Bayesian design and analysis of external pilot trials for complex interventions

Duncan T. Wilson[*] Rebecca E. A. Walwyn[*]
James M.S. Wason[†] Julia Brown[*] Amanda J. Farrin[*]

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 parameters such as recruitment, retention and adherence rates. 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 bad decision. The assessment of such properties is complicated by 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 nuisance parameters.

Methods: We describe a Bayesian approach to analysing pilot data, making progression decisions, and evaluating the resulting statistical properties of proposed pilot trials at the design stage to inform the choice of sample size. 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 two example pilot trials of complex interventions, each estimating multiple parameters which inform the optimal progression decision. We show that reasonable operating characteristics of the pilot trials can be obtained without increasing their sample size beyond typical values, 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 as-

[*]Leeds Institute of Clinical Trials Research, University of Leeds, Leeds, UK
[†]MRC Biostatistics Unit, Cambridge, UK

sessed in pilot trials to be articulated and used to make better progression decisions, 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 factors such as slow patient recruitment, poor levels of adherence to the intervention or trial protocol, and low completeness of follow-up data. To anticipate and rectify these problems we often conduct small trials prior to the main RCT [1]. These 'pilots' typically take the same form as the planned RCT but with a significantly lower sample size [2]. If there is a seamless transition 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. Whereas internal pilots are primarily used to inform sample size calculations, external pilots are used to address other feasibility aspects [3].

The data generated by a pilot trial is used to help 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 National Institute for Health Research ask that the criteria for progression are pre-specified and included in the research plan [4]. A single pilot trial can collect data on several progression criteria, often focused on the aforementioned areas of recruitment, protocol adherence, and data collection [5]. 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" [6]. The traffic light reference relates to the labelling of these three possible decisions as red, amber and green. Although progression criteria are widely used, there is little published guidance about how they should be determined [5, 7].

In addition to pre-specifying progression criteria, another key design decision is the choice of pilot sample size. Several methods for sample size determination aim to provide a sufficiently precise estimate of the varinace in the primary outcome measure to inform the sample size of the main trial [8, 9, 10, 11, 12, 13]. Others have suggested a simple rule of thumb for when the goal is to identify unforseen problems [14]. Using pilot trials to test effectiveness has been repeatedly advised against, as such a test will likely have low power given the small sample size in the pilot [15, 16, 3]. Increasing the type I error rate from the conventional values has been suggested as one approach to address this issue [17]. This is also suggested in [18] where the authors propose a method which, although described in terms of confidence intervals, is equivalent to powering for a one sided test of effectiveness with a type I error rate of 0.5. In practice, simple rules of thumb such as recruiting 30 participants per arm are commonly used [18, 13].

Parameter estimates from pilots can be subject to considerable sampling variation [12, 19], and so it may be that a certain progression criterion is met, or missed, due to chance alone. We should, therefore, evaluate the statistical properties of any proposed pilot design and progression criteria before we implement them. However, despite the important 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 nuisance parameters [20].

In this paper we will describe a method for designing and analysing 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. We define a simple parametric representation of the loss function and show how the parameters can be chosen either through direct elicitation of preferences, or 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 [21]. 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 [22]. 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. In Section 2 we describe the general framework for pilot design and analysis, some operating characteristics used for evaluation, and a routine for optimising the design. Two illustrative examples are then described in Sections 3 and 4. Finally, we discuss implications and limitations in Section 5.

2 Methods

2.1 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 [23] 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 knowledge and uncertainty in the parameters at the design stage. The second prior, $p_A(\theta)$, will be used when analysing the pilot data. Although it may be that $p_A(\theta) = p_D(\theta)$, it has been argued that regulators are unlikely to accept the prior beliefs of the trial sponsor for analysis of the data and as such a weakly or non-informative prior should be used for $p_A(\theta)$ [21, 24].

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 can then 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 decision $j \neq i$, there will be negative consequences. In particular, we may: proceed to a futile main RCT; discard a promising intervention; or make unnecessary adjustments to the intervention or trial design. We denote these errors as E_1, E_2, E_3 respectively, and use $\bar{E}_i$ to denote error i not occurring. We use a loss function to express the preferences of the decision maker(s) on the space of possible events $E_1 \times E_2 \times E_3$ under uncertainty, specified as follows. Firstly, we scale the function by assigning 0 loss to the event of no errors, $(\bar{E}_1, \bar{E}_2, \bar{E}_3)$, and a maximum loss of 1 to the event of all errors occurring, (E_1, E_2, E_3). We then require the decision maker(s) to state a value c_1 such that they are indifferent between the event $(E_1, \bar{E}_2, \bar{E}_3)$ (that is, proceeding to a futile RCT only) with probability 1 and a simple gamble which will result in the event $(\bar{E}_1, \bar{E}_2, \bar{E}_3)$ with probability c_1 and the event (E_1, E_2, E_3) with probability $1 - c_1$. Similarly, we require a value c_2 such that they are indifferent between the event $(\bar{E}_1, E_2, \bar{E}_3)$ (that is, discarding a promising intervention only) with probability 1 and a simple gamble which will result in the event $(\bar{E}_1, \bar{E}_2, \bar{E}_3)$ with probability c_2 and the event (E_1, E_2, E_3) with probability $1 - c_2$. Taking $c_3 = 1 - c_1 - c_2$, we can now compose a piecewise constant loss function $L(d, \Phi_I) : \{r, a, g\} \times \{\Phi_R, \Phi_A, \Phi_G\} \to [0, 1]$ which takes values as given in Table 1. For example, suppose we make a 'green' decision under the 'amber' hypothesis. The subsequent trial will be futile because the necessary adjustments will not have been made; but we have also discarded a promising intervention, since it would have been redeemed had the adjustments been made. The overall loss is therefore $c_1 + c_2$.

Given a loss function with parameters $\mathbf{c} = (c_1, c_2, c_3)$ we follow the principle of maximising expected utility (or in our case, minimising the expected loss) when making a decision. We first use the pilot data in conjugation with the

Table 1: Losses associated with each decision under each hypothesis.

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

analysis prior $p_A(\theta)$ to obtain a posterior $p(\phi \mid x)$, and then choose the decision i^* such that

$$
\begin{aligned}
i^* &= \underset{i\in\{r,a,g\}}{\arg\min}\, \mathbb{E}_{\phi|x}[L(i,\phi)] & (1)\\
&= \underset{i\in\{r,a,g\}}{\arg\min} \int L(i,\phi)p(\phi|x)d\phi. & (2)
\end{aligned}
$$

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 parameters $\mathbf{c}$:

$$
\begin{aligned}
\mathbb{E}_{\phi|x}[L(r,\phi)] &= p_A c_3 + p_G c_3, & (3)\\
\mathbb{E}_{\phi|x}[L(a,\phi)] &= p_R c_1 + p_R c_2 + p_G c_2, & (4)\\
\mathbb{E}_{\phi|x}[L(g,\phi)] &= p_R c_1 + p_A c_1 + p_A c_3. & (5)
\end{aligned}
$$

For some simple models where a conjugate analysis can be done, the posterior probabilities p_I can be obtained exactly. Otherwise, Monte Carlo estimates 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 $\phi^{(1)}, \phi^{(2)}, \dots, \phi^{(M)} \sim p(\phi \mid x)$,

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

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

2.2 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 an error 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 $(\theta^{(1)}, x^{(1)}), (\theta^{(2)}, x^{(2)}), \ldots, (\theta^{(N)}, x^{(N)})$ 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 2.1, using some 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 $p(1-p)/N$.[1]

2.3 Optimisation

Elicitation of the loss function parameters $\mathbf{c} = (c_1, c_2, c_3)$ can be challenging, particularly when multiple decision-makers are involved [25]. One way to potentially overcome this difficulty is to consider the full set of possible parameter values and examine the operating characteristics they lead to. Specifically, we can consider the problem of of finding optimal $\mathbf{c}$ with respect to these operating characteristics. Thinking of operating characteristics as functions of $\mathbf{c}$, 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 $\mathcal{C} = \{c_1, c_2 \in [0,1] \mid c_1 + c_2 \leq 1\}$.

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 $\mathcal{C}^* = \{\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \ldots, \mathbf{c}^{(K)}\}$ such that each member 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 $\mathbf{c}$. In this case, because $\mathbf{c}$ leads to worse (or at least no better) values of all three operating characteristics when compared to $\mathbf{c}'$, we have no reason to include it in

[1] Note that in the case of complex models which do no admit a conjugate analysis, the posterior probabilities obtained using an MCMC analysis will themselves be approximate and as such the optimal decision will be subject to error, which may increase the variance of the operating characteristic estimates. However, this issue can be sidestepped by assuming 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.

our set$\mathcal{C}^*$. Because the search space $\mathcal{C}$ has only two dimensions, problem (7) can be approximately solved simply by generating a random sample of $\mathbf{c}$'s and estimating the operating characteristics for each. Any parameters which are dominated in this set can then be discarded, and the operating characteristics of those which remaining can be 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 $\mathbf{c}$ 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.

3 Illustrative example - Child psychotherapy (TIGA-CUB)

TIGA-CUB (Trial on Improving Inter-Generational Attachment for Children Undergoing Behaviour problems) was a two-arm, individually-randomised, controlled pilot trial informing the feasibility and design of a confirmatory RCT comparing Child Psychotherapy (CP) to Treatment as Usual (TaU), for children with treatment resistant conduct disorders. The trial aimed to recruit $n = 60$ primary carer-child dyads to be randomised equally to each arm. This sample size was chosen to give desired levels of precision in the estimates of the common standard deviation of the primary outcome, the follow-up rate, and the adherence rate. Here, we focus on the latter two parameters and consider how our proposed method could have informed the design of TIGA-CUB.

We model the number of participants successfully followed-up using a binomial distribution with parameter p_f, and similarly the number successfully adhering to the intervention arm with a binomial distribution with parameter p_a. At the design stage, the follow-up rate p_f was thought to range from 62% to 92%, while the adherence rate p_a ranged from 40% to 95%. We reflect these ranges of uncertainty in our design priors by using a beta priors $p_f \sim Beta(40, 10)$ (thus giving a prior mean of 0.8), and $p_a \sim Beta(11.2, 4.8)$ (giving a prior mean of 0.7). We assume that a uniform 'non-informative' prior $Beta(1, 1)$ will be used for each parameter in the analysis.

Quantitative progression criteria were not pre-specified in TIGA-CUB. For the purposes of illustration, we consider the special case where only 'red' and 'green' decisions and hypotheses are considered. We define the hypothesis Φ_G as the subset of the parameter space where $p_f >= 0.8$ and $p_a >= 0.7$, hypothesis Φ_R being its complement. Thus, in this example we do not consider there to be a trade-off between the two parameters of interest. For the main trial to be feasible, both must be above their respective thresholds.

In this special case, $E[L(g,\phi)] = p_R c_1$ and $E[L(r,\phi)] = p_G c_2$, where $p_R + p_G = 1$ and $c_1 + c_2 = 1$. Decision g is therefore optimal whenever $p_G > c_1$. Moreover, the posterior probability p_G can be easily calculated given pilot due to the beta prior distributions being conjugate. Specifically, given a total sample size n and observing x_f participants with follow-up and x_a participants with adherence, the posterior probability $Pr[\phi \in \Phi_G \mid x]$ is given by

$$p_G = [1 - F(0.8; 1 + x_f, 1 + n - x_f)] \times [1 - F(0.7; 1 + x_a, 1 + n/2 - x_a)], \quad (8)$$

where $F(y; \alpha, \beta)$ denotes the cumulative probability function of the beta distribution with parameters α, β.

At the design stage we can calculate the probability of a futile trial (OC_1),

$$Pr[g, \phi \in \Phi_R] = \int_{\Phi_R} Pr[g \mid \phi] p(\phi) d\phi \quad (9)$$

$$= \int_{\Phi_R} \left(\sum_{x_f=0}^{n} \left[\sum_{x_a=0}^{n/2} \mathbb{I}(p_G < c_1 \mid x_f, x_a, n) p(x_a \mid \phi) \right] p(x_f \mid \phi) \right) p(\phi) d\phi, \quad (10)$$

and similarly for the probability of discarding a promising intervention. As these calculations can be computationally expensive for moderate n due to the nested summation term, we use Monte Carlo approximations as described in Section 2 with $N = 10^6$ samples.

Considering a range of possible sample sizes $10 \leq n \leq 100$ for $c_1 \in \{0.25, 0.5, 0.75\}$, we estimated the error probabilities OC_1 and OC_2 along with the expected loss. Figure 1 plots the results, including 95% Monte Carlo error intervals denoted by the shaded areas. Increases to the parameter c_1 reflect increasing cost associated with running a futile trial, and thus lead to a reduction in the probability of this event at the expense of an increase in the probability of discarding a promising intervention. Our results show that, for the design priors and hypotheses used in this example, the chosen sample size in TIGA-CUB of $n = 60$ can provide error rates broadly in line with conventional type I and II error rates under the usual hypothesis testing framework. For example, if $c_1 = 0.5$ and $n = 60$ then $OC_1 \approx 0.05$ and $OC_2 \approx 0.1$.

In Figure 2 we plot the operating characteristics and expected loss obtained as we vary the loss parameter c_1 from 0 to 1, whilst keeping the sample size fixed at $n = 60$.

4 Illustrative example - Physical activity in care homes (REACH)

The REACH (Research Exploring Physical Activity in Care Homes) trial aimed to inform the feasibility and design of a future definitive RCT assessing a com-

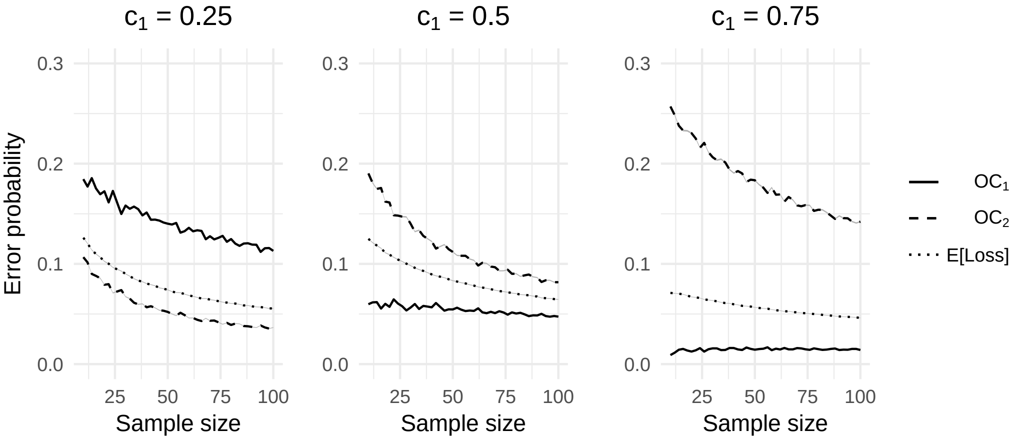


Figure 1: Probabilities of a futile main trial (OC_1) and of discarding a promising intervention (OC_2) for a range of sample sizes and different values of the loss parameter c_1.

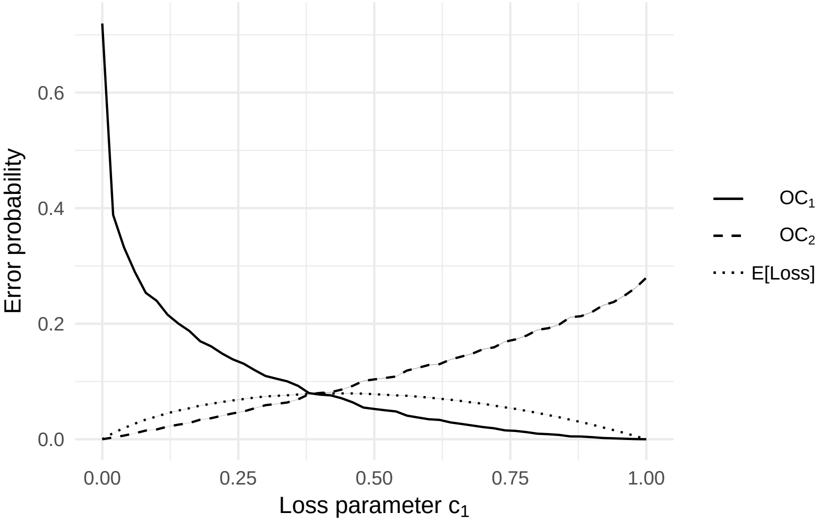


Figure 2: Probabilities of a futile main trial (OC_1) and of discarding a promising intervention (OC_2) for a range of loss parameters c_1 when sample size is fixed at $n = 60$.

Table 2: Pre-specified progression criteria used in the original REACH design.

Outcome	Red	Amber	Green
Recruitment	Less than 8	Between 8 and 10	At least 10
Adherence	Less than 50%	Between 50 and 75%	At least 75%
Follow-up	Less than 65%	Between 65 and 75%	At least 75%

plex intervention designed to increase the physical activity of care home residents [26]. The trial was cluster randomised at the care home level, with twelve care homes randomised equally to treatment as usual (TaU) or the intervention plus TaU.

Data on several feasibility outcomes were collected. Here, we focus on four: 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). Progression criteria using the traffic light system were pre-specified for all of these outcomes except efficacy, as shown in Table 2.

4.1 Model specification

To begin specifying a model for the REACH trial, we first note that the four substantive parameters can be divided into two pairs. Firstly, mean cluster size and follow-up rate relate to the amount of information which a confirmatory trial will gather. Secondly, efficacy and adherence relate to the effectiveness of the intervention, where effectiveness is thought of as the effect which will be obtained in practice when the effect of non-adherence and other pragmatic effects are accounted for. We expect that a degree of trade-off between adherence and efficacy will be acceptable, with a decrease in one being compensated by an increase in the other. Likewise, low mean cluster size could be compensated to some extent by higher follow-up rate, and vice versa.

While there may be trade-offs within these pairs of parameters, we do not expect trade-offs between them. A trial with no effectiveness will be futile regardless of the amount of information collected, and so should not be conducted. 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, not least because they can be easily 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

$$\phi \in \begin{cases} \Phi_R \text{ if } \phi^i \in \Phi^i_R \text{ or } \phi^e \in \Phi^e_R \\ \Phi_G \text{ if } \phi^i \in \Phi^i_G \text{ and } \phi^e \in \Phi^e_G \\ \Phi_A \text{ otherwise.} \end{cases} \tag{11}$$

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

$$\sigma^2 \sim \Gamma^{-1}(\alpha_0, \beta_0),\ \mu_c \sim N(\mu_0, \sigma^2/\nu_0) \tag{12}$$

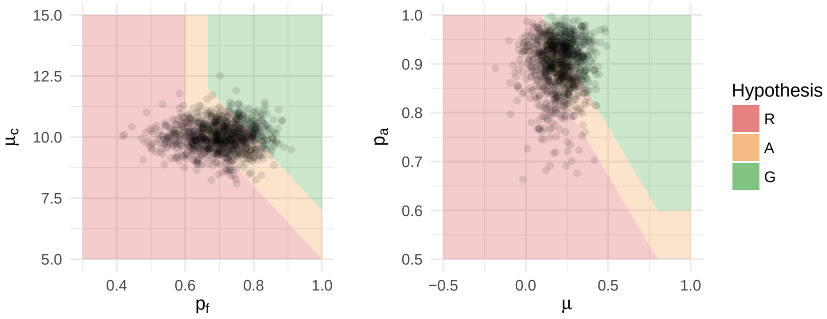
is placed on the mean and variance to allow for prior uncertainty in both parameters. It was anticipated that an average of 8 - 12 residents would be recruited in each care home. To reflect this prior belief we set the hyper-parameters to $\mu_0 = 10, \nu_0 = 6, \alpha_0 = 20, \beta_0 = 39$, giving a prior cluster size of 10 with mean variance 2.05.

We assume that follow up rates are constant across clusters. The number of participants followed-up is assumed to follow a binomial distribution $Bin(\sum_{i=1}^{k} m_i, p_f)$. We take a Beta distribution with hyper-parameters $\alpha_0 = 22.4, \beta_0 = 9.6$ as the prior for p_f. This gives gives a prior with a mean of 0.7 and an effective sample size of 30.

To partition the parameter space into hypotheses, we first consider the case where follow-up is perfect, i.e. $p_f = 1$. Conditional on this, we reason that a mean cluster size of below 5 should lead to a red decision (stop development), whereas a size of above 7 should lead to a green decision (proceed to the main trial). 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 partitioning of the parameter space:

$$(p_f, \mu_c) \in \begin{cases} \Phi^i_R \text{ if } p_f < 0.6 \text{ or } 20 - 15p_f > \mu_c \\ \Phi^i_G \text{ if } p_f > 0.66 \text{ and } 22 - 15p_f < \mu_c \\ \Phi^i_A \text{ otherwise.} \end{cases} \tag{13}$$

The hypotheses are illustrated in Figure 3 (a). Having specified both the hypotheses and the prior distribution for these two parameters, we can obtain prior probabilities of each hypothesis by sampling from the prior and calculating the proportion of these samples falling into the regions Φ^i_R, Φ^i_A and Φ^i_G.



(a) (b)

Figure 3: Marginal hypotheses over parameters for (a) follow-up rate p_f and mean cluster size μ_c; and (b) adherence rate p_a and efficacy μ. Each point is a sample from the joint prior distribution.

We have plotted 1000 samples from the prior in Figure 3 (a), falling into hypotheses Φ^i_R, Φ^i_A and Φ^i_G in proportions 0.354, 0.517, 0.129 respectively. This demonstrates that there is significant prior uncertainty regarding the optimal decision, indicating the potential value of the pilot trial.

4.1.2 Adherence and efficacy

Having defined priors and hypotheses with respect to cluster size and follow-up, we now consider 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 $\beta = 3.2$ giving a prior mean of 0.9 from an effective sample size of 32.

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

$$y_{ij} = X_j \times Y_j \times \mu + u_j + \varepsilon_i. \tag{14}$$

Here, X_j is a binary indicator of care home j being randomised to the intervention arm, Y_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 $\varepsilon_i \sim \mathcal{N}(0, \sigma^2_W)$ is the residual for resident i. We parametrise the model using the intracluster correlation coefficient $\rho = \sigma^2_B/(\sigma^2_B + \sigma^2_W)$, and place priors on

μ, ρ, and σ^2_W in the manner suggested in [27]. Specifically, we choose

$$\mu \sim N(0.2, 0.25^2) \tag{15}$$

$$\sigma^2_W \sim \Gamma^{-1}(50, 45) \tag{16}$$

$$\rho \sim Beta(1.6, 30.4). \tag{17}$$

To reflect prior expectation of an ICC around 0.05 but possibly as large as 0.1, the hyperparameters give a prior mean of 0.05 for the ICC with a prior probability of 0.104 that it will exceed 0.1.

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 suppose 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

$$(p_a, \mu) \in \begin{cases} \Phi^e_R \text{ if } p_a < 0.5 \text{ or } 0.96 - 0.57\mu > p_a \\ \Phi^e_G \text{ if } p_a > 0.6 \text{ and } 1.06 - 0.57\mu < p_a \\ \Phi^e_A \text{ otherwise.} \end{cases} \tag{18}$$

The hypotheses are illustrated in Figure 3 (b). Again, a sample of size 1000 from the joint marginal prior distribution $p(p_a, \mu)$ is also plotted, falling into hypotheses Φ^e_R, Φ^e_A and Φ^e_G in proportions 0.234, 0.470, 0.296 respectively. As before, this indicates substantial prior uncertainty regarding the optimal decision and thus supports the use of a pilot study.

The marginal hypotheses are combined together using equation (11). Considering the same 1000 samples from the design prior plotted in Figure 3, 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 decision r 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[2]. We took the sample size of the

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

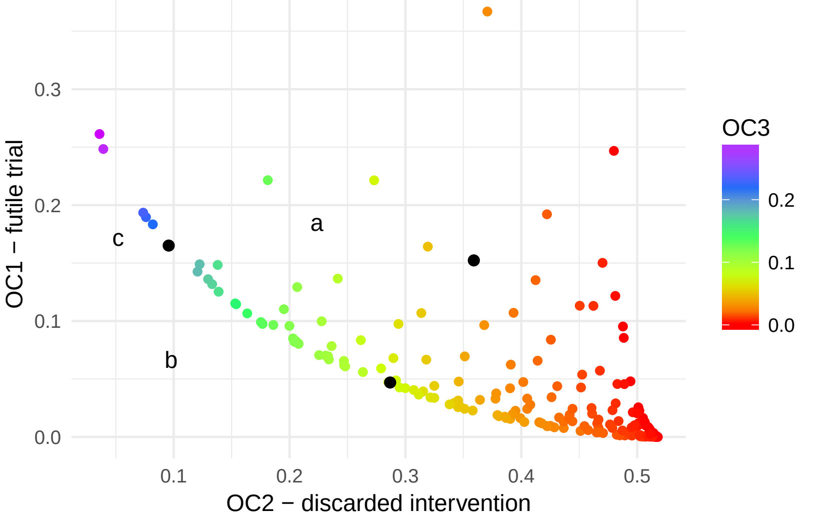


Figure 4: Operating characteristics of the example pilot trial for a range of loss parameter vectors, when a weakly informative analysis prior is used.

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 simulated data set using Stan via the R package rstan [28], in each case 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. This gave a maximum Monte Carlo error of approximately 0.005 when estimating a posterior probability $Pr[\phi \in \Phi_I \mid x]$, which we considered sufficient. These posterior samples were then used to find the posterior probabilities of each hypothesis, for each simulated data set.

We evaluated the operating characteristics for a sample of parameters (c_1, c_2, c_3) as described in Section 2.3. A total of 249 parameter vectors were evaluated, of which 193 led to operating characteristics which were worse in every respect than some other vector (i.e. dominated) and thus were discarded. The operating characteristics of the non-dominated parameters are shown in Figure 4. The three operating characteristics are found to be highly correlated. In particular, changing the parameters to give a lower probability of discarding a promising intervention (OC_2) tends to lead to a reduction in the probability of making an unnecessary adjustment (OC_3). When selecting (c_1, c_2), the key decision appears to be trading off the probability of a futile trial, (OC_1), against OC_2. There is a very limited opportunity to minimise OC_3 at the expense of these. For example, compare points b and c in Figure 4, details of which are given in Table 3. We see that point c reduces OC_3 by 0.049 in comparison to point b, but only at the expense of increase in OC_1 and OC_2 of 0.105 and 0.072 respectively.

We may expect to see a clear relationship between the value of parameters

Table 3: Estimated operating characteristics (with standard errors) of the REACH trial for the three loss parameter vectors highlighted in Figure 4, when a weakly informative analysis prior is used. Costs have been rounded to 2 decimal places; operating characteristics and their errors to 3.

Label	(c_1, c_2, c_3)	OC_1	OC_2	OC_3
a	(0.14, 0.85, 0.02)	0.165 (0.004)	0.096 (0.003)	0.204 (0.004)
b	(0.32, 0.57, 0.10)	0.047 (0.002)	0.287 (0.005)	0.078 (0.003)
c	(0.03, 0.52, 0.45)	0.152 (0.004)	0.359 (0.005)	0.032 (0.002)

c_1, c_2, c_3 and the operating characteristics they relate to. We explore this in Figure 5 with scatter plots of each parameter against each operating characteristic. The results show that there is indeed a strong relationship between the loss assigned to discarding a promising intervention, c_2, and the probability that this event will occur (see bottom right plot). Moreover, c_2 also seems to be the main determinant of operating characteristics OC_1 and OC_3. The implication is that once the $c_2 \in [0, 1]$ has been chosen, the operating characteristics of the trial depend only weakly on the way in which the remaining $1 - c_3$ is allocated to c_1 and c_3. This appears to be due to the fact that, regardless of how errors are weighted, the way we have defined our prior distributions and hypotheses means we are much more likely to make the error of discarding a promising intervention than the other types of error. 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 that 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, repeating the Bayesian analysis using the appropriate analysis prior and obtaining estimated

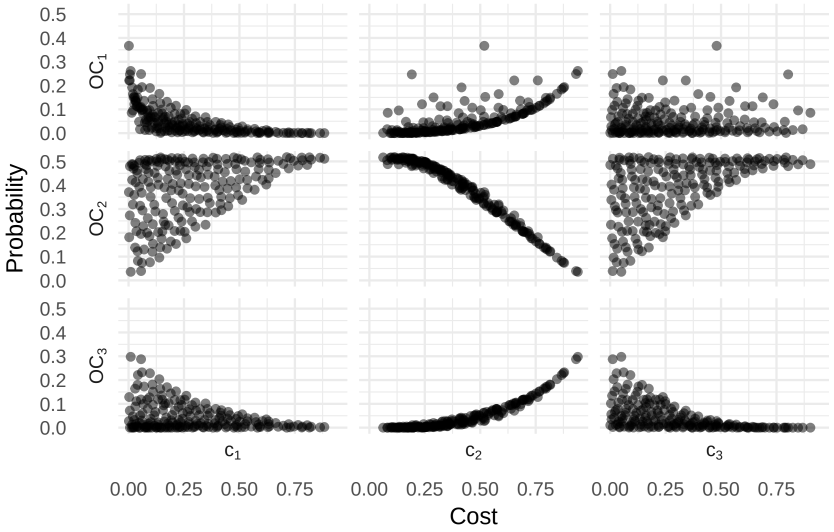


Figure 5: Relationships between the three loss parameters (x axes) and resulting operating characteristics (y axes).

posterior probabilities p_R, p_A and p_G as before. These were used in conjunction with the same set of loss parameter 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 found very little difference (further details are provided in the appendix). When we contrast cases WI and INA, however, there is a clear distinction. Using the INA analysis prior will lead to larger probabilities of a futile trial (OC_1) and of unnecessary adjustment (OC_2), while reducing the probability of discarding a promising intervention (OC_3), for almost all loss parameters. The expected loss is always lower for the INA analysis than for WI, as we would expect. Further details can be found in the appendix.

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 at the design stage, 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, and to inform the choice of sample size.

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 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.

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 [29], with type I and II error rates.

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, for example in systematic reviews or observational studies, expert opinion may 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 [24, 22, 30], 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 [25], 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 [23, 21, 32]. 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 [33, 34]. The computational difficulties will be particularly pertinent when using our approach to choose determine sample size, where several evaluations of different choices will be required. If the choice of sample size can be framed as an optimisation problem, methods for efficient global optimisation of computationally expensive functions such as those described in [35, 36] may be useful [20].

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 loss associated with discarding a promising intervention being independent of how effective the intervention is. An alternative is to try to define a richer representation of the loss function through direct elicitation of the decision makers preferences under uncertainty [37], leading to a fully decision-theoretic approach to design and analysis [38]. However, as previously noted by others [39, 40, 41], implementation of these approaches has been limited in practice and may be indicative of their feasibility.

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 [42, 43, 44].

References

[1] 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.

[2] 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.

[3] Lehana Thabane, Jinhui Ma, Rong Chu, Ji Cheng, Afisi Ismaila, Lorena Rios, Reid Robson, Marroon Thabane, Lora Giangregorio, and Charles Goldsmith. A tutorial on pilot studies: the what, why and how. *BMC Medical Research Methodology*, 10(1):1, 2010.

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

[5] 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.

[6] 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.

[7] 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.

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

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

[10] Julius Sim and Martyn Lewis. The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency. *Journal of Clinical Epidemiology*, 65(3):301–308, mar 2012.

[11] 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.

[12] 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.

[13] 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.

[14] Wolfgang Viechtbauer, Luc Smits, Daniel Kotz, Luc Budé, Mark Spigt, Jan Serroyen, and Rik Crutzen. A simple formula for the calculation of sample size in pilot studies. *Journal of Clinical Epidemiology*, 68(11):1375–1379, nov 2015.

[15] 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.

[16] Mubashir Arain, Michael Campbell, Cindy Cooper, and Gillian Lancaster. What is a pilot or feasibility study? a review of current practice and editorial policy. *BMC Medical Research Methodology*, 10(1):67, 2010.

[17] Ellen Lee, Amy Whitehead, Richard Jacques, and Steven Julious. The statistical interpretation of pilot trials: should significance thresholds be reconsidered? *BMC Medical Research Methodology*, 14(1):41, 2014.

[18] Kim Cocks and David J. Torgerson. Sample size calculations for pilot randomized trials: a confidence intervalÂ approach. *Journal of Clinical Epidemiology*, 66(2):197 – 201, 2013.

[19] Cindy L Cooper, Amy Whitehead, Edward Pottrill, Steven A Julious, and Stephen J Walters. Are pilot trials useful for predicting randomisation and attrition rates in definitive studies: A review of publicly funded trials. *Clinical Trials*, 0(0):1740774517752113, 2018. PMID: 29361833.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

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

[28] Stan Development Team. RStan: the R interface to Stan, 2016. R package version 2.14.1.

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

[30] 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.

[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] 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.

[33] 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.

[34] 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.

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

[36] Olivier Roustant, David Ginsbourger, and Yves Deville. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software*, 51(1):1–55, 2012.

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

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

[39] 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.

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

[41] 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.

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

[43] Heiko Götte, Armin Schüler, 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.

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