Published paper

Assurance calculations for planning clinical trials with time-to-event outcomes

Shijie Ren[a*†] and Jeremy E. Oakley[b]

We consider the use of the assurance method in clinical trial planning. In the assurance method, which is an alternative to a power calculation, we calculate the probability of a clinical trial resulting in a successful outcome, via eliciting a prior probability distribution about the relevant treatment effect. This is typically a hybrid Bayesian-frequentist procedure, in that it is usually assumed that the trial data will be analysed using a frequentist hypothesis test, so that the prior distribution is only used to calculate the probability of observing the desired outcome in the frequentist test. We argue that assessing the probability of a successful clinical trial is a useful part of the trial planning process. We develop assurance methods to accommodate survival outcome measures, assuming both parametric and nonparametric models. We also develop prior elicitation procedures for each survival model so that the assurance calculations can be performed more easily and reliably. We have made free software available for implementing our methods. © 2013 The Authors. Statistics in Medicine published by John Wiley & Sons, Ltd.

Keywords: assurance; elicitation; prior distribution; power; sample size; survival analysis

1. Introduction

Sample size determination is an important part of clinical trial design and conventionally involves power calculations. However, the power of a trial does not necessarily give the probability of the trial demonstrating a treatment effect, as the true treatment effect may be different to that assumed in the power calculation. Several authors have proposed a hybrid classical-Bayesian approach for assessing the probability of a successful trial, given the sample size only, which can then be used to inform sample size decisions.

The hybrid method was first considered by Spiegelhalter and Freedman [1]. They constructed an unconditional probability of having a desired outcome and called this unconditional probability the average power. O'Hagan and Stevens [2] used this method for choosing sample sizes for clinical trials of cost-effectiveness. They referred to the unconditional probability of a successful trial as the 'assurance' of the trial, and we use this term here. O'Hagan et al. [3] extended assurance methods to two-sided testing and equivalence trials, covering the use of non-conjugate prior distributions for uncertain parameters. Chuang-Stein [4] discussed the difference between traditional power calculations and assurance calculations to determine sample sizes, giving an example of planning the next trial based on the results of an early trial. Chuang-Stein and Yang [5] reviewed the concept of assurance and illustrated its use when planning phase III trials. They also applied assurance to study designs when re-estimating a sample size based on an interim analysis.

An assurance calculation requires a prior distribution for the treatment effect, but does not necessarily involve a Bayesian analysis of the trial data. The method of analysis, and in particular the criteria for

[a]School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA, U.K.

[b]School of Mathematics and Statistics, University of Sheffield, The Hicks Building, Hounsfield Road, Sheffield, S3 7RH, U.K.

*Correspondence to: Shijie Ren, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA, U.K.

†E-mail: s.ren@sheffield.ac.uk

which the trial is deemed a 'success', are determined externally, for example, by a regulator. Once the criteria have been specified, a prior distribution is used to assess the probability that these criteria will be met. Typically, the prior distribution will only be used in the design stage and not the analysis. At the design stage, the risk of trial failure is primarily the trial sponsor's, and so it should be uncontroversial for a trial sponsor to use all their prior knowledge in assessing such a risk.

We consider clinical trials in which the endpoint of interest is a survival time. For time-to-event outcome measures, power and sample size calculations have been well studied under various model assumptions. For example, Schoenfeld and Richter [6] developed a power function with a limited recruitment period and a pre-specified follow-up period under the assumption that the survival times in each treatment group follow exponential distributions and patients enter the trial uniformly. Gross and Clark [7] provided a method of calculating sample size by assuming that the sample mean survival time is approximately normally distributed under Weibull models for the survival times. Freedman [8] and Schoenfeld [9] derived sample size formulae under the assumption of proportional hazards based on asymptotic properties of the logrank statistic.

Little has been done in calculating assurance for survival endpoints. Assuming proportional hazards, Spiegelhalter *et al.* [10] derived an assurance formula in the case of equal allocation and follow-up. The only uncertain variable considered was the log hazard ratio, and a normal prior was assumed. In this paper, we extend assurance calculations to accommodate both parametric and proportional hazards models. Under proportional hazards models, we derive an assurance formula assuming uniform patient entry over a limited recruitment period. We consider uncertainty in both the log hazard ratio and the baseline survivor function.

In Section 2, we review how assurance is calculated to determine the unconditional probability of having a desired outcome. In Section 3, we derive assurance calculations for exponential and Weibull survival models and describe the elicitation methods for the required prior distributions. In Section 4, we extend assurance calculations to accommodate proportional hazards models, considering uncertainty in both treatment effect and baseline survivor function. We also describe the procedure of generating the baseline survivor function. Examples are given in Section 5.

2. Assurance and sample size

We now review the concept of assurance. Suppose that a randomised controlled trial is to be conducted to compare an experimental treatment and a standard treatment for a particular disease. A hypothesis test is to be carried out to test the null hypothesis that the treatment effect $\theta = 0$ against the alternative that $\theta \neq 0$. On the basis of a power calculation, the sample size is chosen to solve

$$P(\text{Reject } H_0 | \theta = \theta_A) = \pi^*, \tag{1}$$

for some desired probability π^*.

The power of the test $P(\text{Reject } H_0 | \theta = \theta_A)$ provides the probability of successfully rejecting the null hypothesis if the true value of θ is the specified θ_A. As the true value of θ may be very different to θ_A, the actually probability of successfully rejecting the null hypothesis may be very different to the power.

Assurance is the unconditional probability that the trial will end with the desired outcome, which we derive via

$$P(\text{'Successful trial'}) = \int P(\text{'Successful trial'} | \theta) f(\theta) d\theta, \tag{2}$$

where $f(\theta)$ is the prior distribution for the true treatment effect θ. If a successful trial simply corresponds to rejecting a null hypothesis of no treatment effect, then the assurance in (2) can be thought of as an expected power (interpreting θ_A in (1) as the true value of the treatment effect, rather than some minimum clinically relevant difference).

If our desired outcome is to reject the null hypothesis with data favouring the experimental treatment, then assurance is given by

$$\gamma = P(\text{'Successful trial'})$$
$$= P(\text{Reject } H_0, \hat{\theta} > 0)$$
$$= \int P(\text{Reject } H_0, \hat{\theta} > 0 | \theta) f(\theta) d\theta,$$

with $\hat{\theta} > 0$ indicating that the data favour the experimental treatment. We again emphasise that specifying what constitutes a 'successful trial' is not part of the assurance method; the criteria determining a successful trial are set externally, and the idea of assurance is to use prior information to determine the probability that these criteria are met.

The power of a clinical trial can, in theory, be made as large as desired by increasing the sample size. The same does not hold for assurance. For a large enough sample size, we will 'observe' the true treatment effect, so that the assurance converges to the prior probability that the new treatment is suitably effective. If this prior probability is low, no trial will have a high assurance of success; we cannot 'beat the prior'.

We restrict the scope of this paper to frequentist methods for analysing the trial data, but Bayesian methods could be used and are discussed in depth by Ibrahim *et al.* [11] and Christensen *et al.* [12]. (In particular, Ibrahim *et al.* [11] discussed Bayesian inference for all the survival models considered in this paper.) In this case, there may be a distinction between the prior used in the design stage, and the prior used in the analysis stage, if the regulator is not willing to accept the trial sponsor's prior.

3. Assurance calculations for parametric survival models

We now suppose that, in each of two treatment groups, the outcome variable for each patient is the survival time to some event and consider exponential and Weibull models for the survival times. For each model, we first choose the analysis method, and hence the criteria for a successful trial. We then consider assurance calculations and elicitation methods for the required prior distributions.

3.1. Exponential distribution

We first suppose that the survival times in each treatment group follow an exponential distribution, with hazard rates λ_1 and λ_2 ($i = 1$ for the control group and $i = 2$ for the experimental group) and allow for a limited recruitment period from time 0 to R with uniform patient entry and T as the total trial length. The time origin for survival time is when a patient enters the trial, not when the trial starts. Here, we consider the analysis method based on Schoenfeld and Richter [6].

The null hypothesis is $m_1/m_2 = 1$, where m_i is the median survival time in group i, against the alternative $m_1/m_2 = \phi$, where ϕ is the minimum clinically important difference. Note that assuming an exponential model, the hypotheses stated earlier are equivalent to $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$, where

$$\theta = \log(\lambda_2/\lambda_1). \tag{3}$$

The test statistic is

$$\frac{\hat{\theta} - 0}{\sqrt{\frac{1}{d_1} + \frac{1}{d_2}}} \sim N(0, 1) \text{ under } H_0 : \theta = 0,$$

where d_i is the number of events in group i and $\hat{\theta}$ is the maximum likelihood estimate of θ.

On the basis of the asymptotic properties of the test statistic, the power π^E of a $100\alpha\%$ two-sided test is

$$\pi^E = P(\text{Reject } H_0 | \lambda_1, \lambda_2)$$

$$= \Phi\left(-\frac{\theta}{\sqrt{\frac{1}{N_1 P_{1e}} + \frac{1}{N_2 P_{2e}}}} - Z_{\alpha/2}\right) + \Phi\left(\frac{\theta}{\sqrt{\frac{1}{N_1 P_{1e}} + \frac{1}{N_2 P_{2e}}}} - Z_{\alpha/2}\right), \tag{4}$$

where N_i is the number of patients in group i and P_{ie} is the probability of an individual patient in group i experiencing the outcome event during the trial. Schoenfeld and Richter [6] derived an exact formula for P_{ie}:

$$P_{ie} = 1 - \frac{\exp\{-\lambda_i(T - R)\} - \exp\{-\lambda_i T\}}{\lambda_i R} \quad \text{for } i = 1, 2. \tag{5}$$

The assurance of rejecting the null hypothesis with data favouring the experimental treatment is

$$\gamma^E = P(\text{Reject } H_0, \hat{\theta} < 0)$$

$$= \int \Phi\left(-\frac{\theta}{\sqrt{\frac{1}{N_1 P_{1e}} + \frac{1}{N_2 P_{2e}}}} - Z_{\alpha/2}\right) f(\lambda_1, \lambda_2) \mathrm{d}\lambda_1 \mathrm{d}\lambda_2, \tag{6}$$

where θ and P_{ie} are functions of λ_1 and λ_2.

3.1.1. Constructing the priors. From (6), we see that a joint prior is required for λ_1 and λ_2. Kadane and Wolfson [13] argued that it is better to ask for opinion about observable quantities rather than parameters in statistical models, and we follow their advice here. We elicit $f(\lambda_1, \lambda_2)$ via judgements about survival rates at some specified time.

To construct a prior for each parameter, first note that

$$\lambda_i = \frac{-\log(S_i(t_0))}{t_0}, \quad \text{for } i = 1, 2, \tag{7}$$

with $S_i(t_0)$ the survival rate for group i at time t_0. Hence, to elicit the joint prior distribution, we first elicit judgements about $S_1(t_0)$ instead of eliciting beliefs about λ_1 directly. An expert may judge $S_1(t_0)$ to be informative for $S_2(t_0)$, so that λ_1 and λ_2 are not independent. To elicit this dependence, we propose to elicit judgements about the difference, $\rho = S_2(t_0) - S_1(t_0)$, and assume that ρ is independent of $S_1(t_0)$.

Methods for eliciting univariate distributions are given in Sections 6.3 and 6.4 of O'Hagan *et al.* [14] and can be implemented using the freely available SHELF package [15] and the MATCH online elicitation tool available at http://optics.eee.nottingham.ac.uk/match/uncertainty.php. See also Johnson *et al.* [16] for a systematic review of elicitation methods.

One option is to elicit a beta distribution for $S_1(t_0)$ and a normal distribution for ρ, truncating the normal prior if necessary to ensure $S_2(t_0) \in (0, 1)$. (An alternative would be to use a shifted and scaled beta distribution for ρ, although we have not found the need to truncate to cause significant computational problems.)

For illustration, we describe a 'trial roulette' method proposed by Gore [17] to elicit a normal prior for ρ. The method is based on the fixed interval approach, in which the expert is asked to provide a probability that the unknown quantity of interest will fall in a pre-fixed interval. Using the SHELF package, the facilitator, who conducts the elicitation, firstly elicits from the expert the lower and upper bounds of the range of plausible values for ρ. Then the facilitator divides the range from the lower bound to upper bound in to 10 equal-width 'bins'. The expert is asked to specify his or her probability of ρ lying in a particular bin by placing 'chips' in that bin, with the proportion of chips allocated representing the probability. The number of chips given to the expert is specified by the facilitator. For example, if in total 20 chips are used, then each chip represents a probability of 0.05. The trial roulette method is simple to use and provides the expert with an immediate display of her elicited judgements.

A parametric distributed can be fitted to the elicited judgements using a least squares procedure: the parameters are chosen to make the fitted probabilities as close as possible to the elicited probabilities. Feedback should be provided to the expert for checking the adequacy of the elicited distribution.

One illustration of the use of elicitation in clinical trials is given by Parmar *et al.* [18], who used a questionnaire to elicit the log hazard ratio given in (3) by eliciting a point estimate for $S_1(t_0)$ and a prior distribution for ρ using the roulette method. Tan *et al.* [19] and Hiance *et al.* [20] both adapted this questionnaire and used it for a phase III trial. Our elicitation process is more complicated, as we consider uncertainty in both $S_1(t_0)$ and ρ.

Given the elicited distributions $f(S_1(t_0))$ and $f(\rho)$, we estimate γ^E using Monte Carlo simulation:

$$\gamma^E \simeq \frac{1}{M} \sum_{j=1}^{M} \Phi\left(\frac{\theta^{(j)}}{\sqrt{\frac{1}{N_1 P_{1e}^{(j)}} + \frac{1}{N_2 P_{2e}^{(j)}}}} - Z_{\alpha/2}\right), \tag{8}$$

where $\theta^{(j)}$ and $P_{ie}^{(j)}$ for $i = 1, 2$, are obtained by the following steps.

1. Simulate $S_1^{(j)}(t_0)$ from the elicited prior distribution $f(S_1(t_0))$.
2. Simulate $\rho^{(j)}$ from the elicited prior distribution $f(\rho)$ and calculate $S_2^{(j)}(t_0) = S_1^{(j)}(t_0) + \rho^{(j)}$.
3. Calculate $\lambda_1^{(j)}$ and $\lambda_2^{(j)}$ using (7).
4. Calculate $\theta^{(j)}$ using (3).
5. Calculate $P_{1e}^{(j)}$ and $P_{2e}^{(j)}$ using (5).

The process is computationally quick, so we can make M very large to ensure convergence.

3.2. Supporting software

We have made available software to implement the methods in this paper. The software can be downloaded from www.jeremy-oakley.staff.shef.ac.uk/assurance.zip. For the exponential case, we have a written an interactive elicitation tool for computing assurance. The code is written in R [21] and uses the `rpanel` package of Bowman and Crawford [22] and the `tkrplot` package of Tierney [23] to provide interactive graphics. The tool helps to elicit the prior distributions of the baseline survival rate $S_1(t_0)$ at a specified time t_0, and the survival difference ρ between the experimental group and the control group at time t_0, using the trial roulette method. The tool also provides feedback to check the adequacy of the elicited distributions. Once the priors are specified, the tool draws both power and assurance curves for the corresponding elicited distributions. Users can see immediately how changes in the elicited beliefs affect the assurance.

3.3. Weibull distribution

We now suppose that the survival times of patients receiving the standard and experimental treatment follow Weibull distributions, with scale parameters λ_1, λ_2 and shape parameters κ_1, κ_2, respectively. The probability density function of the Weibull distribution in each treatment group is

$$f(t) = \lambda_i \kappa_i t^{\kappa_i - 1} \exp\left(-\lambda_i t^{\kappa_i}\right), \quad t \geqslant 0, \kappa_i > 0, \lambda_i > 0,$$

for $i = 1, 2$. The method of analysis that we consider here is to compare mean survival times for each group. We assume that the sample mean survival times are approximately normally distributed: $\overline{X}_i \sim N\left(\mu_i, \sigma_i^2/N_i\right)$, with N_i the number of patients in group i ($i = 1$ for the control group, $i = 2$ for the experimental group).

Gross and Clark [7] derived a power function of a $100\alpha\%$ two-sided test of the null hypothesis that the mean survival times are the same, $\mu_1 = \mu_2$, against the alternative $\mu_1 \neq \mu_2$. They used the test statistic

$$\frac{\overline{X}_2 - \overline{X}_1}{\sqrt{\sigma_1^2/N_1 + \sigma_2^2/N_2}} \sim N(0, 1) \text{ under } H_0 : \mu_1 = \mu_2.$$

The power formula is

$$
\begin{aligned}
\pi^W &= P(\text{Reject } H_0 | \mu_1, \mu_2) \\
&= \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{\sigma_1^2/N_1 + \sigma_2^2/N_2}} - Z_{\alpha/2}\right) + \Phi\left(-\frac{\mu_2 - \mu_1}{\sqrt{\sigma_1^2/N_1 + \sigma_2^2/N_2}} - Z_{\alpha/2}\right),
\end{aligned}
\tag{9}
$$

where μ_i and σ_i^2, for $i = 1, 2$, are expressed by

$$\mu_i = \frac{\Gamma(1 + 1/\kappa_i)}{\lambda_i^{(1/\kappa_i)}}, \tag{10}$$

$$\sigma_i^2 = \frac{\Gamma(1 + 2/\kappa_i) - [\Gamma(1 + 1/\kappa_i)]^2}{\lambda_i^{(2/\kappa_i)}}. \tag{11}$$

As the variance parameters σ_1^2 and σ_2^2 are unknown, we switch to using a two-sample t-test in the assurance calculation. The assurance of rejecting the null hypothesis with data favouring the experimental treatment is given by

$$\gamma^W = P(\text{Reject } H_0, \bar{X}_2 > \bar{X}_1)$$

$$= \int P\left(\frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2}} > T_{1-\alpha/2;\nu}\right) f(\kappa 1, \kappa_2, \lambda_1, \lambda_2) \mathrm{d}\kappa_1 \mathrm{d}\kappa_2 \mathrm{d}\lambda_1 \mathrm{d}\lambda_2, \quad (12)$$

where $T_{1-\alpha/2;\nu}$ is the $100 \times (1 - \alpha/2)$ percentile from the t-distribution with degrees of freedom ν calculated according to Welch's t-test. As in the exponential case, we estimate this integral using Monte Carlo simulation.

3.3.1. Constructing the priors. To derive the assurance, a joint prior distribution for $\kappa_1, \kappa_2, \lambda_1$ and λ_2 is needed. Clearly, making judgements directly about these parameters would be too difficult, so we again construct the priors from judgments about survival rates.

Several authors have presented methods for eliciting an expert's opinion for a single Weibull distribution. In Singpurwalla [24], beliefs about the median survival time and shape parameter κ are elicited. Berger and Sun [25] and Kaminskiy and Krivtsov [26] both considered a predictive approach, in which survival rates at two specified times are elicited. We consider a similar approach, allowing for the possibility of dependence between the two uncertain survival distributions.

For each group, the shape and scale parameters can be estimated using the survival rate after two periods. Let $S_i(t_0)$ and $S_i(t_0')$ be the survival rates at time t_0 and t_0', where $t_0' > t_0$ without loss of generality. The Weibull parameters are derived from

$$\kappa_i = \frac{\log\left(\frac{\log(S_i(t_0'))}{\log(S_i(t_0))}\right)}{\log\left(\frac{t_0'}{t_0}\right)}, \quad (13)$$

$$\lambda_i = \frac{-\log(S_i(t_0))}{t_0^{\kappa_i}}, \quad \text{for } i = 1, 2. \quad (14)$$

We could elicit the prior distribution for κ_i and λ_i by eliciting an expert's opinion about $S_i(t_0)$ and $S_i(t_0')$ and then applying (13) and (14), but the expert may judge that $S_i(t_0)$ and $S_i(t_0')$ are dependent. Instead, we suggest eliciting beliefs about the following four observable quantities (assuming independence):

$$S_1(t_0),$$
$$\delta_{11} = S_1(t_0) - S_1(t_0'), \quad (15)$$
$$\delta_{12} = S_2(t_0) - S_1(t_0), \quad (16)$$
$$\delta_{22} = S_2(t_0) - S_2(t_0'), \quad (17)$$

and this will induce a joint prior for $\kappa_1, \kappa_2, \lambda_1$ and λ_2. Another option is to elicit judgments about odds ratios instead of differences.

If beliefs about the differences between two survival rates are elicited, one option is to elicit a beta distribution for $S_1(t_0)$, and a normal prior for δ_{12}, beta priors for δ_{11}, δ_{22}. It may be necessary to truncate the priors to ensure $S_1(t_0'), S_2(t_0)$ and $S_2(t_0')$ all in the range $(0, 1)$, but this is unlikely to be a significant computational problem. If the odds ratios are the uncertain quantities of interests, we could again elicit a beta distribution for $S_1(t_0)$, and lognormal distributions for the odds ratios.

We estimate γ^W using Monte Carlo simulation:

$$\gamma^W \simeq \frac{1}{M} \sum_{j=1}^{M} I\left(\frac{\bar{X}_2^{(j)} - \bar{X}_1^{(j)}}{\sqrt{\hat{\sigma}_1^{2(j)}/N_1 + \hat{\sigma}_2^{2(j)}/N_2}} > T_{1-\alpha/2;\nu}\right), \quad (18)$$

where $I()$ is the indicator function, and the simulation procedure for each j is as follows.

1. Simulate $S_1^{(j)}(t_0)$, $\delta_{11}^{(j)}$, $\delta_{12}^{(j)}$ and $\delta_{22}^{(j)}$ from their elicited prior distribution.
2. Calculate $S_1^{(j)}(t_0')$, $S_2^{(j)}(t_0)$ and $S_2^{(j)}(t_0')$ from the sampled values in step 1 using Equations (15)–(17).
3. Calculate $\kappa_i^{(j)}$ and $\lambda_i^{(j)}$ for $i = 1, 2$, using Equations (13) and (14).
4. Simulate survival times $X_{i1}^{(j)}, \ldots, X_{iN_i}^{(j)}$ for $i = 1, 2$, for the two groups, from Weibull distributions with the parameter values calculated in step 3.
5. Calculate the sample means $\bar{X}_i^{(j)}$ and sample variances $\hat{\sigma}_i^{2(j)}$ for $i = 1, 2$, from the simulated data in step 4.

Again, the process is computationally quick, so M can be chosen to be very large to ensure convergence. The R code to compute the assurance for the Weibull model is also available in the supporting software described in Section 3.2.

3.4. Priors based on historical data

Suitable historical data may be available for informing the prior distributions, particularly with regard to the control arm of a trial. We could then derive a posterior distribution given the historical data (following the approaches described by Ibrahim *et al.* [11]), which could be used as the prior in the assurance calculation. For example, in the exponential case, the distribution for the rate parameter λ_1 in the control arm could be based on the historical data (and perhaps a noninformative prior), and we would then only elicit prior judgements about the difference between the treatments $\rho = S_2(t_0) - S_1(t_0)$.

Ibrahim *et al.* [11]) also described the use of 'power priors'. This also involves deriving a posterior distribution given the historical data, but a posterior in which the likelihood function is downweighted, by raising it to a power between 0 and 1. The downweighting may be used to reflect differences between the study populations in the historical and new trials.

4. Assurance calculation for proportional hazard and nonparametric survivor function models

We now consider a proportional hazards model, with no parametric assumption about the underlying survivor functions, for a two-arm trial with uniform patient entry during the recruitment period 0 to R and total study length T. We suppose that the trial results will be analysed with a logrank test. Let h_i and S_i denote the hazard and survivor function for treatment group i ($i = 1$ for the control group and $i = 2$ for the experimental group), respectively, with ϕ the hazard ratio $h_2(t)/h_1(t)$. A two-sided $100\alpha\%$ logrank test is performed to test the null hypothesis, H_0, that the log hazard ratio, $\theta = \log(\phi)$, is zero against the alternative $\theta \neq 0$.

Assuming equal number of patients per treatment group, the power formula is

$$\pi = P(\text{Reject } H_0|\theta)$$

$$= \Phi\left(-\theta\sqrt{\frac{d}{4}} - Z_{\alpha/2}\right),$$

where d is the total number of events in the trial. Spiegelhalter *et al.* [10] derived an exact assurance formula assuming a normal prior $N(m, v)$ for θ:

$$\gamma = \Phi\left(\frac{-Z_{\alpha/2}\sqrt{\frac{d}{4}} - m\frac{d}{4}}{\sqrt{\frac{d}{4} + \left(\frac{d}{4}\right)^2 v}}\right).$$

This assumes that all patients are monitored until the outcome event with no limitation of the length of the trial. We extend this formula to allow for limited recruitment and follow-up periods.

Our assurance calculation is based on the power formula derived by Schoenfeld [9]. The recruitment rate is uniform over an interval 0 to R, and there is a follow-up period to time T. The time origin for survival time is when a patient enters the trial. Given the total number of patients N, the test statistic is

$$\frac{U - 0}{\sqrt{\frac{NP_e}{\frac{1}{Q_1} \times \frac{1}{Q_2}}}} \sim N(0, 1) \text{ under } H_0 : \theta = 0,$$

where U is the logrank statistic, P_e is the probability that an individual patient will experience the outcome event during the trial and Q_i is the proportion of patients allocated to group i for $i = 1, 2$. Let P_{1e} and P_{2e} denote the probabilities that a patient from treatment groups 1 and 2 will experience the outcome event during the trial, respectively. The power formula for a two-sided $100\alpha\%$ logrank test is

$$\pi^P = P(\text{Reject } H_0 | \theta, P_e)$$

$$= \Phi\left(-\theta\sqrt{\frac{NP_e}{\frac{1}{Q_1} \times \frac{1}{Q_2}}} - Z_{\alpha/2}\right) + \Phi\left(\theta\sqrt{\frac{NP_e}{\frac{1}{Q_1} \times \frac{1}{Q_2}}} - Z_{\alpha/2}\right), \quad (19)$$

where

$$P_e = Q_1 P_{1e} + Q_2 P_{2e}. \quad (20)$$

The assurance, considering uncertainty in θ and P_e, is

$$\gamma^P = P(\text{Reject } H_0, \hat{\theta} < 0)$$

$$= \int \Phi\left(-\theta\sqrt{\frac{NP_e}{\frac{1}{Q_1} \times \frac{1}{Q_2}}} - Z_{\alpha/2}\right) f(\theta, P_e) \mathrm{d}\theta \mathrm{d}P_e, \quad (21)$$

where $\hat{\theta} < 0$ implies that the experimental treatment is better than the standard/placebo.

To compute the assurance γ^P, a joint prior for θ and P_e is required. Depending on whether there are data available, the elicitation procedures for P_e are different. When there are no data available about the standard treatment, we could elicit P_e directly, and the assurance is calculated using Equation (21). When data about the standard treatment are available, we will use the data to learn about $S_1(.)$ and then derive P_e from

$$P_e = Q_1 P_{1e} + Q_2 P_{2e}$$

$$= Q_1 \left(1 - \frac{1}{R}\int_{T-R}^{T} S_1(u)\mathrm{d}u\right) + Q_2 \left(1 - \frac{1}{R}\int_{T-R}^{T} [S_1(u)]^{\exp(\theta)}\mathrm{d}u\right). \quad (22)$$

With both the data available and no data available cases, beliefs about the log hazard ratio θ are required. The model parameter θ can be expressed in terms of the survival rates at a fixed time point t_0 in each group:

$$\theta = \log\left(\frac{\log[S_1(t_0) + \rho]}{\log[S_1(t_0)]}\right), \quad (23)$$

where ρ denotes the difference between the survival rates in the two groups at time t_0 (group 2 minus group 1). We elicit opinion about the survival rates $S_1(t_0)$ and ρ instead of the model parameter θ as in the case of exponential models.

4.1. Constructing the priors with no data available

From (21), a joint prior distribution for θ and P_e is required. Under the proportional hazards model, P_{2e} is given by

$$P_{2e} = 1 - (1 - P_{1e})^{\exp(\theta)}. \quad (24)$$

Hence, to elicit the joint prior distribution for θ and P_e, we can elicit beliefs about P_{1e}, $S_1(t_0)$ and ρ and then apply (24), (23) and (22). As in Section 3.1.1, we could elicit independent beta distributions for P_{1e} and $S_1(t_0)$ and an independent normal distribution for ρ. As before, a Monte Carlo simulation can be used to estimate (21).

4.2. Constructing the priors with available data

Information from a pilot study or historic data for the standard treatment may be available at the planning stage. In this section, we describe a method of incorporating both information from the data and expert opinion to obtain the final joint prior distribution for the assurance calculation.

From (22), P_e is determined by $S_1(.)$ and θ. Hence, we consider a joint distribution for $(\theta, S_1(.))$ rather than (θ, P_e). The integral in the first term of (22) can be estimated numerically, for example, using Simpson's rule:

$$\hat{P}_{1e} = 1 - \frac{1}{3(H-1)} \sum_{k=1}^{H} w_k S_1(u_k),\tag{25}$$

where $w_k = 1, 4, 2, 4, 2 \ldots, 4, 1$ for $k = 1, \ldots, H$, and H is the number of subintervals in the interval $(T - R)$ to T with H an odd number. We now just require a joint prior for $(\theta, S_1(u_1), \ldots, S_1(u_H))$.

Taking into account both uncertainty in the log hazard ratio θ and baseline survival rates $S_1(u_k)$ for $k = 1, \ldots, H$, the (approximate) assurance γ^{P_D} is

$$\begin{aligned}\gamma^{P_D} &= P(\text{Reject } H_0, \hat{\theta} < 0)\\ &= \int P(\text{Reject } H_0, \hat{\theta} < 0 | \theta, S_1(u_1), \ldots, S_1(u_H))\\ &\quad \times f(\theta, S_1(u_1), \ldots, S_1(u_H)) \mathrm{d}\theta \mathrm{d}S_1(u_1) \ldots \mathrm{d}S_1(u_H).\end{aligned}\tag{26}$$

We suppose that survival data (with right censoring at J distinct times, $\tau_1, \ldots, \tau_J$) are available for the standard treatment, and so consider inference for $S_1(u_1), \ldots, S_1(u_H)$ using a Dirichlet distribution, as in Susarla and Ryzin [27]. To handle censoring, we use a Gibbs sampling approach, as suggested by Kuo and Smith [28], to generate $(S_1(u_1), \ldots, S_1(u_H))$ from its posterior distribution, which is another Dirichlet distribution. See also Ibrahim *et al.* [11] for further discussion of inference for survivor functions using Dirichlet distributions.

The problem with censored data is that we do not observe the exact event times. Hence, in the Gibbs sampler, we firstly simulate event times for each censored observation and then update the prior distribution using these simulated times and the observed uncensored event times.

We now describe the procedure for simulating $(S_1(u_1), \ldots, S_1(u_H))$ using the Gibbs sampling approach. The first step is to partition the sample space $[(T - R), T]$ into $(H + J)$ subintervals according to the censored times and the quadrature points in Simpson's rule. The uncertain quantities of interest are $S_1(u_1), \ldots, S_1(u_H)$, and the nuisance parameters are $S_1(\tau_1), \ldots, S_1(\tau_J)$. To simplify the notation, we define

$$S_1(u_1), S_1(u_2), \ldots, S_1(\tau_j), \ldots, S_1(u_H) \equiv S_1(t_1), S_1(t_2), \ldots, S_1(t_{H+J-1}), S_1(t_{H+J}),$$

where t_j is the jth smallest value in the set $\{u_1, \ldots, u_H, \tau_1, \ldots, \tau_J\}$.

We define a vector of probabilities $p_{1:H+J+1} = (p_1, p_2, \ldots, p_{H+J+1})$ of an event occurring in each subinterval:

$$p_{1:H+J+1} \equiv (S_1(0) - S_1(t_1), S_1(t_1) - S_1(t_2), \ldots, S_1(t_{H+J-1}) - S_1(t_{H+J}), S_1(t_{H+J})).$$

We consider a Dirichlet process prior for $S_1(.)$ with parameter function α, of the form $\alpha([t, \infty)) = c_0 G(t)$. The function $G(.)$ represents the beliefs about the shape of $S_1(.)$. The precision parameter c_0 is a positive real number, and it measures how much weight to put on these prior beliefs. The prior distribution of $p_{1:H+J+1}$ is a Dirichlet distribution:

$$p_{1:H+J+1} \sim \mathcal{D}(c_0[G(0) - G(t_1)], c_0[G(t_1) - G(t_2)], \ldots, c_0[G(t_{H+J-1}) - G(t_{H+J})], c_0 G(t_{H+J})).$$

In the Gibbs sampler, we iterate between sampling event times for the censored data conditional on the probabilities $p_{1:H+J+1}$, and sampling a new probability vector $p_{1:H+J+1}$ given the sampled event times.

4.2.1. Sampling the unobserved event times for the censored data. For the Gibbs sampler, conditional on the probabilities $p_{1:H+J+1}$, we need to sample which interval each censored event time occurred in (we do not actually need the precise event time). We introduce variables $Z_{k+1,k}, \ldots, Z_{H+J+1,k}$, which decompose the number of censored observations r_k in the interval $(t_{k-1}, t_k]$, into the number of events that fall in the intervals $(t_k, t_{k+1}], \ldots, (t_{H+J-1}, t_{H+J}], (t_{H+J}, \infty)$, so that $r_k = \sum_{l=k+1}^{H+J+1} Z_{l,k}$. The full conditional distribution of $Z_{k+1,k}, \ldots, Z_{H+J+1,k}$ given the probabilities $p_{1:H+J+1}$ is a multinomial distribution with sample size r_k and probability parameters $\eta_{k+1,k}, \ldots, \eta_{H+J+1,k}$, where

$$\eta_j = \frac{p_j}{\sum_{j=k+1}^{H+J+1} p_j}$$

for $j = k+1, \ldots, H+J+1$.

4.2.2. Sampling the probabilities $p_{1:H+J+1}$. Following the sampling of $Z_{k+1,k}, \ldots, Z_{H+J+1,k}$, define d'_k to be the revised number of events in the interval $(t_{k-1}, t_k]$, which is the sum of the observed events and sampled events:

$$d'_k = d_k + \sum_{j=1}^{H+J} Z_{k,j}, \tag{27}$$

where d_k and $Z_{k,j}$ are the number of observed events and simulated events in the interval $(t_{k-1}, t_k]$, respectively. New probabilities $p_1, p_2, \ldots, p_{H+J+1}$ are sampled from their full conditional distribution, which is another Dirichlet distribution:

$$
\begin{aligned}
(p_1, p_2, \ldots, p_{H+J+1}) | d'_1, \ldots, d'_{H+J+1} \sim \mathcal{D} \big(& c_0[G(0) - G(t_1)] + d'_1, c_0[G(t_1) - G(t_2)] \\
& + d'_2, \ldots, c_0[G(t_{H+J-1}) - G(t_{H+J})] \\
& + d'_{H+J}, c_0 G(t_{H+J}) + d'_{H+J+1} \big).
\end{aligned}
$$

4.2.3. Eliciting the prior for θ and calculating the assurance. To elicit the prior for the log hazard ratio θ, we only need to elicit the prior for the survival difference ρ at time t_0 and then use simulation to obtain the prior for θ using (23). Simulated $S_1(t_0)$ can be obtained when generating the survival rates using the Gibbs sampling approach. As before, the Monte Carlo simulation can be used to estimate (26). The R code to compute the assurance for the proportional hazards models is also available in the supporting software described in Section 3.2.

5. Numerical examples

In this section, we provide examples to illustrate how assurance is computed to inform the sample size choice under different model assumptions. We also present how elicited priors have an effect on the assurance. In each example, we suppose that a randomised controlled trial is going to be conducted to compare two treatment effects with an equal number of patients allocated to each treatment group. The sample sizes determined using power calculations are based on a 5% two-sided hypothesis test.

5.1. Exponential model

We first consider sample sizes based on power. We suppose that the trial has a 3-year recruitment period with a 2-year follow-up period and that 60% of patients receiving the standard treatment are expected to be alive after 5 years. For the power calculation, we consider an absolute 20% increase in patient survival for the experimental group.

Using Equation (7), we have the model parameters $\lambda_1 = 0.102$ and $\lambda_2 = 0.0446$. To achieve a specified power π^*, the required sample size N is determined by solving

$$\pi^* = \pi^E(N),$$

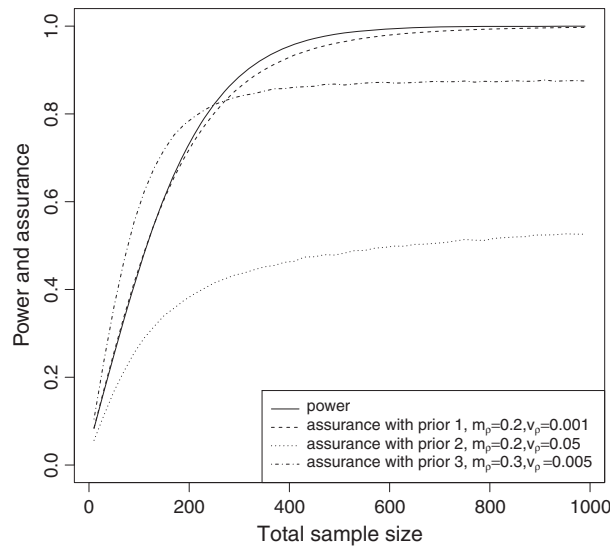where the power function π^E is given in Equation (4).

Figure 1. The comparison between power and assurance γ^E considering $\rho \sim N(m_\rho, v_\rho)$ and $S_1(5) \sim B(60, 40)$.

To calculate assurance γ^E, an expert's judgments about the 5-year survival rate in the control group $S_1(5)$ and the 5-year survival difference ρ are assessed using univariate elicitation methods. Suppose this yields $S_1(5) \sim B(a_s, b_s)$ and $\rho \sim N(m_\rho, v_\rho)$. In the following, we look at three scenarios for the priors.

- Scenario 1: $\rho \sim N(0.2, 0.001)$ and $S_1(5) \sim B(60, 40)$.
- Scenario 2: $\rho \sim N(0.2, 0.05)$ and $S_1(5) \sim B(60, 40)$.
- Scenario 3: $\rho \sim N(0.3, 0.005)$ and $S_1(5) \sim B(60, 40)$.

In scenario 1, the prior $\rho \sim N(0.2, 0.001)$ indicates a strong prior belief that the 5-year survival difference is around 0.2. In scenario 2, $v_\rho = 0.05$ implies that $P(S_2(5) - S_1(5) > 0) = 0.769$. In scenario 3, the prior $\rho \sim N(0.3, 0.005)$ expresses the belief that $P(S_2(5) - S_1(5) < 0) = 0.00003$, that is, the experimenter believes that the experimental treatment has a very high probability of being superior.

Figure 1 shows how the assurances differ given the different joint prior distributions. When an expert has strong beliefs (scenario 1) that the treatment effect will be close to that as specified in the power calculation, the required sample size informed by assurance is similar to that determined by the power calculation. In scenario 2, the assurance cannot exceed 80%, as the prior probability of the new treatment being superior is 76.9%. In scenario 3, a smaller sample size may be required to achieve an 80% probability of having a successful trial given the very high prior probability of the experimental treatment being superior.

5.2. Weibull model

We first consider determining sample sizes using the power function under the Weibull model. Suppose that the 1-year survival rate in the control group is expected to be 20% and to decrease to 10% at the end of the second year. We consider a two-sided hypothesis test of the null hypothesis of no change in the mean survival time between the control and experimental groups. For the power calculation, we suppose that the survival rate in the experimental group at 1 year is 30% and at 2 years is 20%.

Using Equations (13) and (14), the Weibull parameters in each group are $\kappa_1 = 0.52$, $\lambda_1 = 1.61$, $\kappa_2 = 0.42$ and $\lambda_2 = 1.20$. Using Equations (10) and (11), the mean and variance of the survival times in each group are $\mu_1 = 0.75$, $\sigma_1^2 = 2.57$, $\mu_2 = 1.89$, and $\sigma_2^2 = 30.33$. To achieve a specified power π^*, the require sample size N is determined by solving

$$\pi^* = \pi^W(N),$$

where π^W is given in Equation (9).

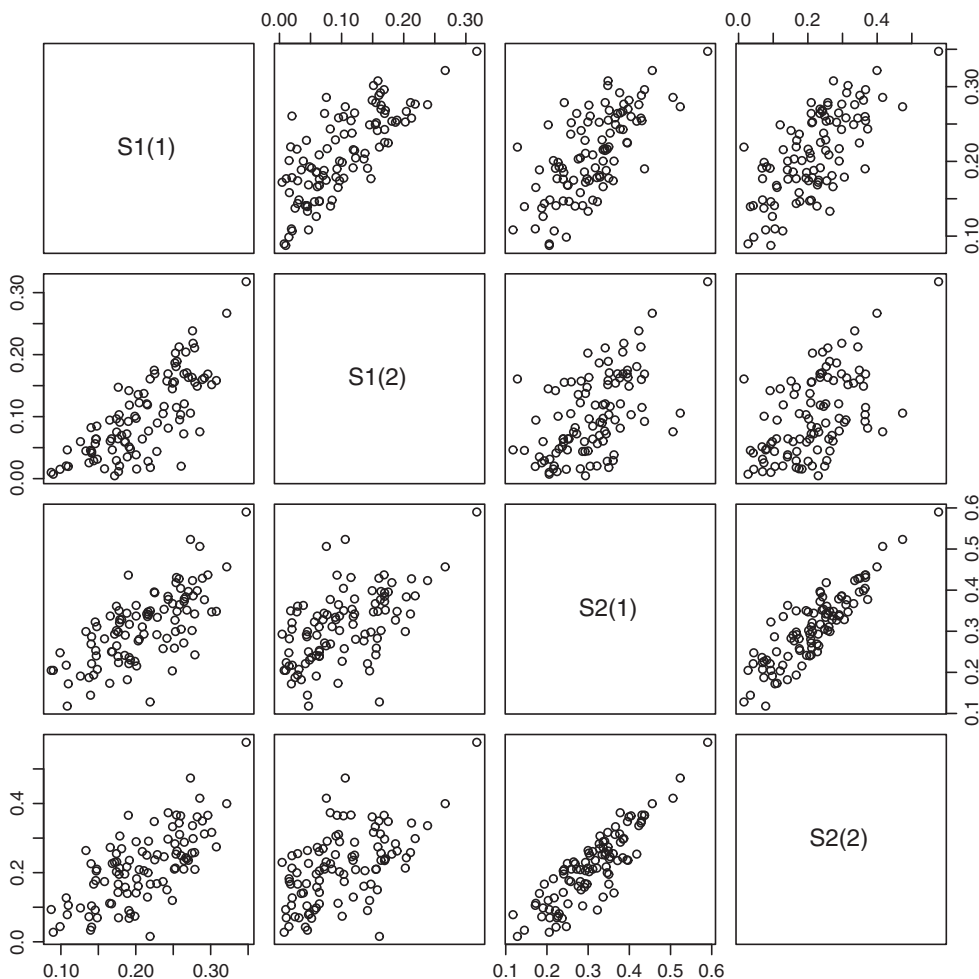| **Table I.** Elicited quartiles of $S_1(1)$, δ_{11}, δ_{12} and δ_{22}.				
	$S_1(1)$	δ_{11}	δ_{12}	δ_{22}
Lower quartile	0.15	0.08	0.05	0.05
Median	0.2	0.11	0.1	0.1
Upper quartile	0.23	0.15	0.14	0.12

Figure 2. Scatterplots of $S_i(1)$ and $S_i(2)$ for $i = 1, 2$, given the elicited quartiles.

To compute the assurance, we consider eliciting an expert's opinion about the 1-year survival rate in the control group $S_1(1)$, the survival difference at 1 year between two groups, δ_{12}, and the difference in survival probability between the experimental and control groups at 1 and 2 years, denoted by δ_{11} and δ_{22}, respectively.

Suppose that we elicited quartiles for the uncertain quantities, as given in Table I. Using the MATCH online elicitation tool, the distributions fitted to the elicited judgements are given later.

$$S_1(1) \sim Beta(8.10, 32.81),$$
$$\delta_{11} \sim Beta(4.36, 32.61),$$
$$\delta_{12} \sim N(0.097, 0.004),$$
$$\delta_{22} \sim Beta(2.23, 20.10).$$

Figure 2 shows how the four survival rates $S_1(1)$, $S_1(2)$, $S_2(1)$ and $S_2(2)$ are correlated given the elicited quartiles listed in Table I. Figure 3 illustrates the comparison between the power and assurance
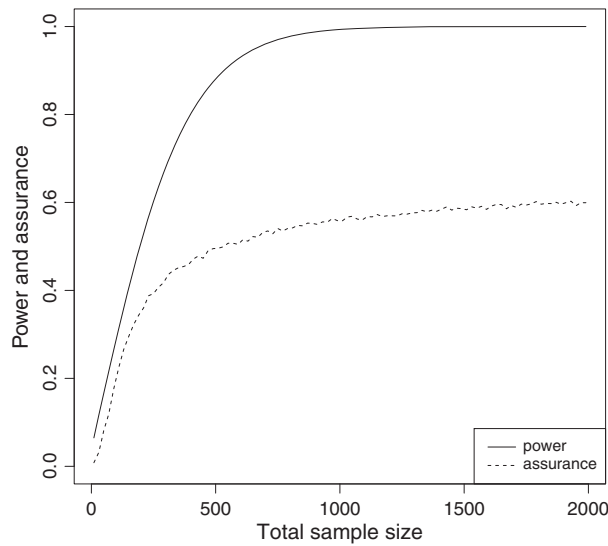
Figure 3. The comparison between power and assurance γ^W considering uncertainty in $S_1(1), \delta_{11}, \delta_{12}$ and δ_{22}.

functions. Given the quartiles of the uncertain quantities in Table I, the prior probability that the experimental treatment is indeed superior is 73.9%, which cannot be exceeded by the assurance. The large difference in terms of power and assurance given a fixed large sample size is because uncertainty in the prior distributions has a large influence on the probability of success.

5.3. Proportional hazard model

Suppose that a trial is planned to have a 5-month recruitment period with a 5-month follow-up period. In our example, we use the data given in Kaplan and Meier [29] as the available information for the standard treatment. The data are 0.8, 1.0*, 2.7*, 3.1, 5.4, 7.0*, 9.2 and 12.1*, where '*' denotes censored observations.

We first consider sample size calculations using the power function. For the power calculation, we consider an absolute 17.5% increase in the 7-month survival rate for the experimental treatment compared with the standard. Using the Kaplan–Meier estimate, the survival rate at 7 months for the standard treatment is $S_1(7) = 0.525$, so the corresponding log hazard ratio θ is -0.591. Using Equation (20), the probability, P_e, that a patient will experience the outcome event during the trial is 0.42. To achieve a specified power π^*, the required sample size N is determined by solving

$$\pi^* = \pi^P(N),$$

where π^P is given in Equation (19).

Considering uncertainty in both the log hazard ratio and survivor function in the control group, the quantities that need to be elicited are the difference ρ in survival probabilities at 7 months between the experimental and standard treatment, and parameter function α of the Dirichlet process prior. Suppose we have already obtained the prior $\rho \sim N(0.175, 0.01)$. Furthermore, an expert proposes the mean of the standard treatment survivor function to be an exponential distribution with a 5-month survival rate $S_1(5)$ of 50%. Hence, the parameter function α of the Dirichlet process is $\alpha([t, \infty)) = c_0 \exp(-\lambda t)$, where λ is given by

$$\lambda = \frac{-\log(0.50)}{5} = 0.139.$$

We fix c_0 at 1, which represents a fairly weak prior for the survival rates $S_1(.)$ with the data being dominant. Figure 4 shows the comparison between the power calculated on the basis of the Kaplan–Meier estimate of $S_1(.)$ and the assurance. The assurance (dashed line) in this case cannot exceed 95.5%, which is the prior probability of the experimental treatment being superior. If we had a stronger prior for ρ, for example, $\rho \sim N(0.175, 0.001)$, then the calculated assurance would be very close to the power (as shown by the dotted line in Figure 4).
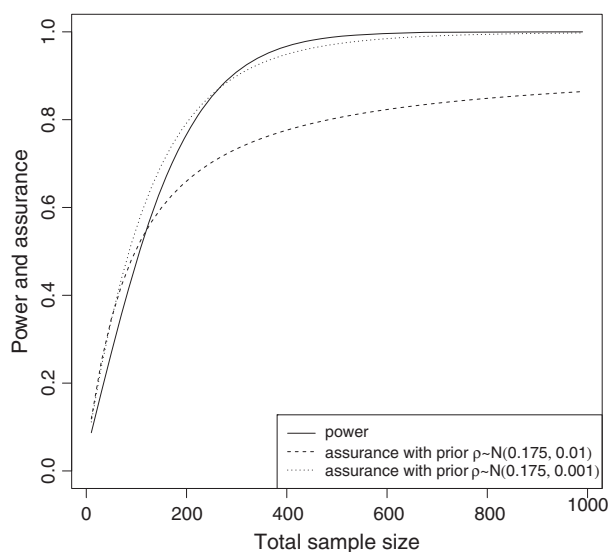
Figure 4. The comparison between power and assurance γ^P considering uncertainty in θ and $S_1(.)$.

6. Summary

We have extended the assurance method to accommodate time-to-event outcomes in clinical trials, assuming one of three analysis methods, and we have made software available for implementing our methods. The reliability of an assurance probability will depend on the reliability of the elicited prior, and so it will be important to check the robustness of assurances to the choice of prior. However, the process of formally assessing the evidence in support of a new treatment and quantifying the attendant uncertainties could itself form a useful part of the trial planning process. Overall, we believe that it is clearly useful to know the probability of a trial producing a successful result, and in the context of clinical trial planning, the extra effort required in using the assurance method is relatively small.

Acknowledgements

References

1. 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.
2. O'Hagan A, Stevens JW. Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making* 2001; **21**:219–230.
3. O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. *Pharmaceutical Statistics* 2005; **4**:187–201.
4. Chuang-Stein C. Sample size and the probability of a successful trial. *Pharmaceutical Statistics* 2006; **5**:305–309.
5. Chuang-Stein C, Yang R. A revisit of sample size decisions in confirmatory trials. *Statistics in Biopharmaceutical Research* 2010; **2**:239–248.
6. Schoenfeld DA, Richter JR. Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics* 1982; **38**:163–170.
7. Gross AJ, Clark VA. *Survival Distributions: Reliability Applications in the Biomedical Sciences*. Wesley: New York, 1975.
8. Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine* 1982; **1**:121–129.
9. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics* 1983; **39**:499–503.
10. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-care Evaluation*. John Wiley and Sons Ltd: England, 2004.
11. Ibrahim JG, Chen M, Sinha D. *Bayesian Survival Analysis*. Springer: New York, 2001.
12. Christensen R, Johnson W, Branscum A, Hanson T. *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. CRC Press: Boca Raton, 2010.

13. Kadane JB, Wolfson LJ. Experiences in elicitation. *Statistician* 1998; **47**:1–20 (with discussion, 55–68).
14. O'Hagan A, Buck CE, Daneshkhah A, Eiser JE, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T. *Uncertain Judgements: Eliciting Expert Probabilities*. John Wiley and Sons Ltd: England, 2006.
15. Oakley JE, O'Hagan A. Shelf: The SHeffield ELicitation Framework (version 2.0), School of Mathematics and Statistics, University of Sheffield, 2010. http://tonyohagan.co.uk/shelf.
16. Johnson SR, Tomlinson GA, Hawker GA, Granton JT, Grosbein HA, Feldman BM. Methods to elicit beliefs for Bayesian priors: a systematic review. *Journal of Clinical Epidemiology* 2010; **63**:355–369.
17. Gore SM. Biostatistics and the medical research council. *Medical Research Council News* 1987; **35**:19–20.
18. Parmar MKB, Spiegelhalter DJ, Freedman LS. The CHART trials: Bayesian design and monitoring in practice. *Statistics in Medicine* 1994; **13**:1297–1312.
19. Tan SB, Chung YFA, Tai BC, Cheung YB, Machin D. Elicitation of prior distributions for a phase III randomized controlled trial of adjuvant therapy with surgery for hepatocellular carcinoma. *Controlled Clinical Trials* 2003; **24**:110–121.
20. Hiance A, Chevret S, Lévy V. A practical approach for eliciting expert prior beliefs about cancer survival in phase III randomized trial. *Journal of Clinical Epidemiology* 2009; **62**:431–437.
21. R Development Core Team. R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2011. http://www.R-project.org/, ISBN 3-900051-07-0.
22. Bowman AW, Crawford E. R package `rpanel`: simple control panels (version 1.0-5), University of Glasgow, UK, 2008. http://www.stats.gla.ac.uk/~adrian/rpanel.
23. Tierney L. *tkrplot: Tk rplot*, 2011. http://CRAN.R-project.org/package=tkrplot, R package version 0.0-23.
24. Singpurwalla ND. An interactive PC-based procedure for reliability assessment incorporating expert opinion and survival data. *Journal of the American Statistical Association* 1988; **83**:43–51.
25. Berger JO, Sun D. Bayesian analysis for the poly-Weibull distribution. *Journal of the American Statistical Association* 1993; **88**:1412–1418.
26. Kaminskiy MP, Krivtsov VV. A simple procedure for Bayesian estimation of the Weibull distribution. *IEEE Transactions on Reliability* 2005; **54**:612–616.
27. Susarla V, Ryzin JV. Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association* 1976; **71**:897–902.
28. Kuo L, Smith AFM. Bayesian computations in survival models via the Gibbs sampler. In *Survival Analysis: State of the Art*, Klein JP, Goel PK (eds). Kluwer Academic: Dordrecht, 1992; 11–24.
29. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**:457–81.