Journal Title
XX(X):1–24

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Optimising error rates in programmes of pilot and definitive trials using Bayesian statistical decision theory

Duncan T. Wilson[1]

Abstract

Pilot trials of complex interventions are often conducted in advance of definitive trials to assess feasibility and to inform their design. Although pilot trials typically collect primary endpoint data, preliminary tests of effectiveness have been discouraged given their low power resulting from the small pilot sample size and assuming a conventional type I error rate. Power could be increased at the cost of a higher type I error rate, but there is little methodological guidance on how to determine the optimal balance between these factors. We consider a Bayesian decision-theoretic approach to this problem, introducing a utility function and defining an optimal pilot and definitive trial programme as that which maximises expected utility. We base utility on changes in average primary outcome, the cost of sampling, treatment costs, and the decision-maker's attitude to risk. We apply this approach to re-design OK-Diabetes, a pilot trial of a complex interventions with a continuous primary outcome with known standard deviation. We then examine how optimal programme characteristics vary with the parameters of the utility function. We find that the conventional approach of not testing for effectiveness in a pilot trial can be considerably sub-optimal.

Keywords

Clinical trial, pilot trial, external pilot, statistical decision theory, optimal design, expected utility

1 Introduction

Pilot trials are a type of feasibility study which take the same form as a planned definitive trial of a complex intervention, but on a smaller scale[1]. Internal pilots constitute the initial phase

[1]Leeds Institute of Clinical Trials Research, University of Leeds, Leeds, UK

Corresponding author:
Duncan T. Wilson, Clinical Trials Research Unit, Leeds Institute of Clinical Trials Research, University of Leeds, Leeds, LS2 9JT, UK
Email: d.t.wilson@leeds.ac.uk

of the definitive trial, with the pilot data being used in the final analysis. In contrast, external pilots are conducted separately to the definitive trial, with a clear gap between the two stages. A key goal of any pilot trial is to guide the decision of whether or not the definitive trial should go ahead, typically with a focus on feasibility issues such as recruitment rates and levels of missing data[2–4].

Although pilot trials will generally collect data measuring the effectiveness of the intervention, assessing effectiveness at the pilot stage has been discouraged due to concerns that the small pilot sample size will provide low power and lead to effective interventions being incorrectly discarded[5–8]. This criticism rests on two assumptions. Firstly, it assumes that the pilot and definitive trials will share a primary endpoint, with the same minimal clinically important difference (MCID). Secondly, it assumes that any pilot trial hypothesis test will be conducted with a significance level in the conventional range of 0.01 - 0.1. For example, an two-arm parallel group external pilot trial with a normally distributed primary endpoint designed using the rule-of-thumb of 35 participants per arm[9] would have a power of 23% (or equivalently, a type II error of $\beta = 0.77$) to detect a standardised effect size of 0.3 when using a one-sided type I error rate of $\alpha = 0.025$.

While the assumption of a constant MCID will often hold, there is no obvious reason for type I error rates in pilots to be constrained at conventionally low levels. Indeed, by not testing at all we effectively obtain a procedure with error rates $\alpha = 1, \beta = 0$. This testing strategy is only optimal if we have an absolute preference for minimising type II errors over type I errors in the pilot, a preference too extreme to be expected in practice. As illustrated in Figure 1, it will often be possible to reduce α considerably (in this case, from 1 to 0.75) at the cost of only a small increase in β (from 0 to 0.027). Although relaxing the type I error rate in a pilot has been suggested before[10,11], there is a lack of methodological guidance for determining exactly how much it should be relaxed by, or for choosing an appropriate pilot sample size.

One possible approach to defining optimal error rates is through Bayesian statistical decision theory. Under this framework we define a suitable utility function which encodes our preferences, and make decisions based on the expected value of this utility with respect to a prior distribution which expresses our uncertainty on the unknown parameters. Although the theory is well established[12–14] and has been proposed in much methodological work around optimal trial design[15], it has been argued that the requirement of specifying a utility function has led to low uptake in practice[16].

In this paper we aim to propose a simple and quite general form for a utility function in clinical trials, making clear the assumptions which are encoded in it and thus allowing its applicability or otherwise to the problem at hand to be judged. The utility we propose is closely related to several existing proposals in the literature[17], but with some key differences. One particular aspect we have considered is the decision-maker's attitude to risk, an issue sidestepped by many exiting proposals which assume, explicitly or implicitly, that the decision-maker is risk-neutral. We will show that the attitude to risk can have a considerable effect on decision making, and is key to answering one of the motivating questions of this paper: in what situations, if any, is it optimal to *not* test effectiveness in a pilot trial?

The remainder of this paper is structured as follows. We define the specific problem under consideration in Section 2, and describe the proposed method in Section 3. In Section 4 we

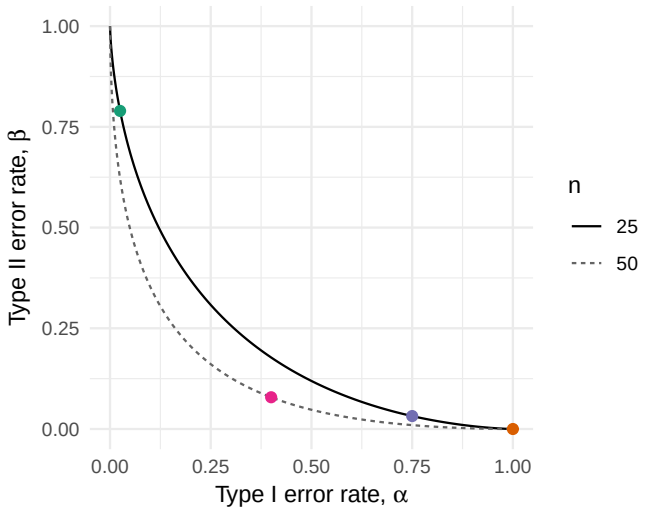

Figure 1. Operating characteristic curves for an external pilot trial testing efficacy.

illustrate the application of the method to design an external pilot of a complex intervention. We evaluate the properties of the method over a range of possible scenarios in Section 5, and then outline some extensions in Section 6. Finally, we conclude with a discussion of the strengths and limitations of the proposed approach in Section 7.

2 Problem

Consider the problem of jointly designing an external pilot trial and subsequent definitive trial. We will denote these, respectively, as stages $i = 1$ and $i = 2$ of the overall programme. We consider the case where both trials are parallel group studies comparing an intervention to control. We assume that the comparison focusses on superiority in terms of the mean difference of a normally distributed primary endpoint with known standard deviation common to each arm, where the same endpoint is used in both the pilot and definitive trials (although see Section 6 for relaxations of some of these assumptions).

We denote the true mean difference by μ, and consider the case where the the primary analysis at each stage will be a test of the null hypothesis $H_0 : \mu = 0$. The test at stage i will compare the sample mean difference between groups, denoted x_i, to a pre-specified critical value, denoted d_i. At the pilot stage, a positive result (i.e. $x_1 > d_1$) will indicate that we should proceed to the definitive trial. At the definitive stage, a positive result (i.e. $x_2 > d_2$) will indicate that the intervention should be recommended for use over the control treatment. The thresholds d_1, d_2, along with the per-arm sample sizes at each stage n_1, n_2, collectively define the design of the overall programme. The design problem we consider in this paper is to optimise $n_i, d_i, i = 1, 2$.

Given some alternative hypothesis $H_1 : \mu = \mu^*$, we define the type I and II error rates for stage i in the usual way, i.e. $\alpha_i = Pr[x_i > d_i \mid \mu = 0]$ and $\beta_i = Pr[x_i \leq d_i \mid \mu = \mu^*]$. An alternative summary of the pilot and definitive trial programme is then $\alpha_i, \beta_i, i = 1, 2$.

3 Maximising expected utility in trial programmes

We consider a Bayesian view of the frequentist design problem, and therefore require a prior distribution for the unknown true mean difference μ. This prior information will be used only to guide the choice of the frequentist design and analysis parameters, and not in any analysis of the trial data itself. As such, a non- or weakly-informative prior is not appropriate; rather, the prior should be a subjective summary of the decision-maker's knowledge and uncertainty about μ. For computational tractability we will assume a conjugate normal prior $p(\mu)$ with mean m_0 and variance s^2.

We define optimal design variables as those which maximise the expectation, with respect to the prior $p(\mu)$, of a utility function. We construct the utility function in three steps, following the procedures described by Keeney and Raiffa [13]. First, we identify the *attributes* which we consider will be of interest to the decision maker. We then define a *value function* over the space of these attributes, which encodes the decision-maker's preferences under conditions of certainty. We then transform this value function into a *utility function* by incorporating the decision-maker's attitude to risk. Throughout, we will make explicit the assumptions which are implied by the parametric forms of the value and utility functions, returning to discuss their implications in Section 7.

3.1 Attributes

The first attribute of interest is the change in average outcome for the patient population, in terms of the primary endpoint, after the trial programme has been conducted. We denote this change by d. It will be determined by three factors: the trial outcomes (in terms of the binary hypotheses test results); the true treatment difference μ; and the manner in which patients adopt the new treatment if the definitive trial concludes it is effective. Considering the latter requires a model predicting how patients will choose treatments. For simplicity, we will use a simple model assuming that the whole population will adopt the new treatment if the result of the definitive trial is positive, leading to a change in outcome of $d = \mu$. Otherwise, the population will retain the control treatment and the change in outcome will be $d = 0$.

The sample size at both stages of the programme is of interest, being associated with both monetary costs and the exposure of patients to research. The total sample size will vary from programme to programme, both by design and through the uncertain outcome of the pilot trial. Assuming that sampling at each stage is equivalent in terms of the costs involved, we can then identify the total sample size $n = n_1 + n_2$ as our second attribute.

While the focus of comparison between the intervention and control is in terms of the primary outcome, the treatments will in general differ in other aspects. We limit ourselves to the case where these differences are deterministic and known, and will refer to them as the treatment costs associated with the intervention. In this case, we can encapsulate these differences in a single

indicator variable, and our third attribute, $C \in \{0, 1\}$ where $C = 0$ if and only if a positive result is obtained in the definitive trial.

3.2 Value

Considering preferences under conditions of certainty, we use the notation $y \prec y'$ to mean y is preferred to y', and $y \sim y'$ to mean indifference between y and y'. We assume that $\prec$ is a weak ordering (i.e. complete and transitive) of all possible values y. We aim to specify a value function $v(n, C, d)$ which will assign to each possible set of attribute values a real number (its *value*) which corresponds with the qualitative preferences of the decision maker. That is, we require $v(n, C, d)$ such that

$$(d, n, C) \prec (d', n', C') \Leftrightarrow v(d, n, C) < v(d', n', C')$$

for all $(d, n, C), (d', n', C')$. We will use a value function with the following additive form:

$$v(d, n, C) = v_d(d) + v_n(n) + v_c(C),$$

where v_d, v_n and v_c are value functions representing preferences over each of the attributes when considered in isolation. It has been shown that such an additive value function exists if and only if each pair of attributes is *preferentially independent* of the remaining attribute.

Definition Preferential independence[13] (page 101). *The pair of attributes X and Y is preferentially independent of attribute Z if $(x', y', z) \prec (x'', y'', z) \Longleftrightarrow (x', y', z') \prec (x'', y'', z')$ for any z, z'.*

This condition requires, for example, that attributes n and C are preferentially independent of attribute d in the sense that preferences on the (n, C) space for a fixed d' do not depend on the specific value of d'.

We impose further structure on the value function by assuming the single-attribute functions $v_d(d)$ and $v_n(n)$ are linear in their arguments. This implies a very specific preference structure, where the value of a change in attribute level from d' to $d' + \Delta$ is independent of the starting level d'; and likewise for attribute n. Attribute C only has two levels and therefore v_c does not require any further specification. For ease of notation we will use the following single-attribute value functions:

$$v_d(d) = k_d d, \;\; v_n(n) = k_n n, \;\; v_C(C) = k_c C.$$

We elicit the scaling parameters k_d, k_n and k_c as follows. First, we ask the decision-maker to specify a change in outcome $\bar{d}$ that would justify increasing the total sample size from 0 to some value n^*. That is, we require $\bar{d}$ such that

$$(d = 0, n = 0, C = 1) \sim (d = \bar{d}, n = n_*, C = 1).$$

Note that the linear nature of v_n and v_d mean the specific choice of n_* is arbitrary. Secondly, we ask for the change in outcome $\hat{d}$ that would justify switching from the current standard treatment to the intervention under study. That is, we require $\hat{d}$ such that

$$(d = 0, n, C = 1) \sim (d = \hat{d}, n, C = 0).$$

Finally, since value functions are invariant under linear transformations, we can set $k_n + k_c + k_d = 1$ and are left with the following system of equations:

$$\begin{aligned} k_d\bar{d} + k_n n_* &= 0 \\ k_d\hat{d} - k_c &= 0 \\ k_d + k_n + k_c &= 1 \end{aligned}$$

This gives the following scaling parameters:

$$k_d = 1/(1 + \hat{d} - \bar{d}/n_*),\ k_n = -k_d\bar{d}/n_*,\ k_c = k_d\hat{d}. \tag{1}$$

3.3 Utility

The value function represents preferences under conditions of certainty, but in reality we are uncertain about the true values of all three attributes. To accommodate this uncertainty, we move from a value function to a utility function. This allows us to compare probability distributions over the attribute space, as opposed to only fixed points, and make decisions by choosing the distribution which has the largest expected utility.

To define the form of our utility function, we first argue that the change in outcome d is *utility independent* of the other two attributes, n and C.

Definition Utility independence[13] (page 226). *Attributes Y is utility independent of attribute Z when conditional preferences for lotteries on Y give z' do not depend on the particular level of z'.*

This means that regardless of how much we have sampled, or whether or not we have incurred the treatment costs associated with the intervention, our preferences for gambles on the change in outcome will be the same. Given this assumption together with the additive form of the value function, the utility function must have one of the following forms[13] (page 330):

$$u(n, C, d) = \begin{cases} 1 - e^{-\rho v(n,C,d)}, & \rho > 0 \\ v(n, C, d), & \rho = 0 \\ -1 + e^{-\rho v(n,C,d)}, & \rho < 0 \end{cases} \tag{2}$$

A value of $\rho = 0$ implies a risk-neutral attitude, whereas ρ greater than (less than) 0 implies a risk-averse (risk-seeking) attitude. To elicit ρ we first note that we can think about gambles on d whilst ignoring the value of the other attributes, since d is utility independent of n and C. We then elicit the value d^* such that we would be indifferent between the following:

1. Obtaining d^* with certainty;
2. A gamble which will result in d_{min} with probability 0.5 and d_{max} with probability 0.5.

That is, we find d^* such that

$$u(n, C, d^*) = 0.5u(n, C, d_{min}) + 0.5u(n, C, d_{max}).$$

In the special case of risk-neutrality, $\rho = 0 \Leftrightarrow d^* = 0.5d_{min} + 0.5d_{max}$. For $\rho \neq 0$ we have

$$d^* = -\frac{1}{\rho} \ln \left(0.5e^{-\rho d_{min}} + 0.5e^{-\rho d_{max}} \right),$$

and so we can determine ρ given d^*. For example, setting (arbitrarily) $d_{min} = 0, d_{max} = 1$, a value of $\rho = 2$ corresponds with $d^* = 0.283$. That is, $\rho = 2$ implies an indifference between obtaining a guaranteed change in outcome of 0.283, and a simple 50/50 gamble between no change at all and a change of 1.

3.4 Expected utility

Denote by $G_i \in \{0, 1\}$ an indicator variable where $G_i = 1$ if there is a positive test result at stage i, and $G_i = 0$ otherwise. For the problem considered here, $G_i = 1 \Leftrightarrow x_i > d_i$. Noting that the attributes d, n and C are completely determined by the fixed programme design z, the realisations of G_1 and G_2, and the true treatment effect μ, we re-write utility as $u(\mu, G_1, G_2 \mid z)$. Focussing on the case where $\rho > 0$ (the other cases will follow), we have

$$\begin{aligned} u(\mu, G_1, G_2 \mid z) = 1 - \exp(& - \rho[k_d\mu + k_n(n_1 + n_2)]G_1G_2 \\ & - \rho[k_n(n_1 + n_2) + k_c]G_1(1 - G_2) \\ & - \rho[k_n n_1 + k_c](1 - G_1)), \end{aligned} \tag{3}$$

The expected utility conditional on μ is

$$\begin{aligned} E_{G_1,G_2,|\mu,z}[u(\mu, G_1, G_2 \mid z)] = & Pr[G_1 = 1, G_2 = 1 \mid \mu] \left(1 - e^{\rho(k_d\mu + k_n(n_1+n_2)}\right) + \\ & Pr[G_1 = 1, G_2 = 0 \mid \mu] \left(1 - e^{-\rho(k_n(n_1+n_2)+k_c)}\right) + \\ & Pr[G_1 = 0 \mid \mu] \left(1 - e^{-\rho(k_n n_1 + k_c)}\right), \end{aligned} \tag{4}$$

Since the sample means are conditionally independent and normally distributed as $x_i \mid \mu \sim N(\mu, 2\sigma^2/n_i)$, the relevant conditional probabilities are easily calculated. We are then left with integrating out the unknown true treatment difference μ:

$$E[u(\mu, G_1, G_2 \mid z)] = \int E_{G_1,G_2,|\mu,z}[u(\mu, G_1, G_2 \mid z)]p(\mu)d\mu. \tag{5}$$

As we are integrating with a normal density weighting function, we use Gauss-Hermite quadrature (implemented in the 'fastGHQuad' R package[18]) to evaluate the integral efficiently and accurately. For the special case where pilot trial operating characteristics are fixed at $\alpha_1 = 1, \beta_1 = 0$, the expected utility of the programme is equivalent to the expected utility of a single definitive trial with set up costs equivalent to the pilot sample size n_1. This expectation can be calculated exactly, meaning that numerical integration can be avoided in this case. Full details are given in the appendix.

3.5 Optimisation

Optimal programme designs can be found by solving the optimisation problem

$$\begin{aligned} \max_{z=(n_1,d_1,n_2,d_2)} \quad & E[u(\mu, G_1, G_2 \mid z)] \\ \text{subject to } & n_i \in \mathbb{N}, i = 1, 2, \\ & d_i \in \mathbb{R}, i = 1, 2. \end{aligned} \tag{6}$$

To solve this problem we use the gradient-assisted local optimisation method of[19] as implemented in the R[20] package 'optimx'[21]. Full details are provided in the supplementary material.

4 Illustration

OK-Diabetes aimed to assess the feasibility of evaluating supported self-management for adults with learning disabilities and type II diabetes[22]. The original target sample size was 30 patients per arm, chosen based on a rule-of-thumb[5] and to allow the feasibility objectives of the study to be addressed. The team were asked by the funder to consider assessing the potential efficacy of the intervention to determine whether a seamless confirmatory trial should go ahead. A continuous measure of participant blood sugar levels (HbA1c) at six months was chosen as the efficacy outcome. The SD of this outcome was identified to be 1.5%[23]. A mean change of 0% was considered to be of no interest, whilst a mean reduction of 0.5% at six months was deemed clinically important.

The target sample size was increased to 56 participants per arm, giving $1 - \beta_1 = 0.82$ power to detect a true mean reduction of 0.5% using a one-sided test with a type I error rate of $\alpha_1 = 0.2$. Although the error rates for the subsequent definitive trial were not specified, we note that a sample size of 190 participants per arm would lead to $1 - \beta_2 = 0.9$ power to detect a true mean reduction of 0.5% using a conventional one-sided type I error rate of $\alpha_2 = 0.025$. Here, we consider how the proposed method could be used to determine optimal choices of $\alpha_i, \beta_i, i = 1, 2$.

4.1 Prior and utility

To apply the proposed method we require a prior distribution on the treatment difference, and a utility function. For the former, we use a conjugate normal prior with parameters $m_0 = 0$ and $s_0 = 0.6$. This represents a sceptical prior, being centred at the null hypothesis of no difference and with a variance corresponding to a prior belief that $\mu > \mu_1 = 0.5$ with probability 0.20.

For the utility function, we first consider the change in outcome which would be enough to justify the costs of switching from the current standard treatment to the new treatment under study. To determine this value we note that a conventional definitive trial design, with a type I error rate of 0.025 and a power of 0.8 to detect $\mu = 0.5$, would lead to 0.5 power when $\mu = 0.3$. This implies an indifference between adopting the new treatment and staying with the current standard if this was the true treatment difference[24], and thus gives a rationale for choosing $\hat{d} = 0.3$. For the cost of sampling, we seek to identify a change in treatment which would justify an increase in sample size from 0 to $n_* = 50$ (recall that the choice of n_* is arbitrary). For

Table 1. Optimal sample size and error rates for the OK-Diabetes external pilots trial ($i = 1$) and subsequent definitive trial($i = 2$), both for the general unrestricted case and where we insist on not testing effectiveness in the pilot trial.

Problem	n_1	n_2	α_1	β_1	α_2	β_2	Expected utility
Unrestricted	41	146	0.39	0.110	0.041	0.132	-0.42874
No pilot test	30	110	1.00	0.000	0.036	0.254	-0.42292

the purposes of illustration we suppose that this leads to $\bar{d} = 0.005$, meaning that we consider an increase in sample size of 5,000 to be worth paying if we obtained a *guaranteed* change in treatment effect of 0.5, the MCID in this problem.

Given these judgements and using equation 1, we have the value function

$$v(d, n, C) = 0.769d - 0.0000769n + 0.231C.$$

Moving to utility, we set $d_{min} = 0$ and $d_{max} = 0.5$ (arbitrarily) and consider the change of treatment we would like to obtain for certain for it to be judged equivalent to a simple 50/50 gamble between d_{min} and d_{max}. We suppose a risk-averse attitude leads to $d^* = 0.19$, corresponding to $\rho = 2$. Our utility function is then

$$u(d, n, C) = 1 - \exp[-2 \times (0.769d - 0.0000769n + 0.231C)].$$

4.2 Optimal design

We consider two variations of the optimal design problem. First, we optimise jointly over the pilot and main trial programme ('unrestricted'). Then, we optimise only the main trial whilst fixing $\alpha_1 = 1, \beta_1 = 0$ ('no pilot test'). In both cases we note that the original OK-Diabetes sample size of 30 per arm was intended to allow feasibility questions to be addressed, and so we set this as a lower limit of n_1. The algorithm takes around 1 second to converge to a solution in the general unrestricted case. For the special case of no pilot test where an exact expression of expected utility is available, convergence takes around 0.003 seconds. The results are given in Table 1.

In the unrestricted case we find that the optimal programme involves an external pilot sample size of $n_1 = 41$ participants per arm, between the initial and revised choices of sample size of 30 and 56 used in OK-Diabetes. The balance of error rates in the pilot is, however, substantially different to those chosen previously. We find that a large type I error rate of $\alpha_1 = 0.39$ (one sided) is used, allowing a high power of $1 - \beta_1 = 0.89$ whilst maintaining a low sample size. Having allowed a large type I error rate in the pilot, the optimal definitive trial uses a lower $\alpha_2 = 0.041$. In isolation this is somewhat higher than the conventional choice of 0.025, but note that when combined with the type I error rate of the pilot trial it leads to an overall type I error rate of $0.39 \times 0.041 = 0.016$. The optimal definitive sample size of 146 per arm then corresponds to a power of 0.868, with an overall type II error rate for the programme of $1 - (0.110 + 0.132 - 0.110 \times 0.132) = 0.773$.

When we insist on not testing in the external pilot we obtain a lower definitive trial sample size of $n_2 = 110$, with type I error rate $\alpha_2 = 0.036$ and power $1 - \beta_2 = 0.746$. The expected utility of this programme is 0.00582 lower than the optimal unrestricted programme. To interpret this,

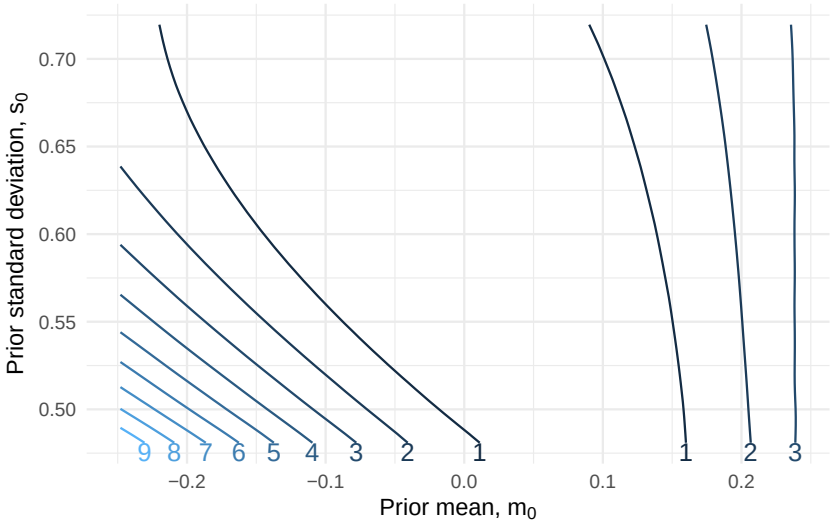

Figure 2. Amount of regret when using the proposed OK-Diabetes programme design as the prior mean m_0 and prior standard deviation s_0 vary.

we can translate utilities back to values and then into attribute units. Specifically, this difference in utility corresponds to a difference in value of

$$\frac{1}{2}\left[\log(1-0.42558)-\log(1-0.42874)\right]=0.005068229,$$

and dividing this by k_n leads to an effective difference of 66 participants. That is, we can consider the unrestricted optimal design to be more efficient than the restricted design by an amount equivalent to recruiting and following up 66 participants. Thus, in this case, the conventional policy of not testing effectiveness in pilot trials is considerably inefficient.

4.3 Sensitivity analysis

The suggested programme design is optimal only for a certain choice of prior and utility parameters, and so it is of interest to assess how robust the design is to deviations from these. To do this we consider a range of alternative parameter values and, for each, determine the optimal programme design. The expected utility of this optimal design can then be compared against that of the proposed design, converted into units of sample size as above. We will refer to this difference as the *regret*. We conducted two sensitivity analyses: first, we varied the prior parameters m_0 and s_0; secondly, we varied the utility parameters ρ and $\bar{d}$. All other parameters were kept at their original values.

Figure 2 plots the regret over a range of prior means m_0 and prior standard deviations s_0. We varied the prior mean from -0.5 to 0.5, moving from extremely sceptical to enthusiastic beliefs. We find that over this range there is little to be gained from moving from the proposed design to

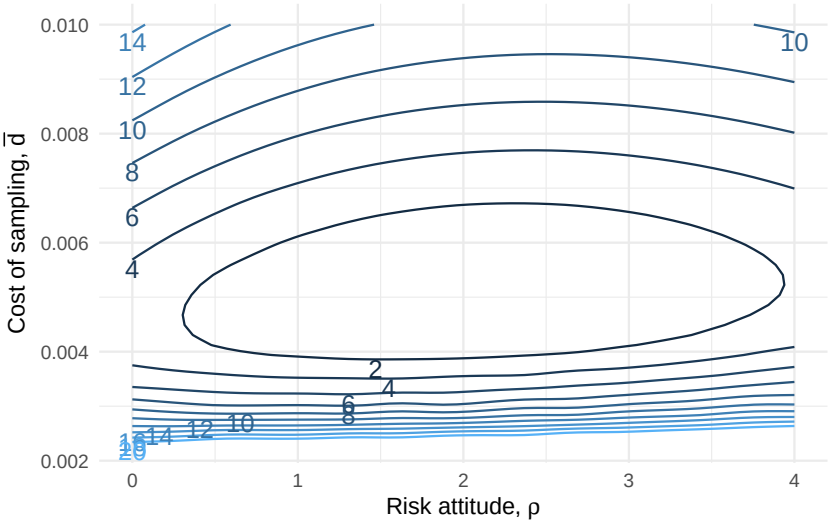

Figure 3. Amount of regret when using the proposed OK-Diabetes programme design as the attitude to risk ρ and cost of sampling $\bar{d}$ vary.

the optimal design, providing the prior standard deviation is equal to or greater than the initial choice of $s_0 = 0.6$. As we decrease s_0 down to 0.48 the penalty of using the proposed design can increase, but the magnitude of these penalties depends on m_0. From these results we can conclude that the proposed design is quite robust to misspecification of the prior distribution, in the sense that if the choices of m_0, s_0 are not quite an accurate reflection of our prior beliefs, the design will still have an expected utility close to that of the true optimal design.

Corresponding results for varying utility parameters ρ and $\bar{d}$ are given in Figure 3. We see that the proposed design is quite robust to misspecification of the attitude to risk, and to underestimation of the cost of sampling. However, if the cost of sampling is initially overestimated, the proposed design can become considerably sub-optimal. For example, maintaining $\rho = 2$ but halving the cost of sampling from 0.005 to $\bar{d} = 0.0025$ means the proposed design is worse than the true optimal design by an amount equivalent to 24 participants. This analysis suggests that the choice of $\bar{d}$, in particular, should be carefully examined to ensure it is a true reflection of the decision-maker's preferences.

5 Evaluation

In the OK-Diabetes example we found that the standard policy of not testing for efficacy in an external pilot trial can be considerably sub-optimal. Here, we consider a range of different utility function parameter values and examine when, if at all, not testing in the pilot trial is optimal. Throughout, we maintain the same archetypal sceptical prior with $m_0 = 0$ and $s_0 = 0.6$. We considered the nine scenarios formed by setting the cost of sampling $\bar{d}$ to one

of $\{0.0025, 0.005, 0.01\}$, and the treatment cost parameter $\hat{d}$ to one of $\{0.1, 0.2, 0.3\}$. For each of the nine scenarios we varied the attitude to risk, with $\rho \in [-5, 5]$, finding optimal programme designs over this range. We did this for two cases: firstly, assuming that a pilot sample size of $n_1 \geq 30$ is required in order to address feasibility questions; and secondly, removing this lower bound.

5.1 The case $n_1 \geq 30$

The results are given in Figure 4, which plots how the error rates of both the pilot ($i = 1$) and definitive ($i = 2$) trials vary with ρ, for each of the nine scenarios. In the scenarios considered, it is always optimal to test for effectiveness in the pilot trial, although the type I error rate used can be quite high. The largest we found was $\alpha_1 = 0.89$, when $\rho = -1.8, \bar{d} = 0.0025$ and $\hat{d} = 0.1$ (top left panel in Figure 4). The trends in Figure 4 suggest that decreasing $\bar{d}$ and/or $\hat{d}$ could potentially lead to higher α_1, but we failed to find any case where $\alpha_1 = 1$. Note that although α_1 is found to be increasing as ρ decreases in all scenarios (for sufficiently low ρ), and so we might expect there to be some negative ρ such that $\alpha_1 = 1$, these scenarios also have $n_2 = 0$. That is, the pilot trial is the only trial being conducted.

The broad trends which emerge from Figure 4 are that optimal type I errors tend to decrease as we become more risk-averse, while optimal type II errors stay relatively stable. As the treatment costs increase (moving from left to right in Figure 4), both type I and II errors tend to decrease. And, as the cost of sampling increases (moving from top to bottom in Figure 4), both type I and II errors tend to decrease. In all nine scenarios we find there is a point where the definitive trial jumps to a design of $n_2 = 0, \alpha_2 = 1, 1 - \beta_2 = 1$, meaning the pilot trial is the only trial which will be run. The point where this happens is always for a negative value of ρ. That is, there is a point where a sufficiently risk-seeking attitude will imply the optimal action is to run only one trial.

5.2 The case $n_1 \geq 0$

We now examine the characteristics of optimal programmes with no lower bound on the sample size at the pilot stage. This will be the case when the purpose of the pilot trial is only to assess effectiveness, as opposed to feasibility, and is similar to the problems considered in related work on optimal pilot design[25]. The results are given in Figure 5, which plots how the error rates of both the pilot ($i = 1$) and definitive ($i = 2$) trials vary with ρ, for each of the nine scenarios.

The trends of how optimal error rates and sample sizes vary with the utility function parameters are broadly similar to those shown in Figure 4. We see similar inflection points, where now a sufficiently risk-seeking attitude will result in an optimal pilot trial sample size of $n_1 = 0$, leaving one one trial to be conducted. Optimal pilot trial type I error rates are only found to be $\alpha_1 = 1$ when $n_1 = 0$. In the scenarios considered here, we again fail to find a situation where it is optimal to run a pilot trial, but not test for effectiveness.

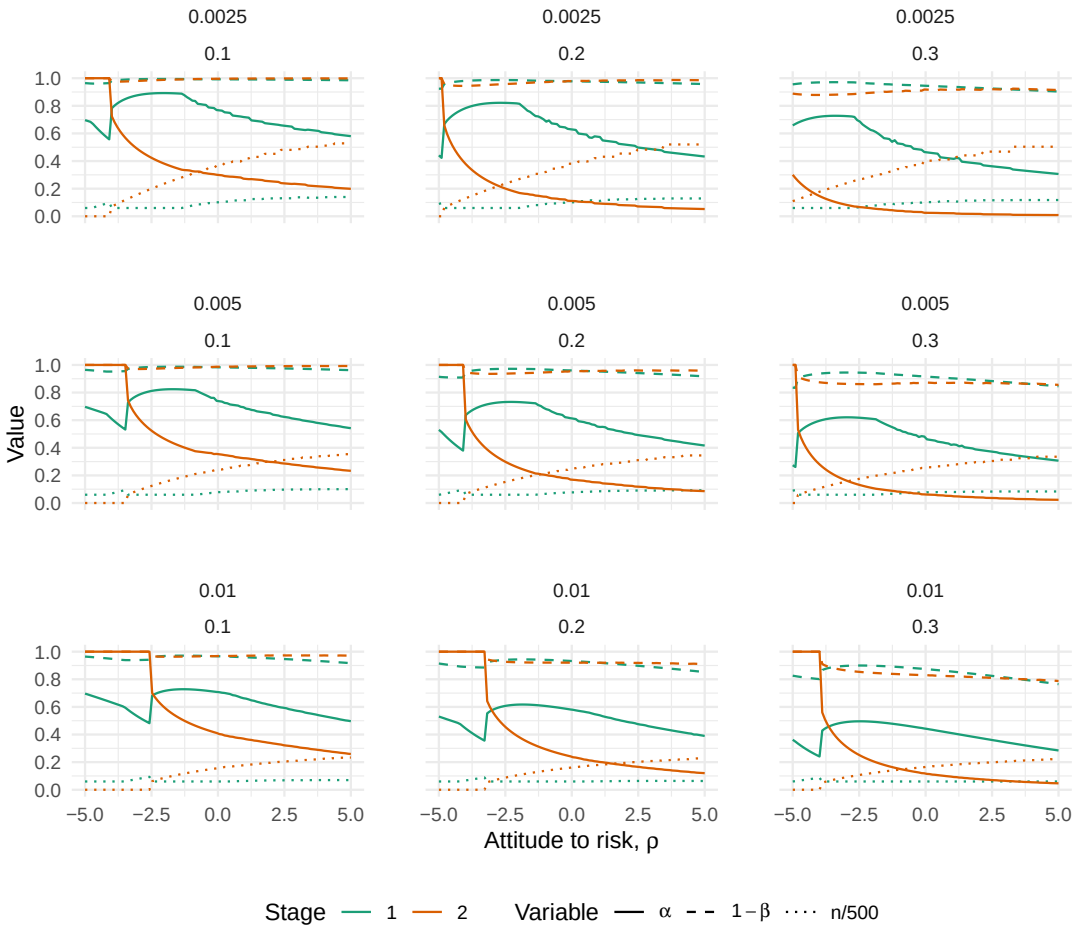

Figure 4. Optimal type I error rates (solid lines), type II error rates (dashed lines) and scaled sample size (dotted lines) for varying values of ρ (the attitude to risk, where higher means more risk-averse), when the pilot sample size is constrained to $n_1 \geq 30$.. Plots vary horizontally with treatment costs, $\hat{d} \in \{0.1, 0.2, 0.3\}$, and vertically with sampling costs, $\bar{d} \in \{0.0025, 0.005, 0.01\}$.

6 Extensions

6.1 Non-inferiority

Although we have focussed on superiority trials, the decision-theoretic approach extends naturally to the non-inferiority setting. In a non-inferiority trial we are interested in testing whether or not the new intervention is not worse than the control by some pre-specified amount (the non-inferiority margin). The proposed method encompasses this setting by allowing for

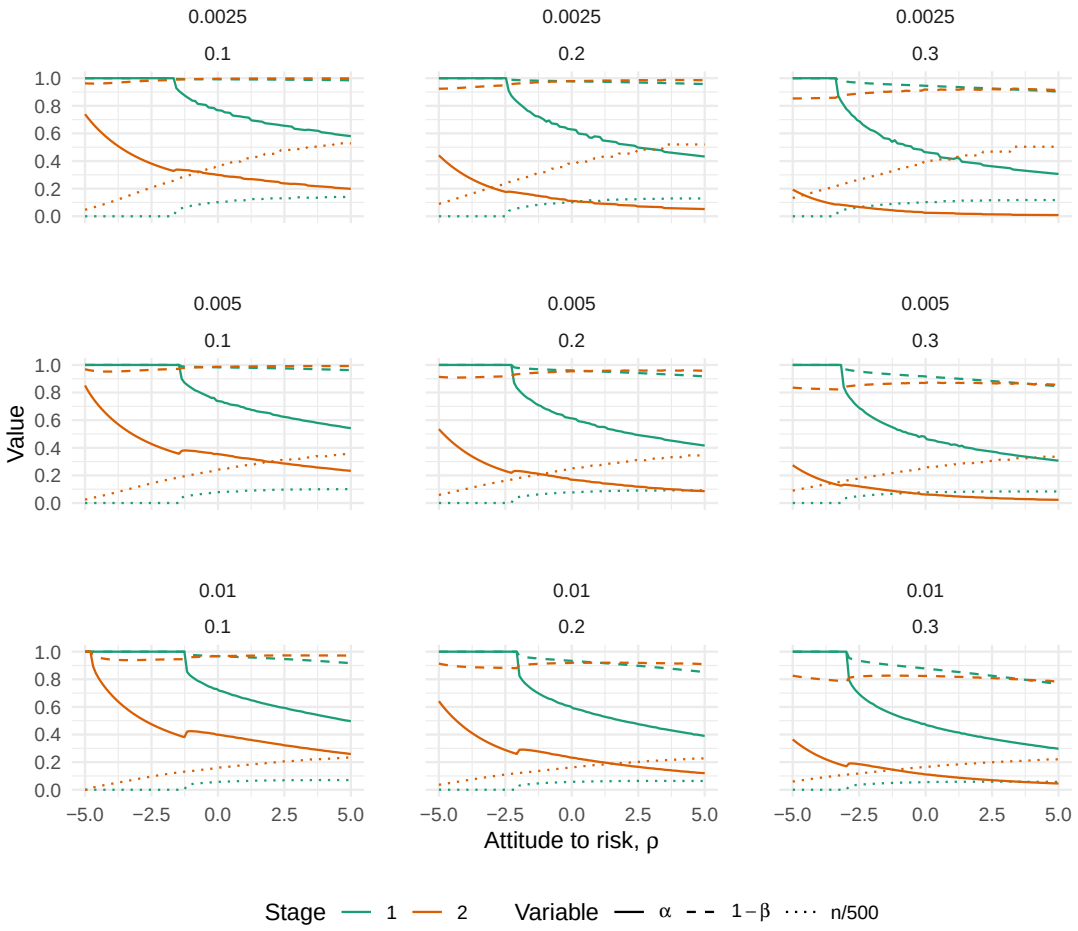

Figure 5. Optimal type I error rates (solid lines), type II error rates (dashed lines) and scaled sample size (dotted lines) for varying values of ρ (the attitude to risk, where higher means more risk-averse), when the pilot sample size is unconstrained. Plots vary horizontally with treatment costs, $\hat{d} \in \{0.1, 0.2, 0.3\}$, and vertically with sampling costs, $\bar{d} \in \{0.0025, 0.005, 0.01\}$.

negative choices of the parameter $\hat{d}$, which denotes the amount of treatment difference we would consider equivalent to the costs of adopting the new treatment. In the non-inferiority setting the new treatment will be considered cheaper (in some sense, not necessarily monetary) than the control, leading to a negative $\hat{d}$.

As an example, we return to the OK-Diabetes trial of Section 4 but now consider tests of non-inferiority, where the non-inferiority margin is $\mu = -0.5$. Keeping all other parameters at their previous values but setting $\hat{d} = -0.3$, the optimal sample size at each stage is now

Table 2. Optimal sample size and error rates for the OK-Diabetes pilot trial ($i = 1$) and subsequent definitive trial($i = 2$), when the pilot is external and internal.

Problem	n_1	n_2	α_1	β_1	α_t	β_t	Expected utility
External	41	146	0.39	0.110	0.016	0.228	-0.42874
Internal	45	121	0.42	0.084	0.016	0.213	-0.42954

$n_1 = 77, n_2 = 430$. The type I error rate is now defined as the probability of a positive test result conditional on the non-inferiority margin $\mu = -0.5$, while power is the probability of a positive test result conditional on no difference, $\mu = 0$. The optimal design then has error rates $\alpha_1 = 0.68, \beta_1 = 0.0057, \alpha_2 = 0.034, \beta_2 = 0.0011$.

6.2 Internal pilots

Internal pilot trials are distinguished from external pilots by their data being used at the final analysis, with a seamless gap between the pilot and definitive trial stages. Extending our problem to the internal pilot setting, we continue to conduct a first test based on the pilot sample mean difference x_1, but now follow this with a test of the overall sample mean difference x_t, where

$$x_t = \frac{n_1 x_1}{n_1 + n_2} + \frac{n_2 x_2}{n_1 + n_2}.$$

We can now apply equation 4 in the internal pilot by defining $G_1 = x_1 > d_1$ and $G_2 = x_t > d_2$. The relevant probabilities can be calculated by noting that the pair x_1, x_t, conditional on μ, follow a bivariate normal distribution. Specifically, the appendix will show that

$$\begin{pmatrix} x_1 \\ x_t \end{pmatrix} \mid \mu \sim BN\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \frac{2\sigma^2}{n_1} & \frac{2\sigma^2}{n_1+n_2} \\ \frac{2\sigma^2}{n_1+n_2} & \frac{2\sigma^2}{n_1+n_2} \end{pmatrix} \right).$$

The probabilities in equation 4 are now with respect to this bivariate normal distribution, and can be calculated using (for example) the R package 'mvtnorm'[26]. Expected utility can then be calculated as before, integrating the conditional expected utility over the normal prior $p(\mu)$ using quadrature.

By way of illustration, we found the optimal internal pilot and definitive trial programme for the OK-Diabetes example. The details are provided in Table 2, where we also include the optimal programme for the external pilot case as found in Section 4. Note that we summarise the programmes here in terms of the error rates of the pilot hypothesis test (α_1, β_1) and the overall error rates of the programme (α_t, β_t), as these are of more interest in the internal pilot setting. We find that the overall type I error rates are approximately equal for both the external and internal pilot cases, and overall type II rates are very similar. The internal pilot programme has a slightly higher expected utility, which we might expect given the fact that all of the data is being utilised in the final analysis.

6.3 Binary endpoint

Suppose we have a binary primary endpoint, with probability of occurrence of τ_C in the control arm and τ_I in the intervention arm. We are interested in the absolute difference $\mu = \tau_C - \tau_I$,

in the sense that this is what features in our utility. We use Beta priors to reflect our prior uncertainty, with $\tau_C \sim Beta(a_C, b_C)$ and $\tau_I \sim Beta(a_I, b_I)$.

We can extend our approach to this scenario by returning to equation 4 and noting that, now conditioning on the pair of probabilities (τ_C, τ_I), the probabilities of test success/failure at each stage can still be easily calculated. Now, rather than integrating the conditional expectation over a single normal prior density, we must integrate over the two independent beta distributions for τ_C and τ_I. We can use general numerical integration methods such as the adaptive procedure implemented in the R package 'cubature'[27] to estimate these integrals, although these will be slower than the Gauss-Hermite method used in the continuous case.

6.4 Unknown variance

When the variance of the primary outcome is unknown but still assumed to be common across arms, the tests at each stage of the design will compare t-statistics

$$t_i = \frac{x_i}{\sqrt{2\hat{\sigma_i}^2/n_i}}$$

for stage $i = 1, 2$, where $\hat{\sigma_i}^2$ are estimates of the variance at each stage. Conditional on μ, σ^2, these t-statistics will follow non-central t-distributions with $2n_i - 2$ degrees of freedom and non-centrality parameter $\mu/\sqrt{2\sigma^2/n_i}$. The probability of test success/failure at each stage in equation 4 (now conditioning on μ and σ^2) can then be readily calculated. Given this conditional utility, we can proceed as before by numerically integrating over the prior distribution, now a joint prior $p(\mu, \sigma^2)$.

7 Discussion

We have explored how Bayesian statistical decision theory can be used to define optimal type I and II error rates for trial programmes involving a pilot trial and a subsequent definitive trial. We have introduced a general utility function, outlining the associated assumptions, and demonstrated how its parameter values can be determined. When evaluating the conventional approach to pilot trial analysis, we found that a policy of not testing effectiveness was consistently sub-optimal. As a result, we recommend that pilot data can and should be used to conduct a preliminary test of effectiveness prior to the definitive trial, when the assumptions described in this paper hold. This would lead to a considerable improvement in the complex intervention evaluation pathway, as more ineffective interventions are identified at the pilot stage.

A key component of the decision-theoretic approach is the utility function. For simplicity, we did not include any set-up costs relating to the pilot or definitive trial. If these are thought to be important, expressing them in units of sample size would allow them to be included in the model easily. In terms of the resulting effect on optimal design characteristics, set-up costs would mean a design with either $n_1 = 0$ or $n_2 = 0$ becoming more attractive. As such, we might expect to see such designs becoming optimal over a larger range of values for ρ in Figures 4 and 5. We did not attempt to predict the number of patients who will be affected by the results of the definitive trial, or the manner in which they will adopt the intervention following a significant

result. Were such a model to be included, the utilities of the people participating in the trial could be included and balanced against the utilities of those who stand to benefit from the trial results. Such considerations will be particularly important in small population contexts, such as with rare diseases, where the trial population can form a considerable fraction of the overall target population[17].

The exponential form of the utility function was based on an additive value function and an assumption of utility independence, in addition to an assumed mutual preferential independence between the three attributes. Although the appropriateness of these assumptions must be judged in light of the problem at hand, we note that an additive utility function is often assumed in related decision-theoretic work[17,24,28–30] . As shown in equation 2, an additive utility entails these assumptions while also assuming risk-neutrality on the part of the decision maker. It could be argued that the function $v_d(d)$ would actually take a non-linear form, with diminishing returns as d increases. This will depend on the specific clinical context, but we note that in practice our assumption only requires linearity over the plausible region of d, as specified by the prior $p(\mu)$, which may be more reasonable.

We have assumed throughout that the effect being measured is the same at both the pilot and definitive trial stages. One argument against assessing effectiveness in a pilot trial is that it may lead to a biased effect estimate[8]. For example, it may be that the treatment providers involved at the pilot stage are likely to be more specialised, or more enthusiastic, than those who will be involved at the definitive stage. The methods proposed should, therefore, be used only when the pilot trial is an accurate reflection of the definitive trial. If it is thought that there will be some systematic difference, this could potentially be modelled in the Bayesian framework by introducing a bias parameter in the model and describing a prior distribution for it. In the extreme case where the amount of bias is completely unknown, the pilot data will be uninformative for the true treatment effect and testing at the pilot stage should be avoided.

We have considered programmes where a hypothesis test is used in the primary analysis of the pilot and definitive trials. Further work could explore how a Bayesian analysis of pilot trial data could be used to update prior beliefs and use the revised knowledge to optimise the subsequent definitive trial. At the programme design stage, the pilot trial sample size could then be determined using value of information methods[24]. A potential difficulty with such an approach is the computational aspect of such calculations, although techniques for enabling fast calculation of the expected value of sample information may be useful in this context[31,32].

We have focussed on using pilot trials to test the efficacy of the intervention. Typically, pilot trials are used to estimate other parameters relating to the feasibility of the definitive trial, such as recruitment, follow-up and adherence rates[33]. Learning about these parameters is clearly of value, and this value could be included in our utility formulation to be offset against the cost of sampling.

The optimisation problem 6 is not trivial, and we found some variability in performance of different optimisation algorithms. The suggested method was found to be robust, but it would be advisable to check for global convergence when applying to a given problem. This could be done by using other algorithms to check they agree or by using different starting points. Alternatively, several closely related problems could be solved and the resulting optimal programme characteristics plotted, much as we have done in the sensitivity analyses of Section

4.3. We would expect to see smooth variation, with any erratic behaviour would suggest some convergence issues. Note this is exemplified in Figure 5, where some small blips in the operating characteristic curves can be seen and would suggest a slight failure in convergence at these points.

The motivation for our approach is the lack of guidance on how to balance the sample size and the type I and II error rates when testing in a pilot trial. The problem is by no means limited to the pilot trial setting, but for larger studies it tends to be 'solved' by constraining type I and II error rates at default levels, typically $\alpha \leq 0.05$ (two-sided) and $\beta \leq 0.2$. Recently, much debate has focussed on if and how the default type I error rate should be changed[34], or abandoned altogether[35–37]; and similarly with the case of power thresholds[38].

Acknowledgements

Acknowledgements.

Declaration of conflicting interests

The Authors declare that there is no conflict of interest.

Funding

This work was supported by the Medical Research Council [grant number MR/N015444/1].

References

1. Eldridge SM, Lancaster GA, Campbell MJ et al. Defining feasibility and pilot studies in preparation for randomised controlled trials: Development of a conceptual framework. *PLOS ONE* 2016; 11(3): e0150205. DOI:10.1371/journal.pone.0150205. URL `http://dx.doi.org/10.1371/journal.pone.0150205`.
2. Craig P, Dieppe P, Macintyre S et al. Developing and evaluating complex interventions: the new medical research council guidance. *BMJ: British Medical Journal* 2008; 337. DOI:10.1136/bmj.a1655.
3. Thabane L, Ma J, Chu R et al. A tutorial on pilot studies: the what, why and how. *BMC Medical Research Methodology* 2010; 10(1): 1. DOI:10.1186/1471-2288-10-1. URL `http://www.biomedcentral.com/1471-2288/10/1`.
4. Eldridge SM, Chan CL, Campbell MJ et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *BMJ* 2016; : i5239DOI:10.1136/bmj.i5239. URL `http://dx.doi.org/10.1136/bmj.i5239`.
5. Lancaster GA, Dodd S and Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *Journal of Evaluation in Clinical Practice* 2004; 10(2): 307–312. DOI: 10.1111/j..2002.384.doc.x. URL `http://dx.doi.org/10.1111/j..2002.384.doc.x`.
6. Arain M, Campbell M, Cooper C et al. What is a pilot or feasibility study? a review of current practice and editorial policy. *BMC Medical Research Methodology* 2010; 10(1): 67. DOI: 10.1186/1471-2288-10-67. URL `http://www.biomedcentral.com/1471-2288/10/67`.
7. Westlund E and Stuart EA. The nonuse, misuse, and proper use of pilot studies in experimental evaluation research. *American Journal of Evaluation* 2016; 38(2): 246–261. DOI:10.1177/1098214016651489.

8. Sim J. Should treatment effects be estimated in pilot and feasibility studies? *Pilot and Feasibility Studies* 2019; 5(1). DOI:10.1186/s40814-019-0493-7.
9. Teare M, Dimairo M, Shephard N et al. Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials* 2014; 15(1): 264. DOI: 10.1186/1745-6215-15-264. URL `http://www.trialsjournal.com/content/15/1/264`.
10. Cocks K and Torgerson DJ. Sample size calculations for pilot randomized trials: a confidence intervalÂ approach. *Journal of Clinical Epidemiology* 2013; 66(2): 197 – 201. DOI:http://dx.doi.org/10.1016/j.jclinepi.2012.09.002. URL `http://www.sciencedirect.com/science/article/pii/S0895435612002740`.
11. Lee E, Whitehead A, Jacques R et al. The statistical interpretation of pilot trials: should significance thresholds be reconsidered? *BMC Medical Research Methodology* 2014; 14(1): 41. DOI: 10.1186/1471-2288-14-41. URL `http://www.biomedcentral.com/1471-2288/14/41`.
12. Raiffa H and Schlaifer R. *Applied Statistical Decision Theory*. Harvard College, 1961.
13. Keeney RL and Raiffa H. *Decisions with multiple objectives: preferences and value tradeoffs*. John Wiley & Sons, 1976.
14. Lindley DV. The choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)* 1997; 46(2): 129–138. DOI:10.1111/1467-9884.00068. URL `http://dx.doi.org/10.1111/1467-9884.00068`.
15. Hee SW, Hamborg T, Day S et al. Decision-theoretic designs for small trials and pilot studies: A review. *Statistical Methods in Medical Research* 2016; 25(3): 1022–1038. DOI: 10.1177/0962280215588245. URL `http://smm.sagepub.com/content/25/3/1022.abstract`. `http://smm.sagepub.com/content/25/3/1022.full.pdf+html`.
16. Joseph L and Wolfson DB. Interval-based versus decision theoretic criteria for the choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)* 1997; 46(2): 145–149. DOI:10.1111/1467-9884.00070. URL `http://dx.doi.org/10.1111/1467-9884.00070`.
17. Pearce M, Hee SW, Madan J et al. Value of information methods to design a clinical trial in a small population to optimise a health economic utility function. *BMC Medical Research Methodology* 2018; 18(1): 20. DOI:10.1186/s12874-018-0475-0. URL `https://doi.org/10.1186/s12874-018-0475-0`.
18. Blocker AW. *fastGHQuad: Fast 'Rcpp' Implementation of Gauss-Hermite Quadrature*, 2018. URL `https://CRAN.R-project.org/package=fastGHQuad`. R package version 1.0.
19. Byrd RH, Lu P, Nocedal J et al. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* 1995; 16(5): 1190–1208. DOI:10.1137/0916069.
20. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL `https://www.R-project.org/`.
21. Nash JC and Varadhan R. Unifying optimization algorithms to aid software system users:optimxforR. *Journal of Statistical Software* 2011; 43(9). DOI:10.18637/jss.v043.i09.
22. Walwyn REA, Russell AM, Bryant LD et al. Supported self-management for adults with type 2 diabetes and a learning disability (OK-diabetes): study protocol for a randomised controlled feasibility trial. *Trials* 2015; 16(1). DOI:10.1186/s13063-015-0832-9. URL `http://dx.doi.org/10.1186/s13063-015-0832-9`.
23. House A, Ajjan R, Bryant L et al. Managing with learning disability and diabetes. URL `http://www.nets.nihr.ac.uk/projects/hta/1010203`. Accesed 6th October 2014.

24. Willan AR and Pinto EM. The value of information and optimal clinical trial design. *Statistics in Medicine* 2005; 24(12): 1791–1806. DOI:10.1002/sim.2069. URL http://dx.doi.org/10.1002/sim.2069.
25. Stallard N. Optimal sample sizes for phase II clinical trials and pilot studies. *Statistics in Medicine* 2012; 31(11-12): 1031–1042. DOI:10.1002/sim.4357. URL http://dx.doi.org/10.1002/sim.4357.
26. Genz A, Bretz F, Miwa T et al. *mvtnorm: Multivariate Normal and t Distributions*, 2017. URL https://CRAN.R-project.org/package=mvtnorm. R package version 1.0-6.
27. Narasimhan B, Johnson SG, Hahn T et al. *cubature: Adaptive Multivariate Integration over Hypercubes*, 2018. URL https://CRAN.R-project.org/package=cubature. R package version 2.0.3.
28. Gittins J and Pezeshk H. A behavioral bayes method for determining the size of a clinical trial. *Drug Information Journal* 2000; 34(2): 355–363. DOI:10.1177/009286150003400204. URL http://dij.sagepub.com/content/34/2/355.abstract. http://dij.sagepub.com/content/34/2/355.full.pdf+html.
29. Kikuchi T and Gittins J. A behavioral bayes method to determine the sample size of a clinical trial considering efficacy and safety. *Statist Med* 2009; 28(18): 2293–2306. URL http://dx.doi.org/10.1002/sim.3630.
30. Hee SW and Stallard N. Designing a series of decision-theoretic phase II trials in a small population. *Statistics in Medicine* 2012; 31(30): 4337–4351. DOI:10.1002/sim.5573. URL http://dx.doi.org/10.1002/sim.5573.
31. Strong M, Oakley JE, Brennan A et al. Estimating the expected value of sample information using the probabilistic sensitivity analysis sample: a fast, nonparametric regression-based method. *Medical Decision Making* 2015; 35(5): 570–583. DOI:10.1177/0272989x15575286.
32. Heath A, Manolopoulou I and Baio G. Estimating the expected value of sample information across different sample sizes using moment matching and nonlinear regression. *Medical Decision Making* 2019; 39(4): 346–358. DOI:10.1177/0272989x19837983.
33. Avery KNL, Williamson PR, Gamble C et al. Informing efficient randomised controlled trials: exploration of challenges in developing progression criteria for internal pilot studies. *BMJ Open* 2017; 7(2): e013537. DOI:10.1136/bmjopen-2016-013537. URL https://doi.org/10.1136%2Fbmjopen-2016-013537.
34. Benjamin DJ, Berger JO, Johannesson M et al. Redefine statistical significance. *Nature Human Behaviour* 2017; 2(1): 6–10. DOI:10.1038/s41562-017-0189-z.
35. Amrhein V and Greenland S. Remove, rather than redefine, statistical significance. *Nature Human Behaviour* 2017; 2(1): 4–4. DOI:10.1038/s41562-017-0224-0.
36. Lakens D, Adolfi FG, Albers CJ et al. Justify your alpha. *Nature Human Behaviour* 2018; 2(3): 168–171. DOI:10.1038/s41562-018-0311-x.
37. Amrhein V, Greenland S and McShane B. Scientists rise up against statistical significance. *Nature* 2019; 567(7748): 305–307. DOI:10.1038/d41586-019-00857-9.
38. Bacchetti P. The other arbitrary cutoff. *The American Statistician* 2019; : 1–3DOI:10.1080/00031305.2019.1654920.

Appendix

Expected utility for a single trial

The utility function is $u(\mu, x)$, where μ is the true difference and x is the sample mean difference. Note that

$$x \mid \mu \sim N\left(\mu, \frac{2\sigma^2}{n}\right).$$

Given a normal prior $\mu \sim N(\mu_0, \sigma_0^2)$, the posterior distribution of μ is also normal. Specifically

$$\mu \mid x \sim N\left(\mu_1 = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{xn}{2\sigma^2}\right),\ \sigma_1^2 = 1/\left(\frac{1}{\sigma_0^2} + \frac{n}{2\sigma^2}\right)\right).$$

The marginal distribution of the sample mean difference is

$$x \sim N\left(\mu_0, \sigma_x^2 = \sigma_0^2 + \frac{2\sigma^2}{n}\right).$$

For the case $\rho = 0$, the utility function is the value function. The expected utility is then

$$\begin{aligned} E[u(\mu, x)] =& E_x\left[E_{\mu|x}[u(\mu, x)]\right] \\ =& E_x\left[I\{x > d\}(k_d\mu_1 + k_n n) + I\{x \le d\}(k_n n + k_c)\right] \\ =& k_n n + Pr(x > d)k_d E_{x|x>d}\left[\mu_1\right] + Pr(x \le d)k_c, \end{aligned}$$

The two probabilities follow from the marginal distribution of the sample mean. For example,

$$Pr[x > d] = 1 - \Phi\left(\frac{d - \mu_0}{\sqrt{\sigma_0^2 + \frac{2\sigma^2}{n}}}\right)$$

The expectation is of μ_1, the mean of the posterior distribution on μ given x, with respect to the distribution of x conditional on $x > d$. This is a truncated normal with mean

$$E_{x|x>d}\left[\mu_1\right] = \mu_0 + \sigma_x\phi\left(\frac{d - \mu_0}{\sigma_x}\right) / \left(1 - \Phi\left(\frac{d - \mu_0}{\sigma_x}\right)\right).$$

For the case $\rho > 0$, we again start with the law of total expectation giving $E[u(\mu, x)] = E_x\left[E_{\mu|x}[u(\mu, x)]\right]$. Then,

$$\begin{aligned} E_{\mu|x}[u(\mu, x)] &= E_{\mu|x}[1 - e^{-\rho(k_d\mu + k_n n)I\{x>d\} - \rho(k_n n + k_c)I\{x<d\}}] \\ &= I\{x > d\}E_{\mu|x}[1 - e^{-\rho(k_d\mu + k_n n)}] + I\{x < d\}E_{\mu|x}[1 - e^{-\rho(k_n n + k_c)}] \end{aligned}$$

For the first term we have:

$$E_{\mu|x}[1 - e^{-\rho(k_d\mu + k_n n)}] = 1 - e^{-\rho k_n n}E_{\mu|x}[e^{-\rho k_d\mu}]$$

This takes the form as a moment generating function: for a normally distributed $Y \sim N(a, b^2)$,

$$E[e^{tY}] = e^{ta + \frac{b^2 t^2}{2}}.$$

This gives us

$$E_{\mu|x}[1 - e^{-\rho(k_d\mu + k_n n)}] = 1 - e^{-\rho k_n n}\left(e^{t\mu_1 + \frac{\sigma_1^2 t^2}{2}}\right),$$

where $t = \rho k_d$. This then all gives an expected utility conditional on x of

$$E_{\mu|x}[u(\mu, x)] = I\{x > d\}\left[1 - e^{-\rho k_n n}\left(e^{t\mu_1 + \frac{\sigma_1^2 t^2}{2}}\right)\right] + I\{x < d\}\left[1 - e^{-\rho(k_n n + k_c)}\right].$$

We now need to take the expectation of this over the marginal distribution for x:

$$\begin{aligned} E[u(\mu, x)] &= E_x\left[E_{\mu|x}[u(\mu, x)]\right] \\ &= E_x\left[I\{x > d\}\left(1 - e^{-\rho k_n n} e^{t\sigma_1^2(\frac{\mu_0}{\sigma_0^2} + \frac{xn}{2\sigma^2}) + \frac{\sigma_1^2 t^2}{2}}\right)\right] + E_x\left[I\{x < d\}\left(1 - e^{-\rho(k_n n + k_c)}\right)\right]. \end{aligned}$$

The first term is equal to

$$Pr[x > d]\left(1 - e^{-\rho k_n n} e^{\frac{\sigma_1^2 t^2}{2}} e^{\frac{t\sigma_1^2 \mu_0}{\sigma_0^2}}\right) E_x\left[e^{\frac{t\sigma_1^2 xn}{2\sigma^2}}\right].$$

For the expectation, we again use a moment generating function. Here, the sample difference x is normally distributed $N(\mu_0, \sigma_x^2)$ but we have conditioned on $x > d$ and so have a truncated normal. The moment generating function for $Y \sim N(a, b^2), Y > d$ is

$$E[e^{rY}] = \left(e^{ar + \frac{b^2 r^2}{2}}\right)\left(\frac{1 - \Phi(\frac{d-a}{b} - br)}{1 - \Phi(\frac{d-a}{b})}\right)$$

So, for our case we have

$$E_x\left[e^{\frac{t\sigma_1^2 xn}{2\sigma^2}}\right] = \left(e^{\mu_0 r + \frac{\sigma_x^2 r^2}{2}}\right)\left(\frac{1 - \Phi(\frac{d-\mu_0}{\sigma_x} - \sigma_x r)}{1 - \Phi(\frac{d-\mu_0}{\sigma_x})}\right),$$

where $r = \frac{t\sigma_1^2 n}{2\sigma^2}$. Finally, putting everything together, we get

$$\begin{aligned} E[u(\mu, x)] =& E_x\left[E_{\mu|x}[u(\mu, x)]\right] \\ =& \left[1 - \Phi\left(\frac{d - \mu_0}{\sqrt{\sigma_0^2 + \frac{2\sigma^2}{n}}}\right)\right]\left[1 - e^{-\rho k_n n} e^{\frac{\sigma_1^2 t^2}{2}} e^{\frac{t\sigma_1^2 \mu_0}{\sigma_0^2}}\left(e^{\mu_0 r + \frac{\sigma_x^2 r^2}{2}}\right)\left(\frac{1 - \Phi(\frac{d-\mu_0}{\sigma_x} - \sigma_x r)}{1 - \Phi(\frac{d-\mu_0}{\sigma_x})}\right)\right] + \\ & \Phi\left(\frac{d - \mu_0}{\sqrt{\sigma_0^2 + \frac{2\sigma^2}{n}}}\right)\left[1 - e^{-\rho(k_n n + k_c)}\right]. \end{aligned}$$

The case $\rho < 0$ follows.

Internal pilot sampling distribution

Recall that we have sample means from both the first and second stage of an internal pilot design, x_1, x_2, with the combined sample mean $x_t = \frac{n_1 x_1}{n_1+n_2} + \frac{n_2 x_2}{n_1+n_2}$ being used at the final testing stage. The distribution of the combined mean, conditional on μ is $x_t \mid \mu \sim N\left(\mu, \frac{2\sigma^2}{n_1+n_2}\right)$. To fully specify the joint distribution of (x_1, x_t) we require the covariance:

$$cov(x_1, x_t) = E[x_1 x_t] - E[x_1]E[x_t].$$

By the law of total expectation,

$$\begin{aligned}
E[x_1 x_t] &= E_{x_1}(E_{x_1 x_t|x_1}[x_1 x_t]) \\
&= E_{x_1}(x_1 E_{x_t|x_1}[x_t]) \\
&= E_{x_1}\left(x_1 E_{x_t|x_1}\left[\frac{n_1 x_1}{n_1+n_2} + \frac{n_2 x_2}{n_1+n_2}\right]\right) \\
&= E_{x_1}\left(\frac{n_1 x_1^2}{n_1+n_2} + \frac{n_2 x_1 \mu}{n_1+n_2}\right) \\
&= \frac{n_1}{n_1+n_2}E[x_1^2] + \frac{n_2}{n_1+n_2}E[x_1]\mu \\
&= \frac{n_1}{n_1+n_2}(var(x_1) + E[x_1]^2) + \frac{n_2}{n_1+n_2}\mu^2 \\
&= \frac{n_1}{n_1+n_2}\left(\frac{2\sigma^2}{n_1} + \mu^2\right) + \frac{n_2}{n_1+n_2}\mu^2.
\end{aligned}$$

Then,

$$\begin{aligned}
cov(x_1, x_t) &= \frac{n_1}{n_1+n_2}\left(\frac{2\sigma^2}{n_1} + \mu^2\right) + \frac{n_2}{n_1+n_2}\mu^2 - \mu^2 \\
&= \frac{2\sigma^2}{n_1+n_2} + \frac{n_1\mu^2}{n_1+n_2} + \frac{n_2\mu^2}{n_1+n_2} - \mu^2 \\
&= \frac{2\sigma^2}{n_1+n_2}.
\end{aligned}$$

Partial derivatives of expected utility

Denoting $Pr[G_i = 1 \mid \mu] = Pr[x_i > d_i 1 \mid \mu] = g(n_i, d_i)$ and $Pr[G_i = 0 \mid \mu] = Pr[x_i \leq d_i 1 \mid \mu] = \bar{g}(n_i, d_i)$ we can re-write the conditional expected utility of equation 4 as

$$\begin{aligned}
f(n_1, n_2, d_1, d_2) =& g(n_1, d_1)g(n_2, d_2) - g(n_1, d_1)g(n_2, d_2)\exp(-\rho(k_d\mu + k_n(n_1+n_2)))+ \\
& g(n_1, d_1)\bar{g}(n_2, d_2) - g(n_1, d_1)\bar{g}(n_2, d_2)\exp(-\rho(k_n(n_1+n_2) + k_c))+ \\
& \bar{g}(n_1, d_1) - \bar{g}(n_1, d_1)\exp(-\rho(k_n n_1 + k_c)).
\end{aligned}$$

Then, using the partial derivatives

$$\frac{d}{dn_i}g(n_i, d_i) = \frac{d}{dn_i}\left[1 - \Phi\left(\frac{d_i - \mu}{\sqrt{2\sigma^2/n_i}}\right)\right] = -\phi\left(\frac{d_i - \mu}{\sqrt{2\sigma^2/n_i}}\right)\frac{d_i - \mu}{2\sqrt{2\sigma^2/n_i}}$$

and

$$\frac{d}{dd_i}g(n_i, d_i) = \frac{d}{dd_i}\left[1 - \Phi\left(\frac{d_i - \mu}{\sqrt{2\sigma^2/n_i}}\right)\right] = -\phi\left(\frac{d_i - \mu}{\sqrt{2\sigma^2/n_i}}\right)\frac{1}{\sqrt{2\sigma^2/n_i}},$$

we can calculate the partial derivatives of the integrand in equation 5. Given the quadrature approximation being used, the partial derivatives of the expected utility can then be calculated as the weighted sum of derivatives at the quadrature points. Full details can be found in the supplementary material.