

Original citation:

Pearce, Michael, Hee, Siew Wan, Madan, Jason, Posch, Martin, Day, Simon, Miller, Frank, Zohar, Sarah and Stallard, Nigel (2018) *Value of information methods to design a clinical trial in a small population to optimise a health economic utility function*. BMC Medical Research Methodology, 18 (1). p. 20. doi:[10.1186/s12874-018-0475-0](https://doi.org/10.1186/s12874-018-0475-0)

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/99080>

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work of researchers of the University of Warwick available open access under the following conditions.

This article is made available under the Creative Commons Attribution 4.0 International license (CC BY 4.0) and may be reused according to the conditions of the license. For more details see: <http://creativecommons.org/licenses/by/4.0/>

A note on versions:

The version presented in WRAP is the published version, or, version of record, and may be cited as it appears here.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

RESEARCH ARTICLE

Open Access



Value of information methods to design a clinical trial in a small population to optimise a health economic utility function

Michael Pearce¹, Siew Wan Hee², Jason Madan³, Martin Posch⁴, Simon Day⁵, Frank Miller⁶, Sarah Zohar⁷ and Nigel Stallard^{2*} 

Abstract

Background: Most confirmatory randomised controlled clinical trials (RCTs) are designed with specified power, usually 80% or 90%, for a hypothesis test conducted at a given significance level, usually 2.5% for a one-sided test. Approval of the experimental treatment by regulatory agencies is then based on the result of such a significance test with other information to balance the risk of adverse events against the benefit of the treatment to future patients. In the setting of a rare disease, recruiting sufficient patients to achieve conventional error rates for clinically reasonable effect sizes may be infeasible, suggesting that the decision-making process should reflect the size of the target population.

Methods: We considered the use of a decision-theoretic value of information (VOI) method to obtain the optimal sample size and significance level for confirmatory RCTs in a range of settings. We assume the decision maker represents society. For simplicity we assume the primary endpoint to be normally distributed with unknown mean following some normal prior distribution representing information on the anticipated effectiveness of the therapy available before the trial. The method is illustrated by an application in an RCT in haemophilia A. We explicitly specify the utility in terms of improvement in primary outcome and compare this with the costs of treating patients, both financial and in terms of potential harm, during the trial and in the future.

Results: The optimal sample size for the clinical trial decreases as the size of the population decreases. For non-zero cost of treating future patients, either monetary or in terms of potential harmful effects, stronger evidence is required for approval as the population size increases, though this is not the case if the costs of treating future patients are ignored.

Conclusions: Decision-theoretic VOI methods offer a flexible approach with both type I error rate and power (or equivalently trial sample size) depending on the size of the future population for whom the treatment under investigation is intended. This might be particularly suitable for small populations when there is considerable information about the patient population.

Keywords: Decision theory, Health economics, Power, Rare disease, Regulator, Type I error, Value of information

*Correspondence: n.stallard@warwick.ac.uk

²Statistics and Epidemiology, Division of Health Sciences, Warwick Medical School, University of Warwick, Coventry, UK
Full list of author information is available at the end of the article

Background

Prior to approval a drug typically goes through various phases of clinical development, beginning with assessing pharmacology in humans (phase I), followed by exploration of therapeutic efficacy (phase II) and finally, confirmation of the effectiveness (phase III). This is not a necessary ordering, for example, prior to presenting overall clinical development, results and issues of the drug's efficacy and safety (this list is long) to regulatory agencies, further investigation of the effect on human pharmacology may be conducted. Based on the submitted information, the regulatory authorities approve the product that has demonstrated safety and effectiveness for the intended population.

Treatments for rare diseases may not go through all the phases of clinical development prior to submission to regulatory authorities. Buckley presented a brief summary of treatments intended for rare diseases and approved by European regulator that did not go through all these phases [1]. The difference is usually attributable to the small population where it is infeasible or impossible to recruit many patients for trials.

The focus of this paper is the design of phase III trials in particular in a small population. One of the fundamental issues in clinical trial design is sample size determination. The most common approach is to perform a power calculation for a hypothesis test. In a two-arm randomised controlled trial aiming to establish superiority of an experimental treatment to the standard treatment, the null hypothesis may state that the experimental treatment is not superior to the control.

The sample size is determined by restricting type I and II error rates. A type I error occurs when a true null hypothesis is rejected and a type II error occurs when a false null hypothesis is not rejected. Typically, the one-sided type I error rate, α , is set to 0.025. The type II error rate, β , is set to 0.10 or 0.20 for some specified alternative hypothesis. Correspondingly, the power to detect the pre-defined value, $1 - \beta$, is 0.90 or 0.80. Values for α and β are customarily set without consideration of the severity or prevalence of the disease. In practice, however, clinical trials in rare diseases have smaller sample sizes than those in more common conditions, indicating that conventional type I and type II error rates are being compromised in this setting [2, 3]. Increasing the type I error rate might be appropriate in a trial for rare disease as the small population means that the number of patients who would benefit from an effective treatment may be small and so it may be justifiable to have alternate levels of type I error so that an effective treatment may be made available more easily.

To determine the sample size it seems reasonable to consider information such as previous trial outcomes, the cost of making type I or type II errors, the number of

people affected, the financial burden of the treatment and other costs and rewards that will result from the treatment being approved and marketed. One model that could be used to optimise the sample size accounting for all the information is the Bayesian decision theoretic approach. In this framework, decisions are made so as to maximise a utility function that quantifies their desirability of an outcome. The utility function may be made up of observed responses from a sample of patients, the costs of treating patients and conducting the trial, and the profits from a successful treatment. Some authors have proposed to include the cost of trials, the profit from a successful trial (or a loss from an unsuccessful trial), two endpoints (e.g. efficacy and adverse events) and the size of the population, N , so that the sample size required for the trial is optimised (see Hee et al. [4]). By including N we also incorporate the potential gain to future patients. Such an approach seems particularly suitable in the setting of a small population, when there is likely to be considerable knowledge of the size of the target population prior to the start of a clinical trial and this is likely to be much smaller than in other settings.

Methods

A Bayesian decision theoretic approach to sample size determination

In our approach we assume that following safety and efficacy exploration studies, once the drug has been shown to be effective in a phase III trial for the intended population, it obtains regulatory approval. The objective of a phase III trial is primarily to confirm effectiveness. The typical design considered to be the gold standard in providing the best evidence in assessing efficacy is the randomised controlled trial (RCT). Let n be the size of the RCT and assume that patients are randomised in a 1:1 ratio to either the experimental or standard arm. Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in/2})$, $i = (E, S)$, denote the outcomes and \bar{Y}_i denote the sample mean from $n/2$ patients from the experimental (E) and standard treatment (S) arms, respectively. Assume \bar{Y}_i is normally distributed with mean θ_i and known variance, $2\sigma_i^2/n$, that is $\bar{Y}_i \sim N(\theta_i, 2\sigma_i^2/n)$. In the classical frequentist setting, we test the null hypothesis that the two population means, θ_E and θ_S , are equal, equivalently $\theta = \theta_E - \theta_S = 0$. The difference of the sample means, denoted by $\bar{X} = \bar{Y}_E - \bar{Y}_S$ is normally distributed with mean θ and variance τ^2/n where $\tau^2/n = 4\sigma^2/n$ when $\sigma_E^2 = \sigma_S^2 = \sigma^2$ (equal variance) or $\tau^2/n = 2(\sigma_E^2 + \sigma_S^2)/n$ when $\sigma_E^2 \neq \sigma_S^2$ (unequal variance).

In the Bayesian setting, θ is assumed to be unknown but to follow a distribution with specified form. Assume that the prior density of θ , which measures the belief regarding the parameter prior to observing any responses, is $N(\mu_0, \sigma_0^2)$. The prior density may be elicited from

previous trials, case studies or experts' opinions [5–9]. Following observations from patients, the prior is updated to give a posterior distribution summarising belief about θ given the observed data.

In the Bayesian decision theoretic framework, a utility function summarises the value of all possible actions given θ . The action that maximises the expected value of this utility over the posterior distribution of θ given the observed data can then be chosen as the optimal action. As with frequentist hypothesis testing, we may consider that at the end of the trial, there are two possible actions, to reject the null hypothesis or not. The utility function depends on the sample size n , and the decision, $d \in \{\text{do not reject } H_0, \text{ reject } H_0\}$, given θ . Let $d(\bar{x}, n)$ denote the action taken at the end of the trial given data \bar{x} with a sample size n and $G(n, d, \theta)$ denote the utility function. We may also choose n optimally. As we will not have observed any responses during the planning stage, the expected utility is obtained from the distribution of \bar{X} given θ and n . As θ is unknown, the expectation of the expected utility is taken over the prior density [4]. The expected utility is then a function of the sample size and decision rule, denoted by

$$\mathcal{G}(n, d) = \int_{\bar{x}} \int_{\Theta} G(n, d(\bar{x}, n), \theta) f(\theta | \bar{x}, n) d\theta f(\bar{x} | n) d\bar{x}, \quad (1)$$

where $f(\theta | \bar{x}, n)$ is the posterior density of θ given \bar{x} and

$$f(\bar{x} | n) = \frac{1}{\sigma_x} \phi\left(\frac{\bar{x} - \mu_0}{\sigma_x}\right)$$

is the prior-predictive density function for \bar{X} before sampling, with $\sigma_x^2 = \sigma_0^2 + \tau^2/n$ the prior predictive variance of \bar{X} and $\phi(\cdot)$ the normal density function. If the action taken at the end of the trial is determined by the outcome of some hypothesis test, the function d and hence the expected utility $\mathcal{G}(n, d)$, will depend on the type I error rate, or equivalently the critical value, for that test, as described in more detail in the next subsection. The optimisation problem then becomes one of choosing the trial sample size and test type I error rate.

Formulation of utility function

We assume that the decision maker for our proposed model represents society. A reward corresponds to improved treatment either of patients in the trial or of future patients if the experimental treatment receives regulatory approval.

The regulator is assumed to approve the experimental treatment if the observed difference, $\bar{X} = \bar{x}$, is greater than a threshold, $z_\alpha \tau / \sqrt{n}$, where z_α is the upper α percentile of the standard normal distribution, so that $d(\bar{x}, n) = \text{'reject'}$

H_0 ' or 'do not reject H_0 ' for \bar{x} above or below this threshold. This is equivalent to a classical frequentist analysis of the primary endpoint using a significance test conducted at (one-sided) level α . This assumption is similar to that proposed by Pezeshk et al. who also assume that the regulatory agency will test the null hypothesis that there is no difference between the outcomes means at the α test size [10]. Since the decision function d then depends only on z_α , we will write $\mathcal{G}(n, z_\alpha)$ for $\mathcal{G}(n, d)$ given by (1) above, and refer to z_α as the significance level threshold.

We assume that the size of the population for the treatments we are testing is known, and denote this by N . The number of patients that can be treated following the trial depends on N and on the trial sample size. We assume that during the trial a proportion ρ of patients are enrolled in the trial, so that there are $n(1 - \rho)/\rho$ concurrent patients not enrolled in the trial and the number of patients remaining to be treated is $N - n/\rho$.

We suppose that the utility given \bar{x} represents the gain from treating patients with the experimental treatment. There are $n/2$ such patients in the trial and, if $\bar{x} \geq z_\alpha \tau / \sqrt{n}$ so that the new treatment is approved, $N - n/\rho$ such patients following the trial. The reward for each treated patient is taken proportional to the true effect size, i.e. the more effective the treatment is the greater the reward will be. Note that it is considered a loss if the treatment effect is < 0 .

The utility function also includes fixed and variable financial costs [11]. Fixed costs are incurred regardless of the size of the trial, denoted by c_f ; these may be the setting up and running cost of the trial. Variable costs depend on the size of the trial. The cost per patient, denoted by c_1 , may be administrative costs to recruit, screen, treat and follow-up patients in the conduct of the trial. Our proposed utility function also includes additional costs, either monetary or harmful effects, of treating a patient with the experimental treatment, denoted by c_2 . All these costs may be scaled to one unit of efficacy such that they are relative to θ . Alternatively, the treatment efficacy may be scaled to one unit of cost such that it is relative to monetary reward and costs [12].

The utility at the end of the trial if the regulator approves the experimental treatment is the reward from treating patients in the trial and the future less the cost of treating them and both the fixed and variable trial costs. If the regulator does not approve the experimental treatment, the utility is the reward from treating patients in the trial minus the cost of treating patients in the trial with the experimental treatment and fixed and variable costs. Possible actions are to approve or not the experimental treatment. Thus, the utility function, taken relative to the baseline of treating all patients with the control treatment, is

$$G(n, d, \theta) = \begin{cases} (\theta - c_2) \left(N - \frac{n}{\rho} + \frac{n}{2} \right) - c_1 n \\ -c_f I_{\{n>0\}} & d = \text{reject } H_0, \\ (\theta - c_2) \frac{n}{2} - c_1 n \\ -c_f I_{\{n>0\}} & d = \text{do not reject } H_0, \end{cases} \quad (2)$$

where $I_{\{n>0\}} = 1$ if $n > 0$ and 0 otherwise. The expected utility given by (1) is then

$$\begin{aligned} \mathcal{G}(n, z_\alpha) &= \iint \left\{ (N - n/\rho)(\theta - c_2) I_{\{\tilde{x} \geq z_\alpha \tau / \sqrt{n}\}} \right. \\ &\quad \left. + \frac{n}{2}(\theta - c_2) - c_1 n - c_f I_{\{n>0\}} \right\} \\ &\quad \times f(\theta|\tilde{x}) f(\tilde{x}|n) d\theta d\tilde{x} \\ &= (N - n/\rho) \left\{ (\mu_0 - c_2) \Phi(-Z) + \frac{\sigma_0^2}{\sigma_x} \phi(Z) \right\} \\ &\quad + \frac{n}{2}(\mu_0 - c_2) - c_1 n - c_f I_{\{n>0\}}, \end{aligned} \quad (3)$$

where $\Phi(\cdot)$ is the normal cumulative distribution function and $Z = (z_\alpha \tau / \sqrt{n} - \mu_0) / \sigma_x$ is the z-score of the significance level threshold, z_α , on the prior predictive distribution of \tilde{X} . The full derivation is presented in the Additional file 1.

Optimisation

The optimal sample size and significance level threshold are obtained by maximising the expected utility, $\mathcal{G}(n, z_\alpha)$. Let z_α^* denote the optimal value for the significance level threshold. This is obtained for any n by differentiating $\mathcal{G}(n, z_\alpha)$ with respect to z_α , to obtain

$$\begin{aligned} \frac{d}{dz_\alpha} \mathcal{G}(n, z_\alpha) &= (N - n/\rho)(\mu_0 - c_2) \frac{d}{dz_\alpha} \Phi(-Z) \\ &\quad + (N - n/\rho) \frac{\sigma_0^2}{\sigma_x} \frac{d}{dz_\alpha} \phi(Z) \\ &= (N - n/\rho) \left[(\mu_0 - c_2) \frac{\tau / \sqrt{n}}{\sigma_x} \phi(Z) \right. \\ &\quad \left. + \frac{\sigma_0^2}{\sigma_x} \frac{\tau (z_\alpha \tau / \sqrt{n} - \mu_0) / \sqrt{n}}{\sigma_x^2} \phi(Z) \right]. \end{aligned} \quad (4)$$

Equating the differentiated expression (4) to zero, for a given sample size n the optimal z_α is

$$z_\alpha^* = c_2 \frac{\sigma_x^2 \sqrt{n}}{\sigma_0^2 \tau} - \mu_0 \frac{\tau}{\sigma_0^2 \sqrt{n}}. \quad (5)$$

Subsequently, the optimal sample size, n^* , is obtained numerically by substituting (5) into (3) (see Additional file 1 for details).

Results

Application to a case study

Abrahamyan et al. presented a decision analytic value of information (VOI) model for assessing evidence on treatments for children with severe haemophilia A, a rare disease [12]. They summarised that there are three types of treatments currently available in various developed countries; alternate day prophylaxis (AP), on-demand (OD) and tailored prophylaxis (TP) of intravenous administration of recombinant Factor VIII (FVIII). They utilised information from US and Canada studies to evaluate whether or not to conduct another trial. They performed three pairwise comparisons; TP vs. AP, OD vs. TP and OD vs. AP, and in each comparison the optimal decision, whether to conduct another trial or to accept one of the treatments as a standard therapy, was given depending on the maximum acceptable price per unit health gain.

We adapted this work to illustrate the application of our proposed model. Table 1 shows the estimated values used in our example. The primary endpoint was binary, whether or not the patient had magnetic resonance imaging (MRI)-detected joint damage. Similar to Abrahamyan et al., we assume large sample approximations to this binary endpoint and so the efficacy mean, $\hat{\mu}_t$, $t = (\text{AP}, \text{OD}, \text{TP})$, is the proportion of patients without MRI-detected joint damage. The sample and prior variances are $\hat{\sigma}_t^2 = \hat{\mu}_t(1 - \hat{\mu}_t)$ and $\hat{\sigma}_{0t}^2 = \text{var}(\hat{\mu}_t) = \hat{\mu}_t(1 - \hat{\mu}_t)/n_{0t}$, respectively where n_{0t} is the number of prior observations for treatment t reported by Abrahamyan et al., that is $n_{0,\text{AP}} = 27$, $n_{0,\text{OD}} = 29$ and $n_{0,\text{TP}} = 24$.

To illustrate our model, consider a two-arm RCT comparing the on-demand (OD), which we assumed to be the standard treatment, with the tailored prophylaxis (TP). The point prevalence of haemophilia A in US is 7.0 per 100,000 which is about 22,400 in the US given the US population of approximately 320 million [13, 14].

Let the measure of efficacy be the absence of MRI-detected haemophiliac joint damage and the unknown parameter measuring treatment benefit be the difference of proportion of patients without MRI-detected joint damage valued in monetary terms using the value the decision-maker places on each unity of efficacy, which will be denoted λ . The prior distribution for the unknown

Table 1 Summary statistics and costs (in dollar, \$) by treatment for haemophilia A, adapted from Abrahamyan et al. [12]

Statistics	Treatment, t		
	AP	OD	TP
Prior mean, $\hat{\mu}_t$	0.9259	0.5517	0.7917
Prior variance, $\hat{\sigma}_{0t}^2$	0.0025	0.0085	0.0069
Sample variance, $\hat{\sigma}_t^2$	0.0686	0.2473	0.1649
Mean cost, c_{2t} (\$)	176,397	56,619	117,651

AP, Alternate day prophylaxis; OD, on-demand; TP, tailored prophylaxis

parameter is then normal with prior mean $\mu_0 = \lambda(\mu_{TP} - \mu_{OD})$ and prior variance $\sigma_0^2 = \lambda^2(\sigma_{0,TP}^2 + \sigma_{0,OD}^2)$. As shown in Table 1, $\sigma_{TP}^2 \neq \sigma_{OD}^2$, therefore, the sample variance is estimated by $\tau^2 = 2\lambda^2(\sigma_{TP}^2 + \sigma_{OD}^2)$. Following Abrahamyan et al., take $\lambda = \$400,000$ leading to the values given in Table 2. Suppose the fixed financial cost incurred from conducting the trial is $c_f = \$1m$, the cost of conducting the trial per patient is $c_1 = \$5,000$ and the cost of treating a patient is $c_2 = c_{2,TP} - c_{2,OD} = \$61,032$ (see, Table 2).

Similar to Abrahamyan et al., we assume annual haemophilia A incidence of 200 with 1/5 of patients recruited to the trial ($\rho = 0.2$). Let the time horizon be 20 years and assume that all new cases will be prescribed with the recommended treatment after the trial. The total number of future patients who will benefit from the new treatment is thus $4000 - 5n$.

Figure 1 shows the expected utility for $z_\alpha = z_\alpha^*$ for a range of values of n . The optimal sample size, corresponding to the value for which the expected utility is highest is shown by the plotted point and is equal to $n^* = 46$ (i.e. 23 patients per arm), $\mathcal{G}(n^*, z_\alpha^*) = \141 million and the threshold to approve TP is $z_\alpha^* = 0.36876$, equivalent to $\alpha = 0.35615$.

Both n^* and z_α^* are far from their conventional values. In particular, except in the cases when $n = 0$, the power is much smaller than a conventional level. Taking $\alpha = 0.025$ and $\beta = 0.8$ for an alternative $\theta_A = \sigma_0/2 = \$24,819$ would require $n = 268$. The expected utility for this design is \$109 million.

Operating Characteristics

Figure 2 shows optimal sample sizes and significance levels obtained from our model for population sizes between 100 and 10,000,000 with other parameters fixed as $\theta \sim N(96,000, (49,638)^2)$, $\tau^2 = (\$363,202)^2$, $c_1 = \$5,000$, $c_2 = \$61,032$ and $c_f = \$1m$. Figure 2a shows that if the population is small ($N < 3000$), the optimal sample size is $n^* = 0$. In this case the optimal significance level threshold to approve TP is $z_\alpha^* \rightarrow -\infty$ (equivalently,

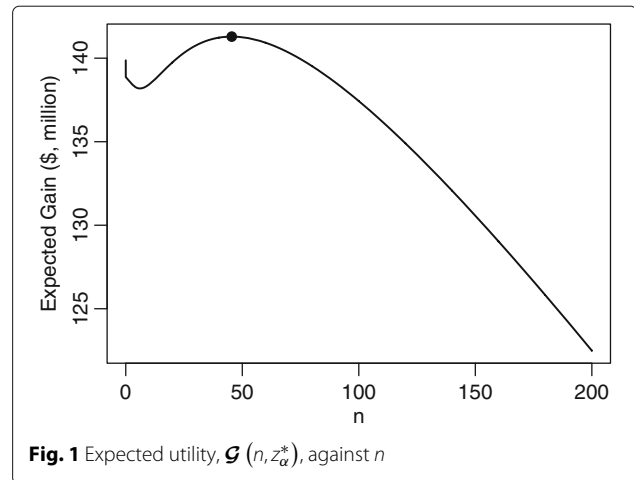


Fig. 1 Expected utility, $\mathcal{G}(n, z_\alpha^*)$, against n

$\alpha \rightarrow 1$, see Fig. 2c and d respectively), corresponding to an optimal decision to approving the experimental treatment based on the prior belief alone.

Other than for very small population sizes ($N > 3,000$) the optimal trial sample size increases with N , with $\log(n^*)$ increasing approximately linearly with $\log(N)$. This reflects work by Cheng et al. and Stallard et al. showing that as N increases $n^* \propto \sqrt{N}$ [15, 16]. The optimal significance level decreases as N increases (Fig. 2d), requiring a stricter level of statistical significance. Figure 2b shows the type II error rate of the test using sample size n^* and significance level threshold z_α^* , to detect an alternative $\theta_A = \sigma_0/2 = \$24,819$. In this case the type II error rate increases to approach 1 as N increases, so that that power approaches 0. This is reasonable since as $\theta_A < c_2$, this true effect difference is insufficient to justify recommendation of the treatment for use in future patients, though is very different to a more conventional approach in which the power will increase with increasing n .

Figure 3 shows optimal sample size (n^*), threshold (z_α^*), type I (α^*), and II (β^*) error rates, against the size of the population for various values of cost, $c_2 = 0, c_2 = \$12,409, c_2 = \$24,819, c_2 = \$61,032, c_2 = \$96,000$ and $c_2 = \$120,819$. Cases include c_2 equal to, above and below θ_A and the prior mean.

When $c_2 = 0$ we have $n^* > 0$ only if $N > 390,000$ (Fig. 3a). If $N < 390,000$ the optimal decision is to approve the experimental treatment without a trial. When $c_2 = \mu_0$, it is optimal not to start the trial if the population size, $N < 200$. In this case prior belief is that the expected utility from the new and standard treatments are equal so that rejection or non-rejection of H_0 lead to the same expected utility so z_α can be taken to be ∞ or $-\infty$, or equivalently α may be taken to be 1 or 0. When $c_2 > \mu_0$ although it is optimal to conduct no trial if the population size is less than 640 it is optimal not to approve TP but to continue with the standard treatment, OD. This

Table 2 Parameter estimates for the pairwise comparison between on-demand (OD) and tailored prophylaxis (TP)

Parameter	Estimates
Population size, N	4000
Prior mean, $\mu_0 = \lambda(\mu_{TP} - \mu_{OD})$, (\$)	96000
Prior variance, $\sigma_0^2 = \lambda^2(\sigma_{0,TP}^2 + \sigma_{0,OD}^2)$, (\$)	$(49638)^2$
Sample variance, $\tau^2 = 2\lambda^2(\sigma_{TP}^2 + \sigma_{OD}^2)$, (\$)	$(363202)^2$
Cost of conducting the trial per patient, c_1 , (\$)	5000
Cost of treating a patient, $c_2 = c_{2,TP} - c_{2,OD}$, (\$)	61032
Fixed financial cost incurred from conducting the trial, c_f , (\$)	1 million

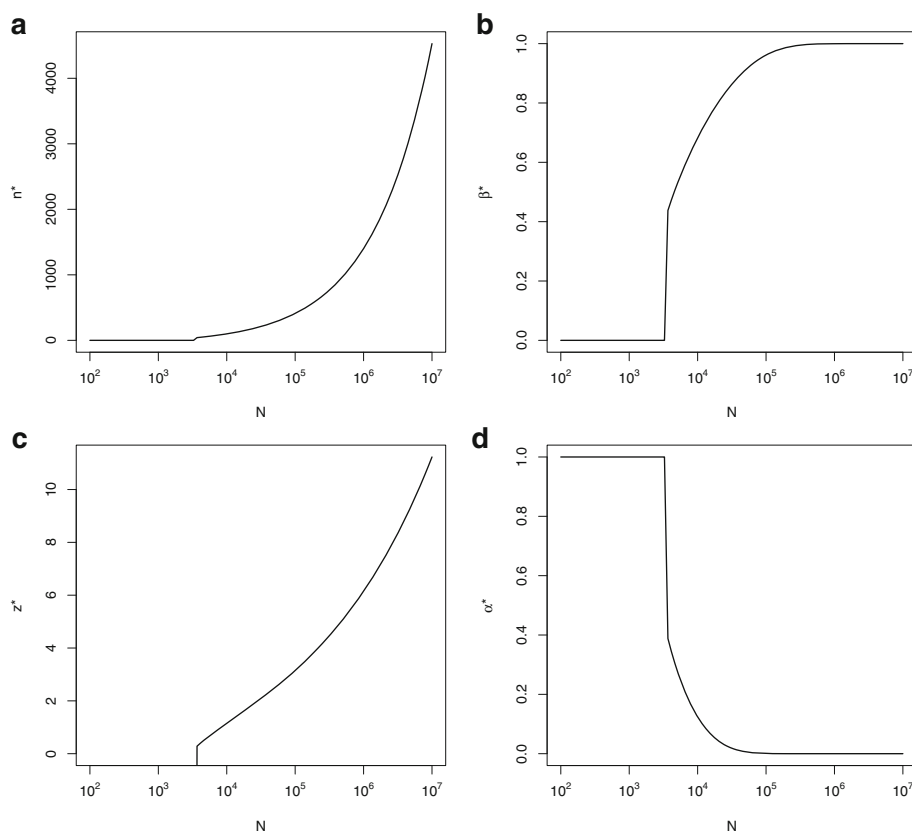


Fig. 2 Optimal (a) sample size, n^* , (b) type II error rate, β^* , in order to detect an alternative $\theta_A = \sigma_0/2 = \24819 (c) z^* and (d) type I error rate, α^* against the size of the population, N , with fixed $\mu_0 = \$96000$, $\sigma_0^2 = (\$49638)^2$, $\tau^2 = (\$363202)^2$, $c_1 = \$5000$, $c_2 = \$61032$ and $c_f = \$1$ million

is shown in Fig. 3d where the optimal type I error rate, $\alpha^* \rightarrow 0$. The optimal type II error rate, β^* approaches 1 (power approaches 0) as N and n^* become large when $c_2 > \theta_A$ and, though this is hard to see from the plot, approaches 0 (power approaches 1) as N becomes large when $c_2 < \theta_A$. Although not shown on these plots, as c_2 becomes very large, the optimal sample size, n^* , decreases, with $n^* = 0$ for largest c_2 values reflecting very strong prior belief that the TP is not sufficiently promising to overcome the cost c_2 .

Additional calculations were performed to evaluate the sensitivity of the results obtained to specification of the prior mean, μ_0 , and variance, σ_0^2 . Figures illustrating the optimal design parameters, n^* and α^* are given in the Additional file 2. These show that increasing or decreasing μ_0 has a similar effect on the choice of n and α as decreasing or increasing c_2 and that increasing or decreasing the prior variance, that is decreasing or increasing the level of prior information assumed, decreases or increases the value of N below which $n^* = 0$.

Discussion

The methodology of decision theory is the foundation of the value of information (VOI) method presented in the

health economics literature. The method is used to assess uncertainty in existing evidence and aid decision-making as to whether to adopt one of the competing treatments or to obtain more information. The decision-theoretic VOI method is based on expected utility maximisation and does not involve constructing and testing of hypotheses. Rather costs of gathering additional information are compared with expected returns in terms of the reduced likelihood of making the wrong decision, and the utility loss from doing so, though the result remains a decision as to whether or not the new treatment should be used and so is analogous to a hypothesis test. This requires specifying a utility function, which will be based on all outcomes (e.g. depending on the decision-maker, this may include health care use, adverse events, quality of life). The underlying methods are consistent in that for reasonable utility functions the decision to approve a treatment will be made so long as it appears sufficiently promising, corresponding to a hypothesis test conducted at some significance level. A number of authors have proposed decision-theoretic approaches to clinical trial design, both in the setting of normally distributed outcomes as considered here and binary outcomes [4]. A distinctive feature of this work is the specific focus on the small population setting. This is

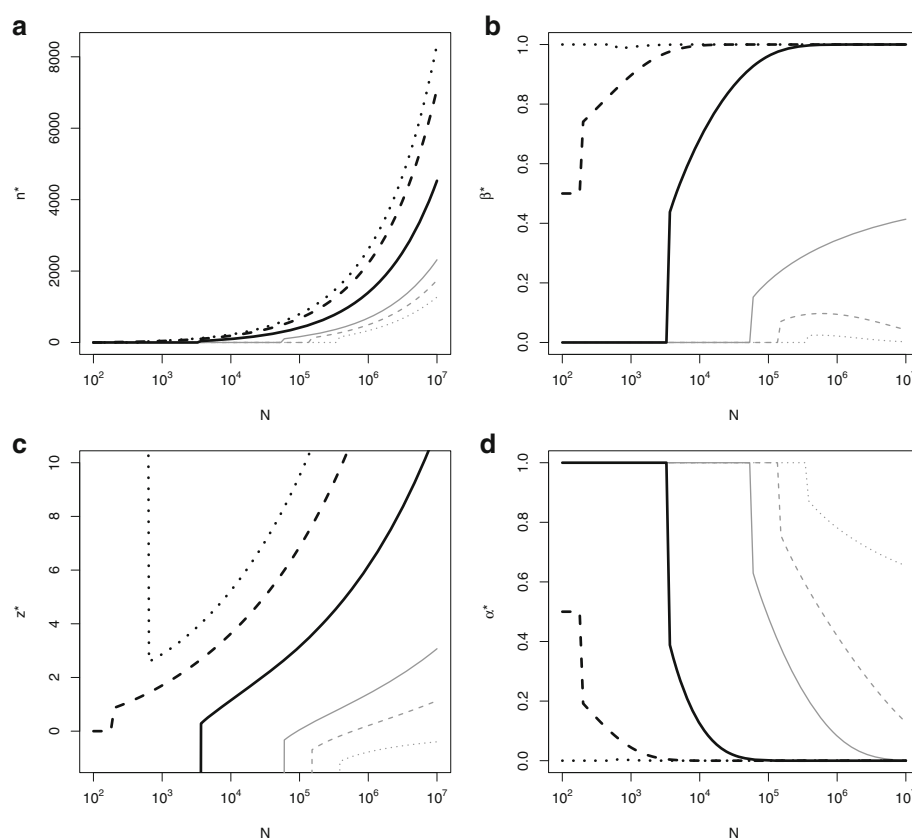


Fig. 3 Optimal (a) sample size, n^* , (b) type II error rate, β^* , (c) z_α^* and (d) type I error rate, α^* in order to detect an alternative $\theta_A = \$24819$ against the size of the population, N , with fixed $\mu_0 = \$96000$, $\sigma_0^2 = (\$49638)^2$, $\tau^2 = (\$363202)^2$, $c_1 = \$5000$ and $c_f = \$1$ million for different values of; $c_2 = 0$ (light grey dotted line), $c_2 = \sigma_0/4 = \$12409$ (light grey dashed line), $c_2 = \sigma_0/2 = \$24819$ (light grey solid line), $c_2 = \$61032$ (heavy black solid line), $c_2 = \mu_0 = \$96000$ (heavy black dashed line) and $c_2 = \sigma_0/2 + \mu_0 = \120819 (heavy black dotted line)

reflected in the construction of the utility function used, for example in the assumption of a fixed finite future population, with this decreasing in size depending on the length of the clinical trial.

Regulatory agencies may prefer the use of classical frequentist methods in the evaluation of pharmaceutical products. In the design of a clinical trial based on the frequentist method, the type I error rate, α , is usually restricted at a one-sided 0.025 level. This value is a conventional, but arbitrary, choice, and there is some indication that in practice there may be some flexibility depending on severity and/or prevalence of the disease. Our approach involves calculating a z-score for a primary clinical outcome that corresponds to maximising expected utility. It therefore allows a hypothesis testing framework to be maintained, but ensures that the error rate for that hypothesis test on the primary outcome is consistent with a utility maximisation framework. Results from Fig. 3 suggest that different α -levels are appropriate for diseases with different prevalence rates. Discussion of choice of α -level depending on prevalence and severity of disease has been considered by Montazerhodjat and Lo

(2015) [17]. In their working paper they also show that the optimal critical value increases with disease prevalence, but find a greater dependence on the severity of the disease with a more severe disease (e.g. pancreatic cancer) requiring a lower critical value (e.g. $z_\alpha = 0.587$, $\alpha = 0.279$) and a less severe disease (e.g. prostate cancer) requiring a higher critical value (e.g. $z_\alpha = 2.252$, $\alpha = 0.012$).

In the example presented above, we made some simplifying assumptions from the Abrahamyan et al. model for tractability, such as assuming independence of costs and clinical outcomes, which lead to slightly different optimal designs. In our model, we assumed that there is a non-trial cost in treating a patient with the experimental treatment, c_2 , that is known and fixed. However, c_2 may also be interpreted as the harm and risk of the experimental treatment. Other authors, for example, have interpreted this cost as critical event, treatment for side effects due to the experimental treatment, the loss from the reversal of a decision which is when after the decision to adopt the new treatment is made but following information from subsequent patients the treatment has to be withdrawn,

or observing adverse events both within and outside the trial which will affect costs and take-up of the new treatment [18–20]. As such the optimisation of sample size may depend on two endpoints; efficacy and safety. This may be simplified by aggregating efficacy and safety in a clinical utility index that could be used as the primary endpoint. Therefore, the unknown parameter, θ , would represent the mean difference in this utility index.

There are two possible decisions in our model where the experimental treatment is approved for treating future patients depending if the observed difference is greater than the optimal threshold. An alternative to defining the threshold at some frequentist test size, we could define it as some monetary value such as willingness to pay for one unit of health, e.g. in the model proposed by Willan and Eckermann where the decision making is from the perspective of the society who is responsible to decide whether or not to reimburse a new intervention at the given price to the company. In their model, the societal's decision is based on the threshold of the willingness to pay at a certain price [21].

We also assume that the size of the population is known. Based on Abrahamyan et al., who estimated the annual incidence to be 200, we took $N = 4000$ corresponding to a 20 year period of market exclusivity (the time during which the new treatment is the sole drug in the market). Alternatively, the population size could be modelled by considering the accrual time, market exclusivity period (the new treatment being the sole drug in the market) and the delay taken from the availability of results to the submission to the regulatory approval, production, marketing, distribution and sales [11, 19, 21, 22]. The inclusion of time or a model of growth and decay in our model is straight forward.

The estimate of the proportion of patients who will benefit from the new treatment if it is approved for routine use may be more complex by including patient's life expectancy and quality of life [23–25]. Or depending on the severity of the disease, the number of future patients who would take up the recommended new treatment depends on the treatment efficacy in a piecewise linear function [10, 26–29]. There may be no future patients who receive the treatment if the effectiveness is lower than a predefined threshold and some maximum number of patients if it is greater than a predefined threshold, i.e. the improvement is sufficiently large. If the efficacy is in between the lower and upper thresholds, the number of patients is a linear function that increases with the efficacy. This could effectively limit the choice of the optimal value α^* as insufficiently strong evidence from the trial of a treatment effect might lead to the new treatment not being used for any patients following the trial.

A challenge associated with implementation of the method proposed is specification of the prior distribution.

A number of authors have discussed elicitation of prior distribution parameters for Bayesian trial design, including in the setting of a rare disease [9, 30]. Using a more informative prior generally results in a smaller optimal trial sample size, sometimes with the optimum to be to conduct no trial at all but to make a decision based on prior data alone. If this is considered to be inappropriate, careful consideration should be given to the prior distribution as well as the specification of parameters of the utility function.

We have assumed that regulatory agencies make a simplistic binary decision, approve or not the submitted product. Agencies such as FDA/EMA may also give “accelerated approval/conditional approval”. Our proposed model can be expanded to include this decision where the utility function for this action may depend on the subsequent actions that the sponsor would take. Those subsequent actions may depend on whether or not the regulatory agencies might take given future observations. This type of sequential decision-theoretic model has been explored by various authors [31–35].

Conclusions

Decision-theoretic VOI analysis provides an alternative to conventional power calculations for the determination of the sample size for a clinical trial. Using such an approach, the final decision at the end of the trial and choice of the trial sample size are made so as to maximise the posterior expected value of a specified utility function.

Although the trial is not designed so as to control frequentist error rates at specified levels, since the decision to approve a new treatment will be taken so long as the observed mean difference is sufficiently large, this is equivalent to a frequentist hypothesis test conducted with a type I error rate corresponding to the optimal decision. The method can thus be seen as providing a flexible approach in which both the type I error rate and the power (or equivalently the sample size) of a trial can reflect the size of the future population for whom the treatment under investigation is intended. We believe such an approach could be particularly valuable in the setting of a small population such as a trial of a rare disease, when there is likely to be considerable knowledge of the size of the target population.

Additional files

Additional file 1: Derivation of expected utility function. Detailed mathematical derivation of results cited in main paper. (PDF 102 kb)

Additional file 2: Additional figures for different prior distributions. Additional figures similar to Fig. 3 in the main paper giving optimal designs and operating characteristics for a range of prior distribution mean and variance values. (PDF 112 kb)

Abbreviations

AP: Alternate day prophylaxis; OD: On demand; RCT: Randomised controlled trial; TP: Tailored prophylaxis; VOI: Value of information

Acknowledgements

The authors are grateful to the Editor and two reviewers for their helpful comments on the manuscript.

Funding

This work was conducted as part of the InSPIRe (Innovative methodology for small populations research) project funded by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement number FP HEALTH 2013-602144.

Availability of data and materials

Not applicable: no data or materials were used in this research.

Authors' contributions

The conception and design of the research study along with the interpretation of results were conducted by M. Pearce, SWH, JM, M. Posch, SD, FM, SZ and NS. Programming was conducted by M. Pearce, SWH and NS, who also drafted the manuscript. All authors were involved in critical revision of the manuscript, approved the final version and have agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Ethics approval and consent to participate

Not applicable: no patient data were used in this research.

Consent for publication

Not applicable: no patient data were used in this research.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Complexity Science, University of Warwick, Coventry, UK. ²Statistics and Epidemiology, Division of Health Sciences, Warwick Medical School, University of Warwick, Coventry, UK. ³Warwick Clinical Trials Unit, Division of Health Sciences, Warwick Medical School, University of Warwick, Coventry, UK. ⁴Section of Medical Statistics, CeMSIS, Medical University of Vienna, Vienna, Austria. ⁵Clinical Trials Consulting and Training Limited, Buckingham, UK. ⁶Department of Statistics, Stockholm University, Stockholm, Sweden. ⁷INSERM, U1138, team 22, Centre de Recherche des Cordeliers, Université Paris 5, Université Paris 6, Paris, France.

Received: 22 August 2017 Accepted: 14 January 2018

Published online: 08 February 2018

References

- Buckley BM. Clinical trials of orphan medicines. *Lancet*. 2008;371:2051–5.
- Bell SA, Tudur Smith C. A comparison of interventional clinical trials in rare versus non-rare diseases: an analysis of clinicaltrials.gov. *Orphanet J Rare Dis*. 2014;9:1–11.
- Hee SW, Willis A, Tudur Smith C, Day S, Miller F, Madan J, Posch M, Zohar S, Stallard N. Does the low prevalence affect the sample size of interventional clinical trials of rare diseases? An analysis of data from the aggregate analysis of clinicaltrials.gov. *Orphanet J Rare Dis*. 2017;12:44.
- Hee SW, Hamborg T, Day S, Madan J, Miller F, Posch M, Zohar S, Stallard N. Decision-theoretic designs for small trials and pilot studies: a review. *Stat Methods Med Res*. 2016;25:1022–38.
- Blanc TJ, Conahan TJ, Merin RG, Prager RL, Richter JJ. Bayesian Methods and Ethics in a Clinical Trial Design. In: Kadane JB, editor. New York: Wiley; 1996. p. 159–62.
- Chaloner K, Church T, Louis TA, Matts JP. Graphical elicitation of a prior distribution for a clinical trial. *J R Stat Soc Ser D Stat*. 1993;42:341–53.
- Kadane J, Wolfson LJ. Experiences in elicitation. *J R Stat Soc Ser D Stat*. 1998;47:3–19.
- Kinnersley N, Day S. Structured approach to the elicitation of expert beliefs for a Bayesian-designed clinical trial: a case study. *Pharm Stat*. 2013;12:104–13.
- O'Hagan A. Eliciting expert beliefs in substantial practical applications. *J R Stat Soc Ser D Stat*. 1998;47:21–35.
- Pezeshk H, Nematollahi N, Maroufy V, Gittins J. The choice of sample size: a mixed Bayesian/frequentist approach. *Stat Methods Med Res*. 2009;18:183–94.
- Patel NR, Ankolekar S. A Bayesian approach for incorporating economic factors in sample size design for clinical trials of individual drugs and portfolios of drugs. *Stat Med*. 2007;26:4976–88.
- Abrahamyan L, Willan AR, Beyene J, McIlmont M, Blanchette V, Feldman BM. Using value-of-information methods when the disease is rare and the treatment is expensive - the example of hemophilia A. *J Gen Intern Med*. 2014;29:767–73.
- Orphadata: Rare Diseases Epidemiological Data. INSERM. 1997. <http://www.orphadata.org>. Accessed 9 May 2016.
- The World Bank: Population total. 2016. <http://api.worldbank.org/v2/en/indicator/SP.POP.TOTL?downloadformat=excel>. Accessed 28 Sept 2016.
- Cheng Y, Su F, Berry DA. Choosing sample size for a clinical trial using decision analysis. *Biometrika*. 2003;90:923–36.
- Stallard N, Miller F, Day S, Hee SW, Madan J, Zohar S, Posch M. Determination of the optimal sample size for a clinical trial accounting for the population size. *Biom J*. 2017;59:609–25.
- Montazerhodjat V, Lo AW. Is the FDA too conservative or too aggressive? A Bayesian decision analysis of clinical trial design. NBER Working Paper No. 21499 2015. 23 September 2016, <http://www.nber.org/papers/w12499>.
- Ades AE, Lu G, Claxton K. Expected value of sample information calculations in medical decision modeling. *Med Dec Making*. 2004;24:207–27.
- Eckermann S, Willan AR. Expected value of information and decision making in HTA. *Health Econ*. 2007;16:195–209.
- Kikuchi T, Gittins J. A behavioral Bayes method to determine the sample size of a clinical trial considering efficacy and safety. *Stat Med*. 2009;28:2293–306.
- Willan AR, Eckermann S. Value of information and pricing new healthcare interventions. *Pharmacoeconomics*. 2012;30:447–59.
- Willan AR. Optimal sample size determinations from an industry perspective based on the expected value of information. *Clin Trials*. 2008;5:587–94.
- Hornberger J, Egtesady P. The cost-benefit of a randomized trial to a health care organization. *Control Clin Trials*. 1998;19:198–211.
- Halpern J, Brown Jr BW, Hornberger J. The sample size for a clinical trial: a Bayesian-decision theoretic approach. *Stat Med*. 2001;20:841–58.
- Willan AR, Pinto EM. The value of information and optimal clinical trial design. *Stat Med*. 2005;24:1791–1806.
- Gittins J, Pezeshk H. A behavioral Bayes method for determining the size of a clinical trial. *Drug Inf J*. 2000;34:355–63.
- Gittins J, Pezeshk H. How large should a clinical trial be? *J R Stat Soc Ser D Stat*. 2000;49:177–87.
- Pezeshk H, Gittins J. A fully Bayesian approach to calculating sample sizes for clinical trials with binary responses. *Drug Inf J*. 2002;36:143–50.
- Maroufy V, Marriott P, Pezeshk H. An optimization approach to calculating sample sizes with binary responses. *J Biopharm Stat*. 2014;24:715–31.
- Hampson LV, Whitehead J, Eleftheriou D, Brogan P. Bayesian methods for the design and interpretation of clinical trials in very rare diseases. *Stat Med*. 2014;33:4186–201.
- Berry DA, Ho CH. One-sided sequential stopping boundaries for clinical trials: a decision-theoretic approach. *Biometrics*. 1988;44:219–27.
- Lewis RT, Berry DA. Group sequential clinical trials: A classical evaluation of bayesian decision-theoretic designs. *J Am Stat Assoc*. 1994;89:1528–1534.
- Mehta CR, Patel NR. Adaptive, group sequential and decision theoretic approaches to sample size determination. *Stat Med*. 2006;25:3250–69.
- Chen MH, Willan AR. Determining optimal sample sizes for multistage adaptive randomized clinical trials from an industry perspective using value of information methods. *Clin Trials*. 2013;10:54–62.
- Yin G, Lam CK, Shi H. Bayesian randomized clinical trials: from fixed to adaptive design. *Contemp Clin Trials*. 2017;59:77–86.