Article:
Wilson, DT, Walwyn, REA, Brown, J et al. (2 more authors) (2016) Statistical challenges in assessing potential efficacy of complex interventions in pilot or feasibility studies. Statistical Methods in Medical Research, 25 (3). pp. 997-1009. ISSN 0962-2802

https://doi.org/10.1177/0962280215589507

Statistical challenges in assessing potential efficacy of complex interventions in pilot or feasibility studies

Duncan T. Wilson*[1], Rebecca E. A. Walwyn[1], Julia Brown[1], Amanda J. Farrin[1], and Sarah R. Brown[1]

[1]Clinical Trials Research Unit, Leeds Institute of Clinical Trials Research, University of Leeds, Leeds, LS2 9JT, UK

Abstract

Early phase trials of complex interventions currently focus on assessing the feasibility of a large RCT and on conducting pilot work. Assessing the efficacy of the proposed intervention is generally discouraged, due to concerns of underpowered hypothesis testing. In contrast, early assessment of efficacy is common for drug therapies, where phase II trials are often used as a screening mechanism to identify promising treatments. In this paper we outline the challenges encountered in extending ideas developed in the phase II drug trial literature to the complex intervention setting. The prevalence of multiple endpoints and clustering of outcome data are identified as important considerations, having implications for timely and robust determination of optimal trial design parameters. The potential for Bayesian methods to help to identify robust trial designs and optimal decision rules is also explored.

1 Introduction

Complex interventions contain several distinct and potentially interacting components, each of which may contribute to the efficacy of an intervention as a whole [21]. For example, psychotherapy may be viewed as being composed of two treatment variables, namely techniques described in a therapy manual together with a therapist delivering these techniques [71]. This contrasts with typical drug treatments, where the drug is the only treatment variable to consider. While drug regimens may be complex [46], randomisation and blinding allow the effects of a drug to be separated from the context in which it is provided. Broad classes of complex interventions include surgical, behavioural, psychological, educational and physical interventions. The evaluation of a complex intervention raises specific challenges, and several frameworks have therefore been proposed to guide this process. These include a widely used framework proposed by the MRC [15, 21]; the IDEAL initiative aimed at surgical interventions [5]; and the Multiphase Optimisation STrategy (MOST) [17, 18]. The most recent MRC framework is summarised in Figure 1.

As shown in Figure 1, 'feasibility and piloting' is identified as one of four key stages in the development and evaluation of complex interventions. While the definitions of, and distinctions between, feasibility and pilot studies are not always clear [3, 65, 7], the MRC guidance states that the purpose of this stage is to inform the design of a subsequent large, definitive trial assessing the effectiveness of the intervention. Several parameters required for designing the definitive trial may be estimated at this stage, including the variance of the proposed outcome measure(s), recruitment and follow up rates, and intra-class correlation coefficients (ICCs) in trials with clustering

*Corresponding author: D.T.Wilson@leeds.ac.uk

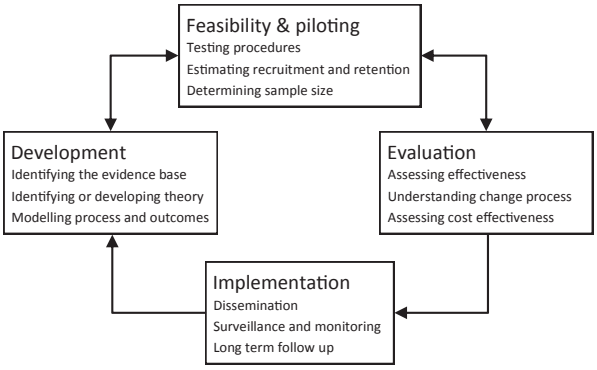


Figure 1: Current MRC guidance on the development and evaluation of complex interventions, adapted from [21].

effects. Characteristics relating more directly to the intervention, such as its acceptability and the level of adherence, may also be assessed. Gathering information relating to these factors reduces the likelihood of the large trial failing due to poor design.

While MRC guidance recommends evaluating a complex intervention following feasibility or pilot work, in practice it is not uncommon for feasibility or pilot studies to include evaluation through hypothesis testing. For example, a recent review found that 21 of 26 feasibility and pilot studies surveyed included a hypothesis test [3]. However, the size of these studies is often derived using generic rules of thumb [32, 10] rather than through formal power calculations, with the review finding that only 9 of the 26 studies reported a sample size calculation [3]. As a result, hypothesis tests are likely to be underpowered [36] and the typical recommendation is that such tests should be de-emphasized, interpreted with extreme caution, or avoided altogether [36, 4, 3, 38].

It could be argued, however, that a formal assessment of potential efficacy or activity *should* be carried out in pilot and feasibility studies, and that such studies should be properly designed to address this objective. In this manner, feasibility or pilot work could not only ensure that subsequent large scale randomised controlled trials (RCTs) of complex interventions are well designed, but could also reduce the rate at which such trials fail due to an inherently ineffective intervention. Moreover, this approach would clearly be more efficient than conducting a feasibility or pilot study and then a separate study assessing only efficacy. To begin developing such designs, one may look to methods developed in the drug setting. There, small 'phase II' trials which focus on making an early assessment of efficacy, identifying promising and discarding unpromising drug treatments, are commonplace.

The application of phase II designs to the complex intervention setting is not straightforward due to challenges that are commonly encountered in complex intervention trials. For example, the assumption that patient outcomes are statistically independent is often violated as a consequence of cluster randomisation [8], a group-based intervention [72], or therapist variation [48]. The associated implications for precision are compounded in cases where only a small number of clusters are available, as is often the case in feasibility or pilot studies [53]. Furthermore, the multi-component nature of complex interventions will mean that an assessment of efficacy will often have to be based on multiple endpoints [21], in contrast to the single binary indicator of 'success' often used in phase II studies.

An example which serves to illustrate each of these points is the OK-Diabetes (OK-D) feasibility trial of a supported self-management intervention for adults with type II diabetes and learning disabilities [28]. The OK-D study involves first developing a manualised intervention and then carrying out a randomised feasibility study whose objectives include estimation of recruitment rates, testing of data collection forms and assessment of the feasibility of delivering the intervention. While the feasibility study individually randomises treatment packages to patients, diabetes specialist nurses provide the intervention and may, therefore, induce a clustering effect in the intervention arm. Furthermore, as the intervention is newly developed only two nurses will be involved. The intervention is targeted at three aspects of poor diabetes self-management, and as a result there are a number of possible outcomes to be considered when assessing efficacy.

It has been proposed that the OK-D feasibility trial be extended to allow for a preliminary assessment of the efficacy of the developed intervention as a formal objective, highlighting the need for appropriate trial design methodology. In this paper we will review the approach to assessing efficacy developed in the context of phase II trials for drug therapies, setting out the key methodological challenges to their application in feasibility and pilot studies of complex interventions, and thereby outlining future directions for methodological research. In Section 2, an overview of phase II designs and their key characteristics will be provided. In Section 3 multiple endpoints and clustering will be discussed in detail, considering the formulation of decision rules, difficulties arising from nuisance parameters, and practical difficulties in determining sample size in a timely manner. Finally, in Section 4 conclusions are drawn and further avenues for future research are suggested.

2 Efficacy evaluation in oncology drug trials

Following the determination of a safe dose in phase I, but before a definitive RCT in phase III, phase II trials typically act as a screening mechanism to screen out ineffective drugs at an early stage and progress only the most promising treatments to phase III. Phase II designs tend to employ a decision-focussed approach to inference, with an emphasis on determining if a subsequent phase III trial is warranted as opposed to estimation of underlying parameters. This approach is typically sustained through the use of Neyman-Pearson hypothesis testing or, alternatively, through Bayesian decision-theoretic methods [41].

In the case of hypothesis testing, trial design focusses on ensuring type I and II error rates remain within pre-specified nominal bounds. Perhaps the simplest phase II design to employ hypothesis testing for a single binary outcome was proposed by Fleming [23], then extended by A'Hern [2] from an approximate to an exact test. To use the design, one must first specify a success rate p_0 which, if true, would mean the new intervention would not be worthy of further investigation. An alternative hypothesis p_A must then be given, corresponding to a success rate which would certainly merit a full evaluation in a definitive RCT. Applying this to the OK-D problem, we could set $p_0 = 0.05$ and $p_A = 0.2$. The A'Hern design for this problem, guaranteeing a type I error rate of 5% or less and a power of at least 90%, would be a single arm trial with a sample size of $n = 38$. Decision making at the end of the trial is then based on counting the number of successes observed, denoted s, and comparing this with the design-derived cut-off point c. In this example, if $s \geq c = 4$ the intervention should proceed to a definitive RCT, otherwise its evaluation should be terminated.

A wide range of alternative phase II designs have been published, accounting for the variety of problems to which they may be applied [9]. Only a brief overview of the main differences between designs, with references to examples, is considered here. One point of differentiation between the designs is in the number of stages. While the A'Hern [2] design described above involved a single decision point, designs such as those proposed by Simon [57] include an additional interim analysis to allow for the phase II trial to terminate early due to futility. Single arm designs may be contrasted with randomised designs [58], which allow a concurrent as opposed to historical control to be used. Multi-arm designs, for cases where multiple treatments are available for evaluation at once, have also been described [33]. While the majority of phase II designs focus on a single endpoint relating to efficacy, several have been proposed which can consider additional measures relating to, for example, toxicity [13, 63] or further aspects of efficacy [56].

In addition to hypothesis testing designs there are also a number which adopt a Bayesian framework. These vary in the extent to which Bayesian methodology is employed, from allowing some prior information to be incorporated in the form of probability distributions [64] to full decision-theoretic frameworks [12, 62]. The multitude of designs available requires a thorough assessment of the key design criteria specific to the trial in question, to ensure an appropriate design is selected.

3 Efficacy evaluation in complex intervention trials

When applying ideas from phase II trials in an early phase complex intervention setting, it is important to take account of complexities relating to (i) prevalence of multiple endpoints and (ii) recruitment- and treatment-related clustering effects.

3.1 Multiple endpoints

Multiple endpoints, on which the decision of proceeding to phase III should be based, arise due to several reasons. In addition to an assessment of efficacy requiring several endpoints, due to the multi-component nature of the intervention, endpoints relating to safety, acceptability and adherence are often required. Further to these, endpoints relating to the feasibility of a phase III trial, such as measurements of the recruitment and follow up rates, should be taken into account. Thus, while phase II drug trials are not always limited to a single endpoint, early phase evaluations of complex interventions may routinely involve more.

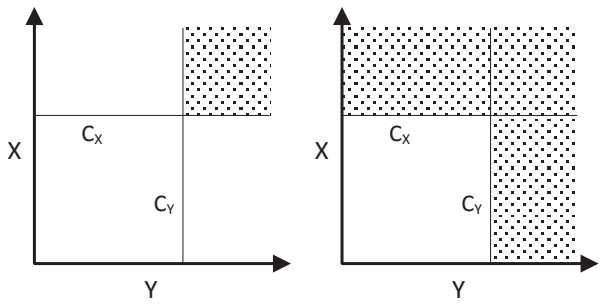
As an example, the original design of the OK-D feasibility study included three feasibility criteria which were to be met to consider progression to phase III. These took the form of threshold values of recruitment rate, numbers lost to follow up, and adherence of participants in the intervention arm. In addition to these three endpoints, a further four endpoints were of interest in terms of assessing efficacy. Specifically, continuous measurements of glycated haemoglobin (HbA1c), blood pressure, total cholesterol and body mass (BMI) are all proposed as potential efficacy endpoints, with no single one anticipated to be sensitive to all components of the intervention.

3.1.1 Decision rules

In the single endpoint case, a decision rule regarding progression to a phase III trial can be defined by a single cut-off point, as was illustrated in the example in Section 2. Where several endpoints are present, specifying the form of the decision rule for progression to phase III becomes more complex. This problem has been addressed to a limited extent in the drug setting. In the case of two binary endpoints describing efficacy and toxicity, phase II designs such as that of Bryant and Day [13] consider separate null and alternative hypotheses for the two endpoints, resulting in four 'states of nature'. Specifically, defining 'unacceptable' and 'acceptable' levels of both efficacy and toxicity as p_{E0}, p_{E1}, p_{T0}, p_{T1}, the four states are defined as $H_{ij} : p_E = p_{Ei}$, $p_T = p_{Tj}$ for $i, j = 0, 1$. The design aims to ensure that the probability of rejecting the drug when it has satisfactory efficacy *and* toxicity, i.e. the type II error, remains within a nominal bound. Two separate type I errors are also kept within nominal bounds, relating to the probability of proceeding to phase III with an ineffective *or* a toxic drug. The resulting design specifies a cut-off point for each endpoint, both of which must be reached for the drug to proceed to phase III. This can be illustrated graphically as an acceptance region, as shown in Figure 2a.

In cases where two binary measures of efficacy are of interest, phase II designs such as that of Sill et al. [56] employ a rule whereby we proceed to phase III if *either* quantity reaches the specified cut-off point. The form of the resulting acceptance region is illustrated in Figure 2b. Again, the form of this rule dictates the possible types of errors, with a single type I error rate in this case and two type II error rates. One advantage of these decision rules is the ability to discriminate between endpoints through their nominal error rates. For example, in the case of the Bryant & Day design, progressing to phase III with a toxic treatment may be considered more of a risk than progressing with an ineffective treatment, and so the nominal type I error rate relating to toxicity could be set to a lower level to ensure this error is less likely. Similarly, in the case of two efficacy endpoints and the Sill et al. design, one could set the nominal type II error of the preferred endpoint to be lower than the other to ensure the trial will be more likely to detect a treatment which is promising in this respect.

Beyond this use of nominal error rates, designs such as those of Bryant and Day [13] and Sill et al. [56] do not provide any means with which to describe any relative preferences between different qualities of the



(a) $x \geq C_x$ AND $y \geq C_y$ (b) $x \geq C_x$ OR $y \geq C_y$

Figure 2: Example decision rules arising from phase II designs for two endpoints, where the shaded area of the sample space represents the decision to progress to phase III.

treatment. With multiple endpoints to consider, it is possible that the decision to progress to phase III could involve trading off one aspect of the treatment against another. For example, one may be happy to accept a slightly toxic treatment if it demonstrated substantial efficacy, but not if efficacy was only moderate. Phase II designs allowing for such a trade-off were proposed by Conaway and Petroni [19, 20]. Considering binary efficacy and toxicity endpoints with parameters p_E and p_T, the authors propose dividing the parameter space $0 \leq p_E, p_T \leq 1$ into two complementary subspaces defining the null and alternative hypotheses. They propose a statistical test based on the 'I-divergence' measure [49], with the statistic being analogous to the distance of the observed sample from the null hypothesis subspace. Type I and II error rates are defined (the latter with respect to a point alternative hypothesis), and the choice of sample size made to ensure error rates remain within prespecified bounds. It is noted that the method may be applied to general specifications of the null hypothesis space, and is suggested that future research consider extending the design to allow for more general loss functions than the $0-1$ loss implicit in the proposed method. While providing more flexibility when specifying trade-offs between endpoints, in comparison to the design of Bryant and Day [13] this design has been shown to lack robustness to misspecification of the degree of correlation between them [67].

An alternative approach to acknowledging the presence of multiple endpoints is proposed by Sargent et al. [51]. In this phase II design, the decision space related to the trial is expanded from $\{stop, go\}$ to include a third, intermediate decision. Considering an explicit primary endpoint, if the corresponding observations are strong enough (in either direction) the trial will lead to one of *stop* or *go*. If the observations are less conclusive, it is suggested that the decision should now be made by considering other endpoints of interest. This design therefore provides a formal mechanism to allow for the inclusion of more than one endpoint without requiring any specification of their nature or relationship to one another at the design stage. All that is assumed is that a partial ordering of preference exists, with the primary endpoint considered more important than all other endpoints. As such, it represents a flexible methodology which could be applied to the complex intervention setting where many endpoints are of interest. The extra complexity of the decision rule does require that two addition nominal probabilities, relating to the minimum probability of making *correct* decisions under the null and alternative hypotheses, are specified. By way of illustration, in the same setting as that described in Section 2 (i.e. $p_0 = 0.05, p_A = 0.2$, nominal type I error and power 0.05 and 90% respectively), a design which guarantees correct decision rates of 0.8 would specify a total of 27 participants. If $s \leq 2$ the decision is made to stop, while if $s \geq 4$ the decision is made to proceed. If $2 < s < 4$ an 'inconclusive' decision is made based on the primary endpoint, and additional endpoints considered.

A further option which should be considered as a means with which to effectively address the challenge of multiple endpoints is to use a Bayesian decision-theoretic framework, as employed by Stallard et al. [63] and others in the drug context. This involves the specification of a utility function $u(d, \theta)$ which assigns a quantitative value to each possible decision d under every state of nature θ. For example, consider the case of two binary endpoints relating to toxicity (p_T) and efficacy (p_E), as discussed by Bryant and Day [13]. We now wish to assign a utility to each of the de-

cisions $\{stop, go\}$ for each value of $(p_T, p_E) \in [0,1]^2$. If beliefs regarding the likely values of parameters p_T and p_E can be specified through probability distributions, it is possible to calculate the expected utility of any decision d by averaging the utility function over the parameter space. Then, when faced with deciding whether or not the intervention should progress to phase III, the decision with Maximum Expected Utility (MEU) can be selected. The same MEU principle can be applied when determining the sample size of the trial in question. In order to do so, prior distributions on the parameters of interest must be elicited, after which the trial design which maximises expected utility over all possible trial outcomes can be found [40]. In the context of the present discussion, the specification of a utility provides a highly flexible means with which to encode the preferences of the decision maker(s). Allowing us to explicitly quantify any acceptable trade-offs between different endpoints, this approach will lead to decisions which are optimal with respect to these preferences. Whilst the specification of an appropriate subjective utility function may be difficult [61], it should be emphasised that Frequentist trial design can also involve subjective judgement when selecting nominal error rates, and that this may be less intuitive than the Bayesian alternative [6]. Moreover, recent examples such as the early phase drug trial described by Thall et al. [66] demonstrate the feasibility of employing Bayesian decision theoretic designs in practice. Where it is not feasible to specify a utility function, alternative Bayesian methods for sample size determination are available [1].

3.1.2 Trial specification

A further difficulty arising from the use of multiple endpoints is encountered when setting the specific design parameters for the trial. As illustrated in Figure 2, increases in the number of endpoints can correspond to increases to the dimensions of space in which the decision rule is defined. Accordingly, the number of potential decision rules which could be considered can increase. This feature can been seen in the phase II context when comparing the two stage design of Simon [57], which accounts for a single endpoint, with its extension to the two endpoint setting proposed by Bryant and Day [13]. In that case, given a proposed maximum sample size for each stage of the trial, n, the two endpoint design will have a factor of n^2 more possible parameterisations than the single endpoint design. As a result, the task of finding the specific 'best' parameterisation becomes more demanding and less amenable to simple, exhaustive search methods. This point is noted by Sill et al. [56], who propose heuristic methods to find good parameterisations of their two-endpoint phase II design.

In the complex intervention setting, the presence of several endpoints will compound this difficulty and lead to more sophisticated optimisation routines being required as standard. The design of any such algorithms will be strongly influenced by the nature of the endpoints considered. Binary endpoints will lead to integer trial design parameters (e.g. the threshold number of observed successes), whereas continuous endpoints will lead to continuous design parameters (e.g. the threshold of a t-test). Optimisation algorithms are typically tailored to specific problem types [74], and so different methods will generally be required to solve different problem types efficiently. Metaheuristic algorithms such as genetic optimization, as implemented in the R package 'rgenoud' [44], may provide a flexible solution methodology to address this difficulty, requiring only the tuning of algorithm parameters to ensure good performance.

Where a Bayesian decision-theoretic framework is employed, a decision rule does not have to be specified in advance. The aforementioned method of MEU does not require one [40], instead determining the decision by choosing that which, conditional on the observed data, has greatest expected utility. As a result, when determining the best specification for a trial one will not need to explore different decision rules. The addition of further endpoints will therefore not lead to a more challenging trial design problem, in contrast with some Frequentist cases.

3.2 Clustering

Clustering is a common feature of complex intervention trials and may arise with or without cluster randomisation [47]. For example, while the OK-D feasi-

bility trial is individually randomised, the assumption that patient outcomes are independent is questionable. This is due to the fact that, in the intervention arm of the trial, each participant is allocated to one of a limited number of trained research nurses whose role is to provide support in the delivery of the intervention. The study design is summarised in Figure 3.

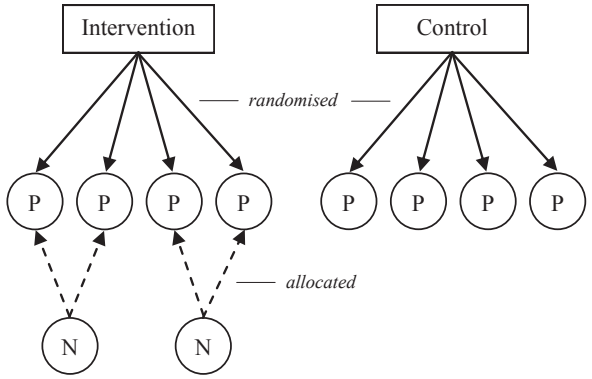


Figure 3: Clustering within the OK-D feasibility study, where patients are randomised to intervention or control and those within the intervention arm are allocated nurses.

The OK-D study may be described as having an *individually randomised, two level, partially nested hierarchical* design and is one of many possible scenarios where one or more sources of clustering are present [72]. By partially nested, we refer to the fact that clustering by research nurse is present in only one of the two arms, and by hierarchical we mean that there is a single research nurse per patient. More generally, the relationship between clusters and patients may be hierarchical, cross-classified (where patients are allocated to more than one type of cluster) or multiple-membership (where patients are allocated to more than one cluster of the same type). In terms of the relationship between treatments and clusters, this could be described as partially or fully nested, partially or fully crossed, or a mix of these for trials with more arms [72]. Specifically, nested designs have different clusters in each arm. For example, Schnurr et al. [52] describe a nested trial comparing Prolonged Exposure to Present-Centred Therapy for women with Posttraumatic Stress Disorder, where each therapist delivered only one of the treatments. Crossed designs have different arms associated with the same clusters [24]. Cohen and Mannarino [16] describe one such trial, comparing Cognitive Behavioural Therapy with Nondirective Supportive Therapy for sexually abused children, where therapists delivered both treatments.

In seeking to apply a phase II design to a problem where clustering is present, the simplest approach would be to ignore the clustering and apply the design 'off-the-shelf' without any modification. However, this can lead to inaccurate estimates of the type I error rate of any proposed trial [73] with the actual rate being higher than that calculated when designing the trial. As such, this approach would lead to ineffective interventions being taken forward for further evaluation in a phase III trial. A phase II design could be extended to account for clustering by including fixed cluster effects. However, such an analysis would imply a restricted focus on the specific clusters considered in that trial, preventing any generalization to a wider population. In the case of the OK-D feasibility study, this would correspond to restricting attention to only those nurse therapists participating in the experiment, as opposed to considering the larger population of therapists from which they are 'sampled' [53, 54, 55]. While it has been argued that this perspective in appropriate in the early phase of development [54], it is possible to improve the generalizability of the analysis by using random cluster effects rather than fixed. This approach has been recommended to account for clustering in individually randomized trials [39, 47], but will lead to a more complex linear mixed effects model.

3.2.1 Complex likelihoods

The hypothesis testing approach typical of phase II trials requires the specification of a test statistic and the derivation of that statistic's sampling distribution under the null and alternative hypotheses. Given analytical formulae describing these distributions, error rates for any decision rule can then be found by examining their tail areas. This approach is feasible in cases such as those considered by Fleming

[23] and A'Hern [2], where the distribution of the test statistic (a count of binary 'successes') is simply the binomial distribution. In multilevel statistical models, as found in trials where clustering is present, statistics such as a mean difference in a linear mixed effects model fitted by maximum likelihood will not necessarily have known analytical sampling distributions [43, 37], particularly in our setting where low sample sizes preclude the use of asymptotic results [22].

When analytical results describing the sampling distribution of the test statistic are not available, Monte Carlo simulation may be employed to estimate type I and II error rates [30, 37]. This involves simulating a number of hypothetical data sets according to a population model which corresponds to either the null or alternative hypothesis and, for each data set, calculating the test statistic. Implementing the proposed decision rule, the resulting action can be compared with the hypothesis used to generate the data and any error, type I or II, counted. This general technique is highly flexible. It can be applied to almost any multilevel structure encountered in practice [11, 27], using any proposed statistic in the analysis. However, this flexibility comes at the expense of a computational burden. The Monte Carlo method can require a significant amount of CPU time in order to perform enough simulations to provide an accurate estimate of error rates. The binary nature of both type I and II errors implies that the width of a confidence interval around an estimated error rate will decrease at a rate of $K/\sqrt{r}$ for a constant K and r simulations. For example, to ensure a 95% confidence interval of ± 0.05 around an estimated type II error rate of 0.2, one would require $r = 24586$ simulation runs. In practice, this may impose a limit on the number of trial specifications which can be considered and evaluated before one is chosen.

The computational burden of simulations may be reduced through their implementation in efficient programming languages such as C++. However, it has been argued that the resulting lack of transparency and difficulties in interpretation, in comparison to popular statistical programming packages such as R, should be taken into account when considering this option [59]. Alternatively, one may expedite the process of selecting an appropriate sample size by simplifying the problem. This technique is used in the freely available MLPowSim [11] software, which identifies a sensible choice of sample size by calculating the power of a restricted grid of designs, incrementing sample size parameters such as the number of clusters and the number of patients per cluster in large steps. By not considering every possible combination of sample size parameters, precision is sacrificed for speed. In the Stata routine SimSam [27], the problem is simplified by assuming all but one sample size parameters are known and fixed. Using heuristics to increase the efficiency of the search process, the optimal value of the remaining parameter (e.g. the number of patients per cluster) can be found in a timely manner. An alternative approach would be to use optimization algorithms which employ surrogate models, such as Efficient Global Optimisation (EGO) [31] and its variants to search over the full space of sample size configurations. These algorithms rely on fitting models, such as Gaussian process, to the simulated data obtained over a limited number of initial sample size configurations. Optimisation then takes place over the surrogate model, increasing efficiency as each evaluation now requires a simple calculation as opposed to a full simulation process. As these algorithms and their components have been implemented in R packages [50] and C++ libraries [42], they can be employed for this purpose without significant difficulty.

The simulation approach may be difficult to implement in cases where 'nuisance parameters' are present in the statistical model. This will often be the case where clustering is present. For example, in a fully nested design one would require a value for the ICC to be used in the population model when generating the data at each step. While it has become increasingly common for ICCs to be reported in the results of trials [14], the early phase context of feasibility and pilot studies implies that little information will be available for the intervention in question. Indeed, gathering information to inform future estimates of ICCs is a common objective of feasibility studies [3]. Thus, calculations of error rates may be dependent on parameter estimates in which there is significant uncertainty. The effect of such uncertainty in ICC

estimates on type II error rates and required sample size has been shown to be considerable [60, 68]. One approach to address this difficulty would be to carry out a sensitivity analysis, using several values of the nuisance parameter covering an appropriate range in order to identify its effect [37]. However, this would further contribute to the computational burden of the simulation approach.

In cases where some information regarding the likely values of nuisance parameters is available, a Bayesian approach would allow for this to be included formally via prior probability distributions [1]. This would fit naturally into the simulation method described thus far, allowing the data generated by the population model to encapsulate uncertainty in the nuisance parameters, leading to more robust estimates of error rates. In the case of ICCs in cluster randomised trials, the use of a prior distribution has been shown to significantly affect both design [69, 70] and analysis [60, 68]. In addition to acknowledging uncertainty in parameters, a Bayesian approach will also facilitate the incorporation of information from other sources. Recent methodology has been developed to allow for the weighting given to such prior beliefs to be adaptively changed in response to the data observed in the current trial [26], where the weighting will decrease as the observed data becomes less commensurate with the historical data [25]. Computationally, the Bayesian approach will require the use of Markov Chain Monte Carlo (MCMC) methods and, as a result, may be present difficulties with respect to timely analysis .

3.2.2 Sample size

In addition to leading to complex statistical models, clustered trial designs present difficulties when interpreting the notion of sample size. In phase II designs, sample size is commonly used as a metric with which to compare the quality of any two trial specifications. Typically, the setting of trial parameters is done in such a way as to minimise sample size subject to type I and II error rates remaining within nominal bounds. Trials with clustering, however, will induce further measures to be minimised by the trial designer. For example, the OK-D study involves k research nurses, each of whom has been assigned m patients. We wish for both k and m to be kept as low as possible whilst ensuring error rates remain within nominal bounds, but these measures are clearly in conflict - reducing one will require increasing the other in order to maintain error rates.

One approach to this problem is to combine the measures into a single weighted combination. This may be achieved through translating each measure to a common scale, such as cost [29, p. 175]. This would then allow one to focus on minimising cost (subject to constrained error rates). Where such a transformation is not available or appropriate, one may still employ a weighted combination method, although it may be challenging to elicit and represent the preferences of the decision maker(s) in this form. An alternative approach would be to set a limit on one measure, so that the other may be minimised subject to this constraint. For example, one could look for the trial with smallest m such that $k \leq 5$ and error rates remain within nominal bounds. Both methods induce an ordering of preference on the set of all possible trial specifications, thus defining the best. An alternative approach would be to consider the minimisation of m and k as independent measures, and attempt to identify a set of trial specifications representing a range of potential trade-offs between them whilst maintaining error rates within nominal bounds. This technique, known as Pareto optimization [45], may be a more realistic reflection of trial design in practice, where it is common for a range of scenarios and options to be explored and presented to the decision maker(s) before a final trial specification is selected. More generally, it should be noted that the error rates of trial configurations are measures which we aim to minimize, and that a constrained approach is typically used (e.g. requiring $\alpha < 0.1$) in addressing them. The benefits of relaxing error rate constraints to encourage the designer to trade-off different performance measures has been illustrated previously [35]. Furthermore, this general framework would extend easily to allow for further objectives to be specified. For example, as illustrated by Jung et al. [34], the specification of a two-stage trial following the Simon [57] design could consider minimising both the expected sample size and the

maximal sample size simultaneously.

3.2.3 Design space

In Section 3.1.2, additional complexity in the specification of decision rules was shown to lead to a more difficult optimisation problem due to an increased number of parameters. Similarly, increasing complexity in terms of multilevel structures due to clustering will also require further parameters or dimensions to be considered when searching for optimal trial specifications [29], and so again it may be beneficial to implement sophisticated optimisation routines rather than exhaustively searching through all possible options. Practically, the impact of increased numbers of design parameters may be limited by bounds on their values. For example, the number of therapists available to deliver an intervention may be fixed, and so when designing the trial one will not have to consider its variation. While such a feature will lead to a simpler optimisation problem, it may also lead to difficulties with regards to parameter estimation and inference.

4 Conclusions and further work

Currently, guidelines for the development and evaluation of complex interventions suggest that early phase experimental work focuses on assessing the feasibility and optimal design of a planned phase III definitive RCT. This contrasts with the drug development setting, where phase II trials are commonly used as a screening mechanism, designed to assess the efficacy of a new treatment and decide if a phase III trial will be worth conducting.

In this paper we have considered how the efficacy of complex interventions could be assessed in the context of current early phase feasibility or pilot studies. With reference to a range of phase II trial designs, challenges to their adaptation to the complex intervention setting have been discussed. The presence of multiple endpoints on which a decision must be based, and the clustering of outcomes in multilevel data structures, have been reviewed in detail. Two recurring themes have emerged. Firstly, the potential benefits of Bayesian methods have been highlighted in the context of decision theoretic approaches to trial design, incorporating uncertainty in trial design parameters and providing robust methods of estimation when only limited numbers of clusters are available. Secondly, we have emphasized the practical need for a sophisticated approach to defining and locating the 'optimal' trial specification for a given problem, in order that the best possible trial specification can be determined in a timely and robust manner.

In addition to difficulties arising from multiple endpoints and clustering, there remain several other features which could be explored in future work. One could consider widening the set of decisions of the study from the simple $\{stop, go\}$ to encompass the refining of the intervention's components or parameters [17], or to include the design specification of the planned phase III study in response to feasibility findings. Further details such as the impact of learning curves could be explored, and the appropriate place of efficacy assessment in the larger development and evaluation framework proposed by the MRC [21] should be considered.

Acknowledgements

Duncan Wilson is funded by a Research Methods Fellowship from the National Institute for Health Research. The authors wish to thank the OK-Diabetes study team (NIHR HTA grant reference 10/102/03) for helpful discussions that shaped the scope of this paper.

References

[1] C. J. Adcock. Sample size determination: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2):261–283, 1997.

[2] R. P. A'Hern. Sample size tables for exact single-stage phase II designs. *Statistics in Medicine*, 20 (6):859–866, 2001.

[3] M. Arain, M. Campbell, C. Cooper, and G. Lancaster. What is a pilot or feasibility study? a review of current practice and editorial policy. *BMC Medical Research Methodology*, 10(1):67, 2010.

[4] D. M. Arnold, K. E. A. Burns, N. K. J. Adhikari, M. E. Kho, M. O. Meade, D. J. Cook, and for the McMaster Critical Care Interest Group. The design and interpretation of pilot trials in clinical research in critical care. *Critical Care Medicine:*, 37(1):S69–S74, 2009.

[5] J. S. Barkun, J. K. Aronson, L. S. Feldman, G. J. Maddern, and S. M. Strasberg. Evaluation and stages of surgical innovations. *The Lancet*, 374 (9695):1089 – 1096, 2009.

[6] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 2nd edition, 1985.

[7] S. Billingham, A. Whitehead, and S. Julious. An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom Clinical Research Network database. *BMC Medical Research Methodology*, 13(1):104, 2013.

[8] J. M. Bland. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Medical Research Methodology*, 4:21, 2004.

[9] S. R. Brown, W. M. Gregory, C. J. Twelves, M. Buyse, F. Collinson, M. Parmar, M. T. Seymour, and J. M. Brown. Designing phase II trials in cancer: a systematic review and guidance. *Br J Cancer*, 105(2):194–199, July 2011.

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

[11] W. J. Browne, M. G. Lahi, and R. M. Parker. *A Guide to Sample Size Calculations for Random Effect Models via Simulation and the MLPowSim Software Package*, 2009.

[12] H. C. Brunier and J. Whitehead. Sample sizes for phase II clinical trials derived from bayesian decision theory. *Statistics in Medicine*, 13(23-24):2493–2502, 1994.

[13] J. Bryant and R. Day. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics*, 51(4):1372–1383, 1995.

[14] M. K. Campbell, D. R. Elbourne, and D. G. Altman. Consort statement: extension to cluster randomised trials. *BMJ*, 328(7441):702–708, 2004.

[15] M. Campbell, R. Fitzpatrick, A. Haines, A. L. Kinmonth, P. Sandercock, D. Spiegelhalter, and P. Tyrer. Framework for design and evaluation of complex interventions to improve health. *BMJ*, 321:694–696, 2000.

[16] J. A. Cohen and A. P. Mannarino. A treatment outcome study for sexually abused preschool children: Initial findings. *Journal of the American Academy of Child & Adolescent Psychiatry*, 35(1):42 – 50, 1996.

[17] L. M. Collins, S. A. Murphy, V. N. Nair, and V. J. Strecher. A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine*, 30(1):65–73, 2005.

[18] L. M. Collins, B. Chakraborty, S. A. Murphy, and V. Strecher. Comparison of a phased experimental approach and a single randomized clinical trial for developing multicomponent behavioral interventions. *Clinical Trials*, 6(1):5–15, 2009.

[19] M. R. Conaway and G. R. Petroni. Bivariate sequential designs for phase II trials. *Biometrics*, 51(2):pp. 656–664, 1995.

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

[21] P. Craig, P. Dieppe, S. Macintyre, S. Michie, I. Nazareth, and M. Petticrew. Developing and evaluating complex interventions: the new medical research council guidance. *BMJ: British Medical Journal*, 337, 9 2008.

[22] A. Donner and N. Klar. *Design and Analysis of Cluster Randomization Trials in Health Research*. London Arnold Publishers, 2000.

[23] T. R. Fleming. One-sample multiple testing procedure for phase II clinical trials. *Biometrics*, 38 (1):143–151, 1982.

[24] H. Goldstein. *Multilevel Statistical Models*. Arnold, 3rd edition, 2003.

[25] B. P. Hobbs, B. P. Carlin, S. J. Mandrekar, and D. J. Sargent. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3):1047–1056, 2011.

[26] B. P. Hobbs, B. P. Carlin, and D. J. Sargent. Adaptive adjustment of the randomization ratio using historical control data. *Clinical Trials*, 10 (3):430–440, 2013.

[27] R. Hooper. Versatile sample-size calculation using simulation. *The STATA Journal*, 13(1):21–38, 2013.

[28] A. House, R. Ajjan, L. Bryant, A. Farrin, E. Graham, C. Hulme, G. Latchford, D. Nagi, D. Riley, and A. Stansfield. Managing with learning disability and diabetes. URL `http://www.nets.nihr.ac.uk/projects/hta/1010203`. Accesed 6th October 2014.

[29] J. Hox. *Multilevel Analysis: Techniques and Applications*. Lawrence Erlbaum Associates, Inc., 2002.

[30] J. Hu and Z. Su. Efficient error determination in sequential clinical trial design. *Journal of Computational and Graphical Statistics*, 17(4): pp. 925–948, 2008.

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

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

[33] S.-H. Jung. Randomized phase II trials with a prospective control. *Statistics in Medicine*, 27 (4):568–583, 2008.

[34] S.-H. Jung, M. Carey, and K. M. Kim. Graphical search for two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 22(4):367 – 372, 2001.

[35] I. Khan, S.-J. Sarker, and A. Hackshaw. Smaller sample sizes for phase II trials based on exact tests with actual error rates by trading-off their nominal levels of significance and power. *British Journal of Cancer*, 107:1801–1809, 2012.

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

[37] S. Landau and D. Stahl. Sample size and power calculations for medical studies by simulation when closed form expressions are not available. *Statistical Methods in Medical Research*, 22(3): 324–345, 2013.

[38] E. Lee, A. Whitehead, R. Jacques, and S. Julious. The statistical interpretation of pilot trials: should significance thresholds be reconsidered? *BMC Medical Research Methodology*, 14(1):41, 2014.

[39] K. J. Lee and S. G. Thompson. The use of random effects models to allow for clustering in individually randomized trials. *Clinical Trials*, 2 (2):163–173, 2005.

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

[41] L. Mariani and E. Marubini. Design and analysis of phase II cancer trials: A review of statistical methods and guidelines for medical researchers. *International Statistical Review / Revue Internationale de Statistique*, 64(1):61–88, 1996.

[42] R. Martinez-Cantin. Bayesopt: A bayesian optimization library for nonlinear optimization, experimental design and bandits. *CoRR*, abs/1405.7430, 2014.

[43] C. E. McCulloch, S. R. Searle, and J. M. Neuhaus. *Generalized, Linear, and Mixed Models*. Wiley, 2nd edition, 2008.

[44] W. R. Mebane, Jr. and J. S. Sekhon. Genetic optimization using derivatives: The rgenoud package for R. *Journal of Statistical Software*, 42 (11):1–26, 6 2011.

[45] K. M. Miettinen. *Nonlinear multiobjemulti optimization*. Kluwer Academic Publishers, 1998.

[46] M. Petticrew. When are complex interventions 'complex'? when are simple interventions 'simple'? *The European Journal of Public Health*, 21(4):397–398, 2011.

[47] C. Roberts. The implications of variation in outcome between health professionals for the design and analysis of randomized controlled trials. *Statistics in Medicine*, 18(19):2605–2615, 1999.

[48] C. Roberts and S. A. Roberts. Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*, 2(2):152–162, 2005.

[49] T. Robertson, F. T. Wright, and R. Dykstra. *Order Restricted Statistical Inference*. New York: John Wiley and Sons, 1988.

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

[51] D. J. Sargent, V. Chan, and R. M. Goldberg. A three-outcome design for phase II clinical trials. *Controlled Clinical Trials*, 22(2):117 – 125, 2001.

[52] P. P. Schnurr, M. J. Friedman, C. C. Engel, and et al. Cognitive behavioral therapy for posttraumatic stress disorder in women: A randomized controlled trial. *JAMA*, 297(8):820–830, 2007.

[53] R. C. Serlin, B. E. Wampold, and J. R. Levin. Should providers of treatment be regarded as a random factor? if it aint broke, dont fix it: A comment on Siemer and Joormann (2003). *Psychological Methods*, 8(4):524–534, 2003.

[54] M. Siemer and J. Joormann. Power and measures of effect size in analysis of variance with fixed versus random nested factors. *Psychological Methods*, 8(4):497–517, 2003.

[55] M. Siemer and J. Joormann. Assumptions and consequences of treating providers in therapy studies as fixed versus random effects: Reply to Crits-Christoph, Tu, and Gallop (2003) and Serlin, Wampold, and Levin (2003). *Psychological Methods*, 8(4):535–544, 2003.

[56] M. W. Sill, L. Rubinstein, S. Litwin, and G. Yothers. A method for utilizing co-primary efficacy outcome measures to screen regimens for activity in two-stage phase II clinical trials. *Clinical Trials*, 9(4):385–395, 2012.

[57] R. Simon. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 10 (1):1 – 10, 1989.

[58] R. Simon, R. E. Wittes, and S. S. Ellenberg. Randomized phase II clinical trials. *Cancer Treatment Reports*, 69(12):1375–81, 1985.

[59] M. K. Smith and A. Marshall. Importance of protocols for simulation studies in clinical drug development. *Statistical Methods in Medical Research*, 2010.

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

[61] D. J. Spiegelhalter, L. S. Freedman, and M. K. B. Parmar. Bayesian approaches to randomized trials. *Journal of the Royal Statistical*

Society. Series A (Statistics in Society), 157(3): 357–416, 1994.

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

[63] N. Stallard, P. F. Thall, and J. Whitehead. Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics*, 55(3):971–977, 1999.

[64] S.-B. Tan and D. Machin. Bayesian two-stage designs for phase II clinical trials. *Statistics in Medicine*, 21(14):1991–2012, 2002.

[65] L. Thabane, J. Ma, R. Chu, J. Cheng, A. Ismaila, L. Rios, R. Robson, M. Thabane, L. Giangregorio, and C. Goldsmith. A tutorial on pilot studies: the what, why and how. *BMC Medical Research Methodology*, 10(1):1, 2010.

[66] P. F. Thall, H. Q. Nguyen, T. M. Braun, and M. H. Qazilbash. Using joint utilities of the times to response and toxicity to adaptively optimize schedule-dose regimes. *Biometrics*, 69(3): 673–682, 2013.

[67] C. Tournoux, Y. D. Rycke, J. Mdioni, and B. Asselain. Methods of joint evaluation of efficacy and toxicity in phase II clinical trials. *Contemporary Clinical Trials*, 28(4):514 – 524, 2007.

[68] R. M. Turner, R. Z. Omar, and S. G. Thompson. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine*, 20(3):453–472, 2001.

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

[70] R. M. Turner, S. G. Thompson, and D. J. Spiegelhalter. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clinical Trials*, 2(2): 108–118, 2005.

[71] R. E. A. Walwyn. *Therapist Variation within Meta-Analyses of Psychotherapy Trials*. PhD thesis, University of Manchester, 2010.

[72] R. E. A. Walwyn and C. Roberts. Therapist variation within randomised trials of psychotherapy: implications for precision, internal and external validity. *Statistical Methods in Medical Research*, 19(3):291–315, 2010.

[73] B. E. Wampold and R. C. Serlin. The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods*, 5(4):425–433, 2000.

[74] D. Wolpert and W. Macready. No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1):67–82, Apr 1997.