RESEARCH ARTICLE

Statistics in Medicine WILEY

Bayesian design and analysis of external pilot trials for complex interventions

Duncan T. Wilson[1] | James M. S. Wason[2,3] | Julia Brown[1] | Amanda J. Farrin[1] | Rebecca E. A. Walwyn[1]

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

[2]Institute of Health and Society, Newcastle University, Newcastle upon Tyne, UK

[3]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

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

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, although the statistical properties of such decision rules are rarely assessed. Such assessment is complicated by several methodological challenges, including the simultaneous evaluation of multiple endpoints, complex multi-level models, small sample sizes, and uncertainty in nuisance parameters. In response to these challenges, we describe a Bayesian approach to the design and analysis of external pilot trials. We show how progression decisions can be made by minimizing the expected value of a loss function, defined over the whole parameter space to 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 piecewise constant parametrization of the loss function, the parameters of which are chosen at the design stage to lead to desirable operating characteristics. We describe a flexible, yet computationally intensive, nested Monte Carlo algorithm for estimating operating characteristics. The method is used to revisit the design of an external pilot trial of a complex intervention designed to increase the physical activity of care home residents.

KEYWORDS
Bayesian, complex interventions, external pilot, pilot trials, sample size

1 | INTRODUCTION

Complex interventions, defined as those comprised of several interacting components,[1] can be challenging to evaluate in randomized controlled trials (RCTs) due to factors such as slow patient recruitment, poor levels of adherence to the intervention, and low completeness of follow-up data. To identify these problems prior to the main RCT we often conduct small trials[1] known as pilots. These typically take the same form as the planned RCT but with a considerably 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 with

a clear gap between the two trials. Pilot trials, which aim to inform the feasibility and optimal design of a subsequent definitive trial,[3] are distinct from phase II trials, which focus instead on assessing potential efficacy and safety.

The data generated by an external pilot trial are 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 United Kingdom, the National Institute for Health Research asks that these *progression criteria* are pre-specified and included in the research plan,[4] and the recent CONSORT extension to randomized pilot trials requires their reporting.[5] A single pilot trial can collect data on several progression criteria, often focused on the aforementioned areas of recruitment, protocol adherence, and data collection.[6] Although they may take the form of single threshold values leading to binary stop/go decision rules, investigators are increasingly using two thresholds to accommodate an intermediate decision between stopping altogether and progressing straight to the main trial, which would allow progression but only after some adjustments have been made.[5] The need for appropriate progression criteria is clear when we consider the consequences of poor post-pilot progression decisions. If the criteria are too lax, there is a greater risk that the main trial will go ahead but found to be infeasible and thus a waste of resources; if the criteria are too strict, a promising intervention may be discarded under the mistaken belief that the main trial would be infeasible. Despite this, there is little published guidance about how they should be determined.[6,7]

In addition to pre-specifying progression criteria, another key design decision is the choice of pilot sample size. Conventional methods of sample size determination, which focus on ensuring the trial will have sufficient power to detect a target difference in the primary outcome, are rarely used since they would lead to a pilot sample size comparable with the main trial sample size. Several methods for pilot sample size determination instead aim to provide a sufficiently precise estimate of the variance in the primary outcome measure to inform the sample size of the main trial.[8-13] Others have suggested a simple rule of thumb for when the goal is to identify unforeseen problems.[14] While some have noted that the low sample size in pilots may lead to a considerable probability that a certain progression criterion will be met (or missed) due to random sampling variation,[12,15] and despite the consequences of making the wrong progression decision, the statistical properties of pilot decision rules are rarely used to inform the choice of sample size. 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.[16]

In this article, we will describe a method for designing and analyzing external pilot trials which addresses these challenges. We take a Bayesian view, allowing for complex models to be estimated in the typically small sample context of pilot trials and for external information to be leveraged.[17] We propose progression decisions should then be made to minimize the expected value of a loss function with respect to a posterior distribution on model parameters. This decision-theoretic approach allows for the various trade-offs between model parameters to be expressed and guide progression decisions. By implicitly defining a pre-specified decision rule, the use of a loss function also ensures operating characteristics can be calculated and used as a basis for pilot trial sample size determination.

We propose a loss function with three parameters whose values can be determined either through direct elicitation of preferences or by considering the pilot trial 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.[18] 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.[19] 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 article is organized 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 optimizing 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 | Prior specification

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 Wang and Gelfand[20] and assume that two joint prior distributions of θ have been specified. First, the *analysis* prior $p_A(\theta)$ is that which will be used when fitting the model once the pilot data is obtained. It has been argued that regulators are unlikely to accept the prior beliefs of the trial sponsor for analysis of the data,[18,21] and as such a weakly or non-informative prior should be used

for $p_A(\theta)$ in order to "let the data drive the inference."[20] The choice of such a prior will depend on the specific model being used, although methodological guidance for various specific cases such as logistic regression[22] and hierarchical models[23] is available. It should be emphasized, however, that the typically small sample size of a pilot trial can mean the effect of the analysis prior is non-negligible. As such, the analysis prior should provide a credible and justifiable representation of prior ignorance, avoiding extreme default choices which may place too much prior weight on infeasible regions of the parameter space.

The *design* prior $p_D(\theta)$ will be used when evaluating the statistical performance of a proposed pilot trial design. It may be considered as purely hypothetical in the spirit of a "what-if" analysis,[20] in which case several candidate design priors may be suggested and performance evaluated under each of these. Alternatively, and as we will assume in the remainder of this article, $p_D(\theta)$ can be a completely subjective prior which fully expresses our knowledge and uncertainty in the parameters at the design stage. Although eliciting such a prior is potentially challenging, many examples describing successful practical applications of expert elicitation for clinical trial design are available,[19,21,24] as are tools for its conduct such as the Sheffield Elicitation Framework (SHELF).[25] From a strictly subjective Bayesian perspective, we can then view the weakly informative analysis prior as representing the beliefs of the person who will analyze the data and who is relatively uninformed with regards to the model parameters.

2.2 | Analysis and progression decisions

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:

- red—discard the intervention and stop all future development or evaluation;
- amber—proceed to the main RCT, but only after some modifications to the intervention, the planned trial design, or both; or
- green—proceed immediately to the main RCT.

In what follows we will denote these decisions by r, a, and g, respectively. We assume that our preferences between the three possible decisions are influenced by ϕ but independent of ψ, formalizing the separation of θ into substantive and nuisance components. We partition the substantive parameter space Φ into three disjoint subspaces Φ_I, for $I = R, A, G$. Each subspace label corresponds to the decision we would make if we knew the true value of ϕ. For example, if $\phi \in \Phi_R$ then the optimal decision is r(ed)-halt development and do not proceed to a definitive trial. We will henceforth refer to these three subsets as *hypotheses*, and to conditioning on the event $\phi \in \Phi_I$ as "under hypothesis Φ_I." Throughout, we will distinguish hypothesis I from the corresponding optimal decision i by using upper 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 make three kinds of mistakes: proceed to an infeasible 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. The occurrence of error j will be denoted by $E_j = 1$, otherwise $E_j = 0$. An error's occurrence will be a function of the decision made d and the true parameter value ϕ, that is, $E_j(d, \phi) : \{r, a, g\} \times \Phi \rightarrow \{0, 1\}$ for $j = 1, 2, 3$. We then 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, defined as

$$L(d, \phi) = c_1 E_1(d, \phi) + c_2 E_2(d, \phi) + c_3 E_3(d, \phi).$$

Note that the additive form of the loss function implies that the our preferences for any one of the attributes E_1, E_2, E_3 are independent of the values taken by the others.[26]

To determine appropriate values of the parameters c_1, c_2, c_3, we first scale the loss function by setting $c_1 + c_2 + c_3 = 1$. Thus, a loss of 0 is obtained if no errors occur, and a loss of 1 is obtained if all errors occur (although note that this is not possible in this setting). We then follow the procedure described by French and Rios Insua (page 99),[26] eliciting some judgments from the decision-maker(s) and using these to determine the values of c_1, c_2, c_3. One such judgment involves a simple gamble of obtaining the event $(E_1 = 0, E_2 = 0, E_3 = 0)$ with probability $1 - p_1$ and the event $(E_1 = 1, E_2 = 0, E_3 = 1)$ with probability p_1. The decision-maker is asked to compare this gamble against an alternative of obtaining the event $(E_1 = 1, E_2 = 0, E_3 = 0)$ for certain, and to adjust the value of p_1 until they feel indifferent between the two options.

TABLE 1 Losses associated with each decision under each hypothesis

		Hypothesis		
		$\phi \in \Phi_R$	$\phi \in \Phi_A$	$\phi \in \Phi_G$
Decision	r	0	c_2	c_2
	a	$c_1 + c_3$	0	c_3
	g	c_1	$c_1 + c_2$	0

Since this indifference implies the expected losses of the two options are equal, we will then have

$$p_1(c_1 + c_3) = c_1.$$

Similarly, we can ask the decision-maker(s) to consider a gamble between the event $(E_1 = 0, E_2 = 0, E_3 = 0)$ with probability $1 - p_2$ and the event $(E_1 = 1, E_2 = 1, E_3 = 0)$ with probability p_2, and compare this against the option of obtaining $(E_1 = 1, E_2 = 0, E_3 = 0)$ for certain. Again, by determining the value of p_2 which corresponds to indifference and thus equal expected loss, we deduce that

$$p_2(c_1 + c_2) = c_1.$$

This gives three equations that can be solved to obtain

$$c_1 = \frac{-p_1 p_2}{p_1 p_2 - p_1 - p_2}, \quad c_2 = \frac{p_1 p_2 - p_1}{p_1 p_2 - p_1 - p_2} \quad c_3 = \frac{p_1 p_2 - p_2}{p_1 p_2 - p_1 - p_2}.$$

Note that the two specific judgments suggested here are only two of many possible similar questions which could be posed to the decision-maker(s). It is recommended that more indifferences are elicited in order to seek out any inconsistencies and further clarify their true preferences.

The loss function will then take values as given in Table 1. For example, suppose we make a "green" decision under the "amber" hypothesis. The subsequent trial will be infeasible 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 maximizing expected utility (or in our case, minimizing the expected loss) when making a progression decision. We first use the pilot data in conjunction with the analysis prior $p_A(\theta)$ to obtain a posterior $p(\phi \mid x)$, and then choose the decision i^* such that

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

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

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

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

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

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

For some simple models that admit a conjugate analysis, 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 $\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(M)} \sim p(\phi \mid x)$,

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

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

2.3 | 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:

- $OC_1 = \mathbb{E}_{p_D}[E_1] = Pr[a \ \& \ \phi \in \Phi_R] + Pr[g \ \& \ \phi \in \Phi_R \cup \Phi_A]$ - probability of proceeding to an infeasible main RCT;
- $OC_2 = \mathbb{E}_{p_D}[E_2] = Pr[r \ \& \ \phi \in \Phi_A \cup \Phi_G] + Pr[g \ \& \ \phi \in \Phi_A]$ - probability of discarding a promising intervention;
- $OC_3 = \mathbb{E}_{p_D}[E_3] = Pr[a \ \& \ \phi \in \Phi_R \cup \Phi_G]$ - probability of making unnecessary adjustments to the intervention or the trial design.

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 dataset, we then apply the analysis and decision-making procedure described in Section 2.2, using some vector **c** to parametrize the loss function. This results in N decisions $i^{(k)}$ which can be contrasted with the corresponding true parameter value $\theta^{(k)}$ and in which hypothesis it resides, noting if any of the three types of errors have 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 $p(1-p)/N$.[*]

2.4 | Eliciting loss parameters through optimization

Elicitation of the loss function parameters $\mathbf{c} = (c_1, c_2, c_3)$ in the manner described in Section 2.2 may be challenging, particularly when multiple decision-makers are involved.[27] An alternative way to determine **c** is through examining the operating characteristics it leads to (for some fixed pilot design). As **c** is adjusted, the balance between the conflicting objectives of minimizing each OC will change, and the task is then to find the **c** which returns the best balance from the perspective of the decision-maker. Formally, and thinking of operating characteristics as functions of **c**, we wish to solve the multi-objective optimization problem

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

where $C = \{c_1, c_2 \in [0, 1] \mid c_1 + c_2 \leq 1\}$.

Since the three objectives are in conflict, there will be no single solution which simultaneously minimizes each one. We would instead like to find a set $C^* = \{\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \ldots, \mathbf{c}^{(K)}\}$ such that each member provides a different balance between minimizing the three operating characteristics. If there exist $\mathbf{c}, \mathbf{c}' \in C^*$ 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 **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 our set C^*. Because the search space C has only two dimensions, problem (7) can be approximately solved by generating a uniform random sample of **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 remain can be illustrated graphically. The decision-maker(s) can then view the range of available options, all providing different trade-offs among the three operating characteristics, and choose from among 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 **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.

---

[*]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 dataset, 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 same MCMC algorithm, using the same seed in the random number generator.

This approach not only ensures that optimization 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 Behavior problems) was a two-arm, individually-randomized, 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 60 primary carer-child dyads, to be randomized 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 (denoted f) using a binomial distribution with parameter p_f, and similarly the number successfully adhering to the intervention (denoted a) with a binomial distribution with parameter p_a. For a fixed pilot trial per-arm sample size n, the parameters of the model are $\phi = (p_f, p_a)$, with no nuisance parameters. Assuming for simplicity that the numbers followed-up and adhering are independent, the likelihood is then

$$p(f, a | p_f, p_a) = \left[\binom{2n}{f} p_f^f (1 - p_f)^{2n-f} \right] \times \left[\binom{n}{a} p_a^a (1 - p_a)^{n-a} \right].$$

At the design stage, the follow-up rate p_f was thought to be somewhere in the range 62% to 92%, while the adherence rate p_a was thought to lie between 40% and 95%. We reflect these ranges of uncertainty in our design priors by using beta distributions $p_f \sim Beta(40, 10)$ (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.

TIGA-CUB's progression criteria included only simple stop/go thresholds, with no intermediate "amber" decisions. As such, in this example, we partition the parameter space into two hypotheses, Φ_G and Φ_R. For the purposes of illustration, 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. The prior distributions on parameters p_f and p_a imply an *a priori* probability of 0.28 that $\phi \in \Phi_G$, that is, that both follow-up and adherence are sufficiently high.

In this special case, the loss function is

$$L(d, \phi) = c_1 E_1(d, \phi) + c_2 E_2(d, \phi)$$

and the expected losses of decisions g and r will be $\mathbb{E}_{\phi|x}[L(g, \phi)] = c_1 p_R$ and $\mathbb{E}_{\phi|x}[L(r, \phi)] = c_2 p_G$, where $p_R + p_G = 1$ and $c_1 + c_2 = 1$. Decision g is therefore optimal whenever $p_G > c_1$. The posterior probability p_G can be easily calculated given the pilot data due to the beta prior distributions being conjugate. Specifically, given a total sample size $2n$ 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

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

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 an infeasible trial (OC_1),

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

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

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.

FIGURE 1 Probabilities of an infeasible 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 $n = 30$
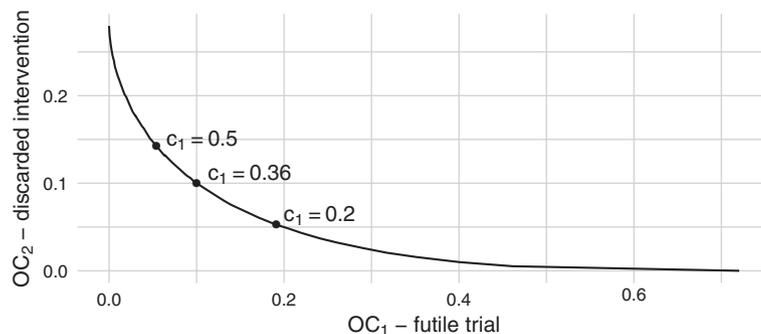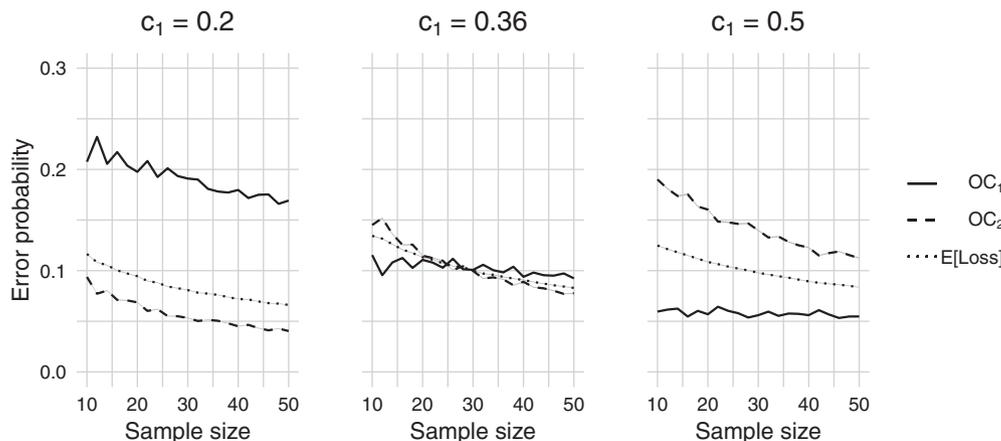

FIGURE 2 Probabilities of an infeasible main trial (OC_1) and of discarding a promising intervention (OC_2) for a range of per-arm sample sizes and different values of the loss parameter c_1

Keeping the sample size fixed at $n = 30$ per arm, we estimated the operating characteristics using a range of cost parameters values $c_1 = 0, 0.02, 0.04, \dots, 1$ using $N = 10^6$ Monte Carlo samples. The results are plotted in Figure 1, with some specific values of c_1 highlighted. The decision-maker can decide which point on the operating characteristic curve best reflects their own priorities in terms of the two types of error. For example, if the consequences of running an infeasible main RCT are considered less important than those of needlessly discarding a potentially effective intervention, the decision-maker may choose to set $c_1 = 0.2$ and would obtain $OC_1 = 0.19, OC_2 = 0.05$.

To examine the effect of adjusting the sample size, we evaluated the operating characteristics obtained for $n = 10, 12, 14, \dots, 50$ per arm whilst setting $c_1 = 0.2, 0.36, 0.5$. The results are shown in Figure 2. Each line includes a shaded area denoting the 95% Monte Carlo error intervals, although these are so small as to be illegible given the high number ($N = 10^6$) of MC samples used for each calculation. Although operating characteristics generally improve as the sample size is increased, we see that for $c_1 = 0.36$ and 0.5 the probability of an infeasible main trial, OC_1, remains flat whilst OC_2 has a downward trend. As we would expect, the expected loss reduces smoothly as n increases in all cases. In contrast, there is some variability beyond that explained by MC error in the OCs. This can be explained by the discrete nature of simulated adherence and follow-up data. Our results show that, for the design priors and hypotheses used in this example, the chosen sample size in TIGA-CUB of $n = 30$ can provide error rates broadly in line with conventional type I and II error rates under the usual hypothesis testing framework.

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 complex intervention designed to increase the physical activity of care home residents.[28] The trial was cluster randomized at the care home level, with twelve care homes in total randomized equally between treatment as usual (TaU) and 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, or average cluster size); adherence (a binary

TABLE 2 Pre-specified progression criteria used in the original REACH design

Outcome	Red	Amber	Green
Recruitment (avg. per care home)	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%

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 potential 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 potential efficacy, as detailed in Table 2.

Denoting the size of the jth cluster by m_j and the number of care homes in each arm by k, we assume that cluster sizes are normally distributed, $m_j \sim N(\mu_c, \sigma^2), j = 1, \dots, 2k$. We further assume that the probability of a participant being followed-up is constant across clusters and arms, and that the total number follows a binomial distribution $f \sim Bin(\sum_{j=1}^{2k} m_j, p_f)$. The number of care homes which successfully adhere to the intervention is assumed to binomially distributed, $a \sim Bin(k, p_a)$.

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_{ij} of resident i in care home j is

$$y_{ij} = X_j \times Y_j \times \mu + u_j + \varepsilon_i. \tag{11}$$

Here, X_j is a binary indicator of care home j being randomized to the intervention arm, Y_j is a binary indicator of care home j successfully adhering to the intervention, μ is the mean treatment effect, $u_j \sim \mathcal{N}(0, \sigma_B^2)$ is the random effect for care home j, and $\varepsilon_i \sim \mathcal{N}(0, \sigma_W^2)$ is the residual for resident i. We parametrize the model using the intracluster correlation coefficient, $\rho = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$.

The parameters describing average cluster size, follow-up, and adherence rates, and mean treatment effect are of substantive interest when making progression decisions, giving $\phi = (\mu_c, p_f, p_a, \mu)$. The remainder are nuisance parameters, $\psi = (\sigma^2, \rho, \sigma_W^2)$.

4.1 | Prior and hypothesis specification

To begin specifying a model for the REACH trial, we first note that the four substantive parameters can be divided into two pairs. First, mean cluster size and follow-up rate relate to the amount of information which a confirmatory trial will gather. Second, potential 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 is accounted for. We expect that a degree of trade-off between adherence and potential 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 potential efficacy. Having specified hypotheses Φ^i_I, Φ^e_I for $I = R, A, G$, we then have

$$\phi \in \begin{cases} \Phi_R & \text{if } \phi^i \in \Phi^i_R \text{ or } \phi^e \in \Phi^e_R \\ \Phi_G & \text{if } \phi^i \in \Phi^i_G \text{ and } \phi^e \in \Phi^e_G \\ \Phi_A & \text{otherwise.} \end{cases} \tag{12}$$
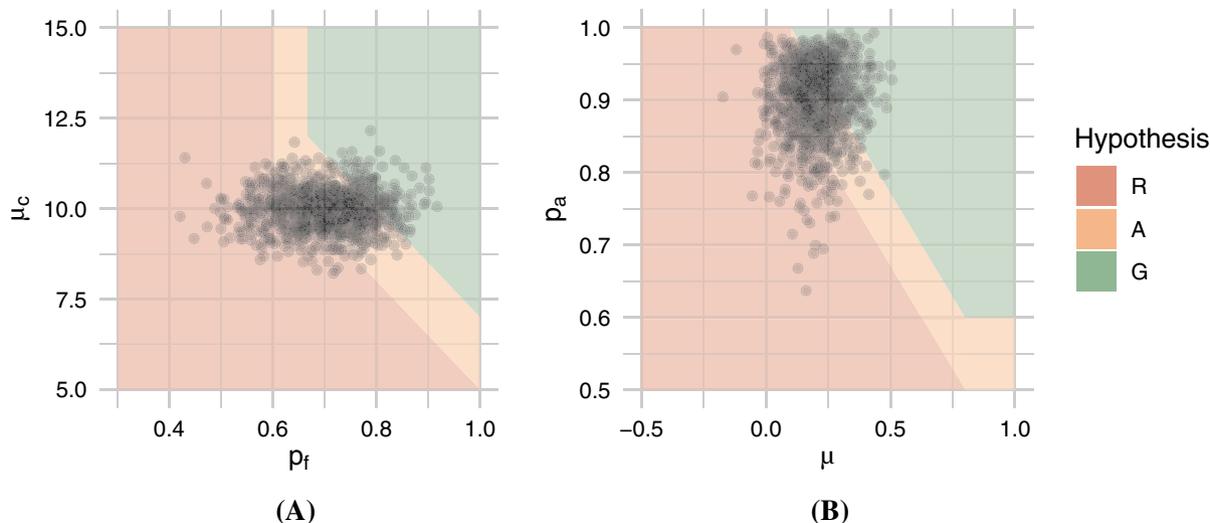
FIGURE 3 Marginal hypotheses over parameters for, A, follow-up rate p_f and mean cluster size μ_c; and, B, adherence rate p_a and potential efficacy μ. Each point is a sample from the joint prior distribution [Colour figure can be viewed at wileyonlinelibrary.com]

4.1.1 | Follow-up and cluster size

Recall that cluster sizes are assumed to be normally distributed with mean μ_c and variance σ^2. A normal-inverse-gamma prior

$$\sigma^2 \sim \Gamma^{-1}(\alpha_0, \beta_0), \ \mu_c \sim N(\mu_0, \sigma^2/\nu_0) \tag{13}$$

is placed on the mean and variance to allow for prior uncertainty in both parameters. It was anticipated that an average of 8 to 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.

For the probability of successful follow-up, p_f, we take a Beta distribution with hyper-parameters $\alpha_0 = 22.4, \beta_0 = 9.6$ as the prior. This gives a prior with a mean of 0.7 and a standard deviation of 0.08.

To partition the parameter space into hypotheses, we first consider the case where follow-up is perfect, that is, $p_f = 1$. 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 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:

$$(p_f, \mu_c) \in \begin{cases} \Phi_R^i & \text{if } p_f < 0.6 \text{ or } 20 - 15p_f > \mu_c \\ \Phi_G^i & \text{if } p_f > 0.66 \text{ and } 22 - 15p_f < \mu_c \\ \Phi_A^i & \text{otherwise.} \end{cases} \tag{14}$$

The hypotheses are illustrated in Figure 3A. 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_R^i, Φ_A^i and Φ_G^i. We have plotted 1000 samples from the prior in Figure 3A, falling into hypotheses Φ_R^i, Φ_A^i, and Φ_G^i in proportions 0.354, 0.517, and 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 potential efficacy

Having defined priors and hypotheses with respect to cluster size and follow-up, we now consider adherence and potential efficacy. Recall that 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 and a standard deviation of 0.05.

For the continuous measure of physical activity, we place priors on the mean effect μ, the intracluster correlation coefficient ρ, and the within-cluster variance σ_W^2 in the manner suggested by Spiegelhalter.[23] Specifically, we choose

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

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

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

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.

While there is potential for adherence to be improved after the pilot, we assume there will be little opportunity to improve the potential efficacy of the intervention. Moreover, we suppose an absolute improvement in adherence of up to around 0.1 is feasible. To define the hypotheses in this subspace, we first set a minimal level of potential 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 potential 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 potential 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

$$(p_a, \mu) \in \begin{cases} \Phi_R^e & \text{if } p_a < 0.5 \text{ or } 0.96 - 0.57\mu > p_a \\ \Phi_G^e & \text{if } p_a > 0.6 \text{ and } 1.06 - 0.57\mu < p_a \\ \Phi_A^e & \text{otherwise.} \end{cases} \tag{18}$$

The hypotheses are illustrated in Figure 3B. Again, a sample of size 1000 from the joint marginal prior distribution $p(p_a, \mu)$ is also plotted, falling into hypotheses $\Phi_R^e, \Phi_A^e,$ and Φ_G^e in proportions 0.234, 0.470, and 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 (12). Considering the same 1000 samples from the design prior plotted in Figure 3, these now fall into the regions $\Phi_R, \Phi_A,$ 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

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.[†] We took the sample size of the trial to be $k = 6$ clusters per arm. For calculating operating characteristics we generated $N = 10^4$ samples from the joint distribution $p(\theta, x) = p(x|\theta)p_D(\theta)$. We analyzed each simulated dataset using Stan via the R package rstan,[29] in each case generating 5000 samples in four chains and discarding the first 2500 samples

---

[†]Full details of the weakly informative prior are given in the supplementary material (see data availability statement).

FIGURE 4 Operating characteristics of the example pilot trial for a range of loss parameter vectors, when a weakly informative analysis prior is used [Colour figure can be viewed at wileyonlinelibrary.com]
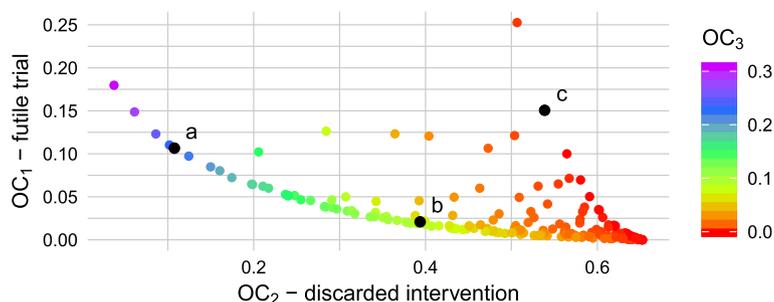

TABLE 3 Estimated operating characteristics (with standard errors) of the REACH trial for the three loss parameter vectors highlighted in Figure 4, when a weakly informative analysis prior is used

Label	(c_1, c_2, c_3)	OC_1	OC_2	OC_3
a	(0.07, 0.9, 0.03)	0.107 (0.003)	0.108 (0.003)	0.232 (0.004)
b	(0.18, 0.58, 0.24)	0.021 (0.001)	0.394 (0.005)	0.08 (0.003)
c	(0.01, 0.29, 0.7)	0.151 (0.004)	0.539 (0.005)	0.002 (0)

Note: Costs have been rounded to 2 decimal places; operating characteristics and their errors to 3.

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 dataset.

We evaluated the operating characteristics for a sample of parameters (c_1, c_2, c_3) as described in Section 2.4. A total of 254 parameter vectors were evaluated, of which 62 led to operating characteristics which were worse in every respect than some other vector (ie, dominated) and 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 an infeasible trial, (OC_1), against OC_2. There is a very limited opportunity to minimize 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 0.078 in comparison to point b, but only at the expense of increase in OC_1 and OC_2 of 0.13 and 0.145, respectively.

We would expect to see a clear relationship between the value of parameters 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, OC_2 (see center 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 $1 - c_2$ 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.

To illustrate the effect of varying sample size in the REACH trial, we set the loss function parameters to that of point a in Figure 4 and Table 3, $(c_1, c_2, c_3) = (0.07, 0.9, 0.03)$. We then estimated the operating characteristics obtained for $k = 6, 12, 18$ clusters per arm. Note that we considered only three choices of sample size due to the significant computational burden of each evaluation. The results are plotted in Figure 6. Increasing the sample size appears to have little effect on OC_1 and OC_3, while leading to a decrease in OC_2, the probability of discarding a promising intervention. This behavior reflects the priorities encoded by the costs parameter, where $c_2 = 0.9$.

4.2.2 | Incorporating subjective priors

Rather than use weakly or non-informative priors when analyzing the pilot data, we may instead want to make use of the (subjective) elicited knowledge of parameter values described in the design prior $p_D(\theta)$. Anticipating criticisms of a fully
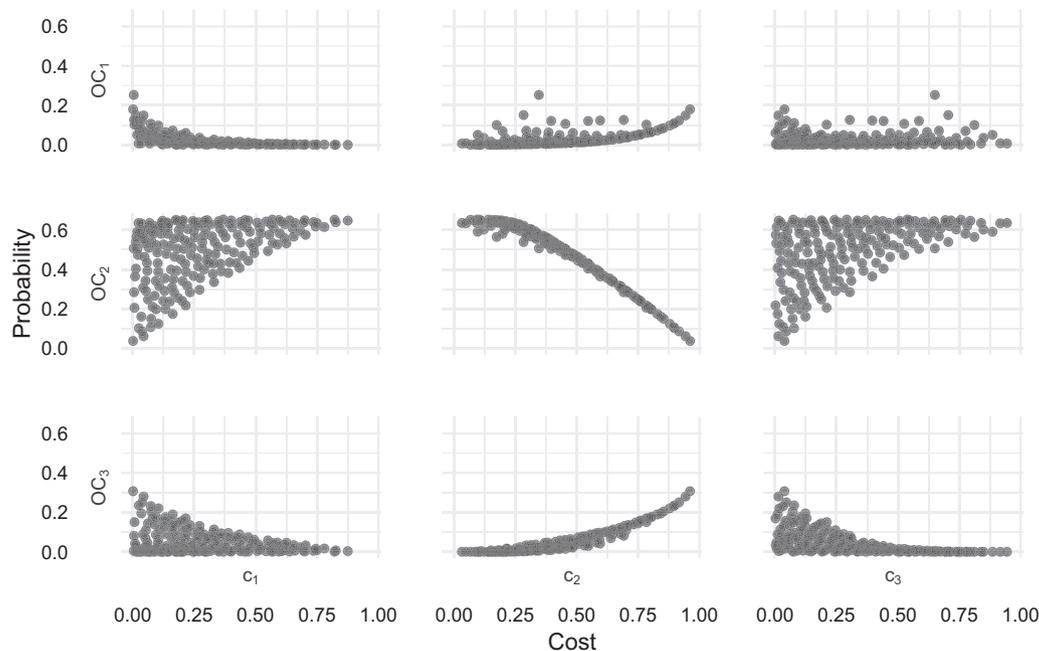
FIGURE 5 Relationships between the three loss parameters (x axes) and resulting operating characteristics (y axes)
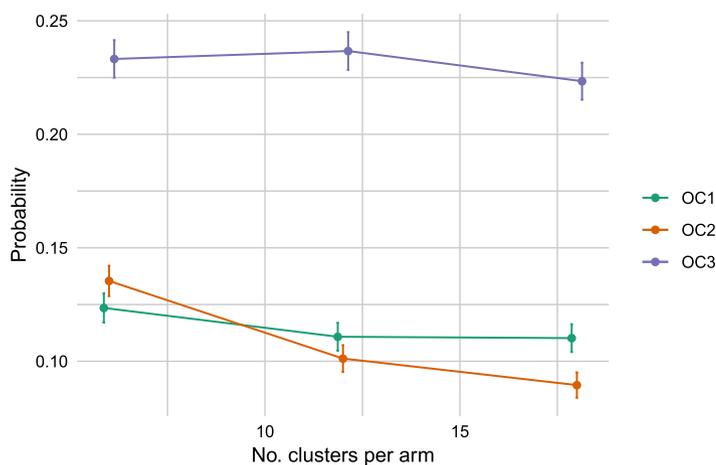

FIGURE 6 Operating characteristics of the REACH trial for per-arm sample sizes $k = 6, 12, 18$ and setting $(c_1, c_2, c_3) = (0.069, 0.116, 0.815)$. Error bars denote 95% confidence intervals. All points have been adjusted horizontally to avoid overlap [Colour figure can be viewed at wileyonlinelibrary.com]

subjective analysis, we can envisage two particular cases where this might be appropriate. First, using the components of the design prior which describe the nuisance parameters ψ while maintaining weakly informative priors on substantive parameters ϕ. Second, 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 at the care home level and only 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 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 (Figure 7).

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 supplementary material). When we contrast cases WI and INA, however, there is

FIGURE 7 Operating characteristics and expected utilities for weakly (WI) and partially informative (INA) [Colour figure can be viewed at wileyonlinelibrary.com]
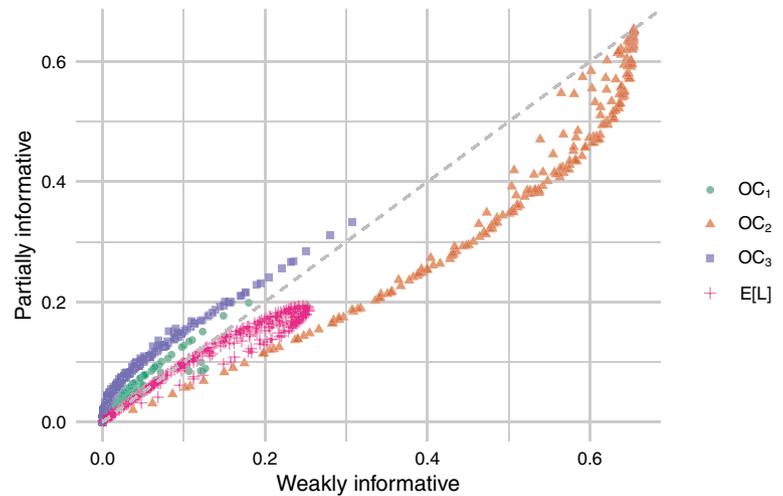
a clear distinction. Using the INA analysis prior will lead to larger probabilities of an infeasible trial (OC_1) and of unnecessary adjustment (OC_3), while reducing the probability of discarding a promising intervention (OC_2), for almost all loss parameters. The expected loss is always lower for the INA analysis than for WI, as we would expect.

5 | DISCUSSION

When deciding if and how a definitive RCT of a complex intervention should be conducted, and basing this decision on an analysis of data from a small pilot trial, there is a risk we will inadvertently make the wrong choice. A Bayesian analysis of pilot data followed by decision-making based on a loss function can help ensure this risk is minimized. 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, 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 point estimates.

Our proposed design is related to the literature on assurance calculations for clinical trials,[18] applying the idea of using unconditional event probabilities as operating characteristics to the pilot trial setting. In doing so we have shown how assurances can be defined for multiple substantive parameters with trade-offs between them, and with respect to the "traffic light" red/amber/green decision structure commonly found in pilot trials. The multi-objective optimization framework we have used to inform trial design allows the decision-maker to explicitly consider the different trade-offs between operating characteristics which are available, and select that which best reflects their own preferences. A similar approach has been taken in the context of phase II trials using the statistical concept of admissible designs.[30,31] This can be contrasted with the conventional and much criticized approach common in the frequentist context, where arbitrary constraints are placed on type I and II error rates in order to define a single optimal design.[32]

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,[19,21,24] as are tools for its conduct such as the Sheffield Elicitation Framework (SHELF).[25] 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 elicitation, such as those set out by Keeney and Raiffa,[27] 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 conjugate analysis to be used when sampling from the posterior distribution. We have used a nested Monte Carlo sampling scheme to estimate operating characteristics, as seen elsewhere.[18,20,33] 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.[34,35] The computational difficulties will be particularly pertinent when using our approach to determine sample size, as several evaluations of different sample size choices will be required. If the choice of sample size can be framed as an optimization problem, methods for efficient global optimization of computationally expensive functions such as those described by Jones[36] and Roustant et al.[37] may be useful.[16] Alternatively, one of several rules-of-thumb for choosing pilot sample size[3,9,11,13] could be used, with the resulting operating characteristics evaluated using the proposed method. Volumes.

We have defined our procedure in terms of a loss function, where the decision-making following the pilot will minimize the 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 exactly 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,[26] leading to a fully decision-theoretic approach to design and analysis.[38] However, as previously noted by others,[39-41] implementation of these approaches has been limited in practice and this may be indicative of their feasibility.

The proposed method could be extended in several ways. More operating characteristics could be defined and used in design optimization, 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 and how it should relate to the observed pilot data. Methodology for jointly modeling a pilot and subsequent main RCT in this manner could be informed by developments for designing phase II/III programs in the drug setting.[42-45]

DATA AVAILABILITY STATEMENT

All simulated data used in this manuscript, together with the code used to generate it, is available at https://github.com/DTWilson/Bayesian_pilot.

ORCID

Duncan T. Wilson https://orcid.org/0000-0001-7949-8718

REFERENCES

1. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new medical research council guidance. *BMJ*. 2008;337:a1655. https://doi.org/10.1136/bmj.a1655.
2. Eldridge SM, Lancaster GA, Campbell MJ, et al. Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. *PLoS One*. 2016;11(3):e0150205. https://doi.org/10.1371/journal.pone.0150205.
3. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *J Eval Clin Pract*. 2004;10(2):307-312. https://doi.org/10.1111/j..2002.384.doc.x.

4. National Institute for Health Research Research for Patient Benefit (RfPB) programme guidance on applying for feasibility studies; 2017.

5. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *BMJ*. 2016;355:i5239. https://doi.org/10.1136/bmj.i5239.

6. Avery KNL, Williamson PR, Gamble C, et al. Informing efficient randomised controlled trials: exploration of challenges in developing progression criteria for internal pilot studies. *BMJ Open*. 2017;7(2):e013537. https://doi.org/10.1136/bmjopen-2016-013537.

7. Hampson LV, Williamson PR, Wilby MJ, Jaki T. A framework for prospectively defining progression rules for internal pilot studies monitoring recruitment. *Stat Methods Med Res*. 2017;27(12):0962280217708906. https://doi.org/10.1177/0962280217708906.

8. Browne RH. On the use of a pilot sample for sample size determination. *Stat Med*. 1995;14(17):1933-1940. https://doi.org/10.1002/sim.4780141709.

9. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. *Pharm Stat*. 2005;4(4):287-291. https://doi.org/10.1002/pst.185.

10. Sim J, Lewis M. The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency. *J Clin Epidemiol*. 2012;65(3):301-308. https://doi.org/10.1016/j.jclinepi.2011.07.011.

11. Teare M, Dimairo M, Shephard N, Hayman A, Whitehead A, Walters S. Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials*. 2014;15(1):264. https://doi.org/10.1186/1745-6215-15-264.

12. Eldridge SM, Costelloe CE, Kahan BC, Lancaster GA, Kerry SM. How big should the pilot study for my cluster randomised trial be? *Stat Methods Med Res*. 2015;25(3):1039–1056. https://doi.org/10.1177/0962280215588242.

13. Whitehead AL, Julious SA, Cooper CL, Campbell MJ. 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. *Stat Methods Med Res*. 2015;25(3):1057–1073. https://doi.org/10.1177/0962280215588241.

14. Viechtbauer W, Smits L, Kotz D, et al. A simple formula for the calculation of sample size in pilot studies. *J Clin Epidemiol*. 2015;68(11):1375-1379. https://doi.org/10.1016/j.jclinepi.2015.04.014.

15. Cooper CL, Whitehead A, Pottrill E, Julious SA, Walters SJ. Are pilot trials useful for predicting randomisation and attrition rates in definitive studies: a review of publicly funded trials. *Clin Trials*. 2018;15(2):189–196. https://doi.org/10.1177/1740774517752113.

16. Wilson DT, Walwyn RE, Brown J, Farrin AJ, Brown SR. Statistical challenges in assessing potential efficacy of complex interventions in pilot or feasibility studies. *Stat Methods Med Res*. 2015;25(3):997-1009. https://doi.org/10.1177/0962280215589507.

17. Willan AR, Thabane L. Bayesian methods for pilot studies. *Clin Trials*. 2020;17(4):414-419. https://doi.org/10.1177/1740774520914306.

18. O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. *Pharm Stat*. 2005;4(3):187-201. https://doi.org/10.1002/pst.175.

19. Crisp A, Miller S, Thompson D, Best N. Practical experiences of adopting assurance as a quantitative framework to support decision making in drug development. *Pharm Stat*. 2018;17(4):317–328. https://doi.org/10.1002/pst.1856.

20. Wang F, Gelfand AE. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Stat Sci*. 2002;17(2):193-208.

21. Walley RJ, Smith CL, Gale JD, Woodward P. Advantages of a wholly Bayesian approach to assessing efficacy in early drug development: a case study. *Pharm Stat*. 2015;14(3):205-215. https://doi.org/10.1002/pst.1675.

22. Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. *Ann Appl Stat*. 2008;2(4):1360-1383. https://doi.org/10.1214/08-aoas191.

23. Spiegelhalter DJ. Bayesian methods for cluster randomized trials with continuous responses. *Stat Med*. 2001;20(3):435-452. https://doi.org/10.1002/1097-0258(20010215)20:3%3C435::AID-SIM804%3E3.0.CO;2-E.

24. Dallow N, Best N, Montague TH. Better decision making in drug development through adoption of formal prior elicitation. *Pharm Stat*. 2018;17(4):301–316. https://doi.org/10.1002/pst.1854.

25. O'Hagan A, Buck CE, Daneshkhah A, et al. *Uncertain Judgements: Eliciting Experts' Probabilities*. Hoboken, NJ: John Wiley and Sons; 2006.

26. French S, Insua DR. *Statistical Decision Theory. No. 9 in Kendall's Library of Statistics*. Oxford, UK: Oxford University Press; 2000.

27. Keeney RL, Raiffa H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Hoboken, UK: John Wiley & Sons; 1976.

28. Forster A, Airlie J. Research exploring physical activity in care homes (REACH): study protocol for a randomised controlled trial. *Trials*. 2017;18(1):182. https://doi.org/10.1186/s13063-017-1921-8.

29. Stan Development Team RStan: the R interface to Stan. R package version 214.1; 2016.

30. Jung SH, Lee T, Kim KM, George SL. Admissible two-stage designs for phase II cancer clinical trials. *Stat Med*. 2004;23(4):561-569. https://doi.org/10.1002/sim.1600.

31. Mander AP, Wason JM, Sweeting MJ, Thompson SG. Admissible two-stage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. *Pharm Stat*. 2012;11(2):91-96. https://doi.org/10.1002/pst.501.

32. Bacchetti P. Current sample size conventions: flaws, harms and alternatives. *BMC Med*. 2010;8(1):17. https://doi.org/10.1186/1741-7015-8-17.

33. Sutton AJ, Cooper NJ, Jones DR, Lambert PC, Thompson JR, Abrams KR. Evidence-based sample size calculations based upon updated meta-analysis. *Stat Med*. 2007;26(12):2479-2500. https://doi.org/10.1002/sim.2704.

34. Strong M, Oakley JE, Brennan A. Estimating multiparameter partial expected value of perfect information from a probabilistic sensitivity analysis sample. *Med Decis Mak*. 2014;34(3):311-326. https://doi.org/10.1177/0272989x13505910.

35. Strong M, Oakley JE, Brennan A, Breeze P. Estimating the expected value of sample information using the probabilistic sensitivity analysis sample: a fast, nonparametric regression-based method. *Med Decis Mak*. 2015;35(5):570-583. https://doi.org/10.1177/0272989x15575286.

36. Jones DR. A taxonomy of global optimization methods based on response surfaces. *J Glob Optim*. 2001;21(4):345-383. https://doi.org/10.1023/A:1012771025575.

37. Roustant O, Ginsbourger D, Deville Y. DiceKriging, DiceOptim: two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *J Stat Softw*. 2012;51(1):1-55.

38. Lindley DV. The choice of sample size. *J R Stat Soc Ser D (Stat)*. 1997;46(2):129-138. https://doi.org/10.1111/1467-9884.00068.

39. Joseph L, Wolfson DB. Interval-based versus decision theoretic criteria for the choice of sample size. *J R Stat Soc Ser D (Stat)*. 1997;46(2):145-149. https://doi.org/10.1111/1467-9884.00070.

40. Bacchetti P, McCulloch CE, Segal MR. Simple, defensible sample sizes based on cost efficiency. *Biometrics*. 2008;64(2):577-585. https://doi.org/10.1111/j.1541-0420.2008.01004_1.x.

41. Whitehead J, Valdés-Márquez E, Johnson P, Graham G. Bayesian sample size for exploratory clinical trials incorporating historical data. *Stat Med*. 2008;27(13):2307-2327. https://doi.org/10.1002/sim.3140.

42. Stallard N. Optimal sample sizes for phase II clinical trials and pilot studies. *Stat Med*. 2012;31(11-12):1031-1042. https://doi.org/10.1002/sim.4357.

43. Wason JMS, Jaki T, Stallard N. Planning multi-arm screening studies within the context of a drug development-program. *Stat Med*. 2013;32(20):3424-3435. https://doi.org/10.1002/sim.5787.

44. Götte H, Schüler A, Kirchner M, Kieser M. Sample size planning for phase II trials based on success probabilities for phase III. *Pharm Stat*. 2015;14(6):515-524. https://doi.org/10.1002/pst.1717.

45. Kirchner M, Kieser M, Götte H, Schüler A. Utility-based optimization of phase II/III programs. *Stat Med*. 2015;35(2):305-316. https://doi.org/10.1002/sim.6624.