Bayesian design of pilot trials for complex interventions

April 30, 2018

Abstract

Background: **Methods**: **Illustration**: **Conclusions**:

1 Introduction

External pilot trials are a common component of the complex intervention development and evaluation pathway [4]. They are used to gather information regarding the feasibility and design of a larger, confirmatory trial of the intervention. Some common aims of pilot trials include the assessment of recruitment rates, data completeness, adherence to interventions, and compliance with randomisation [2]. They generally take a similar form to the planned confirmatory trial, although with a significantly reduced sample size [7].

Formal statistical approaches to the design and analysis of pilot trials are not commonly used. The sample size is often chosen using a simple rule-of-thumb, such as recruiting at least 30 participants to each arm [11, 3][1]. Although hypothesis testing has been advised against due to concerns about underpowered analyses [11, 1], pre-specified progression criteria are often used to determine whether a confirmatory trials should be undertaken and, if so, whether any modification to the intervention or the trial design should be undertaken to ensure its feasibility. These progression criteria specify thresholds against which a parameter estimate is to be compared. For example, consider a pilot trial which will estimate the probability that a participant will be lost to follow-up, $\hat{p}$. An associated progression criteria could be: if $\hat{p} > 0.35$, do not proceed; if $\hat{p} < 0.25$; and if $0.25 < \hat{p} < 0.35$, proceed but only after making modifications to the intervention or trial design aimed at improving data collection.

Despite the critical role progression criteria play in decision making, there is relatively little methodological guidance regarding how they should be determined [2]. If poor criteria are used, the probability of making an incorrect decision may be unacceptably high. A more formal approach to choosing progression criteria, involving statistical analysis of the probability of making such

[1] As previously discussed, this rule seems to have been born out of a mis-reading of the Browne paper - he actually advised against using this rule-of-thumb.

incorrect decisions, may help avoid this. However, such an approach will incur a number of statistical challenges [19]. Multiple outcomes must be considered, data will often be of a multilevel nature due to cluster randomisation or clustering due to treatment provision, the sample size will necessarily be small, and there may be significant uncertainty regarding the value of nuisance parameters (e.g. the ICC in a clsuetr randomised trial [6]). In this paper we will develop and illustrate a method which can overcome each of these challenges. We propose a Bayesian approach, allowing for any external information or knowledge to be formally incorporated, and uncertainty in nuisance parameters to be fully expressed, via prior distributions.

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. The methodology, including details on design optimisation, is then described in Section 3. We illustrate the proposed approach in Section 4, before discussing implications and limitations in Section 5.

2 Motivating example

The REACH (Research Exploring Physical Activity in Care Homes) trial [9] 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. 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.

The trial will collect data pertaining to the recruitment of care home residents to the trial, the successful delivery of the intervention, the completeness of follow-up data collection, and the effectiveness of the intervention.

[TO COMPLETE]

3 Methods

3.1 Design and analysis

Consider a pilot trial which will produce data x according to model $p(x|\theta)$. We assume the design of the trial, including its sample size, is known and fixed. The parameters of substantive interest are denoted by ϕ, nuisance parameters by ψ, with $\theta = (\phi, \psi)$. We assume that a joint prior distribution $p(\theta)$ has been specified. We do not impose any particular restrictions regarding the form of the model or the prior distribution, other than that samples from the prior and the posterior $p(\theta|x)$ can be generated.

Following the observation of the pilot data x, a decision must be made regarding whether or not to progress to a confirmatory RCT. Following the convention in pilot trials, we consider three possible actions. We may: discard the intervention and stop all future development or evaluation; proceed immediately to a confirmatory RCT; or proceed to a confirmatory RCT, but only after some

modifications to the intervention, the planned trial design, or both, have been made. We denote these decisions as r(ed), g(reen) and a(mber) respectively. We make two assumptions regarding our preferences amongst these three decisions. Specifically, for any $i, j \in \{r, a, g\} = \mathcal{D}, i \neq j$,

1. Our preference between i and j, given ϕ, is independent of ψ;
2. For any ϕ, either i is strictly preferred to j, or vice versa.

Assumption (1) formalises the separation of θ into substantive and nuisance components, while assumption (2) guarantees the existence of a unique optimal decision. We denote by Φ_I the set of substantive parameter values for which decision i is preferred to the alternatives, $\Phi_I = \{\phi \in \Phi \mid i \succ j \ \forall \ j \in \mathcal{D}\}$. By assumption (2) the entire substantive parameter space is partitioned as $\Phi = \Phi_R \cup \Phi_A \cup \Phi_G$. We will henceforth refer to these three subsets as *hypotheses*. Throughout, we will distinguish hypothesis I from the corresponding decision i by using capital and lower case letters respectively.

Given the observed data x a decision is made by first calculating the marginal posterior probabilities

$$p_R = Pr[\phi \in \Phi_R \mid x] \tag{1}$$
$$p_G = Pr[\phi \in \Phi_G \mid x]. \tag{2}$$

Note that $p_A = 1 - p_R - p_G$, and that $(p_R, p_G) \in \{[0,1]^2 \mid 0 \leq p_R + p_G \leq 1\} = \mathcal{P}$. A decision rule $a_\delta(p_R, p_G) : \mathcal{P} \to \mathcal{D}$ is then applied. The rules we consider partition the posterior probability space $\mathcal{P}$ into exactly three subsets corresponding to the three decisions, are are parametrised by defined by parameter $\delta = (\delta_1, \ldots, \delta_4) \in [0,1]^2 \times [0,\pi]^2 = \Delta$ as illustrated in Figure 1. The boundary between red and amber regions is a straight line intercepting the y-axis at δ_1 at an angle of δ_3. Similarly, the boundary between amber and green regions is a straight line intercepting the x-axis at δ_2 at an angle of δ_4. A rule is considered valid only if it prevents decision i from being made when the posterior probability of hypothesis Φ_I is zero; and if the intersection of the two boundary lines (shown as a cross in Figure 1) is not in the interior of $\mathcal{P}$. The latter condition ensures that the red and green regions are never adjacent. We assume that the decision rule has been specified in advance, and will discuss how it may be chosen in Section 3.3.

3.2 Evaluation

Given a proposed pilot trial design and associated model, a prior distribution for the parameters of the model, and a decision rule, we wish to define some appropriate operating characteristics which can be used to evaluate and compare different designs. We propose that three events are of particular interest, and define the operating characteristics as their unconditional probabilities. Specifically, we consider the probability of: running a futile confirmatory trial (OC_1); making unnecessary adjustments to the trial design or the intervention (OC_2);

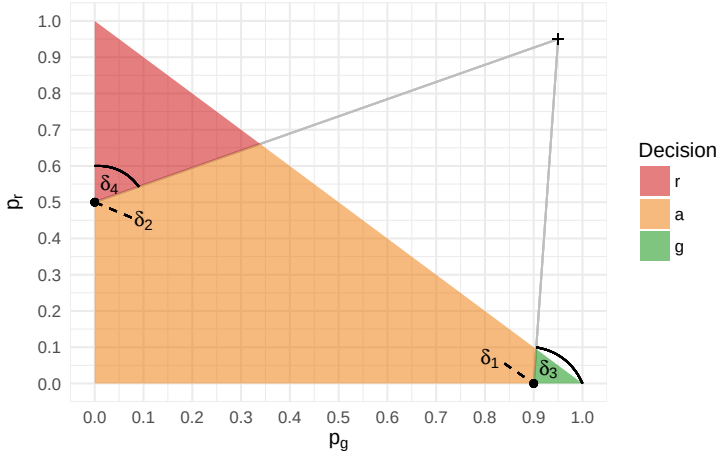

Figure 1: Example decision rule.

		Truth			
		R	A	G	
Decision	r	$p_{R,r}$	$p_{A,r}$	$p_{G,r}$	A_r
	a	$p_{R,a}$	$p_{A,a}$	$p_{G,a}$	A_a
	g	$p_{R,g}$	$p_{A,g}$	$p_{G,g}$	A_g
		Φ_R	Φ_A	Φ_G	1

Table 1: Probabilities of decisions and hypotheses.

and discarding a promising intervention (OC_3). Each of these can be specified in terms of the probabilities of making decision j when $\phi \in \Phi_I$, as set out in Table 2:

1. OC_1 - running a futile confirmatory trial ($p_{R,a} + p_{R,g} + p_{A,g}$);
2. OC_2 - making unnecessary adjustments ($p_{R,a} + p_{G,a}$);
3. OC_3 - discarding a promising intervention ($p_{A,r} + p_{G,r} + p_{A,g}$).

The probability of making decision j when $\phi \in \Phi_I$ is

$$p_{I,j} = \mathbb{E}_\theta[\mathbb{I}(\phi \in \Phi_I \ \& \ a_\delta(p_R, p_G) = j)], \tag{3}$$

where the expectation is with respect to the prior distribution $p(\theta)$ and $\mathbb{I}(.)$ is the indicator function. We can compute Monte Carlo estimates of these probabilities by sampling $\theta^{(1)}, \ldots, \theta^{(N)}$ from the prior distribution $p(\theta)$, from which we extract $\phi^{(1)}, \ldots, \phi^{(N)}$. For each $\theta^{(k)}$, data $x^{(k)} \sim p(x|\theta^{(k)})$ is then simulated. A Bayesian analysis is conducted, from which the posterior probabilities

$p_R^{(k)}, p_G^{(k)}$ are extracted. We then have

$$p_{i,j} \approx \frac{1}{N} \sum_{k=1}^{N} \mathbb{I}(\phi^{(k)} \in \Phi_I, a(p_R^{(k)}, p_G^{(k)}) = j). \tag{4}$$

When (MC)MC methods are required to perform the Bayesian analysis the posterior probability $p_R^{(k)}$ (likewise $p_G^{(k)}$) is itself a Monte Carlo estimate, based on M samples $\phi^{(m)}$ from the marginal posterior distribution $p(\phi|x^{(k)})$. Specifically,

$$p_R^{(k)} = \mathbb{E}_{\phi|x^{(k)}}[\mathbb{I}(\phi \in \Phi_R)] \tag{5}$$

$$\approx \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}(\phi^{(m)} \in \Phi_R). \tag{6}$$

We assume that M MCMC samples will be used when calculating the posterior probabilities p_R, p_G given the actual observed pilot data, and thus that the process being simulated corresponds exactly to that which will be carried out in practice. Any error in the estimation of $p_{I,j}$ therefore stems only from the finite number of samples used in the outer loop, N. For large N the error of the unbiased MC estimate of any probability p, including each of the operating characteristics, will be normally distributed with variance approximately equal to $p(1-p)/N$.

3.3 Optimisation

Evaluation of a trial design is computationally expensive, requiring the generation of $N \times M$ samples. If any changes are made to the sample size of the design, the entire process must be repeated. However, if changes are only made to the decision rule, the same set of N simulated data sets and corresponding Bayesian inferences can be re-used when evaluating the new design. This suggests that, for a fixed sample size, optimisation over the possible decision rules could find that which delivers the best operating characteristics.

Recall from Section 3.1 that Δ is the space of all decision rules. We wish to find the $\delta \in \Delta$ which minimises the operating characteristics $OC_i(\delta)$, $i = 1, 2, 3$. Note, however, that these will typically be in conflict in the sense that any decrease in one will be accompanied by an increase in another. As such, there is no single optimal decsiion rule. However, we can use the notion of Pareto dominance to compare any two rules and identify if one is superior to the other. Formally, we say that rule δ_* dominates δ, denoted $\delta_* \prec \delta$, if $OC_i(\delta_*) \leq OC_i(\delta)\ \forall\ i$ and $OC_j(\delta_*) < OC_j(\delta)$ for some specific j. In this case, rule δ_* should always be preferred to rule δ because it is at least as good in all respects, and strictly better in at least one. We can then define the *Pareto set* as the set of rules which are not dominated by any other rule. Ideally the optimisation procedure would return the Pareto set, but in practice we will find some approximation of it.

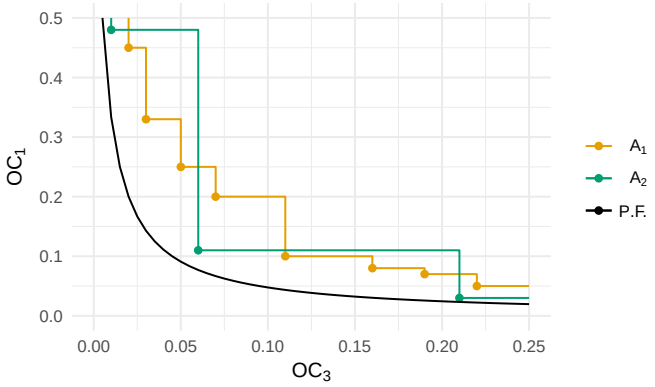

Figure 2: Example Pareto and approximation sets.

An *approximation set* $\mathcal{A}$ is a set of rules which are not dominated by any other rule in $\mathcal{A}$ [8]. Optimising over the space of decision rules aims to find an approximation set $\mathcal{A}$ which is as close as possible to the Pareto set. Clearly, we wish for $\mathcal{A}$ to contain rules which lead to operating characteristics close to those obtained by the Pareto set. However, $\mathcal{A}$ should also provide a sufficient range of options regarding the balance between the three operating characteristics. In this way, the decision maker can view the approximation set, assess the relative trade-offs between the three operating characteristics which are available, and choose the decision rule which best reflects their priorities. An illustration of a Pareto set and two approximation sets is given in Figure 2, where we have considered only operating characteristics OC_1 and OC_3. Approximation set $\mathcal{A}_2$ contains only three rules, although these are all of high quality in terms of both operating characteristics, being close to the Pareto set. However, the larger number of rules in set $\mathcal{A}_2$ provides more opportunity for the decision-maker to select a rule which balances their priorities.

An approximation set can be obtained by applying a multi-objective optimisation algorithm such as the benchmark algorithm NSGA-II [5], as implemented in the R package [12].

3.4 Utility

Recall that the analysis of the trial will come down to computing the three posterior probabilities $Pr[\phi \in \Phi_i|x]$. When making a decision based on these summaries, we are concerned with balancing the costs of each type of error and the probabilities that these will be made. If we assign a fixed cost to our three errors, c_1, c_2 and c_3, we can define a loss function which will take the following values:

Now, assuming that this loss function has been defined in the proper manner and accounts for attitude to risk, we know that the best decsion is that which

		Truth		
		R	A	G
Decision	r	0	c_2	c_2
	a	$c_1 + c_3$	0	c_3
	g	c_1	$c_1 + c_2$	0

Table 2: Loss function values.

minimises the posterior loss. That is, we would choose

$$d = \arg \min_{d \in \{r,a,g\}} \mathbb{E}_{\phi|x}[L(d, \phi)]. \tag{7}$$

Note that this can be translated into a rule based on posterior probabilities as previously discussed, but that this formulation has less parameters. In fact, since loss is invariant under a linear transformation, the costs parameters can be scaled so $c_1 + c_2 + c_3 = 1$ and thus only two parameters must be chosen (compared with 4 in the previous model).

What assumptions have we encoded in using a loss function of this form? We can consider the problem as one of three binary attributes - for each type of error, either it is made or it isn't. We have then assumed that the losses of each attribute are additive - the cost of making one type of error is not affected by whether or not we are making another type of error also. Since each attribute is binary, there is no scope or need to consider attitude to risk separately to the value of a deterministic outcome. How can we validate/verify the function? The parameters will imply the equivalence of a set of gambles. For example, suppose $c_1 = 0.5, c_2 = 0.4$ and $c_3 = 0.1$. Then we must be unable to choose between gamble A: type I error with probability 0.5, or no error with probability 0.5; and gamble B: a type II with probability 0.375, and a type III error with probability 0.625 [they both give an expected loss of 0.25]. To elicit the cost parameters we can consider a gamble of all errors with probability p and no errors with probability $(1 - p)$, and compare this with the certainty of just one error. Then, for example, the loss of a type I error will be the chosen value of p. Doing this for two of the cost parameters, we can then extract the third. Are these types of judgements cognitively feasible?

If we have properly parametrised our loss function, we can consider computational issues again. Recall that we can simulate an arbitrarily large data set with the true hypothesis and a set of statistics summarising the corresponding trial data. That is, we can simulate from the joint distribution of $p(\phi, x)$. We can consider a model $\varphi(x)$ which returns a decision d based on the supplied trial data (or summary stats of it). Any classifier built using any algorithm will have this form. Ideally, we would like to construct a model such that

$$\varphi(x) = \arg \min_{d \in \{r,a,g\}} \mathbb{E}_{\phi|x}[L(d, \phi)]. \tag{8}$$

This is as good as a model could be, representing the same as a full posterior

inference and minimisation of expected loss. It make the best decision possible for any given data x. Any classifier built on the simulated data set will therefore provide a kind of upper bound on the performance of our actual intended procedure. This bound will be in the form of expected loss, not the unconditional error rate operating characteristics defined above. So, we may find a model with $OC_1 = 0.2$ say - that does not mean that the real inference procedure will have $OC_1 < 0.2$. But we can be sure that the expected loss will be lower, and regardless of the operating characteristics, the inference will be of better quality (according to our subjective utility). Thus, if we were to use these models to help judge sample size, they would lead us to a conservative estimate - we will never choose something too small. The better the classification model we can construct, the closer the bound will be. So this seems a good way to approach design - take the default sample size, get a (quick) upper bound on performance, and estimate the OCs. Adjust the sample size if needed. And if things look tight, i.e. sample size is uncomfortably large considering the performance metrics, a full nested MC procedure can be used to get a more accurate evaluation.

For now, leave computational issues for a future piece of work where we can do a detailed study of the types of algorithms which are most appropriate and an empirical evaluation to see how close the bounds can be. Now, consider an alternative to eliciting the costs of the utility function. We can instead take random values, and study the resultant operating characteristics. Then we can decide on the parameters based on measures which are more directly meaningful, and specific to our actual trial design problems as opposed to completely general. Note that this is now substantially simpler than optimisation in the case of the decision rules above, as we have only two dimensions in the parameter space. We can, therefore, generate the posterior samples as before and do a simple grid search over loss functions. We can still assess all results in terms of Pareto dominance. We see that the vast majority of solutions are efficient - not surprising given the more restrictive parametrisation.

3.5 Elicitation

Elicitation of prior distributions is challenging, but there are established methods available - e.g. SHELF. We can at least comment on the feasibility of these, e.g. note that we might get 2 or three parameters elicited in a day involving several experts. Refer to methods for building priors form data, e.f. for ICCs. Note that it will provide a foundation for any future trials.

Defining hypotheses in the way we have described is new, so we need some guidance on how to do it. First thing is to recognise that we are dealing with a multi-attribute problem. We want to identify attributes that are mutually preferentially independent. For example, in REACH we can consider two groups - the cluster size and follow up attributes, and the adherence and efficacy attributes. We want both a logistically feasible trial AND a intervention that will deliver benefit. But each of these goals can be achieved by a mixture of their components - a low cluster sample size would be compensated by a high follow-

up rate and vice versa; a low adherence rate would be compensated by a high level of efficacy and vice versa. But these trade-offs will only work to an extent - e.g. very low adherence will not be acceptable, regardless of the efficacy.

If we can find groups like these, we can then construct hypotheses independently and combine together with AND operations. So find the attributes where we have trade-offs, and elicit these trade-offs. In our example we find two 2D spaces to define trade-offs in, which is ideal as we can simply draw the boundaries. We have used some crude linear boundaries, which might be fine, but it would be easy to extend to non-linear boundaries (e.g. fit a non-parametric model to drawn data).

Note that this is all done in a deterministic manner, i.e. we are in the realm of value functions rather than utility functions (at least until we specify the costs components

We can also consider active learning methods for when we need to define trade-offs in higher dimensions, where we use an algorithm to query the user for classification of points and iteratively fit a model. On the above, note that this can all be done after the nested MC algorithm has been run. The algorithm will return for iteration in the outer loop a prior and posterior point in the parameter space. We can then take that data and classify into whichever hypothesis we like.

[**?**] - Gaussian processes with monotonicity information

[**?**] - Gaussian Process Preference Elicitation

[**?**] - Real-time Multiattribute Bayesian Preference Elicitation with Pairwise Comparison Queries

For two attributes x, y, first define the marginal boundaries - if attribute x crosses the green-amber boundary, we won't make decision g. Likewise with the amber-red boundary. Denote these boundaries as $x_{r,a}, x_{a,g}, y_{r,a}, y_{a,g}$.

Following this we need to consider trade-offs. To do so, assume that $x = x_{r,a}$. Then elicit the value $y^c_{r,a}$ such that $y < y^c_{r,a} \to r$ is optimal. That is, if attribute x is on the decision boundary, what values of y is needed to put us on one side of it? In the same manner we can elicit $x^c_{r,a}, y^c_{a,g}$ and $x^c_{a,g}$. We can then plot the hypotheses along with a sample form the marginal prior, and adjust as we feel necessary.

Note that this method allows us to state the method precisely without excessive notation, but that we could easily define hypotheses by fitting a piecewise linear model to any number of points in the marginal parameter space. That is, a completely non-parametric method can be used.

4 Illustration

4.1 Hypothetical example

Consider a simplification of the REACH cluster randomised pilot trial described in Section 2, where interest lies only in the average cluster size (i.e. the average number of participants recruited to the trial in each care home), and in the

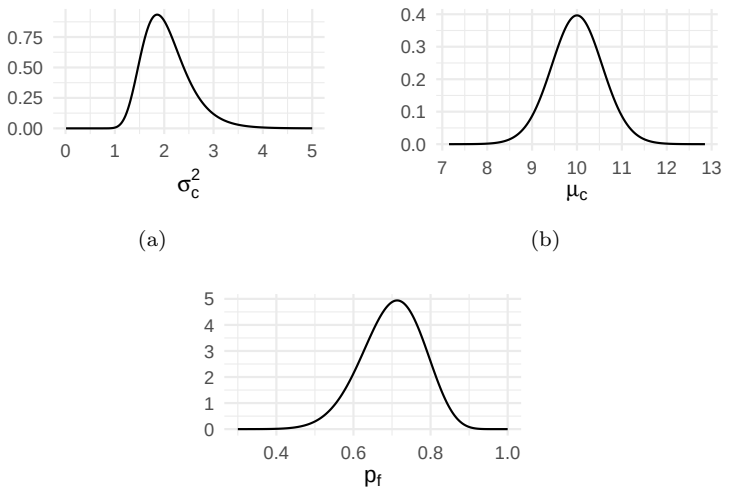

Figure 3: Prior distributions for the mean and variance of cluster size and the probability of successful follow-up.

rate of successful follow-up. To apply the methods of Section 3, we first define the model $p(x|\theta, n)$. We assume that cluster sizes are normally distributed, $m_i \sim \mathcal{N}(\mu_c, \sigma^2), i = 1 \dots k$. A normal-inverse-gamma prior is placed on the mean and variance to allow for prior uncertainty in both parameters:

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

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 3.

After observing the k cluster sizes m_i with mean $\bar{m}$, the posterior distribution is given by the hyper-parameters

$$\mu_1 = \frac{\nu_0\mu_0 + k\bar{m}}{\nu_0 + k} \tag{10}$$

$$\nu_1 = \nu_0 + k \tag{11}$$

$$\alpha_1 = \alpha_0 + k/2 \tag{12}$$

$$\beta_1 = \beta_0 + \frac{1}{2}\sum_{i=1}^{k}(m_i - \bar{m})^2) + \frac{k\nu_0}{\nu_0 + k}\frac{(\bar{m} - \mu_0)^2}{2}. \tag{13}$$

The number of participants followed-up 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 3. After observing f participants being followed-up, the updated hyper-parameters for

Figure 4: Samples from the joint prior distribution of the hypothetical example, where the parameter space is coloured to denote the three hypotheses Φ_R, Φ_A and Φ_G.

the posterior distribution of p_f are

$$\alpha_1 = \alpha_0 + f \tag{14}$$

$$\beta_1 = \beta_0 + \sum_{i=1}^{k} m_i - f. \tag{15}$$

The hypotheses for the parameters of interest, $\phi = (\mu_c, p_f)$, are as follows:

$$\phi \in \begin{cases} \Phi_R \text{ if } \mu_c < 9 \text{ or } p_f < 0.6 \text{ or } 1.6 - 0.1\mu_c > p_f \\ \Phi_A \text{ if } \phi \notin \Phi_R \ \& \ [\mu_c < 10 \text{ or } p_f < 0.7 \text{ or } 1.75 - 0.1\mu_c > p_f] \\ \Phi_G \text{ otherwise.} \end{cases} \tag{16}$$

A random sample from the prior distribution of ϕ is shown in Figure 4, with the corresponding hypotheses highlighted.

Upon obtaining the pilot data and determining the posterior distribution of ϕ by using the conjugate relations described above, we use the rule illustrated in Figure 1 to determine our decision. To estimate the operating characteristic of this trial, we used $N = 10^4$ and $M = 10^4$ samples. As generating $M = 10^4$ samples from each posterior distribution is not computationally expensive in this conjugate setting, we repeat the analysis using $k = 6, 12, 18, 24, 30$ clusters. We keep the decision rule constant throughout. The estimated operating characteristics for each sample size are plotted in Figure 5.

The error bars in Figure 5 denote 2 standard errors, illustrating that the number of samples used gives a reasonable degree of precision in our estimates. For all values of k considered, operating characteristics OC_1 and OC_3 are considerably lower than OC_2. As sample size increases, substantial improvements

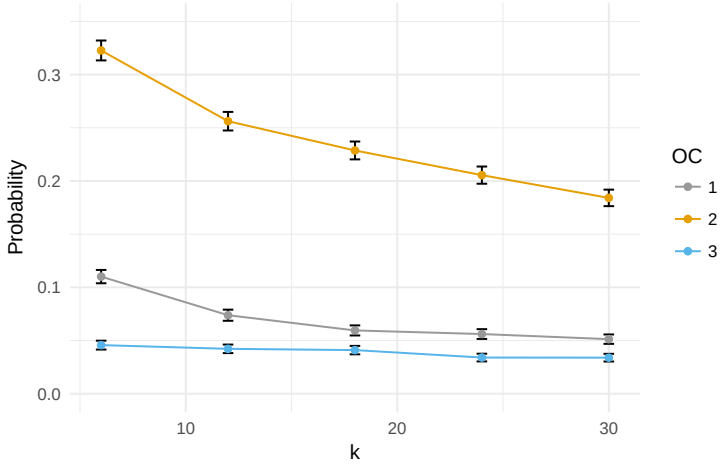

Figure 5: Estimated operating characteristics for the hypothetical example as sample size increases. Error bars denote two standard errors.

are seen in OC_2 with only moderate improvement in OC_1. In contrast, OC_3 appears to be unaffected by changes in sample size. The hypotheses boundaries are shown as dotted lines as before. In Table 3 we present the individual posterior probability terms for the cases $k = 6$ and $k = 30$.

To further understand the obtained results, consider the composition of each operating characteristic. We would expect the extreme probabilities of $p_{R,g}$ and $p_{G,r}$ to be low, and these are indeed found to be approximately 0. Errors are therefore largely confined to the terms $p_{R,a}, p_{A,r}, p_{A,g}$ and $p_{G,a}$. Of these, Table 3 shows that the probability of an error when $\phi \in \Phi_A$ is very low. This can be explained by the fact that the prior mean is located at the point $\phi = (10, 0.7) \in \Phi_A$. As Bayesian inference will lead to a posterior distribution

			Truth			
			R	A	G	
Decision	r	$k = 6$	0.098	0.044	0.001	0.142
		$k = 30$	0.157	0.032	0.000	0.189
	a	$k = 6$	0.108	0.524	0.214	0.847
		$k = 30$	0.049	0.545	0.135	0.729
	g	$k = 6$	0.000	0.002	0.010	0.011
		$k = 30$	0.000	0.002	0.080	0.082
			0.206	0.579	0.215	1

Table 3: Probabilities of decisions and hypotheses.

between the prior and the likelihood function of the observed data, the posterior will generally be pulled back towards this point. For the posterior probabilities to favour hypotheses Φ_R or Φ_G we would require quite extreme data to be observed. Moreover, the majority of the prior mass in Φ_A is in the centre of the area, as opposed to at the borders as is the case in Φ_R and Φ_G. Thus, the low values of $p_{A,r}$ and $p_{A,g}$ are not surprising. The majority of the error lies in the terms $p_{R,a}$ and $p_{G,a}$. That the latter is larger than the former can be expected given that the prior probabilities of the two hypotheses are similar, while the decision rule allocates a significantly smaller region of the posterior probability space to decision g than r. Note that the sum of these probabilities is exactly operating characteristic OC_2, which explains its relatively high magnitude. Table 3 shows that both of these probabilities are reduced by increasing k, leading to large reductions in OC_2 and moderate reductions in OC_1.

4.2 REACH

We now return to the full motivating example. Recall that the outcomes of interest are mean cluster size, probability of successful follow-up, probability of adherence, and the efficacy of the intervention. The former two outcomes are taken to be as in the preceding example, with the same models and prior distributions. Conditional on recruiting n_E participants into the intervention arm of the pilot, the number of participants who adhere to the intervention is assumed to be binomially distributed with probability p_a. We assume that adherence is absolute, in the sense that a participant who does not adhere to the intervention will not receive the treatment effect. We place a Beta prior on p_a with hyper-parameters $\alpha = 28.8$ and $\beta = 3.2$.

Efficacy outcomes are expected to be correlated within care homes. We model this by including a random effect u_j for each care home, $j = 1, \ldots, k$. The outcome $y_{i,j}$ of participant i in care home j is modelled as

$$y_{i,j} = \beta_i^t \times \beta_i^a \times \mu + u_j + e_i. \tag{17}$$

Here, β_i^t is a binary indicator of participant i receiving the intervention, β_i^a is a binary indicator of participant i adhering to the intervention, $u_j \sim \mathcal{N}(0, \sigma_B^2)$ is the random effect for care home j and $e_i \sim \mathcal{N}(0, \sigma_W^2)$ is the random effect for participant 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 [15]. Specifically, we choose

$$\mu \sim \mathcal{N}(0.2, 0.25^2) \tag{18}$$

$$\sigma_W^2 \sim \Gamma^{-1}(50, 45) \tag{19}$$

$$\rho \sim Beta(1.6, 30.4) \tag{20}$$

These prior distributions are illustrated in Figure 6. We assume that all parameters are independent, giving a joint prior distribution which is the product of each individual prior distribution.

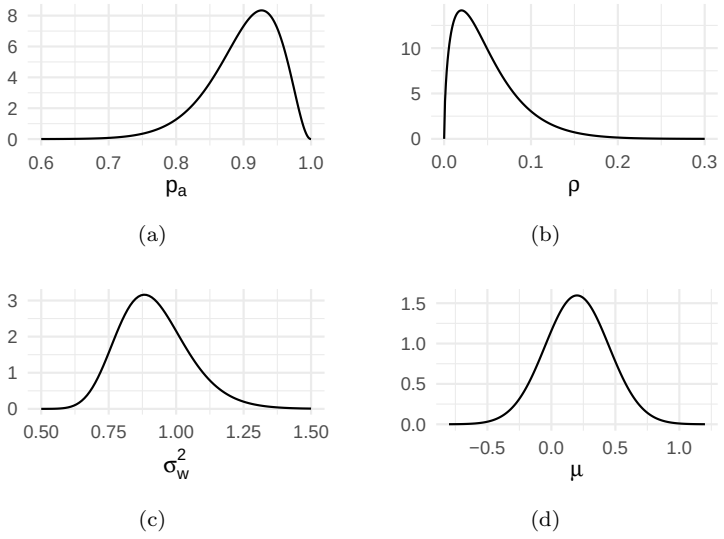

Figure 6: Second group of subfigures.

Incorporating adherence and efficacy parameters into the hypotheses defined in Equation 16, these are now defined as

$$\phi \in \begin{cases} \Phi_R \text{ if } \mu_c < 9 \text{ or } p_f < 0.6 \text{ or } 1.6 - 0.1\mu_c > p_f \text{ or } p_a < 0.8 \text{ or } \mu < 0 \\ \Phi_A \text{ if } \phi \notin \Phi_R \ \& \ [\mu_c < 10 \text{ or } p_f < 0.7 \text{ or } 1.75 - 0.1\mu_c > p_f \text{ or } p_a < 0.9 \text{ or } \mu < 0.2] \\ \Phi_G \text{ otherwise.} \end{cases} \tag{21}$$

Given the random effects and the interaction between adherence and efficacy, a conjugate posterior analysis is no longer possible. Instead, posterior samples are generated using MCMC. We use Stan (via rstan) to do so, sampling 5000 samples in four chains. Discarding the first 2500 samples in each chain to allow for burn-in gives $M = 10^4$ samples in total. This is done for each of the $N = 10^4$ samples in the outer loop. Given the significant computational expense, we consider only the case of $k = 12$ clusters (in total). After generating the samples we apply the optimisation routine described in Section 3.3. We use the NSGA-II algorithm with the default settings implemented in the mco package, which returns an approximation set of 100 decision rules. The operating characteristics of these rules are plotted in Figure 7. For comparison we also conducted a simple grid search across the space of decision rules and evaluated the operating characteristics of each. These are plotted in Figure 8, where all dominated rules are plotted in grey and non-dominated rules in colour. Labelled in Figure 8 are the decision rules which always lead to the same decision, regardless of the posterior probabilities. These are labelled r, a and g.

From Figure 7 we immediately see the range of operating characteristics which are feasible, given the fixed sample size, and the trade-offs between these

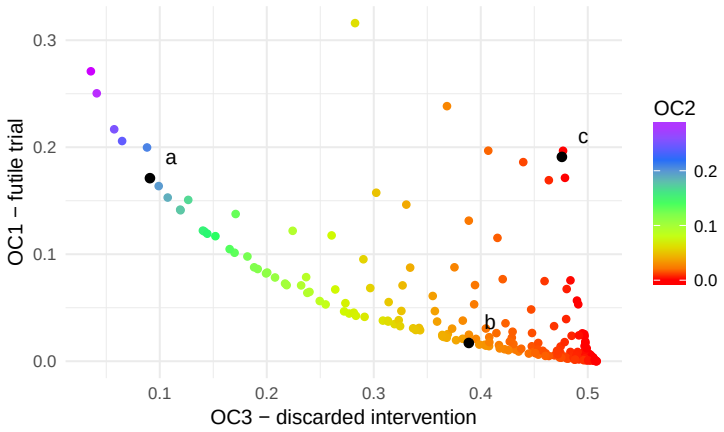

Figure 7: Operating characteristics of efficient decision rules for the REACH trial with $k = 12$ clusters.

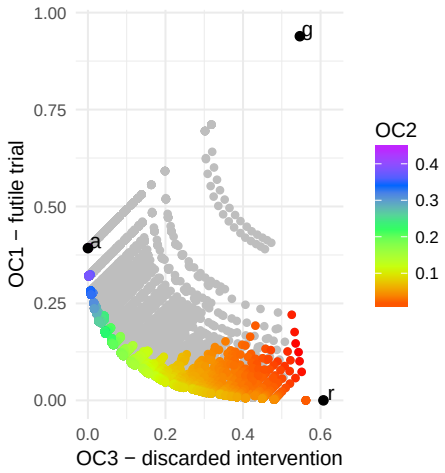

Figure 8: Operating characteristics of efficient decision rules for the REACH trial with $k = 12$ clusters - a grid search, where grey solutions are dominated.

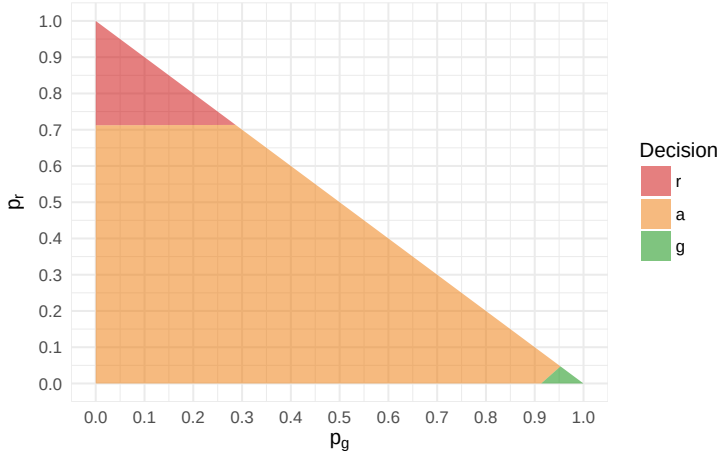

Figure 9: Example efficient decision rule (a).

which we must consider. For example, the results suggest that we need not consider values of OC_1, OC_2, and OC_3 greater than 0.38, 0.44, and 0.56 respectively. Moreover, rules which reduce one of the operating characteristics to 0 are found for each case.

In the rules of the approximation set we see an almost linear inverse relationship between OC_2 and OC_3. To understand why this is, note that if we assume the term $p_{G,r}$ in OC_3 is approximately 0, we are left with $OC_2 = p_{R,a} + p_{G,a}$ and $OC_3 \approx p_{A,r} + p_{A,g}$. Thus, these criteria are almost opposite: OC_2 is the probability of incorrectly making the decision a; OC_3 is the probability of incorrectly *not* making decision a. The larger the area corresponding to decision a in the decision rule, the larger OC_2 and the smaller OC_3 will be. This is illustrated when comparing the decision rules labelled (a) and (b) in Figure 7, shown in Figures 9 and 10 respectively. The former rule achieves a low $OC_3 \approx 0.025$ by having a relatively high threshold for making decision r. In contrast, rule (b) leads to decision r more easily, resulting in higher OC_3.

While the opportunity to trade-off performance between OC_1 and OC_3 is clear, there appears to be limited opportunity to improve in either of these at the expense of an increase in OC_2. This is particularly so for rules leading to a OC_3 value of less than around 0.2. Rule (c) in Figure 7 represents an opportunity to reduce OC_2 at the expense of one or both of OC_1 and OC_3. As illustrated in Figure 11, this is achieved by having a relatively small area corresponding to decision a, allowing larger areas for decisions r and g.

5 Discussion

The benefits brought by the Bayesian approach must be set against the challenges it brings, particularly in terms of computation time and operational feasi-

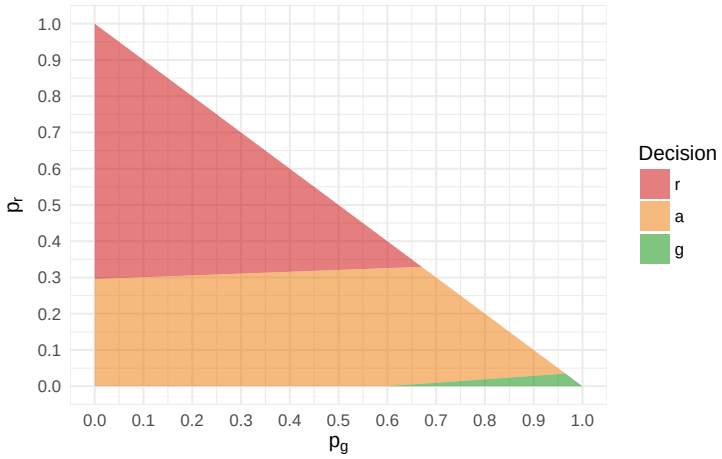

Figure 10: Example efficient decision rule (b).

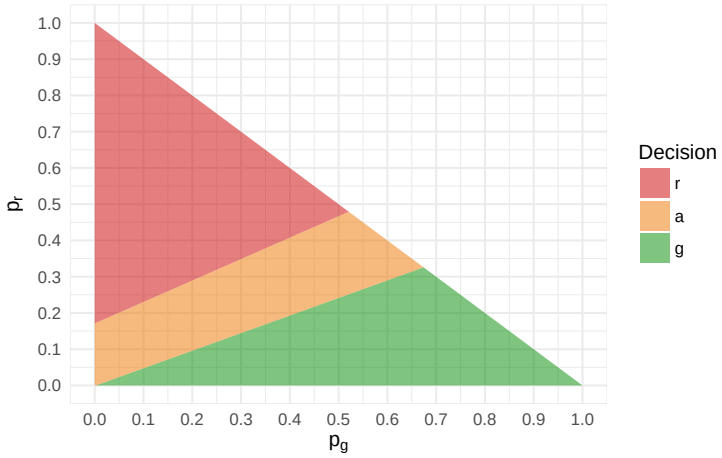

Figure 11: Example efficient decision rule (c).

		$\phi \in \Phi_R$	$\phi \in \Phi_A$	$\phi \in \Phi_G$
Decision	r	-	-	-
	a	-	-	-
	g	-	-	-

Table 4: Latent utility function (to complete).

bility. 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. Moreover, a decision rule is also required if one does not wish to optimise over these.

The specification of a joint prior distribution may be a particularly challenging and time-consuming task. Although some relevant data relating to similar parameters in different trials may be available, expert opinion will still be required to articulate the relevance of such data to the problem at hand. When no data at all are available, which is not unlikely given the early phase nature of pilot studies, expert opinion will be the only source of information. Elicitation of expert opinion can be conducted using an appropriate tool, such as the Sheffield Elicitation Framework (SHELF) [14]. The expense incurred through formal elicitation may make it hard to justify, particularly at the grant application stage. Where possible, the statistical model should be constructed with an aim to ensuring the parameters can be considered independent [13], thus allowing the full joint distribution to be defined as the product of the individual priors. Although we used the same prior distribution when simulating the trial data and when conducting the Bayesian inference on that data in our examples, An alternative approach would be to use separate priors for each stage [18]. For example, it may be that for the analysis to be accepted as 'objective', a vague or non-informative prior must be used at the analysis. Such a strategy is simple to carry out in the proposed method.

In addition to specifying priors, the parameter space must also be divided into three hypotheses. This may also prove challenging in practice, particularly when the number of parameters to be considered is greater than two. Further methodological research will be required to understand how best to elicit this information.

Although we have not described a fully decision-theoretic approach to the design and analysis of pilot trials, any decision rule of the form described in Section 3 may have an equivalent utility function such that using the rule will always lead to the same decision as maximising the utility function. For example, the decision rule illustrated in Section 1 is equivalent to choosing $i \in \mathcal{D}$ such that $\mathbb{E}_\phi[u(i, \phi)]$ is maximised, where $u(i, \phi)$ is a discrete function taking values as given in Table 4.

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. One potential approach

to improve efficiency is to use non-parametric regression to predict the result of a Bayesian inference, given some simulated data, thus bypassing the need to undertake a full MCMC analysis for each of the N samples in Equation 4. This approach has been applied successfully in the context of expected value of information calculations [16, 17]. In terms of optimal design, we have not considered here the question of finding the smallest sample size such that leads to sufficient operating characteristics. 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 [10].

A number of modifications could be made to the proposed methodology. In particular, any operating characteristic could be defined and used in addition to, or instead of, the three proposed in this paper, providing they can be defined in terms of the probabilities of Table 2. The multi-criteria optimisation methodology we have described will extend easily as the number of criteria are increased, although it should be noted that the presentation of solutions will become challenging.

Comparison with the frequentist approach. If we can get the WP1 paper published, we can take the same example and compare. Use a uniform prior over the frequentist hypothesis space?

Think about using independence ideas to help construct the hypotheses - e.g. preference independence, where our preferences between y and y' doesn't change as z changes to z'. That is not always the case here - it is something we have to assume in the frequentist case. But we can break the problem down using this - e.g. we might have trade-offs between follow-up and cluster size, and trade-offs between efficacy and adherence, but these two groups are preferentially independent from each other.

Connection to multiple testing literature as discussed in WP1. [**?**] - Challenge of multiple co-primary endpoints: a new approach. Discusses the 'reverse multiple testing problem', where we want all the endpoints to be good rather than any of them. Notes the problem with doing multiple univariate tests and proceeding if all are positive is that it can give low overall power. Argues that instead of controlling maximum type I error, we should control average type I error over the null space - so a kind of hybrid Bayesian approach.

Our design considers the probability of making an amber decision when in the amber hypothesis. We do not consider what that decision will actually look like. For example, the type of amber adjustments needed are very different in different parts of the hypothesis, so we might end up adjusting recruitment and leaving adherence when it is the other way round, and our design does not consider that as an error. This points to a significant challenge in pilot trials - the difficulty in modelling post-trial action, the richness of the decision space. Contrast with the drug setting where nothing much changes, and we just decide on a phase III sample size. Hence plenty papers looking at optimising the whole process, not just the phase II trial itself.

References

[1] 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.

[2] 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.

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

[4] 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.

[5] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, apr 2002.

[6] 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.

[7] 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.

[8] Michael TM Emmerich, André H Deutz, and Jan Willem Klinkenberg. Hypervolume-based expected improvement: Monotonicity properties and exact computation. In *Evolutionary Computation (CEC), 2011 IEEE Congress on*, pages 2147–2154. IEEE, 2011.

[9] 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.

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

[11] 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.

[12] Olaf Mersmann. *mco: Multiple Criteria Optimization Algorithms and Related Functions*, 2014. R package version 1.0-15.1.

[13] 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.

[14] 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*. 2006.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.