Bayesian sample size determination for cost-effectiveness studies with censored data

Daniel P. Beavers[1]*, **James D. Stamey**[2]

1 Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC, United States of America, **2** Department of Statistical Science, Baylor University, Waco, TX, United States of America

* dbeavers@wakehealth.edu

Abstract

Cost-effectiveness models are commonly utilized to determine the combined clinical and economic impact of one treatment compared to another. However, most methods for sample size determination of cost-effectiveness studies assume fully observed costs and effectiveness outcomes, which presents challenges for survival-based studies in which censoring exists. We propose a Bayesian method for the design and analysis of cost-effectiveness data in which costs and effectiveness may be censored, and the sample size is approximated for both power and assurance. We explore two parametric models and demonstrate the flexibility of the approach to accommodate a variety of modifications to study assumptions.

Introduction

Cost and effectiveness are often modeled jointly in biomedical studies, and there is typically interest in quantities that are functions of both of these outcomes. Comparing treatments on the basis of cost-effectiveness is increasingly important for clinical trials, and recent publications have considered approaches to assess power for cost-effectiveness studies [1–3]. Often such power calculations assume normally distributed costs and effectiveness, which is convenient as it allows for closed form formulae for sample size determination [4–6]. The bivariate normal approach has been utilized in recent methodological work allow for novel study designs, such as the case of large multicenter trials with or without cluster randomization, where approaches such as maximin have been utilized to obtain optimal sample sizes for cost-effectiveness studies [5, 6]. However, for a variety of reasons, observations are often censored, most commonly as a result of the data originating from a survival study. Cost-effectiveness models for survival data have recently been considered [7, 8]. The case of normally distributed costs and effectiveness with censored cost data is discussed in [8]. However, it is well known that costs are often skewed and gamma and log-normal distributions are used. A wide array of cost-effectiveness models and Bayesian analysis of each is described in [9]. One of the models they consider is a Weibull model for effectiveness and a log-normal model for costs. They then extend the log-normal cost model to a non-parametric model. Skewness and censoring in a cardiovascular trial is accounted for in [7]. However, to our knowledge at this point no studies

have proposed sample size determination and power estimation for cost-effectiveness studies with right-censored outcomes.

In this paper we seek to provide a simulation-based algorithm for determining the sample size for a cost-effectiveness study for survival data in which both costs and effectiveness are censored. We consider both traditional power calculations and assurance [10], and we utilize the Bayesian paradigm for at least three reasons. First, numerous authors have noted that when data are not normally distributed, the Bayesian approach is preferable due to the flexibility of parametric assumptions [7, 9, 11], and thus a Bayesian design is applicable. Secondly, in most frequentist designs, nuisance parameters are considered known at the design stage. This is true both in cases where a closed form is used or in the approach of [12] where simulation is used. The Bayesian design approach allows prior information to be utilized to adequately account for uncertainty in nuisance parameters, while also accounting for the uncertainty inherent in these estimates. The simulation-based approach has been increasingly utilized for study planning [13, 14]. Finally, the Bayesian approach allows for parameter estimation by sampling directly from the posterior distribution using Markov chain Monte Carlo methods. We focus our interest on the incremental net monetary benefit (INMB) but our method would be equally useful for incremental cost-effectiveness ratio, average cost-effectiveness ratio, or other measures. We make the code we use available in the appendix so that users would be able to make appropriate changes for alternative criteria or parameters of interest.

We focus our sample size determination scheme on both power and assurance. Assurance, which will be more clearly defined in Section 1, has been defined as the unconditional probability that a trial achieves a positive outcome in a statistical sense [10], such as statistical significance, and can be useful to account for uncertainty in outcomes derived from multiple parameters. It can be thought of as expected power over all likely values of a parameter rather than for a fixed value under the alternative hypothesis. It has been applied in a variety of models in clinical trials [10, 15–17], including trials with time-to-event outcomes [18], but we extend this prior work to time-to-event cost-effectiveness trials with censored data.

Our paper is organized as follows: we present the two parametric approaches for analysis in Section 1, and we discuss Bayesian considerations to the design of cost-effectiveness studies in Section 2. In Section 3 we discuss priors for the design stage and the algorithm for the sample size determination. In Section 4 we consider examples to demonstrate the method, utilizing both cost-effectiveness models from Section 1. Finally, we discuss our approach and consider new directions for research in Section 5.

1 Methods

Cost-effectiveness models in which effectiveness is measured by survival time are generally defined as the bivariate relationship between survival time, T, and treatment cost, C. It is of interest to model

$$f(t, c) = f^T(t)f^{C|T}(c|t). \tag{1}$$

When censoring occurs survival times are not fully observed, and cost data at the time of censoring underestimate the true cost had survival times been fully observed. Using censored values as replacements for fully observed survival time and costs will lead to biased estimates of cost-effectiveness, and estimation for a model of this complexity makes a parametric form useful, especially if the proportional hazards assumption is restrictive. In particular, a Bayesian approach using Markov chain Monte Carlo (MCMC) estimation can be easily implemented to flexibly model such a scenario. Determining clinical effectiveness in the presence of censored survival times is common. However, accounting for the accrued costs in the presence of

incomplete survival time is far less common, complicated by a reliance on cost estimates conditioned on latent time quantities. Fortunately the MCMC approach makes this type of estimation very straightforward. Missing time values are imputed at each iteration of the chain from the conditional posterior distribution, and the imputed observations can be used to estimate the total costs and variability at the estimated failure time.

Here, we apply parametric models for both costs and effectiveness. We proceed under the context of clinical trials, although the results apply generally to any cost-effectiveness study in which treatment alternatives are being compared. We assume the ith subject is randomly assigned to one of two protocols such that $z_i = 1$ for treatment and $z_i = 0$ for control. Effectiveness is defined based on survival time, which we denote T_i, such that incremental treatment effectiveness is $E(T|Z = 1) - E(T|Z = 0)$. Similarly, we define cost as C_i, and the incremental treatment cost is $E(C|Z = 1) - E(C|Z = 0)$.

1.1 Outcome

Our focus is on the outcome of the incremental net monetary benefit (INMB), which is the average per-patient benefit of treatment where benefit is defined in financial terms [19]. The INMB is a function of the maximum willingness to pay for a unit of health (WTP), which varies based on a payer's financial ability and willingness. Put simply, the INMB is the price of improved or sustained health (unit of health × WTP) minus the cost. In terms of the variables above, we define INMB as

$$INMB(WTP) = [E(T|Z = 1) - E(T|Z = 0)] \times WTP$$
$$-[E(C|Z = 1) - E(C|Z = 0)].$$

(2)

If the INMB is determined to exceed 0 for a given WTP then the new treatment ($z = 1$) is assumed to be cost effective compared to the alternative ($z = 0$). Thus, we are interested in testing the hypotheses H_0: $INMB = 0$ versus H_1: $INMB > 0$, and we conclude that the INMB is significant by requiring that for the posterior distribution of the INMB, $Pr(INMB > 0) \geq 1 - \phi$, where ϕ is the probability of falsely concluding H_1 is true. Common values for ϕ include 0.05, 0.025, and 0.01.

Therefore, for the purposes of study planning, we seek the total sample size, n, to achieve 1) a prespecified posterior power, or probability of a successful trial under the chosen alternative hypothesis, for fixed parameter values, or 2) assurance for variable parameter values. Furthermore, we wish to vary the potential details of the study, such as study length and assumed distribution for $f(t, c)$, as well as incorporate the uncertainty of values such that the resulting power conditions on a range of possible scenarios for both the parameters of the model and values for WTP.

1.2 Model 1: The Weibull-gamma cost-effectiveness model

The first model we employ has been described previously [7]. For the joint distribution in (1), $T_i > 0$ and is assumed to have a Weibull(α_i, λ_i) distribution, where the parameterization of the Weibull is

$$f^T(t) = \alpha(t/\lambda)^{\alpha-1} exp[-(t/\lambda)^\alpha]/\lambda.$$

Furthermore, C_i conditioned on time is assumed to follow an alternately parameterized gamma($\mu(t_i)$, v_i) distribution, where $C_i > 0$ and the pdf of the gamma distribution is

$$f^{C|T}(c|T_i = t_i) = (v/\mu(t_i))^v c^{v-1} exp[-(v/\mu(t_i))c]/\Gamma(v)$$

such that the mean $E[C|T_i = t_i] = \mu(t_i)$ depends on the survival time for the ith participant t_i, and similarly $Var[C|T_i = t_i] = \mu(t_i)^2/v$.

We allow for each randomization arm to have unique parameters, allowing for flexibility in both effectiveness and cost models. For the survival/effectiveness model, we define shape parameter

$$\alpha_i = (1 - z_i)\alpha_0 + z_i\alpha_1$$

and scale parameter

$$\lambda_i = (1 - z_i)\lambda_0 + z_i\lambda_1,$$

where $\alpha_i > 0$ and $\lambda_i > 0$. We define the expected cost of treatment as being linearly related to survival time, such that the conditional mean cost

$$\mu(T_i) = [\beta_0 + \gamma_0 T_i](1 - z_i) + [\beta_1 + \gamma_1 T_i]z_i,$$

and shape parameter

$$v_i = (1 - z_i)v_0 + z_i v_1$$

where $v_i > 0$. We constrain $\mu(T_i) > 0$ by defining $\beta_k > 0$ and $\gamma_k > 0$ where $k \in \{0, 1\}$. This somewhat unusual constraint allows us to interpret the intercept as the "start-up costs" of treatment, and the slope as the average cost accrual of treatment per unit time (ie month, year, etc). Costs can only accrue in a nonnegative direction as time increases.

Using this definition, (2) is redefined as

$$INMB(WTP) = [\lambda_1\Gamma(1 + 1/\alpha_1) - \lambda_0\Gamma(1 + 1/\alpha_0)] \times WTP$$

$$-[\beta_1 + \gamma_1\lambda_1\Gamma(1 + 1/\alpha_1) - \beta_0 - \gamma_0\lambda_0\Gamma(1 + 1/\alpha_0)].$$

(3)

Our approach uses the gamma distribution to model right-skewed costs, but other distributions have been employed successfully as well, including the lognormal and inverse-gamma distributions [9, 20]. Generally it is suggested that sensitivity analyses be performed to consider the robustness of the results to alternate parameterizations of costs [11]. The choice of prior distributions for survival and cost model parameters depends on the available knowledge at the study origin. In the absence of information, we assume diffuse gamma prior distributions on all parameters as elaborated in Section 4.

1.3 Model 2: The normal-normal cost-effectiveness model

An alternate parameterization utilizes the bivariate normal distribution for cost-effectiveness, similar to that published in sample size determination methods for cluster-randomized and multicenter cost-effectiveness trials [5, 6]. Although costs and effectiveness are often non-normal, the convenience of modeling and interpretability has made the bivariate normal a common and robust tool for assessing cost-effectiveness [21], particularly for large trials where asymptotic properties can be invoked [22]. Furthermore, parameters are much more easily interpreted, and Bayesian approaches that rely on MCMC methods for estimation are faster. We extend the model of [8] to allow for censored costs and effectiveness. In doing so, we again utilize (1), so that we begin with $f(t, c) = f(t)f(c|t)$ where

$$T_i \sim Normal(\mu_{T_i}, \tau_{T_i}^2)$$

where

$$\mu_{T_i} = \mu_0(1 - z_i) + \mu_1 z_i$$

and

$$\tau^2_{T_i} = \tau^2_{T_0}(1 - z_i) + \tau^2_{T_1} z_i.$$

Next, the distribution of costs given $T_i = t_i$ is

$$C_i | T_i \sim Normal(\mu_{C_i | T_i}, \tau^2_{C_i})$$

where

$$\mu_{C_i | T_i} = (\theta_1 + \theta_2 t_i)(1 - z_i) + (\theta_3 + \theta_4 t_i) z_i$$

and

$$\tau^2_{C_i} = \tau^2_{C_0}(1 - z_i) + \tau^2_{C_1} z_i.$$

Eq (2) is thus redefined as

$$INMB(WTP) = [\mu_1 - \mu_0]WTP - [\theta_3 + \theta_4 \mu_1 - \theta_1 - \theta_2 \mu_0]. \tag{4}$$

The use of the normal-normal model assumes relatively bell-shaped and symmetric bivariate distributions for survival times and costs. Users of the model should carefully inspect data to ensure this is a justifiable assumption, with the added caveat that in our scenario we might have right-censored data. Furthermore, poorly chosen simulation parameters could yield negative survival times or costs, so in our programs we omit negative values and resample the observation in the rare event of negative times or costs. We assume diffuse normal prior distributions for μ_0, μ_1, θ_1, θ_2, θ_3, and θ_4 and diffuse inverse gamma distributions for τ^2_{T0}, τ^2_{T1}, τ^2_{C0}, and τ^2_{C1}. Further information is presented in Section 4.

1.4 Data likelihood

As previously presented by [7], if we define the death indicator for the ith participant d_i and t_i the minimum of the censoring or survival time, then the log-likelihood is the sum of

$$L_i = d_i \left[\ln f^T(t_i) + \ln f^{C|T}(c_i | t_i) \right] + (1 - d_i) \ln \int_{t_i}^{\infty} f^T(u) \left(\int_{c_i}^{\infty} f^{C|T}(x|u)dx \right) du.$$

The normal-normal model has a similar structure, except with different underlying distributions for the time and cost models.

1.5 Assurance

Traditional statistical power is an estimate of the probability of rejecting a null hypothesis under a fixed alternative hypothesis, which in a clinical trial is usually the treatment effect. Assurance therefore is the probability of rejecting a null hypothesis under the *distribution* of likely values of a treatment effect. Therefore, the assurance is frequently described as the unconditional power of the desired outcome because the probability of a trial success (typically defined as rejection of the null hypothesis based on external criteria) does not assume a fixed alternative hypothesis value [18]. Assurance is often estimated using Monte Carlo simulations to draw a random value for the treatment effect its distribution at each iteration, which will be explained in detail in Sections 2 and 3. For the cost-effectiveness case, we will need to consider

both cost and effectiveness for assurance estimates. While the approach does not necessarily require a Bayesian data analysis, a Bayesian approach helps with the estimation of INMB under censored observations.

2 Bayesian design

The sample size/power determination algorithm approach of Wang and Gelfand requires two sets of prior distributions. The first set is referred to as the design (or sampling) prior. This prior is generally concentrated on the part of the parameter space where interest lies, and it is used to generate data in the simulation algorithm to determine the model's operating characteristics such as power. The second prior is the analysis (or fitting) prior and is used in the data analysis of each simulation-generated data set. Typically, these are also used in the final analysis when the actual data is collected. The design and analysis priors can be the same, though often the analysis priors will be less informative than the design priors.

A common concern in Bayesian analyses regards the informativeness of the prior distributions, in this case the analysis prior, particularly regarding the level of information required to be known prior to the conduct of the study. If prior information regarding the parameters is available, this can easily be represented in the analysis prior distribution. However, if concerns that subjective or biased information could guide the analytic results of the study due to overly influential prior information, noninformative priors should be sufficient to minimize most concerns. In the current approach, we utilize noninformative priors in the sense that they are typically assigned means in line with the null hypothesis (i.e. zeros for regression parameters) and with large variances to allow the data likelihood to dominate the posterior distribution, comparable to a frequentist approach. Nothing about the method prohibits or demands noninformative priors; this choice is available to the analyst, and full disclosure should be provided whichever approach is chosen. The model is relatively robust to misspecification of priors when using noninformative priors, but a misspecified informative prior could lead to posterior means that are overly influenced or restricted by incorrect prior information.

Traditional sample size determination methods typically rely on fixed quantities rather than design prior distributions based on available evidence. The use of design prior distributions allows for researchers to represent uncertainty about the prior evidence for the parameters; furthermore, the traditional approach can be considered a special case of our proposed method in which the design prior distribution has zero variance and all probability mass on the chosen fixed quantity. The use of fixed values, while a common practice, may lead to overstatement of model power due to an investigator being forced to select a value that is unlikely to be observed in the actual study. We consider both cases here in order to directly compare the differences. Our approach allows for more uncertainty, and in Section 4 we demonstrate the difference between assurance, where we use a flexible design prior distribution, compared to posterior power, where we use a fixed values for the design priors (i.e. zero variance distributions).

We utilize a simulation-based approach to determine the required sample size [12]. We denote the data vector for a sample size n as $\mathbf{d}^n = (\mathbf{t}, \mathbf{c}, \mathbf{z})$, where $\mathbf{t} = (t_1, \ldots, t_n)$, $\mathbf{c} = (c_1, \ldots, c_n)$, and treatment indicator vector $\mathbf{z} = (z_1, \ldots, z_n)$. We denote the vector of parameters $\theta = (\alpha_0, \lambda_0, \nu_0, \beta_0, \gamma_0, \alpha_1, \lambda_1, \nu_1, \beta_1, \gamma_1)$ for the Weibull-gamma model and $\theta = (\mu_0, \tau_{T0}, \theta_1, \theta_2, \tau_{C0}, \mu_1, \tau_{T1}, \theta_3, \theta_4, \tau_{C1})$ for the normal-normal model.

One important assumption that is particularly applicable to the design phase that is easily overlooked is the choice of a censoring distribution. Previous literature has demonstrated the censoring distribution can affect the power of a chosen model [23, 24], and although inference on the censoring distribution is rarely performed during analyses, it is crucial for the censoring

distribution to be realistic and relatively robust to the analytical approach. This is particularly important because censored costs are estimated based on the data augmented survival times. Censoring distributions with higher variances can lead to higher variability in INMB estimates, resulting in loss of power. This is considered in Section 4.3.

Statistical assurance is the unconditional probability that a trial will yield a specified outcome, often success [9]. We distinguish this from power by noting that assurance is the expected power with respect to the design prior. For the present problem note that INMB is a function of all model parameters. We seek a sample size n for our study, incorporating uncertainty about the paramter values, that yields an effect size of INMB that is desired to detect success ($Pr(INMB > 0)$) with power $\pi(\theta)$. The power function is then defined as $\pi(\theta) = P(R|\theta)$ where R denotes rejection of the hypothesis. Assurance is then the unconditional probability of rejection

$$P(R) = E(\pi(\theta)),$$

where this expectation is with respect to the design priors. Unlike power which generally goes to 1 as the sample size increases, assurance is dependent on the design prior, and depending on the informativeness and location of the sampling prior, the maximum assurance that a model can attain is less than 1.

3 Description of sample size determination algorithm

The sample size determination method uses a Monte Carlo simulation approach to generate data randomly selected from the design priors for a fixed n and to analyze the data using the proposed model. This is repeated for multiple sample sizes and in most cases the smallest sample size for which some optimality criteria has been achieved is selected as the preferred sample size. In our case, we focus on the probability that the INMB for our active treatment exceeds the control at a statistically significant threshold, such as 95% posterior probability, at a prespecified WTP value, $Pr(INMB > 0|WTP) > 0.95$. In our simulation, we would look for this threshold to be met or exceeded 80% or more for datasets generated from our design priors. The smallest sample size that meets these conditions will be chosen, as detailed below.

Prior to beginning the algorithm, specify the design priors, analysis priors, censoring distribution, and the WTP values of interest. Furthermore, being a simulation-based algorithm, a reasonable set of possible sample sizes should be preidentified to explore the properties of the model given n. Then, execute the following steps for the sample size determination algorithm.

1. Generate values for θ from the design priors. This is particularly important for assurance, but when the desired outcome is power, the parameters are usually fixed.

2. For a reasonable sample size n, generate values for **t** and **c** of length n from the distributions with θ generated in step 1. Generate censored observations as appropriate, and when censoring exists, segregate censored observations for time and cost from the fully observed times and costs.

3. Fit the Bayesian model defined by Eqs 3 and 4 to the simulated data generated in steps 1 and 2 using the analysis priors and approximate the posterior distribution of the INMB, particularly the appropriate alternative hypothesis posterior probability $Pr(INMB > 0|WTP) > 0.95$ for each WTP value. Code for fitting each of the models discussed in the paper in WinBUGS/ OpenBUGS are presented in the Appendix.

4. Repeat steps 1–3 B times at each sample size value under consideration.

5. Calculate the probability of rejecting the null hypothesis $r^{(n)}$ by calculating the proportion of the B iterations where the posterior $Pr(INMB > 0|WTP) > 0.95$ via the formula

$$r^{(n)} = 1/B \sum_{k=1}^{B} I\{Pr(INMB > 0|WTP, \mathbf{d}^{(\mathbf{n},\mathbf{k})}) > 0.95\}.$$

When the design priors are fixed, $r^{(n)}$ is equal to the power, and when the design priors are variable, this proportion is the assurance.

6. Finally, repeat steps 1–5 for a range of sample sizes and plot $r^{(n)}$ by WTP and n to find a sample size that achieves a desired level of power or assurance. We approach this two ways in our examples.

In our case we wish to obtain n^* such that $Pr(Pr(INMB > 0|n^*) \geq 0.95) \geq 0.8$; that is, at least 80% of simulated data sets yield an INMB with at least 95% posterior probability density greater than 0, which is roughly analogous to 80% power or assurance. Our choice for B was 300, although reduced Monte Carlo error can be achieved at with higher values of B, at the expense of time. Alternate values of posterior probability than 0.95 can easily be substituted in steps 3, 5, and 6.

4 Results

As a demonstration of the method, we organize the results section as follows: in Section 4.1 we assume the data arise from the Weibull-gamma model using fixed quantities for design parameters, which allows for estimation of posterior power; in Section 4.2 we assume the data arise from the Weibull-gamma model using variable quantities for design parameters, which allows the determination of assurance; in Section 4.3 we modify the assumed censoring distribution to determine the impact of an altered censoring distribution on model power; and finally, in Section 4.4 we modify the parametric form of the study to explore the normal-normal cost-effectiveness model for both design and analysis.

We generated 300 iterated data sets for each sample size using R v2.12.2, and we fit the data in WinBUGS using the R2WinBUGS library. For each data set, we generated 3 chains using a 2000 iteration burn-in and keeping a 10, 000 iteration sample. Of primary interest was $Pr(INMB > 0|\mathbf{c}, \mathbf{t})$ for an array of different WTP quantities ranging from \$100k to \$350k by \$50k increments. The following examples demonstrate the performance of the method, and we further show how modifications to the study design impact the power and sample size, including the duration of the study and the assumed censoring distribution. The proposed method is extremely flexible and, being simulation-based, is far more adaptable to a variety of inputs than standard closed-form methods. The WinBUGS/OpenBUGS code is available in the appendix as S1 and S2 Files, while the R programs used to simulate the data are presented as appendix S3 and S4 Files.

4.1 Bayesian power using the Weibull-gamma cost-effectiveness model

For the determination of power, we assume fixed quantities for the design priors. We present the chosen parameter values in Table 1 for both the design and analysis portions. We consider two potential lengths for the trial; first, we assume that censoring times are distributed uniformly from 2.5 to 3.5 years, which would be realistic for a study with a 12-month recruitment period that accrues survival times for an average of 3 years with a maximum of approximately 3.5 years. The second scenario assumes that censoring times are distributed uniformly from 3.5 to 4.5 years, which is realistic for a study with a 12-month recruitment period that accrues

Table 1. Design and analysis prior distributions for Weibull-gamma models in Sections 4.1, 4.2, and 4.3.

Parameters	Power Design Priors	Assurance Design Priors	Analysis Prior Distribution
α_0	0.75	$normal(0.75, 0.1^2)$	$gamma(0.1, 0.1)$
λ_0	0.9	$normal(0.9, 0.1^2)$	$gamma(0.1, 0.1)$
ν_0	1.2	$normal(1.2, 0.15^2)$	$gamma(0.1, 0.1)$
α_1	0.5	$normal(0.5, 0.05^2)$	$gamma(0.1, 0.1)$
λ_1	1.5	$normal(1.5, 0.2^2)$	$gamma(0.1, 0.1)$
ν_1	3	$normal(3, 0.2^2)$	$gamma(0.1, 0.1)$
β_0	50	$normal(50, 10^2)$	$gamma(10, 0.1)$
γ_0	75	$normal(75, 15^2)$	$gamma(10, 0.1)$
β_1	100	$normal(100, 18^2)$	$gamma(10, 0.1)$
γ_1	60	$normal(60, 15^2)$	$gamma(10, 0.1)$

survival times for an average of 4 years and no more than 4.5 years. It is expected that a longer study will allow for a more precise estimate of INMB due to fewer censored observations, although whether the added effort yields superior power is answered by the simulation results. These same parameter values are used for the first three examples, and based on the chosen parameter values in each case the benefit expectation $E[INMB|WTP] = 1.93 \times WTP - 150$.

In Fig 1 we present the posterior power for detecting $INMB > 0$ presented as functions of sample size grouped by WTP values and alternately as WTP values grouped by sample size. We observe that the power for a trial of longer duration yields higher posterior power due to the reduced proportion of censored observations, and that the chosen WTP can have a very large impact on the preferred sample size, with larger WTP values requiring smaller sample sizes to achieve 80% power. However, we know that these values could potentially overstate our knowledge of the randomization effect, particularly if prior data are ambiguous or absent. We feel it is appropriate to place less stringent assumptions on our knowledge of cost-effectiveness by contrasting these results to those of an assurance analysis.

4.2 Bayesian assurance using the Weibull-gamma cost-effectiveness model

In the second example we modify the previous simulation except we utilize the variable design priors to provide estimates of assurance for which traditional fixed value inputs are utilized. Estimation of assurance can furthermore demonstrate the inflation in power if we naïvely assume fixed values are "correct" rather than providing variable quantities to capture our uncertainty about the true state of nature. For the subsequent output, we repeat the scenarios from Section 4.1 with the sole exception that at each iteration of the program, we generate a design prior value for each parameter from the distributions indicated in the "Assurance Design Priors" column of Table 1.

We observe the posterior assurance output in Fig 2. There is a marked reduction in posterior assurance compared to the power output in Fig 1. Importantly, we can quantify the penalty for using more stringent assumptions in Section 4.1. For example, in the upper half of Fig 2, corresponding to a trial of 3 years, the power for $250k WTP at $n = 200$ is approximately 70%, while in Fig 1, the posterior power is approximately 80%. Our added uncertainty about the true value of the input parameters that combine to produce the INMB has led to a 10% reduction in the probability a successful trial, and to achieve 80% assurance, a larger sample size is required than for 80% power. Furthermore, we note that in Fig 1 the curves clearly asymptotically approach 1.0, while in Fig 2 the asymptotic trend seems to be toward a value somewhat less than 1.0.
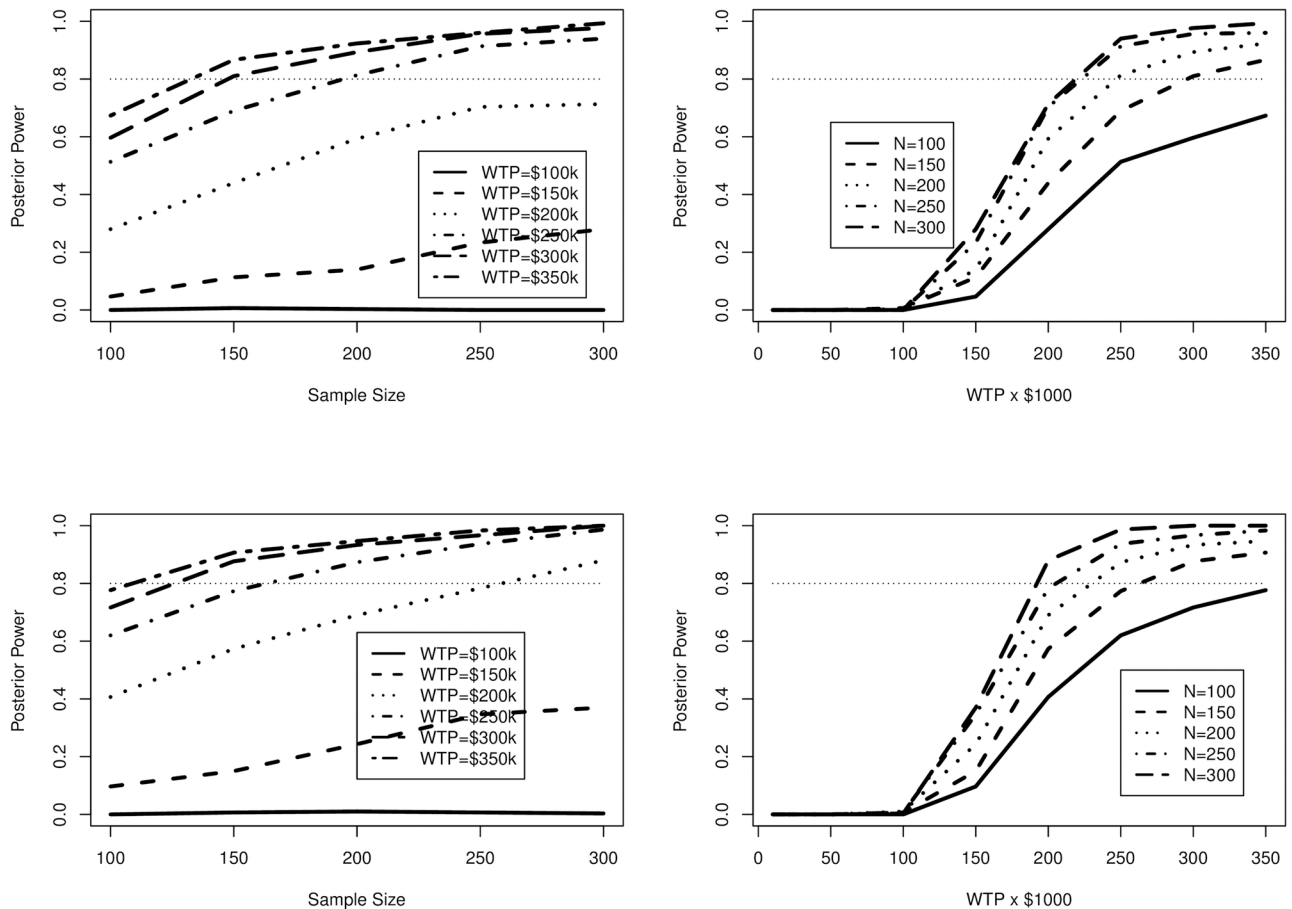
Fig 1. Posterior power estimates for Section 4.1. Average censoring time is 3 years for upper figures and 4 years for lower.

4.3 Effects of modified censoring distribution on Weibull-gamma power

We modify the simulation in Section 4.1 to allow for a different censoring function. Most survival studies experience censored observations due to the study conclusion and due to non-death attrition prior to the study termination. To accomodate this, we allow for a minority of participants to withdraw from the study prior to the termination of the study. In our case, for each participant, we generate a binary indicator with probability 0.8 such that a 1 indicates the censoring value is generated as described in Section 4.1 and a 0 indicates the participant was participant's censoring distribution is randomly sampled from a uniform distribution from 0 to 2.5 or 3.5, depending on whether the mean study duration is 3 years or 4 years, respectively. If the observed death time is greater than the generated censoring time, then the participant is considered censored at the censoring time. This is a far more variable censoring distribution than in Section 4.1; it is much more likely to censor individuals at relatively small follow-up times (i.e. $T < 1$ yr). The net result is that we can compare the power to that of Section 4.1 and determine how sensitive our model is to the assumptions governing the censoring distribution.

In both the cases for the 3 year proposed study and 4 year proposed study, as seen in Fig 3, the added variability of the exponential censoring function slightly reduces the power compared to the censoring mechanism in Section 4.1. This is important because all other
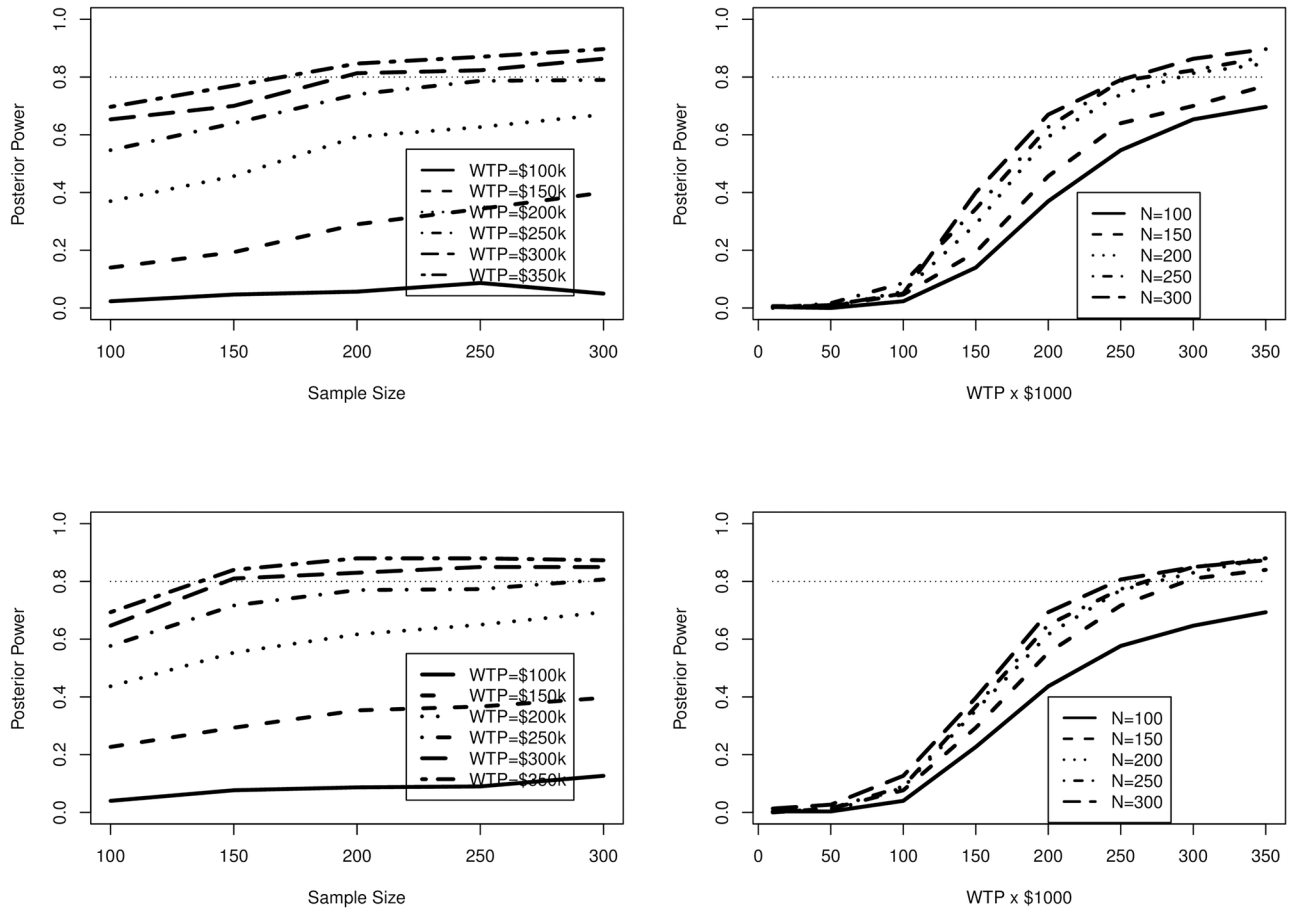
Fig 2. Posterior assurance estimates for Section 4.2. Average censoring time is 3 years for upper figures and 4 years for lower.

components are the same; increased variability in censoring times lead to reduced power. While this finding is expected, we can quantify the impact of different censoring types; for studies in which participants may drop from participation prior to the completion of the follow-up time, we can quantify the difference this type of censoring as opposed to the type defined in Section 4.1.

4.4 Bayesian power using the normal-normal cost-effectiveness model

The final approach assumes that costs and survival times are distributed bivariate normal as proposed in Section 1.3, except in this case time is defined as months rather than years. The outcome of interest is Bayesian power or sample size, and thus design priors are assumed to be fixed quantities. The design and analysis priors are presented in Table 2, which corresponds to a bivariate normal distribution for treatment $Z = 0$ of

$$\begin{pmatrix} T_0 \\ C_0 \end{pmatrix} \sim N\left(\begin{pmatrix} 33.5 \\ 700 \end{pmatrix}, \begin{pmatrix} 8.5^2 & 0.4(8.5)(101.5) \\ 0.4(8.5)(101.5) & 101.5^2 \end{pmatrix} \right)$$
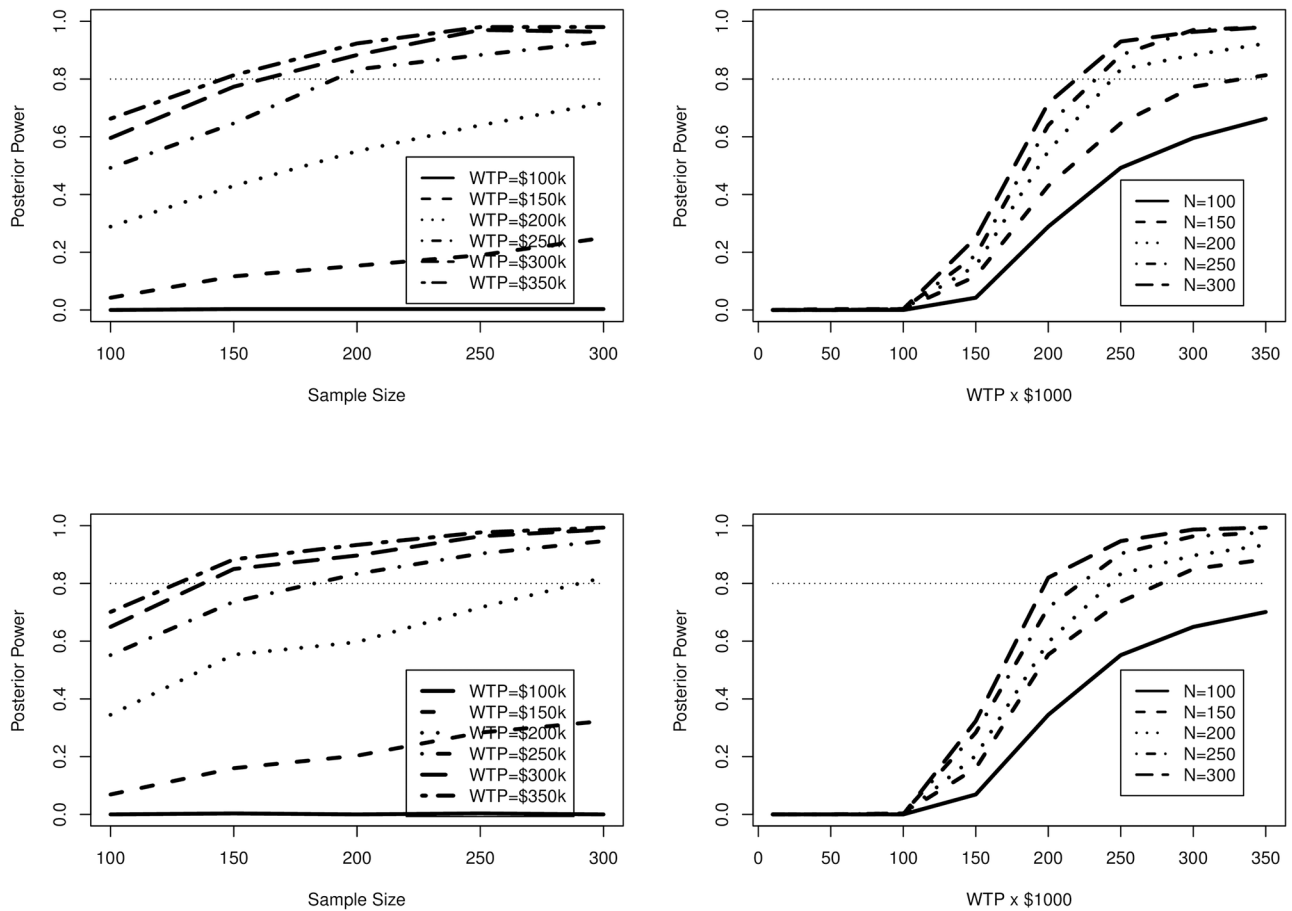
Fig 3. Posterior power estimates from Section 4.3. Average study duration time is 3 years for upper figures and 4 years for lower.

and for treatment $Z = 1$

$$\begin{pmatrix} T_1 \\ C_1 \end{pmatrix} \sim N\left(\begin{pmatrix} 25.0 \\ 130 \end{pmatrix}, \begin{pmatrix} 8.5^2 & 0.3(8.5)(30) \\ 0.3(8.5)(30) & 30^2 \end{pmatrix} \right).$$

Table 2. Design and analysis prior distributions for normal-normal model in Sections 4.4.

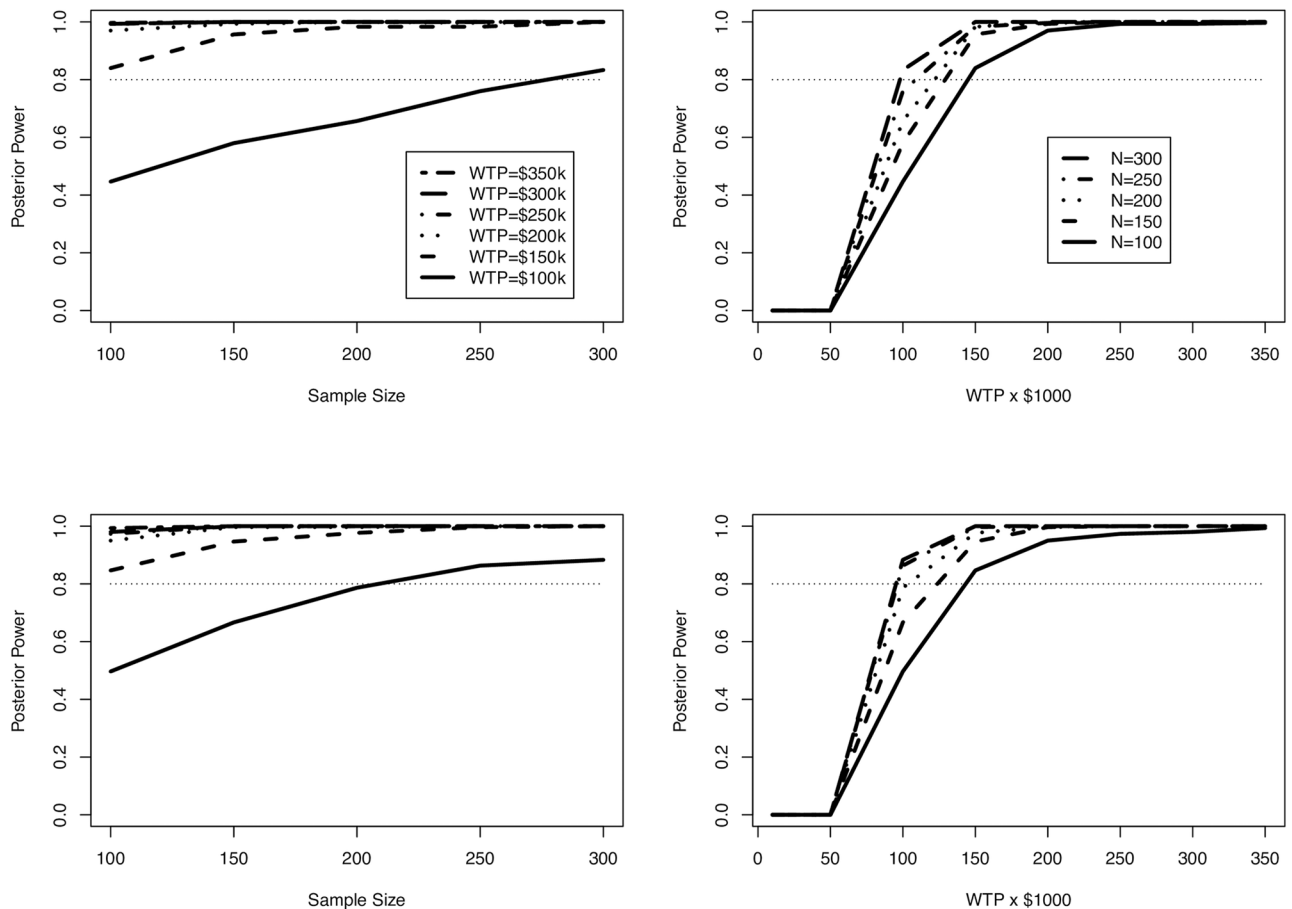| Parameters	Power Design Prior Values	Analysis Prior Distribution
μ_0	25.0	normal(0, 10 000)
τ_{T0}^2	8.5^2	InvGamma(0.01, 0.01)
μ_1	33.5	normal(0, 10 000)
τ_{T1}^2	8.5^2	InvGamma(0.01, 0.01)
θ_1	103.5	normal(0, 10 000)
θ_2	1.06	normal(0, 10 000)
τ_{C0}^2	28.6^2	InvGamma(0.001, 0.001)
θ_3	539.9	normal(0, 10 000)
θ_4	4.78	normal(0, 10 000)
τ_{C1}^2	93.0^2	InvGamma(0.001, 0.001)

Fig 4. Posterior power estimates from Section 4.4. Average censoring time is 3 years for upper figures and 4 years for lower.

The values for the cost model parameters in Table 2 are conditional on *T*. For the fourth example, based on the chosen parameter values the benefit expectation $E[INMB|WTP] = 8.5 \times WTP - 570$.

For this simulation, we generate censoring values from a normal distribution with mean 36 months and standard deviation 2.4 for the top row of plot in Fig 4 and with mean 48 and standard deviation 2.4 for the bottom row, corresponding to censoring times of an average of 3 years and 4 years, respectively (although the data were generated with survival in months rather than years). We observe the prior information used in both the design and analysis phase in Table 2. While a direct comparison with the Weibull-gamma approach is not valid due to the differing parameterization of this model compared to preceding sections, the parameters were selected to at least mimic the previous models, and the data were generated ensuring no negative values existed for time or costs.

5 Discussion

The Bayesian power and sample-size determination approach for cost-effectiveness studies with censored data offers key advantages over existing cost-effectiveness power calculation methods. Users can represent unknown quantities as design prior distributions with variances rather than fixed quantities; survival times and costs can be fully simulated to calculate the INMB rather than relying on previous INMB estimates; and the method is extremely flexible

and adaptable for a variety of distributions and assumptions, utilizing freely available statistical software packages.

The motivating example was inspired by clinical trials in which there is a high death rate, such as cancer or heart failure. Subsequent applications in which there is a lower censoring rate should monitor the imputed observations to ensure the distributional assumptions do not yield unrealistic estimates. Most participants expired prior to censoring in our motivating examples, but in scenarios in which both event times and censoring times can be highly variable, such as incident cardiovascular disease, WinBUGS or OpenBUGS can easily provide posterior estimates of the censored participants' failure time distributions left-truncated at the censoring time. Furthermore, the current work focuses on simple parallel randomized controlled trials, but future work should consider alternate design approaches including cluster randomized trials and minmax designs.

Supporting information

S1 File. The first OpenBUGS program was used to produce INMB estimates for the gamma-Weibull cost-effectiveness model.
(TXT)

S2 File. The second OpenBUGS program was used to produce INMB estimates for the normal-normal cost-effectiveness model.
(TXT)

S3 File. The wg_simulation.r program was used to simulate data for the Weibull-gamma cost effectiveness models.
(R)

S4 File. The nn_simulation.r program was used to simulate data for the normal-normal cost effectiveness models.
(R)

Acknowledgments

We would like to thank the editor and anonymous reviewers whose comments and suggestions helped improve and clarify this manuscript.

Author Contributions

Conceptualization: Daniel P. Beavers, James D. Stamey.

Formal analysis: Daniel P. Beavers, James D. Stamey.

Methodology: Daniel P. Beavers, James D. Stamey.

Project administration: Daniel P. Beavers.

Software: Daniel P. Beavers.

Supervision: James D. Stamey.

Validation: Daniel P. Beavers.

Visualization: Daniel P. Beavers.

Writing – original draft: Daniel P. Beavers, James D. Stamey.

Writing – review & editing: Daniel P. Beavers, James D. Stamey.

References

1. Gafni A, Walter S, Birch S, Sendi P. An opportunity cost approach to sample size calculation in cost-effectiveness analysis. Health Economics. 2008; 17(1):99–107. https://doi.org/10.1002/hec.1244 PMID: 17497751

2. Boyd KA, Briggs AH, Fenwick E, Norrie J, Stock S. Power and sample size for cost-effectiveness analysis: fFN neonatal screening. Contemporary Clinical Trials. 2011; 32:893–901. https://doi.org/10.1016/j.cct.2011.07.007 PMID: 21782976

3. Willan AR. Sample size determination for cost-effectiveness trials. Parmacoeconomics. 2011; 29 (11):1–17.

4. Sarker SJ, Whitehead A, Khan I. A C++ program to calculate sample sizes for cost-effectiveness trials in a Bayesian framework. Computer Methods and Programs in Biomedicine. 2013; 110(3):471–489. http://dx.doi.org/10.1016/j.cmpb.2013.01.008. PMID: 23399102

5. Manju MA, Candel MJJM, Berger MPF. Sample size calculation in cost-effectiveness cluster randomized trials: optimal and maximin approaches. Statistics in Medicine. 2014; 33(15):2538–2553. https://doi.org/10.1002/sim.6112 PMID: 25019136

6. Manju MA, Candel MJ, Berger MP. Optimal and maximin sample sizes for multicentre cost-effectiveness trials. Statistical Methods in Medical Research. 2015; 24(5):513–539. https://doi.org/10.1177/0962280215569293 PMID: 25656551

7. Heitjan DF, Kim CY, Li H. Bayesian estimation of cost-effectiveness from censored data. Statistics in Medicine. 2004; 23:1297–1309. https://doi.org/10.1002/sim.1740 PMID: 15083484

8. Lambert PC, Billingham LJ, Cooper NJ, Sutton AJ, Abrams KR. Estimating the cost effectiveness of an intervention in a clinical trial when partial cost information is available: a Bayesian approach. Health Economics. 2008; 17(1):67–81. https://doi.org/10.1002/hec.1243 PMID: 17533622

9. O'Hagan A, Stevens JW. Bayesian assessment of sample size for clinical trials of cost-effectiveness. Medical Decision Making. 2001; 21:219–230. PMID: 11386629

10. O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. Pharmaceutical Statistics. 2005; 4(3):187–201. https://doi.org/10.1002/pst.175

11. Thompson SG, Nixon RM. How sensitive are cost-effectiveness analyses to choice of parametric distributions? Medical Decision Making. 2005; 25(4):416–423. https://doi.org/10.1177/0272989X05276862 PMID: 16061893

12. Wang F, Gelfand AE. A Simulation-based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models. Statistical Science. 2002; 17(2):193–208. https://doi.org/10.1214/ss/1030550861

13. Cheng D, Branscum AJ, Johnson WO. Sample size calculations for ROC studies: parametric robustness and Bayesian nonparametrics. Statistics in Medicine. 2011; 31(2):131–142. https://doi.org/10.1002/sim.4396 PMID: 22139729

14. Beavers DP, Stamey JD. Bayesian sample size determination for binary regression with a misclassified covariate and no gold standard. Computational Statistics and Data Analysis. 2012; 56:2574–2582. https://doi.org/10.1016/j.csda.2012.02.014

15. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. John Wiley & Sons; 2004.

16. Hee SW, Stallard N. Designing a series of decision-theoretic phase II trials in a small population. Statistics in Medicine. 2012; 31(30):4337–4351. https://doi.org/10.1002/sim.5573 PMID: 22927289

17. Ambrosius WT, Polonsky TS, Greenland P, Jr DCG, Perdue LH, Fortmann SP, et al. Design of the Value of Imaging in Enhancing the Wellness of Your Heart (VIEW) trial and the impact of uncertainty on power. Clinical Trials. 2012; 9(2):232–46. https://doi.org/10.1177/1740774512436882 PMID: 22333998

18. Ren S, Oakley JE. Assurance calculations for planning clinical trials with time-to-event outcomes. Statistics in Medicine. 2014; 33(1):31–45. https://doi.org/10.1002/sim.5916 PMID: 23861270

19. Stinnett AA, Mullahy J. Net Health Benefits: A New Framework for the Analysis of Uncertainty in Cost-Effectiveness Analysis. Medical Decision Making. 1998; 18(2_suppl):S68–S80. https://doi.org/10.1177/0272989X98018002S09 PMID: 9566468

20. Ng ESW, Diaz-Ordaz K, Grieve R, Nixon RM, Thompson SG, Carpenter JR. Multilevel models for cost-effectiveness analyses that use cluster randomised trial data: An approach to model choice. Statistical Methods in Medical Research. 2016; 25(5):2036–2052. https://doi.org/10.1177/0962280213511719 PMID: 24346164

21. Briggs A, Nixon R, Dixon S, Thompson S. Parametric modelling of cost data: some simulation evidence. Health Economics. 2005; 14(4):421–28. https://doi.org/10.1002/hec.941 PMID: 15685641

22. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annual Review of Public Health. 2002; 23:151–69. https://doi.org/10.1146/annurev.publhealth.23.100901.140546 PMID: 11910059

23. Heller G, Simonoff JS. A comparison of estimators for regression with a censored response variable. Biometrika. 1990; 77:515–520. https://doi.org/10.1093/biomet/77.3.515

24. Heller G, Simonoff JS. Prediction in censored survival data: A comparison of the proportional hazards and linear regression models. Biometrics. 1992; 48:101–115. https://doi.org/10.2307/2532742 PMID: 1581480