A Bayesian design for pilot trials for complex interventions

June 5, 2018

Abstract

Background: External pilot trials of complex interventions are used to help determine if and how a confirmatory trial should be undertaken, providing estimates of several parameters relating to trial feasibility and to the quality of the intervention. The decision to progress to the confirmatory trial is typically made by comparing these estimates to pre-specified thresholds known as progression criteria. However, the statistical properties of pilot designs and progression criteria are rarely assessed, leading to a poor understanding of the risks of making a poor decision.
Methods: We describe a Bayesian approach to analysing pilot data, making progression decisions, and evaluating the resulting statistical properties of proposed pilot designs. Decisions are based on minimising the expected value of a loss function. By defining loss over the whole parameter space we allow for preferences and trade-offs between multiple parameters to be articulated and used in the decision making process. The assessment of preferences is kept feasible by using a simple piecewise constant parametrisation of the loss function, the parameters of which are chosen to lead to desirable operating characteristics. We describe a flexible, yet computationally intensive, nested Monte Carlo algorithm for estimating operating characteristics.
Illustration: The method is applied to an example pilot trial of a complex intervention in the care home setting, involving multiple parameters which inform the optimal progression decision. We show that reasonable operating characteristics of the pilot trial can be obtained, even when external information is ignored in the analysis through the use of weakly informative prior distributions.
Conclusion: A Bayesian approach to design and analysis can provide a way for preferences and trade-offs between the multiple parameters assessed in pilot trials to be articulated and used to make better progression decisions. Proposed pilot designs can be evaluated in terms of several operating characteristics, although the computation time required to do so may be restrictive in practice.

1 Introduction

Randomised clinical trials (RCTs) of complex interventions can be compromised by facors such as slow patient recruitment, poor levels of adherence to the intervention or trail protocol, and low completeness of follow-up data. To anticipate these problems we often conduct small pilot trials prior to the main RCT [9]. These trials typically take the same form as the planned RCT but with a significantly lower sample size [15]. If there is a seamless gap between the pilot and the main RCT, with all data being pooled and used in the final analysis, they are known as internal pilots. External pilots, in contrast, are carried out separately to the main RCT using a separate protocol.

Pilot trials use progression criteria to decide if the main RCT should go ahead, and if so, whether the intervention or the trial design should be adjusted to ensure success. In the UK, the the National Institute for Health Research ask that these criteria are pre-specified and included in the research plan [29]. A single pilot trial can collect data on several progression criteria, often focused on the aforementioned areas of recruitment, protocol adherence, and data collection [1]. Reporting these criteria is required by the recent CONSORT extension to randomised pilot trials, which notes that investigators increasingly "use a traffic light system for criteria used to judge feasibility, whereby measures (eg, recruitment rates) below a lower threshold indicate that the trial is not feasible, above a higher threshold that it is feasible, and between the two that it might be feasible if appropriate changes can be made" [13]. The traffic light reference relates to the labelling of these three possible decisions as red, amber and green.

Parameter estimates from pilots can be subject to considerable error [13], and so it may be that a certain progression criteria is met, or missed, due to sampling variation alone. We should, therefore, evaluate the statistical properties of any proposed pilot design and progression criteria before we implement them. However, despite the consequences of making incorrect progression decisions, such a statistical approach is not common practice in pilot trials. This may be due to the methodological challenges commonly found in pilot trials of complex interventions, including the simultaneous evaluation of multiple endpoints, complex multi-level models, small sample sizes, and prior uncertainty in nuisnace parameters [47].

In this paper we will describe a method for evaluating proposed pilot trial designs which addresses these challenges. We take a Bayesian view, where progression decisions are made to minimise the expected value of a loss function. In contrast to a fully decision-theoretic approach such as that outlined in [28], we do not consider the direct elicitation of a loss function. Instead, we define a simple parametric representation of it and show how the parameters can be chosen by considering the operating characteristics they lead to. The operating characteristics we propose are all unconditional probabilities (with respect to a prior distribution) of making incorrect decisions, also known as assurances [32]. Using assurances rather than the analogous frequentist error rates brings several benefits, including the ability to make use of existing knowledge whilst allowing for any uncertainty, and a more natural interpretation [10]. As we will show, assurances are also useful when our preferences for different end-of-trial decisions are based on several attributes in a complex way that involves trading off some against others.

The remainder of this paper is organised as follows. A motivating example of a pilot trial in a care home setting is introduced in Section 2. In Section 3 we describe the general framework for pilot design and analysis, some operating characteristics used for evaluation, and a routine for optimising the design. We return to the example in Section 4 to illustrate the method's application, before discussing implications and limitations in Section 5.

2 Motivating example

The REACH (Research Exploring Physical Activity in Care Homes) trial aims to inform the feasibility and design of a future definitive RCT assessing a complex intervention designed to increase the physical activity of care home residents [16]. The trial is cluster randomised due to the whole-home nature of the intervention, which is designed to assist care-home staff to make step-by-step changes in their approach to working with residents. Twelve care homes are randomised equally to treatment as usual (TaU) or the intervention plus TaU.

The trial will collect data on a number of measures relating to the feasibility of a future RCT, four of which we focus on here: Recruitment (measured in terms of the average number of residents in each care home who participate in the trial); adherence (a binary indicator at the care home level indicating if the intervention was fully implemented); data completion (a binary indicator for each resident of successful follow-up at the planned primary outcome time of 12 months); and efficacy (a continuous measure of physical activity at the resident level). Each of these four attributes are to be considered together when deciding if and how a confirmatory trial should be undertaken. The physical activity measure is assumed to be normally distributed with equal variance in each arm, and we expect clustering effects due to the intervention being delivered at the care home level. No precise estimates of either the total variance or the intra-cluster correlation coefficient are available, although studies in similar areas do provide some indication of likely values. The pilot trial will lead to small samples of data pertaining to both recruitment and delivery, as each is measured at the care home level. In particular, successful delivery can only be assessed at the six care homes in the intervention arm of the trial.

Given the proposed pilot design and sample size, we require a method specifying how progression decisions are to be made once the data has been obtained. Given this, we would like to understand the resulting statistical properties of the trial, in terms of how likely we are to make incorrect decisions, and to judge whether or not the pilot will be sufficiently reliable.

3 Methods

3.1 Analysis and progression decisions

Consider a pilot trial which will produce data x according to model $p(x|\theta)$. We decompose the parameters into $\theta = (\phi, \psi)$, where ϕ denotes the parameters of substantive interest and ψ the nuisance parameters. We follow [44] and assume that two joint prior distributions of θ have been specified, one each for the design and analysis stages of the trial. The first, denoted $p_D(\theta)$, is a completely subjective prior which fully expresses our current knowledge and uncertainty in the parameters. This will be used when evaluating the design of the pilot. The second prior, $p_A(\theta)$, will be used when analysing the pilot data. Although it may be that $p_A(\theta) = p_D(\theta)$, separating the two will let us use a less informative prior in the analysis.

After observing the pilot data x, we must decide whether or not to progress to the main RCT. We consider three possible actions following the aforementioned 'traffic light' system commonly used in pilot trials:

- $r$ed - discard the intervention and stop all future development or evaluation;
- $a$mber - proceed to the main RCT, but only after some modifications to the intervention, the planned trial design, or both, have been made; or
- $g$reen - proceed immediately to the main RCT.

We assume that our preferences between the three possible decisions are influenced by ϕ but independent of ψ, formalising the separation of θ into substantive and nuisance components. We also assume that for any value of ϕ there will be a decision which is considered preferable to the others. That is, there are no parameter values for which we are completely indifferent between the possible decisions. Under this assumption we can partition the substantive parameter space Φ into three subsets where each decision is optimal. We denote these subsets by $\Phi_I = \{\phi \in \Phi \mid i \succ j \ \forall \ j \neq i \in \mathcal{D}\}$, for $I = R, A, G$. We will henceforth refer to these three subsets as *hypotheses*. Throughout, we will distinguish hypothesis I from the corresponding optimal decision i by using capital and lower case letters respectively.

When $\phi \in \Phi_I$ and we choose a decisions $j \neq i$, there will be negative consequences. In particular, we may: proceed to a futile main RCT; make unnecessary adjustments to the intervention or trial design; or discard a promising intervention. We assume that for each of these consequences there is a fixed cost, denoted c_1, c_2 and c_3 respectively. The total costs associated with making decision j under hypothesis I are given by a piecewise constant loss function $L(d, \Phi_I) : \{r, a, g\} \times \{\Phi_R, \Phi_R, \Phi_R\} \rightarrow [0, 1]$ which takes values as given in Table 1. For example, making decision a when $\phi \in \Phi_R$ will lead to both a futile trial and unnecessary adjustments, and so will incur a cost $c_1 + c_2$.

Given a loss function with cost parameters $\mathbf{c} = (c_1, c_2, c_3)$, we follow the principle of maximising expected utility (or in our case, minimising the expected

		Hypothesis		
		Φ_R	Φ_A	Φ_G
Decision	r	0	c_3	c_3
	a	$c_1 + c_2$	0	c_2
	g	c_1	$c_1 + c_3$	0

Table 1: Costs of decisions under hypotheses.

loss) when making a decision. To do so, we first use the pilot data in conjugation with the analysis prior $p_A(\theta)$ to obtain a (marginal) posterior $p(\phi \mid x)$, and then choose the decision i^* such that

$$i^* = \underset{i \in \{r,a,g\}}{\arg\min}\, \mathbb{E}_{\phi|x}[L(i, \phi)] \tag{1}$$

$$= \underset{i \in \{r,a,g\}}{\arg\min} \int L(i, \phi) p(\phi|x) d\phi. \tag{2}$$

We can simplify this expression by noting that, given the piecewise constant nature of the loss function, the expected loss of each decision depends only on the posterior probabilities $p_I = Pr[\phi \in \Phi_I \mid x]$ for $I = R, A, G$ and the costs $\mathbf{c}$:

$$\mathbb{E}_{\phi|x}[L(r, \phi)] = p_A c_3 + p_G c_3, \tag{3}$$

$$\mathbb{E}_{\phi|x}[L(a, \phi)] = p_R c_1 + p_R c_2 + p_G c_2, \tag{4}$$

$$\mathbb{E}_{\phi|x}[L(g, \phi)] = p_R c_1 + p_A c_1 + p_A c_3. \tag{5}$$

Monte Carlo (MC) estimates of the posterior probabilities p_I can be computed based on the samples from the joint posterior distribution generated by an MCMC analysis of the pilot data. Specifically, given M samples $\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(M)} \sim p(\phi \mid x)$,

$$p_I \approx \frac{1}{M} \sum_{k=1}^{M} \mathbb{I}(\phi^{(k)} \in \Phi_I), \tag{6}$$

where $\mathbb{I}(.)$ is the indicator function.

3.2 Operating characteristics

Defining a loss function and following the steps of the preceding section effectively prescribes a decision rule mapping the pilot data sample space $\mathcal{X}$ to the decision space $\{r, a, g\}$. To gain some insight at the design stage into the properties of this rule, we propose to calculate some trial operating characteristics. These take the form of unconditional probabilities of making a mistake when following the rule, calculated with respect to the design prior $p_D(\theta)$. We consider the following:

- OC_1 - probability of proceeding to a *futile main RCT*;

- OC_2 - probability of making *unnecessary adjustments* to the intervention or the trial design;
- OC_3 - probability of *discarding a promising intervention.*

These operating characteristics can be estimated using simulation. First, we draw N samples $(\theta^{(1)}, x^{(1)}), (\theta^{(2)}, x^{(2)}), \dots, (\theta^{(N)}, x^{(N)})$ from the joint distribution $p(\theta, x) = p(x|\theta)p_D(\theta)$. For each data set we then apply the analysis and decision making procedure described in Section 3.1, using some cost vector $\mathbf{c}$ to parametrise the loss function. This results in N decisions $i^{(k)}$ which can be contrasted with the corresponding true parameter value $\theta^{(k)}$, noting if any of the three types of errors had been made. MC estimates of the operating characteristics can then be calculated as the proportion of occurrences of each type of error in the N simulated cases.

Assuming that N is large, the unbiased MC estimate of an operating characteristic with true probability p will be approximately normally distributed with variance $p(1-p)/N$. This assumes that, for each data set, the analysis that is simulated corresponds exactly to the analysis that would be carried out in practice. In particular, we assume that exactly M posterior samples will be generated by the MCMC algorithm and that the same number will be used to seed the random number generator.

3.3 Optimisation

In some cases the cost parameters $\mathbf{c} = (c_1, c_2, c_3)$ could be chosen so as to accurately reflect the preferences, under uncertainty, of the decision maker. However, the elicitation of these preferences can be difficult [24], and so we propose an alternative approach here where we try to optimise the value of $\mathbf{c}$ with respect to the operating characteristics it results in (assuming the pilot design is fixed). Thinking of operating characteristics as functions of costs, we wish to solve the multi-objective optimisation problem

$$\min_{\mathbf{c} \in \mathcal{C}} \; OC_1(\mathbf{c}),\; OC_2(\mathbf{c}),\; OC_3(\mathbf{c}), \tag{7}$$

where $\mathcal{C} = \{c_1, c_2 \in [0,1] \mid c_1 + c_2 \leq 1\}$. This restriction of the search space is possible because a loss function is strategically equivalent to any linear transformation of itself, allowing the costs parameters to be scaled to $c_1 + c_2 + c_3 = 1$.

The three operating characteristics are in conflict in the sense that a decrease in one will generally be accompanied by an increase in another. We would therefore like to find a set of costs $\mathcal{C}^* = \{\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \dots, \mathbf{c}^{(K)}\}$ such that each one provides a different balance between minimising the three operating characteristics. If there exist $\mathbf{c}, \mathbf{c}' \in \mathcal{C}^*$ such that $OC_i(\mathbf{c}') \leq OC_i(\mathbf{c})$ for all $i \in \{1,2,3\}$ and $OC_i(\mathbf{c}') < OC_i(\mathbf{c})$ for some $i \in \{1,2,3\}$, we say that $\mathbf{c}'$ dominates $\mathbf{c}$ and can discard the latter from $\mathcal{C}^*$. Because the search space $\mathcal{C}$ has only two dimensions, problem (7) can be solved simply by generating a random sample of costs and estimating the operating characteristics for each. Any costs which are dominated in this set can then be discarded, and the operating characteristics of the remaining costs illustrated graphically. The decision maker(s) can then view the range of available options, all providing different trade-offs amongst the three operating characteristics, and choose from amongst them.

To solve the problem in a timely manner we must be able to estimate operating characteristics quickly. Noting from equation (3) that the expected loss of each decision depends only on the costs and the posterior probabilities p_R, p_A and p_G, we first generate N samples of these posterior probabilities and then use this same set of samples for every evaluation. This approach not only ensures that optimisation is computationally feasible, but also means that differences in operating characteristics are entirely due to differences in costs, as opposed to differences in the random posterior probability samples.

4 Illustration

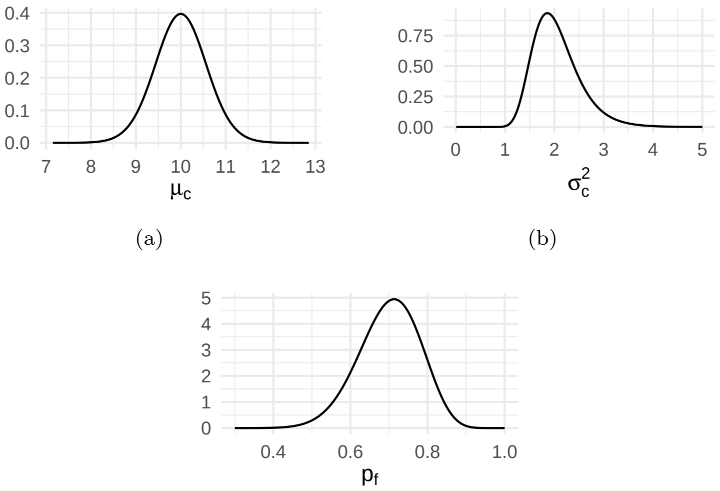
4.1 Model specification

The four substantive parameters of the REACH trial can be divided into two pairs: those relating to the amount of information which a confirmatory trial will gather, (mean cluster size and follow-up rate); and those relating to the effectiveness of the intervention, (adherence and efficacy). Although we assume that the parameters are *probabilistically* independent, we do not assume they are *preferentially* independent. Rather, we expect that a degree of trade-off between the mean cluster size and the follow-up rate will be acceptable, with a decrease in one being compensated by an increase in the other. Likewise, low adherence could be compensated to some extent by improved efficacy, and vice versa.

While there may be trade-offs within these pairs of parameters, we do not expect trade-offs between them. If the effectiveness of the intervention is very low the confirmatory trial will be futile and should not be conducted, regardless of how much information it would be able to gather. Similarly, a confirmatory trial should not be conducted if it is highly unlikely to produce enough information for the research question to be adequately answered. We therefore consider the sub-spaces of Φ formed by these parameter pairs, partition these into hypotheses, and combine these together. Constructing hypotheses in these two-dimensional spaces is cognitively simpler than working in the original four dimensional space, and can be illustrated graphically.

Formally, let Φ^i be the sub-space of mean cluster size and follow-up rate, and Φ^e be that of adherence and efficacy. Having specified hypotheses Φ^i_I, Φ^e_I for $I = R, A, G$, we then have

$$\phi \in \begin{cases} \Phi_R \text{ if } \phi^i \in \Phi^i_R \text{ or } \phi^e \in \Phi^e_R \\ \Phi_G \text{ if } \phi^i \in \Phi^i_G \text{ and } \phi^e \in \Phi^e_G \\ \Phi_A \text{ otherwise.} \end{cases} \tag{8}$$



(a)

(b)

(c)

Figure 1: Prior distributions for the mean and variance of cluster size and the probability of successful follow-up.

4.1.1 Follow-up and cluster size

We assume cluster sizes are normally distributed, $m_i \sim N(\mu_c, \sigma^2),\ i = 1 \dots k$. A normal-inverse-gamma prior

$$\sigma^2 \sim \Gamma^{-1}(\alpha_0, \beta_0),\ \mu_c \sim N(\mu_0, \sigma^2/\nu_0) \tag{9}$$

is placed on the mean and variance to allow for prior uncertainty in both parameters. We set hyper-parameters to $\mu_0 = 10, \nu_0 = 6, \alpha_0 = 20, \beta_0 = 39$, resulting in the marginal prior distributions illustrated in Figure 1.

We assume that follow up rates are constant across clusters. The number of participants followed-up, f is assumed to follow a binomial distribution, $f \sim Bin(\sum_{i=1}^{k} m_i, p_f)$. We take a Beta distribution as the prior for p_f, with hyper-parameters $\alpha_0 = 22.4, \beta_0 = 9.6$, illustrated in Figure 1.

To partition the parameter space into hypotheses, we first consider the case where $p_f = 1$. Conditional on this, we reason that a mean cluster size of below 5 should lead to decision r, whereas a size of above 7 should lead to decision g. As the probability of successful follow-up decreases, we suppose that this can be compensated for by an increase in mean cluster size. We assume the nature of this trade-off is linear and decide that if p_f were reduced to 0.8, we would want to have a mean cluster size of at least 8 to consider decisions a or g. We further decide that a follow-up rate of less than $p_f = 0.6$ would be critically low, regardless of the mean cluster size, and should always lead to decision r. Similarly, a follow-up rate of $0.6 \leq p_f < 0.66$ should lead to modification of the intervention or trial design. Together, these conditions lead to the following

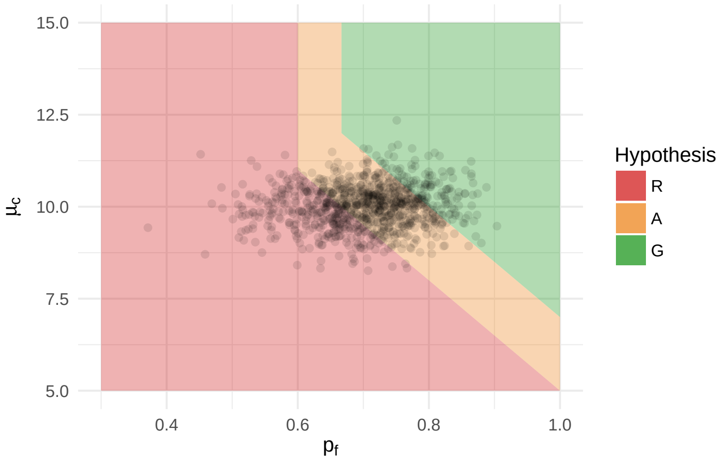


Figure 2: Marginal hypotheses over parameters for follow-u, p_f, and mean cluster size, μ_c. Each point is a sample from the joint prior distribution.

partitioning of the parameter space:

$$(p_f, \mu_c) \in \begin{cases} \Phi^i_R \text{ if } p_f < 0.6 \text{ or } 20 - 15p_f > \mu_c \\ \Phi^i_G \text{ if } p_f > 0.66 \text{ and } 22 - 15p_f < \mu_c \\ \Phi^i_A \text{ otherwise.} \end{cases} \tag{10}$$

The hypotheses are illustrated in Figure 2. A sample of size 1000 from the joint marginal design prior is also plotted in Figure 2, falling into hypotheses Φ^i_R, Φ^i_A and Φ^i_G in proportions 0.354, 0.517, 0.129 respectively.

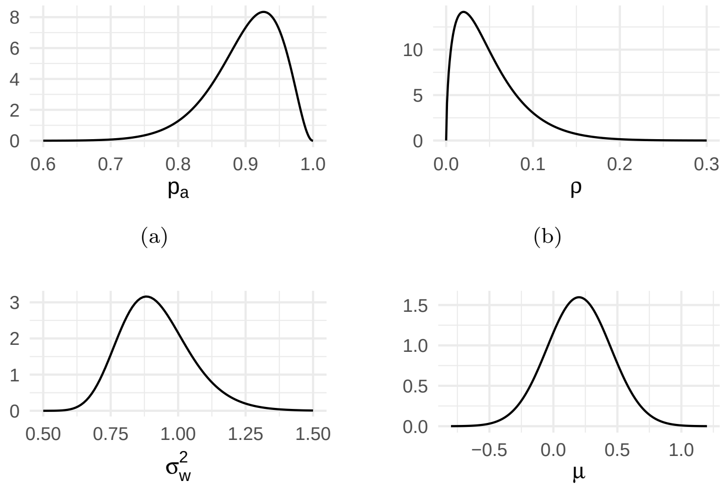
4.1.2 Adherence and efficacy

The number of care homes which successfully adhere to the intervention delivery plan is assumed to be binomially distributed with probability p_a. We assume that adherence is absolute in the sense that all residents in a care home which does not successfully deliver the intervention will not receive any of the treatment effect. We place a Beta prior on p_a, with hyper-parameters $\alpha = 28.8$ and $\beta = 3.2$, as illustrated in Figure 3.

The continuous measure of physical activity is expected to be correlated within care homes. We model this using a random intercept, where the outcome $y_{i,j}$ of resident i in care home j is

$$y_{i,j} = \beta^t_j \times \beta^a_j \times \mu + u_j + e_i. \tag{11}$$

Here, β^t_j is a binary indicator of care home j being randomised to the intervention arm, β^a_j is a binary indicator of care home j succesfully adhering to the intervention, $u_j \sim \mathcal{N}(0, \sigma^2_B)$ is the random effect for care home j and $e_i \sim \mathcal{N}(0, \sigma^2_W)$



(a)

(b)

(c)

(d)

Figure 3: Prior distributions for the probability of adherence, and the mean, total variance, and ICC for efficacy.

is the random effect for resident i. We parametrise the model using the intracluster correlation coefficient $\rho = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$, and place priors on μ, ρ, and σ_W^2 in the manner suggested in [35]. Specifically, we choose

$$\mu \sim N(0.2, 0.25^2) \tag{12}$$

$$\sigma_W^2 \sim \Gamma^{-1}(50, 45) \tag{13}$$

$$\rho \sim Beta(1.6, 30.4) \tag{14}$$

These prior distributions are illustrated in Figure 3.

We consider that while there is potential for adherence to be improved after the pilot, there will be little opportunity to improve the efficacy of the intervention. Moreover, we believe an absolute improvement in delivery of up to around 0.1 is feasible. To define the hypotheses in this subspace we first set a minimal level of efficacy to be 0.1, and decide that we would be happy to make decision g at this point if and only if adherence is perfect. As p_a reduces from 1, a corresponding linear increase in efficacy is considered to maintain the overall effectiveness of the intervention. The rate of substitution for this trade-off is determined to be approximately 0.57 units of efficacy per unit of adherence probability. We consider an absolute lower limit in adherence of $p_a = 0.5$, below which we will always consider decision r to be optimal. Taking these considerations together, the marginal hypotheses are defined as

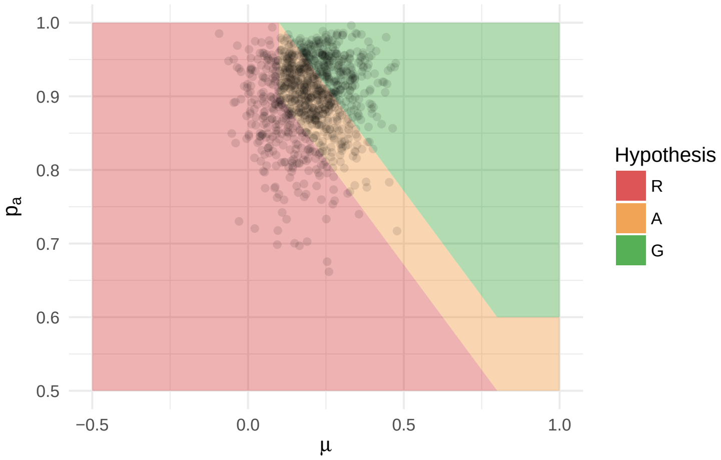


Figure 4: Marginal hypotheses over parameters for adherence, p_a, and efficacy, μ. Each point is a sample from the joint prior distribution.

$$(p_a, \mu) \in \begin{cases} \Phi_R^e \text{ if } p_a < 0.5 \text{ or } 0.96 - 0.57\mu > p_a \\ \Phi_G^e \text{ if } p_a > 0.6 \text{ and } 1.06 - 0.57\mu < p_a \\ \Phi_A^e \text{ otherwise.} \end{cases} \tag{15}$$

The hypotheses are illustrated in Figure 4. Again, a sample of size 1000 from the joint marginal prior distribution $p(p_a, \mu)$ is also plotted, falling into hypotheses Φ_R^e, Φ_A^e and Φ_G^e in proportions 0.234, 0.470, 0.296 respectively.

The marginal hypotheses are then combined together using equation (8). Considering the same 1000 samples form the design prior plotted in Figures 2 and 4, these now fall into the regions Φ_R, Φ_A and Φ_G in proportions 0.507, 0.458, and 0.035 respectively. Note that the prior probabilities of these overall hypotheses are quite different to those of the marginal hypotheses. In particular, there is a considerable increase in the probability that a $r$ed decision will be optimal, and a considerable decrease that decision g will be.

4.2 Evaluation and optimisation

4.2.1 Weakly informative analysis

We applied the proposed method assuming that a weakly informative joint prior distribution will be used at the analysis stage[1]. We took the sample size of the trial to be $k = 12$ clusters, randomised equally between arms. For calculating operating characteristics we generated $N = 10^4$ samples from the joint distribution $p(\theta, x) = p(x|\theta)p_D(\theta)$. We analysed each sample using Stan (via rstan),

[1] Full details of the weakly informative prior are given in the appendix.

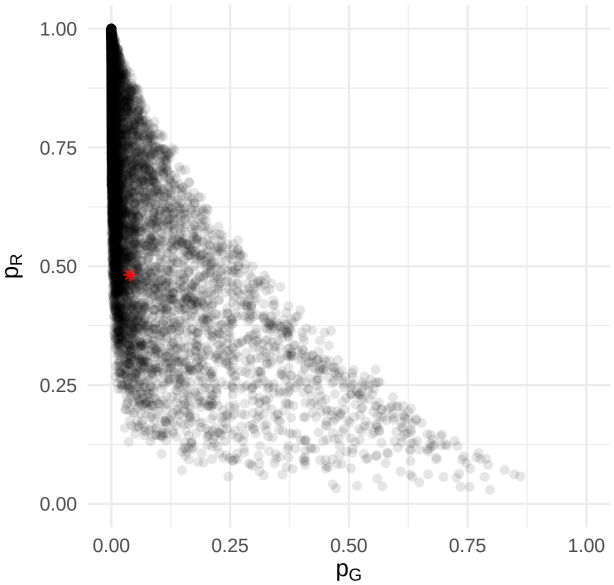


Figure 5: Posterior probabilities (p_R, p_G) of $N = 10^4$ simulated pilot data sets. The prior probability taken from the design prior p_D is shown in red.

generating 5000 samples in four chains and discarding the first 2500 samples in each to allow for burn-in, leading to $M = 10^4$ posterior samples in total.

For each of the N samples in the outer loop we extracted the estimated posterior probabilities p_R and p_G, plotted in Figure 5. We observe a large mass of points with very low values of p_G and high values of p_R. These can be contrasted with the prior probability of each hypothesis, with respect to the design prior $p_D(\theta)$. The majority of simulated data sets lead to a movement from the prior point to a larger value of p_R, suggesting the analysis is being conservative.

We evaluated the operating characteristics for a sample of cost parameters (c_1, c_2) as described in Section 3.3. A total of 249 costs were evaluated, of which 193 were non-dominated. The operating characteristics of these non-dominated costs are shown in Figure 6. As may be expected, the three operating characteristics are highly correlated. In particular, changing the cost vector to give a lower value of OC_3 tends lead to a reduction in OC_2. When selecting a cost vector, the key decision appears to be trading off the probability of a futile trial, OC_1, against the probability of discarding a promising intervention, OC_3. There is a very limited opportunity to minimise the probability of unnecessary adjustments, OC_2, at the expense of these. For example, compare points b and c in Figure 5, details of which are given in Table 2. We see that point c reduces OC_2 by 0.039 in comparison to point b, but only at the expense of increase in OC_1 and OC_3 of 0.147 and 0.055 respectively.

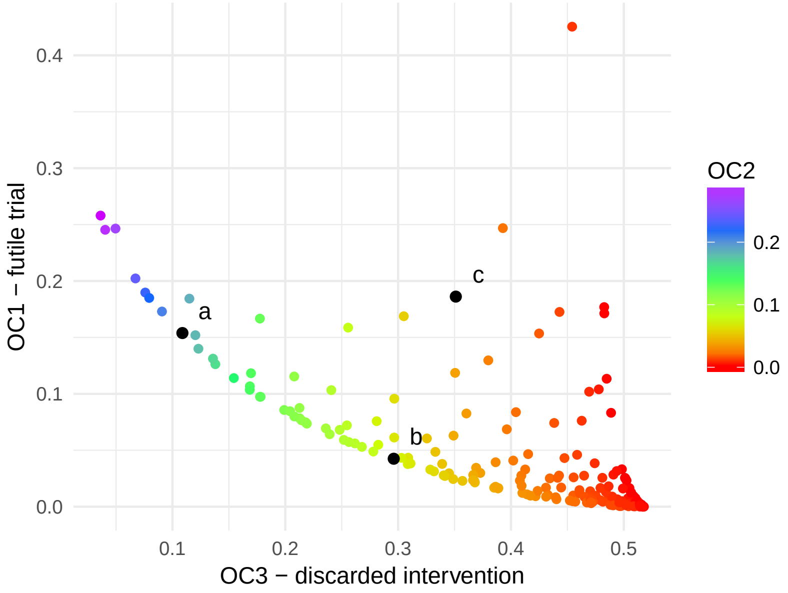


Figure 6: Operating characteristics of the example pilot trial for a range of cost vectors, when a weakly informative analysis prior is used.

Label	(c_1, c_2, c_3)	OC_1	OC_2	OC_3
a	(0.10, 0.07, 0.83)	0.154 (0.004)	0.192 (0.004)	0.109 (0.003)
b	(0.44, 0.01, 0.55)	0.043 (0.002)	0.073 (0.003)	0.296 (0.005)
c	(0.15, 0.76, 0.09)	0.19 (0.004)	0.034 (0.002)	0.351 (0.005)

Table 2: Estimated operating characteristics (with standard errors) of the example pilot trial for the three cost vectors highlighted in Figure 6, when a weakly informative analysis prior is used. Costs have been rounded to 2 decimal places; operating characteristics and their errors to 3.

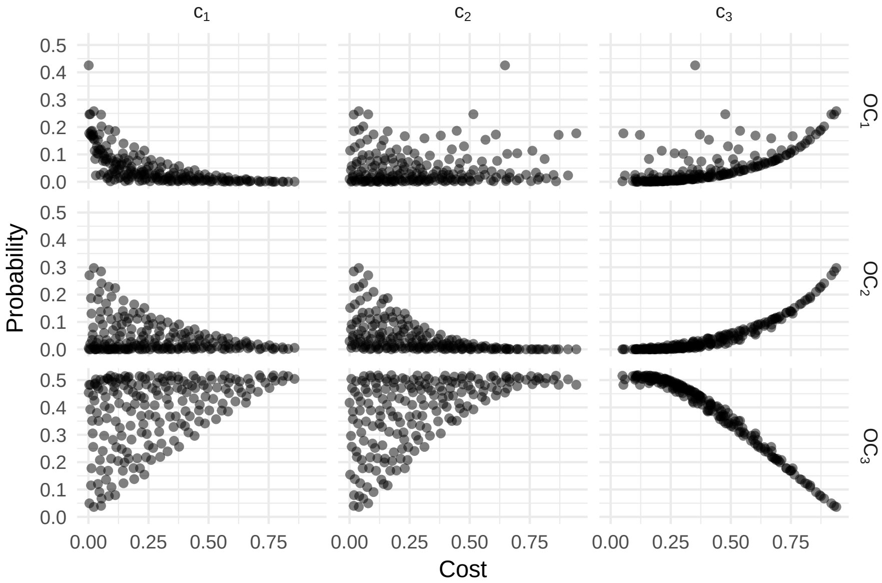


Figure 7: Relationships between the three cost parameters (x axes) and resulting operating characteristics (y axes).

We may expect to see a clear relationship between the value of cost parameters c_1, c_2, c_3 and the operating characteristics they relate to. We explore this in Figure 7 with scatter plots of each cost parameter against each operating characteristic. The results show that there is indeed a strong relationship between the cost assigned to the discarding a promising intervention, c_3, and the probability that this event will occur (see bottom right plot). Moreover, c_3 also seems to be the main determinant of operating characteristics OC_1 and OC_2. The implication is that once the cost $c_3 \in [0, 1]$ has been chosen, the operating characteristics of the trial depend only weakly on the way in which the remaining cost of $1 - c_3$ is allocated to c_1 and c_2. This appears to be due to the fact that, regardless of how errors are weighted, we are much more likely to make the error or discarding a promising intervention than the other types of error (see appendix for further detail). The cost we assign to this error is therefore more influential on the overall operating characteristics than the other costs.

4.2.2 Incorporating subjective priors

Rather than use weakly or non-informative priors when analysing the pilot data, we may instead want to make use of the (subjective) knowledge of parameter values what was elicited and described in the design prior $p_D(\theta)$. Anticipating criticisms of a fully subjective analysis, we can envisage two particular cases where this might be appropriate. Firstly, using the components of the design prior which describe the nuisance parameters ψ while maintaining weakly informative priors on substantive parameters ϕ. Secondly, when very little data on a specific substantive parameter is going to be collected in the pilot, using the informative design prior for that parameter could substantially improve operating characteristics.

We replicated the above analysis for these two scenarios. For the second, we used informative priors for all nuisance parameters and for the probability of adherence, p_a. Recall that this is informed by a binary indicator for the 6 care homes in the intervention arm, and will therefore have very little pilot data bearing on it. For each case we used the same N samples of parameters and pilot data which were used in the weakly informative case, re-doing the Bayesian analysis using the appropriate analysis prior and obtaining estimated posterior probabilities p_R, p_A and p_G as before. These were used in conjunction with the same set of cost vectors $\mathcal{C}$ to obtain corresponding operating characteristics.

For brevity we will refer to the three cases as Weakly Informative (WI), Informative Nuisance (IN), and Informative Nuisance and Adherence (INA). Comparing the operating characteristics of cases WI and IN, we find very little difference. Although the posterior distributions $p(\theta|x)$ are qualitatively different when we examine an individual sample of pilot data, these differences do not translate to the posterior probabilities of hypotheses. Further details are provided in the appendix. When we contrast cases WI and INA, however, there is a clear distinction. As shown in Figure 8, for almost all choices of the cost vector, using the INA analysis prior will lead to larger values of OC_1 and OC_2, while reducing OC_3. Also shown in Figure 8 is the expected loss for both cases. This is consistently lower for the INA analysis than for WI, as we would expect.

5 Discussion

When deciding if and how a definitive RCT of a complex intervention should be conducted, and basing this decision on an analysis of data from a small pilot trial, there is a risk we will inadvertently make a poor choice. A Bayesian analysis of pilot data followed by decision making based on a loss function can help ensure this risk is minimised. The expected results of such a pilot can be evaluated through simulation, producing operating characteristics which help us understand the potential for the pilot to lead to better decision making. These evaluations can in turn be used to find the loss function which leads to the most desirable operating characteristics, avoiding the need for any direct elicitation of decision maker preferences.

Our proposal has been motivated by some salient characteristics of complex intervention pilot trials, and offers several potential benefits over standard pilot trial design and analysis techniques. The Bayesian approach to analysis means that complex multi-level models can be used to describe the data, even when the sample size is small. In contrast to the usual application of independent progression criteria for several parameters of interest, we provide a way for preferential relationships between parameters to be articulated and used when

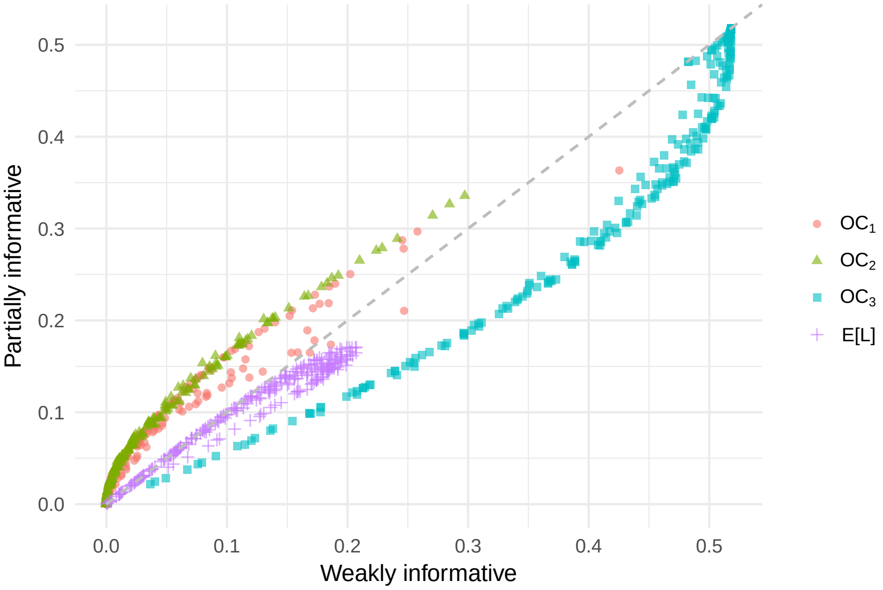


Figure 8: Operating characteristics and expected utilities for weakly (WI) and partially informative (INA) analyses. Each point represents a different cost vector.

making decisions. Using a subjective prior distribution on unknown parameters at the design stage allows both our knowledge and our uncertainty to be fully expressed, meaning we can leverage external information whilst also avoiding decisions which are highly sensitive to imprecise parameter estimates.

The benefits brought by the Bayesian approach must be set against the challenges it brings, particularly in terms of computation time and implementation. In terms of the latter, we are required to specify a joint prior distribution over the parameters θ and a partitioning of the parameter space into the three hypotheses. The specification of the prior distribution may be a challenging and time-consuming task. Although some relevant data relating to similar contexts may be available, expert opinion will still be required to articulate the relevance of such data to the problem at hand. When no data are available, which is not unlikely given the early phase nature of pilot studies, expert opinion will be the only source of information. Although potentially challenging, many examples describing successful practical applications of elicitation for clinical trial design are available [43, 10, 11], as are tools for its conduct such as the Sheffield Elicitation Framework (SHELF) [31]. Dividing the parameter space into three hypotheses may also prove challenging in practice, particularly when trade-offs between more than two parameters are to be elicited. There is a need for methodological research investigating how methods for multi-attribute preference elicitiation, such as those set out in [24], can be applied in this context.

The computational burden of the proposed method is significant, particularly when the model is too complex to allow a simple conjugate analysis to be used when sampling from the posterior distribution. We have used a nested Monte Carlo sampling scheme to estimating operating characteristics, as seen elsewhere [44, 32, 39]. One potential approach to improve efficiency is to use non-parametric regression to predict the expected losses of Equation (3) based on some simulated data, thus bypassing the need to undertake a full MCMC analysis for each of the N samples in the outer loop. This approach has been shown to be successful in the context of expected value of information calculations [37, 38].

We have not considered how to choose the optimal sample size of a pilot trial in this paper. The standard approach is to refer to one of many rules-of-thumb, such as those described in [26, 23, 40], although the shortcomings of such a one-size-fits-all approach have been highlighted [45]. While our framework could be used to assess the operating characteristics of several choices of sample size, the aforementioned computational burden of each assessment will make this difficult in practice. However, the small samples typically available for pilot trials suggests that such optimisation may not be crucial. When it is deemed necessary, methods developed for the design and analysis of expensive computer experiments may provide a way for it to be conducted in an efficient manner [21].

The standard approach to the analysis of pilot trials involves pre-specifying some progression criteria for the parameters of interest, and basing decisions (at least in part) on whether or not parameter estimates satisfy these criteria. This framework specifies a decision rule mapping the pilot data to decisions, as we have done in this paper, but does so implicitly and with no statistical analysis of the resulting properties. An alternative approach would be to employ formal hypothesis tests under a frequentist view. This would have the advantage of not requiring a joint prior distribution to be specified when designing the pilot. However, there has been limited reaseach into how preferential relations between parameters can be incorporated into such testing procedures, typically focussing on the case of two parameters [8, 41]. Moreover, testing multiple parameters can lead to multiplicity. In the case of union tests, type I error rate can be inflated; whilst in the case of intersection tests, type II error rate can be inflated [33]. Hybrid approaches which assume a freqentist analysis but take a Bayesian view to design by averaging type I [7] or type II [34] error rates have been proposed, and may go some way to addressing these issues.

Our proposed design is related to the literature on assurance calculations for clinical trials, applying the idea of using unconditional event probabilities as operating characteristics to the pilot trial setting. In doing so we have defined assurances on multiple substantive parameters and with respect to the 'traffic light' red/amber/green decision structure. The multi-objective optimisation framework we have used to inform trial design avoids setting arbitrary thresholds for operating characteristics as are commonly seen, and criticised [2], with type I and II error rates.

We have defined our procedure in terms of a loss function, where the decision making following the pilot will minimise expected loss. However, the piecewise constant loss function we have proposed may not adequately represent the preferences of the decision maker. For example, we may object to the cost of discarding a promising intervention being independent of how effective the intervention is. The loss function is parametrised with respect to the operating characteristics it produces. While this is a common strategy when implementing decision-theoretic approaches to trial design [27], an alternative is to define the loss function through direct elicitation of the decision makers preferences under uncertainty [17]. This leads to a fully decision-theoretic approach to design and analysis [28]. However, as previously noted by others [22, 3, 46], implementation of these approaches has been limited in practice and may be indicative of the feasibility of this approach.

The proposed method could be extended in several ways. For example, more operating characteristics could be defined and used in design optimisation, more complicated trade-off relationships between multiple parameters could be addressed, or the hypotheses could be expanded to include nuisance parameters which would be used as part of the sample size calculation in the main RCT. A particularly interesting avenue for future research is to consider how to model post-pilot trial actions in more detail. For example, while we allow for the possibility of making an 'amber' decision, indicating that modifications to the intervention or trial design should be made, we do not model what that decision will actually look like. The type of amber adjustments needed are very different in different parts of the hypothesis space, and so we may attempt to improve recruitment rate (for example) when in fact it is adherence which requires improvement. Methodology for jointly modelling a pilot and subsequent main RCT in this manner could be informed by developments for designing phase II/III programs in the drug setting [36, 18, 25].

References

[1] Kerry N L Avery, Paula R Williamson, Carrol Gamble, Elaine O'Connell Francischetto, Chris Metcalfe, Peter Davidson, Hywel Williams, and Jane M Blazeby. Informing efficient randomised controlled trials: exploration of challenges in developing progression criteria for internal pilot studies. *BMJ Open*, 7(2):e013537, feb 2017.

[2] Peter Bacchetti. Current sample size conventions: Flaws, harms, and alternatives. *BMC Medicine*, 8(1):17, Mar 2010.

[3] Peter Bacchetti, Charles E. McCulloch, and Mark R. Segal. Simple, defensible sample sizes based on cost efficiency. *Biometrics*, 64(2):577–585, jun 2008.

[4] Richard H. Browne. On the use of a pilot sample for sample size determination. *Statistics in Medicine*, 14(17):1933–1940, 1995.

[5] Matteo Cellamare and Valeria Sambucini. A randomized two-stage design for phase II clinical trials based on a bayesian predictive approach. *Statistics in Medicine*, 34(6):1059–1078, dec 2014.

[6] Ming-Hui Chen, Joseph G. Ibrahim, Peter Lam, Alan Yu, and Yuanye Zhang. Bayesian design of noninferiority trials for medical devices using historical data. *Biometrics*, 67(3):1163–1170, mar 2011.

[7] Christy Chuang-Stein, Paul Stryszak, Alex Dmitrienko, and Walter Offen. Challenge of multiple co-primary endpoints: a new approach. *Statistics in Medicine*, 26(6):1181–1192, 2007.

[8] Mark R. Conaway and Gina R. Petroni. Designs for phase II trials allowing for a trade-off between response and toxicity. *Biometrics*, 52(4):pp. 1375–1386, 1996.

[9] Peter Craig, Paul Dieppe, Sally Macintyre, Susan Michie, Irwin Nazareth, and Mark Petticrew. Developing and evaluating complex interventions: the new medical research council guidance. *BMJ: British Medical Journal*, 337, 9 2008.

[10] Adam Crisp, Sam Miller, Douglas Thompson, and Nicky Best. Practical experiences of adopting assurance as a quantitative framework to support decision making in drug development. *Pharmaceutical Statistics*, 0(0), 2018.

[11] Nigel Dallow, Nicky Best, and Timothy H Montague. Better decision making in drug development through adoption of formal prior elicitation. *Pharmaceutical Statistics*, 0(0), 2018.

[12] Sandra M Eldridge, Deborah Ashby, and Sally Kerry. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology*, 35(5):1292–1300, 2006.

[13] Sandra M Eldridge, Claire L Chan, Michael J Campbell, Christine M Bond, Sally Hopewell, Lehana Thabane, and Gillian A Lancaster. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *BMJ*, page i5239, oct 2016.

[14] Sandra M Eldridge, Ceire E Costelloe, Brennan C Kahan, Gillian A Lancaster, and Sally M Kerry. How big should the pilot study for my cluster randomised trial be? *Statistical Methods in Medical Research*, 2015.

[15] Sandra M. Eldridge, Gillian A. Lancaster, Michael J. Campbell, Lehana Thabane, Sally Hopewell, Claire L. Coleman, and Christine M. Bond. Defining feasibility and pilot studies in preparation for randomised controlled trials: Development of a conceptual framework. *PLOS ONE*, 11(3):e0150205, mar 2016.

[16] Anne Forster, , Jennifer Airlie, Karen Birch, Robert Cicero, Bonnie Cundill, Alison Ellwood, Mary Godfrey, Liz Graham, John Green, Claire Hulme, Rebecca Lawton, Vicki McLellan, Nicola McMaster, and Amanda Farrin. Research exploring physical activity in care homes (REACH): study protocol for a randomised controlled trial. *Trials*, 18(1), apr 2017.

[17] Simon French and David Rios Insua. *Statistical Decision Theory.* Number 9 in Kendall's Library of Statistics. Oxford University Press, 2000.

[18] Heiko Gtte, Armin Schler, Marietta Kirchner, and Meinhard Kieser. Sample size planning for phase II trials based on success probabilities for phase III. *Pharmaceutical Statistics*, 14(6):515–524, sep 2015.

[19] Lisa V Hampson, Paula R Williamson, Martin J Wilby, and Thomas Jaki. A framework for prospectively defining progression rules for internal pilot studies monitoring recruitment. *Statistical Methods in Medical Research*, 0(0):0962280217708906, 2017. PMID: 28589752.

[20] Joseph G. Ibrahim, Ming-Hui Chen, Mani Lakshminarayanan, Guanghan F. Liu, and Joseph F. Heyse. Bayesian probability of success for clinical trials using historical data. *Statistics in Medicine*, 34(2):249–264, oct 2014.

[21] Donald R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21(4):345–383, 2001.

[22] Lawrence Joseph and David B. Wolfson. Interval-based versus decision theoretic criteria for the choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2):145–149, 1997.

[23] Steven A. Julious. Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics*, 4(4):287–291, 2005.

[24] Ralph L. Keeney and Howard Raiffa. *Decisions with multiple objectives: preferences and value tradeoffs.* John Wiley & Sons, 1976.

[25] Marietta Kirchner, Meinhard Kieser, Heiko Gtte, and Armin Schler. Utility-based optimization of phase II/III programs. *Statist. Med.*, 35(2):305–316, aug 2015.

[26] Gillian A. Lancaster, Susanna Dodd, and Paula R. Williamson. Design and analysis of pilot studies: recommendations for good practice. *Journal of Evaluation in Clinical Practice*, 10(2):307–312, 2004.

[27] Roger J Lewis, Ari M Lipsky, and Donald A Berry. Bayesian decision-theoretic group sequential clinical trial design based on a quadratic loss function: a frequentist evaluation. *Clinical Trials*, 4(1):5–14, 2007. PMID: 17327241.

[28] Dennis V. Lindley. The choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2):129–138, 1997.

[29] National Institute for Health Research. Research for patient benefit (rfpb) programme guidance on applying for feasibility studies, 2017.

[30] Anthony OHagan. Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux. *Environmental Modelling & Software*, 36:35 − 48, 2012. Thematic issue on Expert Opinion in Environmental Modelling and Management.

[31] Anthony O'Hagan, Caitlin E. Buck, Alireza Daneshkhah, J. Richard Eiser, Paul H. Garthwaite, David J. Jenkinson, Jeremy E. Oakley, and Tim Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley and Sons, 2006.

[32] Anthony O'Hagan, John W. Stevens, and Michael J. Campbell. Assurance in clinical trial design. *Pharmaceutical Statistics*, 4(3):187–201, 2005.

[33] Stephen Senn and Frank Bretz. Power and sample size when multiple endpoints are considered. *Pharmaceut. Statist.*, 6(3):161–170, 2007.

[34] D. J. Spiegelhalter and L. S. Freedman. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, 5(1):1–13, 1986.

[35] David J. Spiegelhalter. Bayesian methods for cluster randomized trials with continuous responses. *Statistics in Medicine*, 20(3):435–452, 2001.

[36] Nigel Stallard. Optimal sample sizes for phase II clinical trials and pilot studies. *Statistics in Medicine*, 31(11-12):1031–1042, 2012.

[37] Mark Strong, Jeremy E. Oakley, and Alan Brennan. Estimating multiparameter partial expected value of perfect information from a probabilistic sensitivity analysis sample. *Medical Decision Making*, 34(3):311–326, apr 2014.

[38] Mark Strong, Jeremy E Oakley, Alan Brennan, and Penny Breeze. Estimating the expected value of sample information using the probabilistic sensitivity analysis sample: a fast, nonparametric regression-based method. *Medical Decision Making*, 35(5):570–583, 2015.

[39] Alexander J. Sutton, Nicola J. Cooper, David R. Jones, Paul C. Lambert, John R. Thompson, and Keith R. Abrams. Evidence-based sample size calculations based upon updated meta-analysis. *Statistics in Medicine*, 26(12):2479–2500, 2007.

[40] M Teare, Munyaradzi Dimairo, Neil Shephard, Alex Hayman, Amy Whitehead, and Stephen Walters. Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials*, 15(1):264, 2014.

[41] Peter F. Thall. Some geometric methods for constructing decision criteria based on two-dimensional parameters. *Journal of Statistical Planning and Inference*, 138(2):516–527, feb 2008.

[42] Rebecca M. Turner, A. Toby Prevost, and Simon G. Thompson. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Statistics in Medicine*, 23(8):1195–1214, 2004.

[43] Rosalind J. Walley, Claire L. Smith, Jeremy D. Gale, and Phil Woodward. Advantages of a wholly bayesian approach to assessing efficacy in early drug development: a case study. *Pharmaceutical Statistics*, 14(3):205–215, apr 2015.

[44] Fei Wang and Alan E. Gelfand. A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, 17(2):pp. 193–208, 2002.

[45] Amy L Whitehead, Steven A Julious, Cindy L Cooper, and Michael J Campbell. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Statistical Methods in Medical Research*, 2015.

[46] John Whitehead, Elsa Valdés-Márquez, Patrick Johnson, and Gordon Graham. Bayesian sample size for exploratory clinical trials incorporating historical data. *Statistics in Medicine*, 27(13):2307–2327, 2008.

[47] D. T. Wilson, R. E. Walwyn, J. Brown, A. J. Farrin, and S. R. Brown. Statistical challenges in assessing potential efficacy of complex interventions in pilot or feasibility studies. *Statistical Methods in Medical Research*, 25(3):997–1009, jun 2015.