $\delta \sim N(0.5, 1)$, then we have a mode at 0.5 and $Pr(\delta > 0) \simeq 0.7$: the $N(0.5, 1)$ distribution would appear to describe well the two judgements made by the expert. However, this distribution would also imply $Pr(\delta > 1) \simeq 0.3$, but the expert may not judge a treatment effect at least twice as high as her

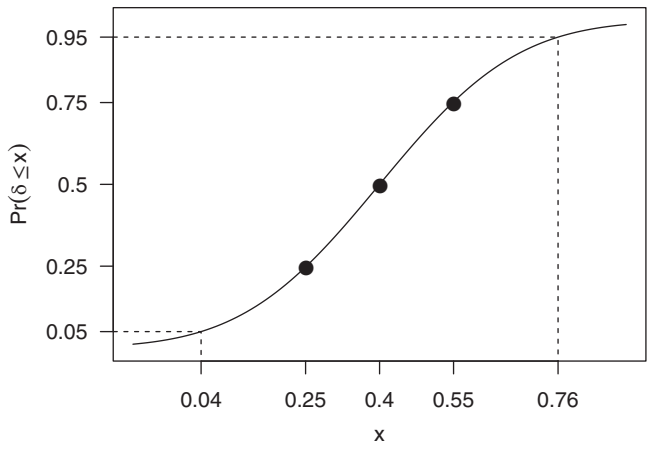

FIGURE 1 An example of eliciting a prior distribution for the treatment effect δ. We suppose the three quartiles have been elicited: these are shown as the black dots. We then fit a parametric distribution (in this example a normal distribution, shown by the solid line) to these points using a least squares approach. To check that the fit is acceptable to the expert, some percentiles from this fitted distribution are fed back to the expert. Here, the dashed lines show the 5th and 95th percentiles from the fitted normal distribution

"most likely value" to be very plausible, and giving such weight to larger values of δ may adversely effect the assessment of the required sample size. Using the mixture approach, we could instead set $Pr(\delta = 0) = 0.3$, and then, conditional on $\delta \neq 0$, set $\delta \sim N(0.5, v)$, with a smaller, more appropriate choice of v.

4 | ELICITING A DISTRIBUTION FOR A VARIANCE

To the best of our knowledge, there has been little work on eliciting beliefs about variances. One existing approach that can be used is based on eliciting beliefs about parameters in linear regression models. Kadane et al[20] and Al-Awadhi and Garthwaite[21] consider elicitation for the parameters $(\mu, \beta_1, ..., \beta_p, \sigma^2)$ in regression models of the form

$$X_i = \mu + \sum_{j=1}^{p} \beta_j z_{ij} + \varepsilon_i, \text{ for } i = 1, 2, ..., n,$$

where $\varepsilon_1, ..., \varepsilon_n \overset{i.i.d}{\sim} \mathrm{N}(0, \sigma^2)$. By setting $\beta_j = 0$ for all j, this would reduce to our case. Al-Awadhi and Garthwaite[21] proposed an elicitation method for quantifying opinions about the parameters of a multivariate normal distribution; the same elicitation method could be used for quantifying beliefs about a univariate normally distributed population. These methods require the expert to update her judgements in light of hypothetical data, under the assumption that the expert updates her beliefs using Bayes' theorem. We think this is a difficult task: the expert may not view hypothetical data as credible and behave the same way had she observed real data, and it is unlikely that the expert would weight prior knowledge and hypothetical data precisely according to Bayes' theorem in any case. The expert may be insensitive to the sample size, for example in accounting for the variability in a sample mean.[22] We think it desirable to have alternative elicitation methods available to the expert.

Kadane et al[20] and Al-Awadhi and Garthwaite[21] infer judgements about the parameters μ and σ^2 from judgements about the observable quantities X_i, by eliciting summaries from the expert's predictive distribution. For example, suppose we wish to elicit an expert's opinion about the variance parameter σ^2 of a random variable X that follows a normal distribution with a known mean μ. Since σ^2 is not directly observable then the expert is asked to make judgements about the observable quantity X, and we infer $p(\sigma^2)$ from these judgements. Any choice of $p(\sigma^2)$ implies a distribution

$$p_X(x) = \int_{\mathbb{R}^+} p_X(x \mid \sigma^2) p(\sigma^2) d\sigma^2,$$

and we suppose that a particular choice of $p(\sigma^2)$ will result in the above integral (approximately) matching the expert's beliefs about X, so that this choice of $p(\sigma^2)$ describes the expert's underlying beliefs about σ^2. A concern here is whether an expert really is able to account for her uncertainty about σ^2 when making judgements about X. A possibility is that the expert instead only makes judgements about X conditional on some estimate of σ^2.

Kadane et al[20] use conjugate priors for μ and σ^2 which force the expert's opinion about the two parameters to be dependent. However, it is possible in reality that knowledge of one parameter would not change the expert's opinion about the other. Al-Awadhi and Garthwaite[23] argued that, unless mathematical tractability is required, then it can be better to assume independence between the two parameters, and that this helps the expert focus on the assessments of each parameter separately. They proposed an elicitation method for the multivariate normal distribution where the mean vector and covariance matrix are assumed to be independent, though their method also asks the expert to update her judgements in the light of hypothetical data.

We argue that the better informed the expert, the less likely a judgement of dependence between the two parameters would be required. For example, consider the distribution of running times for an individual over a distance of 5 km. With no information about the ability of the runner, one might have considerable uncertainty about the mean, for example, an interval of 15 minutes to 1 hour may be judged plausible, with smaller variances of running times associated with smaller means within this interval. But if one already has good knowledge about the particular runner's ability, a much smaller interval may be judged plausible for the mean, and one's beliefs about the variance may not change appreciably given different plausible means.

We propose a new elicitation method for quantifying opinions about an uncertain population mean and variance. Our method does not elicit judgements using hypothetical data and Bayes' theorem, it does not use predictive elicitation and it assumes independence between the mean and variance.

4.1 | The proposed elicitation method

We first ask the expert to suppose that the treatment works and that the effect is "as expected": more precisely, we suppose that $\delta = m$, where m could be the mean or median of the expert's distribution for δ (where, in the mixture case, the mean/median would be conditional on $\delta \neq 0$). The expert can choose her own value for m if she wishes. We also ask the expert to propose a value for the control group mean μ_c; we would expect appropriate values to be readily available from the literature. Without loss of generality, we will suppose in the following that $\mu_c = 0$.

We then ask the expert to propose an interval on the outcome response scale $[k_1, k_2]$, that has meaningful interpretation in terms of patient outcomes. One possibility is to return to the notion of a minimum clinically relevant difference, which we denote by δ^*, and then consider the interval $(-\infty, \delta^*]$. Hence (assuming $\mu_c = 0$), any patient with an observation in this interval could be interpreted as *not* having benefited from the treatment, even though the treatment is assumed to be effective "on average."

Finally, we define ω to be the proportion of patients who would have outcomes in the interval $[k_1, k_2]$, and we ask the expert to consider her uncertainty about ω. In summary, if we have chosen the interval to be $(-\infty, \delta^*]$, we are asking: "Suppose the treatment does have the expected effect (on average). What proportion of patients might, nevertheless, *not* achieve the desired response given the treatment?"

We have

$$\omega = \Phi\left(\frac{k_2 - m}{\sigma_t}\right) - \Phi\left(\frac{k_1 - m}{\sigma_t}\right),$$

Shortly, we will infer the expert's judgement about σ_t via her judgements about ω. To do this, we require σ_t to be a montonic function of ω. This requirement is not met for all possible intervals $[k_1, k_2]$, but will be met if the interval is in the form of one of the following: $(-\infty, k_2]$, $[k_1, m]$, $[m - k, m + k]$, $[m, k_2]$ or $[k_1, \infty)$. For example, if we choose k_1 to be $-\infty$, we have $\Phi((k_1 - m)/\sigma_t) = 0$ and so

$$\sigma_t = \frac{k_2 - m}{\Phi^{-1}(\omega)}. \tag{1}$$

As a simple example to visualise this, suppose we were to elicit an expert's beliefs about students' marks for an undergraduate statistics module, for a large population of students. Suppose the marks are normally distributed with a mean of 60. Then there is a true proportion of students who will get marks between 60 and 70. If this expert, having been told that the mean is 60, is certain this proportion would be less than 0.45 and more than 0.25, this would imply she is certain σ is between 6 and 15. This is illustrated in Figure 2.

Probability judgements about ω can be converted to probability judgements about σ_t, and so we can use the same approach as described in Section 3. However, given the somewhat abstract nature of ω, we would suggest asking the expert for "lower" and "upper" bounds, which we would then interpret as 5th and 95th percentiles, and denote by $\omega_{0.05}$ and $\omega_{0.95}$, respectively.

If for example, the expert is considering the interval $(-\infty, \delta^*]$, with $\delta^* < m$, then it may help her to also consider a second interval $(\delta^*, m]$ and then consider how the population is distributed between these two intervals (noting that 50% of the population must have observations in the interval $(-\infty, m]$). For example, she might judge a split of 2% to 48% across the two intervals highly unlikely, which can help prompt judgements of more plausible allocations (though one should be cautious of anchoring effects).

Given $\omega_{0.05}$ and $\omega_{0.95}$, we can infer the corresponding quantiles of her distribution for the variance, which we denote by $\sigma^2_{t,0.95}$ and $\sigma^2_{t,0.05}$ (eg, using Equation 1 if the interval was of the form $(-\infty, k_2]$). The facilitator now chooses a parametric family of distributions, and obtains the parameter values θ within that family by minimising (numerically)

$$G(\theta) := \left(F_\theta\left(\sigma^2_{t,0.05}\right) - 0.05\right)^2 + \left(F_\theta\left(\sigma^2_{t,0.95}\right) - 0.95\right)^2 \tag{2}$$

where $F_\theta(.)$ is the cumulative distribution function from the chosen family, with parameter values θ.

We suggest fitting a distribution to the precision σ^{-2}, and choosing either a log-normal distribution or a gamma distribution. As only two judgements have been elicited, the log-normal and gamma distributions will fit these two judgements precisely. We can check to see if the assurance changes with either distribution; in our experience, there is little

difference. In our software, we display density estimates of σ for each of these distributions, so that the different fits can be visualised and a preferred fit chosen, if the assurance is sensitive to the choice.

4.2 | Distribution for the control group variance

There are a number of options that could be used for the control group variance σ_c^2:

1 use a point estimate or a distribution based on historical data;
2 assume that $\sigma_c^2 = \sigma_t^2$;
3 assume that σ_c^2 and σ_t^2 are independent and identically distributed;
4 elicit a separate distribution for σ_c^2.

(A fifth option could involve a hierarchical model for σ_c^2 and σ_t^2, so that learning about one could update beliefs about the other, but this would make the elicitation task considerably more difficult).

We think the first option would be the most commonly used in practice. Note that, if a mixture prior has been used for δ, one might then also consider a mixture prior for σ_t^2: if $\delta = 0$, and assuming that in that case, the treatment is no different to the control, one might also suppose $\sigma_t^2 = \sigma_c^2$.

5 | COMPUTING ASSURANCES

Given the elicited prior, we can now compute the assurance for any choice of sample sizes, using the following algorithm.

Algorithm 1 estimating an assurance

Inputs: sample sizes n_t and n_c, the elicited prior $\pi\left(\delta, \sigma_t^2, \sigma_c^2\right)$, and the number of iterations N.
For $i = 1, ..., N$:

1 sample $\delta_i, \sigma_{t,i}^2$ and $\sigma_{c,i}^2$ from $\pi\left(\delta, \sigma_t^2, \sigma_c^2\right)$;
2 sample $x_{1,i}, ..., x_{n_t,i}$ from $N\left(\delta_i, \sigma_{t,i}^2\right)$ and $y_{1,i}, ..., y_{n_c,i}$ from $N\left(0, \sigma_{c,i}^2\right)$;
3 calculate $\bar{x}_i, s_{x,i}^2$ as the sample mean and sample variance of $x_{1,i}, ..., x_{n_t,i}$, and $\bar{y}_i, s_{y,i}^2$ as the sample mean and sample variance of $y_{1,i}, ..., y_{n_c,i}$;
4 calculate the test statistic T_i and degrees of freedom ν_i:

$$
\begin{aligned}
T_i &= \frac{\bar{x}_i - \bar{y}_i}{\sqrt{\frac{s_{x,i}}{n_t} + \frac{s_{y,i}^2}{n_c}}}, \\
\nu_i &= \frac{\left(\frac{s_{x,i}}{n_t} + \frac{s_{y,i}^2}{n_c}\right)^2}{\frac{\left(s_{x,i}^2/n_t\right)^2}{n_t - 1} + \frac{\left(s_{y,i}^2/n_c\right)^2}{n_c - 1}};
\end{aligned}
$$

5 define $R_i = 1$ if $T_i > t_{0.025,\nu_i}$ and 0 otherwise, with $t_{0.025,\nu_i}$ the 97.5th percentile of the Student-t distribution with ν_i degrees of freedom. (We assume here that we require $\bar{x}_i > \bar{y}_i$ for the treatment effect to be beneficial).

The assurance is then estimated as

$$\hat{P}(R) = \frac{1}{N}\sum_{i=1}^{N} R_i.$$

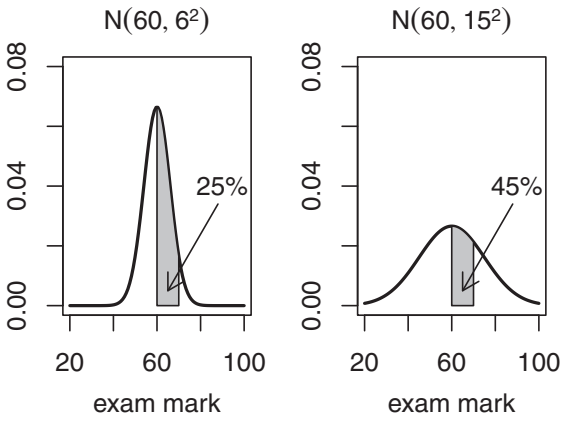

FIGURE 2 Density plots for exam marks assumed to be normally distributed with mean 60. The grey area represents the true proportion of students who get marks between 60 and 70. If the expert is certain this proportion is between 0.25 and 0.45, she is certain σ is between 6 and 15

5.1 | Example

We illustrate the method with a (fictitious) example based on Example 3 from O'Hagan et al[3] (with slight modifications to their numerical values). In their example, they consider a Phase 2 superiority trial to assess the effect of a new drug in reducing C-reactive protein (CRP) in patients with rheumatoid arthritis. Their outcome variable is a patient's reduction in CRP after 4 weeks relative to baseline, and the analysis to be performed is a two-sided test of superiority at the 5% significance level. They considered power calculations assuming $\delta = 0.2$, so we will suppose that 0.2 is the minimum clinically relevant treatment effect.

5.1.1 | The prior distribution for the treatment effect

We suppose that the expert judges a non-zero probability that the treatment will have no effect, but we will consider two scenarios: in the first, the expert judges $Pr(\delta = 0) = 0.5$, and in the second, the expert is more optimistic with $Pr(\delta = 0) = 0.1$. In both scenarios, conditional on $\delta \neq 0$, we suppose that the expert provides three quartiles from her distribution for δ: 0.25, 0.4 and 0.55. A normal distribution with mean 0.4 and standard deviation 0.22 is fitted to these judgements.

5.1.2 | The prior distribution for the variances

The expert is asked to assume that (a) the treatment is effective, with δ equal to 0.4; (b) in the control group, the mean reduction in CRP would be 0; (c) individual patients with reduction from baseline of 0.2 or less would not be judged to have received a clinically meaningful benefit. She is then asked to consider, under these assumptions, what proportion ω of patients in the treatment group would not benefit: the proportion of patients with reductions less than 0.2.

We suppose she judges that this proportion will be between 20% and 40%, which we judge to be the 5th and 95th percentiles of her distribution for ω. These correspond to 5th and 95th percentiles of her distribution for σ_t of 0.24 and 0.8, respectively. We choose to fit a gamma distribution to the precision σ_t^{-2}: minimising Equation (2) numerically, we obtain a shape parameter 2.27 and rate parameter 0.29 in the fitted gamma distribution.

Finally, we suppose that the expert judges that the same distribution will be appropriate for σ_c^2. In the case that $\delta \neq 0$, she judges σ_c^2 and σ_t^2 to be independent, and if $\delta = 0$ then she judges $\sigma_c^2 = \sigma_t^2$.

5.1.3 | Estimating assurances

Assurances are estimated using Algorithm 1. We illustrate the results that would be presented to the trial planners in Table 1, assuming equal numbers in the treatment and control arms.

TABLE 1 Estimated assurances for the two prior elicited prior distributions

Sample size per arm	10	20	50	100	1000
$Pr(\delta = 0) = 0.5$	0.28	0.36	0.42	0.45	0.49
$Pr(\delta = 0) = 0.1$	0.48	0.62	0.74	0.79	0.86

Note: There is the same "diminishing return" from increasing the sample size as one would see in a power curve, but the assurance converges (approximately) to the prior probability that the treatment is effective.

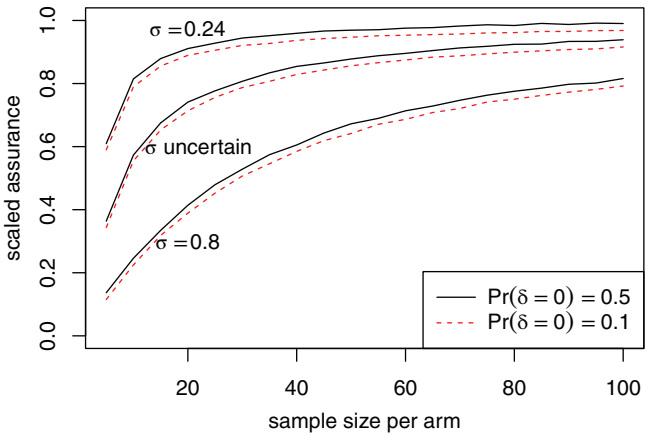

FIGURE 3 Comparing the scaled assurance for fixed variance parameters (upper and lower pairs of curves), and with the elicited prior distribution for the variances (middle pair of curves), for the two scenarios corresponding to different probabilities of no treatment effect. In this example, the scaled assurance is sensitive to different fixed values of the variance, but less sensitive to the mass placed on the event of no treatment effect

If using power to choose a sample size, we might identify the smallest sample size that can achieve some relatively high power, say in the range 80% to 90%; such a power is always achievable, given a sufficiently large sample size. With assurance, however, there will be an upper bound close to the prior probability that the treatment is effective (but not necessarily equal to this probability, given the possibility of Type I errors). If the prior probability that the treatment is effective is only 50%, there is no sample size that can achieve an assurance of 90%.

We define a "scaled assurance" as $Pr(R)/Pr(\delta > c)$, with $c = 0$, or set to some minimum level of effectiveness, if one is necessary. Though this merely re-scales the y-axis in a plot of the assurance against the sample size, it does help to emphasise the "secondary" and "tertiary" risks discussed in Section 2.1: the risks associated with an inappropriate sample size, given that the treatment *is* effective. For example, a scaled assurance close to 1 would mean that, if the treatment is effective, then there is negligible risk that the sample size is too small in the planned study. (The primary risk of an ineffective treatment would remain, of course).

One might choose to target a scaled assurance in the range 80% to 90%, just as one would do with a power. We plot the scaled assurance in Figure 3, for the two scenarios described in Section 5.1.1 (middle set of curves). We can see that for the two scenarios, the scaled assurances are fairly similar. In particular, if targeting a scaled assurance of 80%, the sample size would be similar for either prior probability that the treatment is effective.

We also plot, in Figure 3, scaled assurances where σ_t and σ_c are held equal and fixed at a common value σ (we try both the 5th and 95th percentiles of the elicited distribution for σ_t). As these parameters are varied, there is now a considerable change in the sample size required for any particular scaled assurance.

6 | MULTI-STAGE TRIALS

Given the elicited distributions, we can then investigate what information a proposed trial would provide about δ. Here, we consider a scenario in which a small trial will be conducted, and then a decision will be made whether to commit to a larger trial. The trial planner would want to know how informative the small trial would be; whether it resolve uncertainty about δ sufficiently to make the decision about the larger trial easier.

We suppose the trial sponsor chooses some threshold c of interest. From the expert's elicited distribution we will have $Pr(\delta > c) = x$, which is prior to a proposed small trial. We now consider whether the small trial would resolve this

uncertainty: given the data D that the trial would produce, whether $Pr(\delta > c|D)$ would be close to either 0 or 1. Before the study is conducted, we do not yet know what the data D would be, and so we think of $Pr(\delta > 0|D)$ as a random variable: a function of the unknown data D.

The expert's prior distribution $\pi(\delta, \sigma_c^2, \sigma_t^2)$ will imply a predictive distribution for D, given specified numbers of patients n_t and n_c in the treatment and control arms. Without loss of generality, we can assume $\mu_c = 0$ so that $\mu_t = \delta$. We can use the following simulation algorithm to explore the distribution of $Pr(\delta > 0|D)$.

Algorithm 2 simulating the information gained from a trial

Inputs: sample sizes n_t and n_c, the elicited prior $\pi(\delta, \sigma_t^2, \sigma_c^2)$, and the number of iterations N. For $i = 1, ..., N$:

1 sample $\delta_i, \sigma_{t,i}^2$ and $\sigma_{c,i}^2$ from $\pi(\delta, \sigma_t^2, \sigma_c^2)$;
2 sample $x_{1,i}, ..., x_{n_t,i}$ from $N(\delta_i, \sigma_{t,i}^2)$ and $y_{1,i}, ..., y_{n_c,i}$ from $N(0, \sigma_{c,i}^2)$;
3 define $D_i = (x_{1,i}, ..., x_{n_t,i}, y_{1,i}, ..., y_{n_c,i})$;
4 using Markov chain Monte Carlo, generate a sample $\delta_{i,1}, ..., \delta_{i,M}$ from the posterior distribution of $p(\delta|D_i)$;
5 estimate $Pr(\delta > c|D_i)$ by

$$\hat{Pr}_i = \frac{1}{M} \sum_{j=1}^{M} I(\delta_{i,j} > c),$$

where $I()$ is the indicator function. This produces an (approximate) sample $\hat{Pr}_1, ..., \hat{Pr}_N$ from the distribution of $Pr(\delta > c|D)$. We can then inspect the sample to see how many probabilities are close to either 0 or 1. We use `rjags`[24] to implement the MCMC sampling, and will comment on the choice of N and M in the following example.

6.1 | Example

We now give an illustration, continuing the example from Section 5.1. We consider the prior given by $Pr(\delta = 0) = 0.5$, with conditional distribution $\delta|\delta \neq 0 \sim N(0.4, 0.22^2)$ and $\sigma_t^{-2}, \sigma_c^{-2} \overset{i.i.d}{\sim} \text{Gamma}(\text{shape} = 2.27, \text{rate} = 0.29)$. Hence, prior to the small study, we have

$$Pr(\delta > 0) = Pr(\delta > 0|\delta \neq 0) Pr(\delta \neq 0) = (1 - \Phi(-0.4/0.22)) \times 0.5 = 0.48.$$

We consider a study with n patients per arm, and wish to assess how much more confident we would be that $\delta > 0$, given the study data D. For illustration, we will classify a study as "informative" if, once the study has produced data D, we would have either $Pr(\delta > 0|D) > 0.95$ or $Pr(\delta > 0|D) < 0.05$.

We implement Algorithm 2 with $N = 500$ simulated studies, and $M = 1000$ generated values of δ from each Markov chain. The total computation time (for four different values of n) was approximately 5 minutes on a desktop computer, using a single core (parallel computation could have been used here). This gives estimates of the probability of an "informative" study that are accurate to the first decimal place (compared with larger choices of N and M), which in this context, and noting the reliance on elicited judgements, is likely to be sufficient.

In Table 2, we illustrate the information that could be presented to a decision-maker, to enable a quick comparison of different choices of n. A more detailed summary, for $n = 20$ is presented in Figure 4.

This analysis could also be used to provide feedback about the elicited priors: in some cases certain results may be judged implausible, suggesting a problem with the choice of prior. Specifically, we can investigate the probability that a study with one or two patients per arm would be "informative": one might judge that such a probability should be close to 0. For illustration, we compare two priors:

TABLE 2 The probability that, following a study producing data D with the specified number of patients per arm, we would either have $Pr(\delta = 0|D) > 0.95$ or $Pr(\delta = 0|D) < 0.05$

Number of patients per arm	5	10	20	40
Probability of "informative" study	0.2	0.5	0.7	0.9

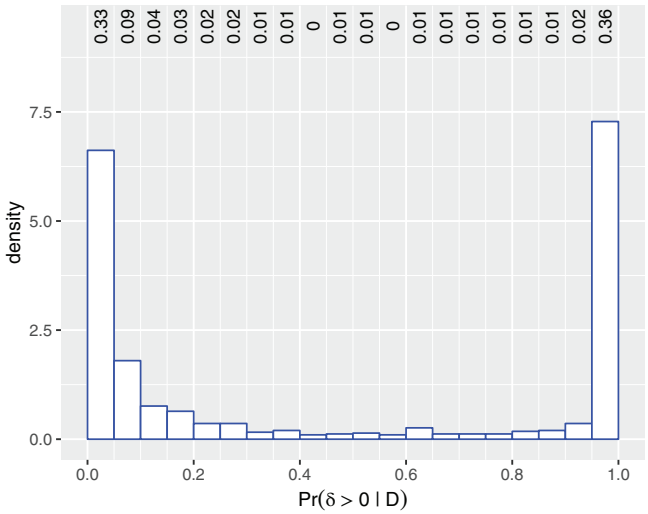

FIGURE 4 Distribution of $Pr(\delta = 0|D)$, as a function of the unknown data D resulting from a trial with 20 patients in each arm. The numbers at the top give the probabilities of $Pr(\delta = 0|D)$ lying in the respective bins. The distribution suggests that, following such a trial, it is likely we would have little uncertainty as to whether $\delta > 0$ or not

TABLE 3 The probability that, following a study producing data D with the specified number of patients per arm, we would either have $Pr(\delta = 0|D) > 0.95$ or $Pr(\delta = 0|D) < 0.05$

Number of patients per arm	1	2	5
Prior 1: probability of "informative" study	0.00	0.01	0.02
Prior 2: probability of "informative" study	0.41	0.50	0.70

Note: Arguably, the probabilities are implausibly high for Prior 2, suggesting a problem with the elicited prior distribution.

- Prior 1: $Pr(\delta = 0) = 0.5$, with conditional distribution $\delta|\delta \neq 0 \sim N(0.4, 0.22^2)$ and $\sigma_t^{-2}, \sigma_c^{-2} \overset{i.i.d}{\sim} \text{Gamma}(\text{shape} = 2.27, \text{rate} = 0.29)$.
- Prior 2: $Pr(\delta = 0) = 0.5$, with conditional distribution $\delta|\delta \neq 0 \sim N(0.4, 0.22^2)$ and $\sigma_t^{-2}, \sigma_c^{-2} \overset{i.i.d}{\sim} \text{Gamma}(\text{shape} = 43.86, \text{rate} = 0.82)$.

The first prior is the same as that used in the previous example. The gamma prior for σ_t^{-2} in Prior 2 results from using the same method in Section 4.1, but now supposing that the proportion of patients would not benefit from the treatment would be between 0.05 and 0.1, implying smaller values of σ_t^2.

Table 3 shows the estimated probabilities of an "informative" study, for each prior and different (small) values of n. Under the belief that σ_t^2 will be small, it would only take one observed response in the treatment group moderately above μ_c to "persuade" us that $\delta > 0$. Assuming this result is implausible, we would then revisit the elicited priors.

7 | SUMMARY

We have expanded the toolkit of assurance methods, to include the case of normally distributed data with uncertain population variances, including scenarios where an intermediate study is planned, to guide the final decision to proceed with a larger study. Assurance values may not be robust to changes in values of variance parameters; plugging in a single point estimate can be misleading if the point estimate is inaccurate. In such cases, it is important to first quantity the uncertainty about the variances. This in itself may add value to the trial planning process: quantifying the

uncertainty about the variance would typically involve discussion of patient heterogeneity and what is known about why some patients may benefit from the treatment more than others.

We have proposed a method for eliciting a distribution about a variance parameter, and have illustrated how to incorporate this within an assurance calculation in clinical trial planning. Making judgements about variability within a population is likely to be difficult, but our method does at least avoid asking an expert to update her beliefs given hypothetical data, or to provide summaries from her predictive distribution which would require "mentally integrating out" uncertain parameters. We have provided software for implementing our methods, which we hope will make the methodology easy to implement. Feedback and suggestions for improvements in the software will be welcome.

ACKNOWLEDGEMENT

Ziyad A. Alhussain would like to thank Deanship of Scientific Research at Majmaah University for supporting this work under Project Number R-1441-126.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Jeremy E. Oakley https://orcid.org/0000-0002-9860-4093

REFERENCES

1. Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial based on subjective clinical opinion. *Stat Med*. 1986;5:1-13.
2. O'Hagan A, Stevens JW. Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Med Decis Making*. 2001;21:219-230.
3. O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. *Pharm Stat*. 2005;4:187-201.
4. Crisp A, Miller S, Thompson D, Best N. Practical experiences of adopting assurance as a quantitative framework to support decision making in drug development. *Pharm Stat*. 2018;17:317-328.
5. Dallow N, Best N, Montague TH. Better decision making in drug development through adoption of formal prior elicitation. *Pharm Stat*. 2018;17:317-328.
6. Ren S, Oakley JE. Assurance calculations for planning clinical trials with time-to-event outcomes. *Stat Med*. 2014;33(1):31-45. https://doi.org/10.1002/sim.5916.
7. Gasparini M, Di Scala L, Bretz F, Racine-Poon A. Some uses of predictive probability of success in clinical drug development. *Epidemiology Biostatistics and Public Health*. 2013;10:e8760-1-e8760-14. https://doi.org/10.2427/8760.
8. Nixon RM, O'Hagan A, Oakley J, et al. The rheumatoid arthritis drug development model: a case study in Bayesian clinical trial simulation. *Pharm Stat*. 2009;8(4):371-389.
9. Hampson LV, Whitehead J, Eleftheriou D, et al. Elicitation of expert prior opinion: application to the MYPAN trial in childhood Polyarteritis Nodosa. *PLoS One*. 2015;10(3):e0120981. https://doi.org/10.1371/journal.pone.0120981.
10. Dias LC, Morton A, Quigley J, eds. *Elicitation: The Science and Art of Structuring Judgement*. New York: Springer; 2018.
11. EFSA (European Food Safety Authority). Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. *EFSA Journal*. 2014;12(6):3734. https://doi.org/10.2903/j.efsa.2014.3734
12. O'Hagan A, Buck CE, Daneshkhah A, et al. *Uncertain Judgements: Eliciting Expert Probabilities*. Chichister: John Wiley and Sons Ltd.; 2006.
13. Cooke R. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. New York: Oxford University Press; 1991.
14. Morgan MG, Henrion M. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge: Cambridge University Press; 1990.
15. Oakley JE, O'Hagan A. SHELF: The Sheffield Elicitation Framework (version 4.0); 2010. School of Mathematics and Statistics, University of Sheffield, UK. http://tonyohagan.co.uk/shelf
16. Kadane JB, Wolfson LJ. Experiences in elicitation. *Journal of the Royal Statistical Society. Series D (The Statistician)*. 1998;47(1):3-19.
17. Oakley JE. Eliciting univariate probability distributions. In: Böcker K, ed. *Rethinking Risk Measurement and Reporting: Volume I Uncertainty, Bayesian Analysis and Expert Judgement*. London: Risk Books; 2010:155-177.
18. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2019.
19. Oakley J. *SHELF*: *Tools to Support the Sheffield Elicitation Framework*; 2019. R Package Version 1.5.0.9000.
20. Kadane J, Dickey J, Winkler R, Smith W, Peters S. Interactive elicitation of opinion for a normal linear model. *J Am Stat Assoc*. 1980;75(372):845-854.
21. Al-Awadhi S, Garthwaite P. An elicitation method for multivariate normal distributions. *Commun Stat*. 1998;27(5):1123-1142.
22. Tversky A, Kahneman D. Belief in the law of small numbers. *Psychol Bull*. 1971;76(2):105-110.
23. Al-Awadhi S, Garthwaite P. Prior distribution assessment for a multivariate normal distribution: an experimental study. *J Appl Stat*. 2001;28(1):5-23.

24. Plummer M. *rjags*: *Bayesian Graphical Models using MCMC*; 2018. R Package Version 4-8.
25. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. *shiny*: *Web Application Framework for R*; 2019. R Package Version 1.3.2.

How to cite this article: Alhussain ZA, Oakley JE. Assurance for clinical trial design with normally distributed outcomes: Eliciting uncertainty about variances. *Pharmaceutical Statistics*. 2020;19:827–839. https://doi.org/10.1002/pst.2040

APPENDIX

An R package, `assurance`, for implementing the methods described in this article is available on GitHub, at `https://github.com/OakleyJ/assurance`. The website also includes an illustration of using the package to replicate the examples in this article.

This package currently requires the `SHELF` R package, available on CRAN. These packages can be installed in R with the commands.

```
install.packages(c("devtools", "SHELF"))
devtools::install_github("OakleyJ/assurance")
```

For non-R users, an app for implementing these methods, produced with `shiny`,[25] can be used online at `https://jeremy-oakley.shinyapps.io/assurance-normal/`. A version of the app for offline use is included in the `assurance` package.