Wilson *et al.*
BMC Medical Research Methodology (2024) 24:226
https://doi.org/10.1186/s12874-024-02351-x

BMC Medical Research Methodology

RESEARCH

Open Access

Three-outcome designs for external pilot trials with progression criteria

Duncan T. Wilson[1*], Eleanor Hudson[1] and Sarah Brown[1]

Abstract

Background Whether or not to progress from a pilot study to a definitive trial is often guided by pre-specified quantitative progression criteria with three possible outcomes. Although the choice of these progression criteria will help to determine the statistical properties of the pilot trial, there is a lack of research examining how they, or the pilot sample size, should be determined.

Methods We review three-outcome trial designs originally proposed in the phase II oncology setting and extend these to the case of external pilots, proposing a unified framework based on univariate hypothesis tests and the control of frequentist error rates. We apply this framework to an example and compare against a simple two-outcome alternative.

Results We find that three-outcome designs can be used in the pilot setting, although they are not generally more efficient than simpler two-outcome alternatives. We show that three-outcome designs can help allow for other sources of information or other stakeholders to feed into progression decisions in the event of a borderline result, but this will come at the cost of a larger pilot sample size than the two-outcome case. We also show that three-outcome designs can be used to allow adjustments to be made to the intervention or trial design before commencing the definitive trial, providing the effect of the adjustment can be accurately predicted at the pilot design stage. An `R` package, `tout`, is provided to optimise progression criteria and pilot sample size.

Conclusions The proposed three-outcome framework provides a way to optimise pilot trial progression criteria and sample size in a way that leads to desired operating characteristics. It can be applied whether or not an adjustment following the pilot trial is anticipated, but will generally lead to larger sample size requirements than simpler two-outcome alternatives.

Keywords Pilot trial, Progression criteria, Sample size

Introduction

When there is some uncertainty about the feasibility of a planned randomised clinical trial (RCT), an external pilot trial can be conducted in advance [1, 2]. External pilots take the form of a smaller version of the main trial [3], and can be used to estimate various parameters of interest when deciding if (and how) to progress to the main study. Investing in a pilot trial can identify potential issues at an early stage, making a successful main trial more likely and reducing overall research waste [4].

Progression decisions are often guided by so-called *progression criteria* [5]. A single two-outcome progression criterion specifies a decision rule which maps the pilot data to a *stop* or *go* outcome. Specifically, the pilot data are used to calculate a statistic, typically an estimate of a parameter of interest, and this statistic is compared against a threshold value. If the statistic exceeds the

*Correspondence:
Duncan T. Wilson
D.T.Wilson@leeds.ac.uk
[1] Clinical Trials Research Unit, Leeds Institute of Clinical Trials Research, University of Leeds, Leeds LS2 9JT, UK

threshold, the suggested decision is to *go* forward to the main trial; otherwise, to *stop* on the grounds of infeasibility. When progression criteria are specified for several parameters, these can be combined by proceeding to the main trial only if all of the estimates exceed their respective thresholds [6] or through more complex decision rules [7]. It has been recommended that, in addition to being reported in the pilot study manuscript, progression criteria are pre-specified at the protocol stage in agreement with the study funder [8, 9].

Progression criteria are often based on a three-outcome 'traffic light' system [10]. These criteria stipulate two threshold values for a given parameter of interest. If the estimate falls below the lower of these, the decision is to stop (red); if the estimate falls above the higher threshold, the decision is to proceed immediately to the main trial (green); and if the estimate falls between the two thresholds, an intermediate decision is reached (amber). The specific purpose and interpretation of this intermediate decision can vary, and will depend on the motivation for using the three-outcome system. Three such motivations can be found in the methodological literature.

The CONSORT 2010 statement argues that basing strict stop/go decisions on a single threshold may lead to an unacceptably high chance of making the wrong decision as a result of sampling variability [5]. Allowing for an intermediate result between *stop* and *go* may reduce the probability of incorrect decisions. A second motivation stems from the fact that many aspects (quantitative and qualitative) being studied in a pilot trial are potentially relevant to the progression decision, and that this decision will be jointly made by several stakeholders (such as the trial team, the trial steering committee, the funder, and patients). A three-outcome system will allow immediate *stop* or *go* decisions to be made if the evidence is sufficiently strong with respect to a handful of key parameters, whilst allowing, in the event of a borderline result, the decision to be informed by other data based on the differing perspectives of all decision makers. Sargent et al. argued this better represents what happens in practice even when a two-outcome process is nominally being followed. In that case, although a borderline result will technically dictate a firm *stop/go* decision, this may be overridden in light of other information [11].

A final reason for an intermediate outcome is to provide the flexibility needed to make some adjustment to the intervention or trial design in an attempt to improve the parameter in question and ensure the feasibility of the main trial. For example, after observing a mediocre follow-up rate in a pilot trial, the trial designers might consider moving from a postal follow-up strategy to one based on contacting the participants over the phone. A three-outcome approach could facilitate this by prescribing an 'adjustment' decision to the intermediate outcome, whilst still allowing for immediate stopping or progression when obtaining 'stop' or 'go' outcomes.

Despite the prevalence of quantitative three-outcome progression criteria [12], there is little statistical guidance to help researchers decide how to specify them. The related question of determining the pilot trial sample size is also undeveloped, with work in this area typically focusing on pilot trials where the primary objective is to estimate the primary outcome variance to inform the main trial sample size calculation. These methods are nevertheless used when this is not the main purpose of the pilot, often in the form of simple 'rules-of-thumb' [13–15]. Three-outcome designs have, however, been proposed in the setting of phase II trials of cancer treatments [16]. Some of these designs were motivated by the same factors given above, and so may provide a useful framework for the design and analysis of pilot trials with three-outcome progression criteria.

In this paper we consider if, and how, three-outcome phase II designs can be used to determine optimal progression criteria and sample size in pilot trials. We begin by introducing a simple example, before arguing that quantitative progression criteria are mathematically equivalent to hypothesis tests, and are best viewed as such. We then review relevant three-outcome phase II trial designs and extend these to the pilot trial setting. Finally, we examine the statistical properties of these pilot trial designs and consider whether or not they can help achieve any of the three motivating goals before concluding with a discussion.

An example

Throughout this article we will refer to a simple example of a pilot trial assessing the probability that a participant in the intervention arm of the main trial will adhere to their prescribed treatment. Specifically, we consider adherence to be measured as a binary outcome and denote the probability of adherence by ρ. Given n patients in the pilot trial's intervention arm, we then model the number of adherers using a binomial distribution with parameters ρ and n. We denote the pilot estimate by $\hat{\rho}$.

We will consider both two- and three-outcome versions of progression criteria. In the two-outcome case, the progression decision is defined by a threshold value x, such that

$$\text{Decision} = \begin{cases} go & \text{if } \hat{\rho} \geq x \\ stop & \text{if } \hat{\rho} < x. \end{cases} \quad (1)$$

In the three-outcome case, we allow for an additional intermediate result and require two thresholds, x_0 and x_1:

$$\text{Decision} = \begin{cases} go & \text{if } \hat{\rho} \geq x_1 \\ pause & \text{if } x_0 < \hat{\rho} < x_1 \\ stop & \text{if } \hat{\rho} < x_0. \end{cases} \quad (2)$$

The specific meaning of the intermediate *pause* result will vary depending on the purpose and context of the pilot trial.

Progression criteria as hypothesis tests

In order to apply the two-outcome progression criteria of Eq. 1, we must choose the sample size n and the threshold x. One way to do so is though constructing a hypothesis test using the approach of A'Hern [17], as follows. First, we identify a parameter value ρ_0 such that if $\rho \leq \rho_0$ we would like to limit the probability of incorrectly making a *go* decision (a type I error) to at most α^*. Similarly, we identify ρ_1 such that if $\rho \geq \rho_1$ we would like to limit the probability of incorrectly making a *stop* decision (a type II error) to at most β^*. For example, we could choose adherence rates of $\rho_0 = 0.5$ and $\rho_1 = 0.7$ to represent poor and promising values respectively, and then use the standard choices of $\alpha^* = 0.05, \beta^* = 0.1$ for our nominal error rates. We then choose values of n and x which minimise n whilst satisfying the type I and II error rate constraints

$$\begin{aligned} \alpha &= \max_{\rho \leq \rho_0} \Pr[\hat{\rho} > x \mid \rho] \\ &= \Pr[\hat{\rho} > x \mid \rho = \rho_0] \leq \alpha^* \end{aligned} \quad (3)$$

$$\begin{aligned} \beta &= \max_{\rho \geq \rho_1} \Pr[\hat{\rho} \leq x \mid \rho] \\ &= \Pr[\hat{\rho} \leq x \mid \rho = \rho_1] \leq \beta^*, \end{aligned} \quad (4)$$

where we have used the monotonicity of power as a function of ρ to note that the type I and II error rates will be maximised when $\rho = \rho_0$ and $\rho = \rho_1$ respectively.

Alternatively, we can work backwards and take any given choice for n and x and calculate the resulting error rates for some hypotheses ρ_0, ρ_1. In particular, whenever a pilot trial progression criteria is specified in the form of Eq. 1, it is mathematically equivalent to a hypothesis test. For example, consider a pilot trial with $n = 15$ participants in the intervention arm and a *stop/go* progression criteria with threshold $x = 9/15$. If we suppose that the null and alternative hypotheses are $\rho_0 = 0.5, \rho_1 = 0.7$, this design will give type I and II error rates of $\alpha = 0.28$ and $\beta = 0.28$. If we instead constrain the error rates to, for example, $\alpha^* = 0.05$ and $\beta^* = 0.1$, the smallest possible sample size satisfying these constraints is $n = 53$ with a corresponding progression threshold is $x = 32/48$.

The equivalence of two-outcome progression criteria and hypothesis tests suggests the latter can provide a statistical framework for determining the former [6]. This will allow us to use hypotheses to express what parameter values would lead to errors of each type, and then subsequently to control the probability of these errors by choosing a sufficient sample size and associated progression threshold.

Extending three-outcome phase II trial designs

Just as standard hypothesis testing can be used as a framework for two-outcome progression criteria, three-outcome extensions of it can be used for the three-outcome progression criteria of Eq. 2. We will consider two such extensions proposed for phase II trials by Sargent et al. [11] and by Storer [18].

The design of Sargent et al. defines four operating characteristics relevant to the three-outcome setting. Firstly, a measure akin to the type I error rate, denoted α_a, is defined as the probability under the null hypothesis $\rho = \rho_0$ that the parameter estimate will exceed the upper threshold x_1 and thereby lead to an incorrect *go* decision. Similarly, a type II error rate β_a is given as the probability, under the alternative hypothesis $\rho = \rho_1$, of the parameter estimate falling below the lower threshold x_0 and leading to an incorrect *stop* decision. Two further operating characteristics relating to the intermediate outcome are then defined: the probability of obtaining a *pause* decision under the null hypothesis, denoted λ, and again under the alternative hypothesis, denoted δ. These operating characteristics are summarised in Table 1 and illustrated in Fig. 1. The authors propose to set constraints on these four operating characteristics and choose n, x_0, and x_1 to minimise n whilst satisfying these constraints. They argue that their designs will lead to a lower sample size requirement than standard two-outcome alternatives by reducing the probabilities of type I and II errors (α_a and β_a) through increasing the probabilities of *pause* decisions (λ and δ).

An alternative three-outcome design proposed by Storer [18] takes the same basic approach, but with a different set of four operating characteristics. Here, the type I error rate α_b is taken to be the probability of exceeding the lower threshold, x_0, under the null; and similarly the type II error rate β_b is now the probability of failing to exceed the upper threshold under the alternative. The remaining two operating characteristics are the probabilities of incorrectly obtaining a *stop* or a *go* decision when the true parameter is at some midpoint $\rho_m \in (\rho_0, \rho_1)$. These operating characteristics, denoted by γ_L and γ_U respectively, reflects the motivation of this

design to *encourage* an intermediate outcome when the true parameter value is between the null and alternative. The operating characteristics are summarised in Table 1 and illustrated in Fig. 2, where we follow the author's suggestion to set $\rho_m = (\rho_1 + \rho_0)/2$.

Considering the proposal of Sargent et al., we note that the measure α_a does not fully capture the probability of making a type I error since a decision to progress to the main trial can be arrived at in two ways: directly, by obtaining $\hat{\rho} > x_1$; or indirectly, by first obtaining a *pause* outcome $x_0 < \hat{\rho} < x_1$ and then deciding to proceed. To capture these situations, we define the probabilities of making incorrect decisions following a *pause* outcome under the null and alternative hypotheses:

$$\eta_0 = \Pr[\text{decide to } go \mid \rho = \rho_0, x_0 < \hat{\rho} \leq x_1] \tag{5}$$

Table 1 Operating characteristics for Sargent et al. and Storer's three-outcome designs [11, 18]

	Symbol	Equation	Description
Sargent	α_a	$\Pr[\hat{\rho} > x_1 \mid \rho = \rho_0]$	Probability of an immediate *go* decision under the null hypothesis
	β_a	$\Pr[\hat{\rho} \leq x_0 \mid \rho = \rho_1]$	Probability of an immediate *stop* decision under the alternative hypothesis
	λ	$\Pr[x_0 < \hat{\rho} \leq x_1 \mid \rho = \rho_0]$	Probability of a *pause* decision under the null hypothesis
	δ	$\Pr[x_0 < \hat{\rho} \leq x_1 \mid \rho = \rho_1]$	Probability of a *pause* decision under the alternative hypothesis
Storer	α_b	$\Pr[\hat{\rho} > x_0 \mid \rho = \rho_0]$	Probability of not obtaining an immediate *stop* decision under the null hypothesis
	β_b	$\Pr[\hat{\rho} \leq x_1 \mid \rho = \rho_1]$	Probability of not obtaining an immediate *go* decision under the alternative hypothesis
	γ_L	$\Pr[\hat{\rho} \leq x_0 \mid \rho = \rho_m]$	Probability of an immediate *stop* decision when $\rho = \rho_m$
	γ_U	$\Pr[x_1 < \hat{\rho} \mid \rho = \rho_m]$	Probability of an immediate *go* decision when $\rho = \rho_m$

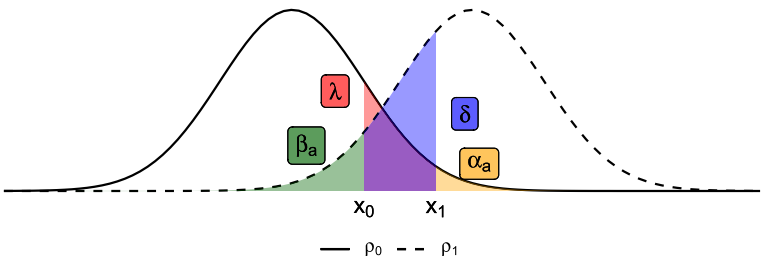

Fig. 1 Graphical illustration of the operating characteristics for Sargent et al.'s three-outcome design [11]. The curves represents the sampling distribution of the estimate under the null hypothesis $\rho = \rho_0$ (solid line) and the alternative hypothesis $\rho = \rho_1$ (dashed line)

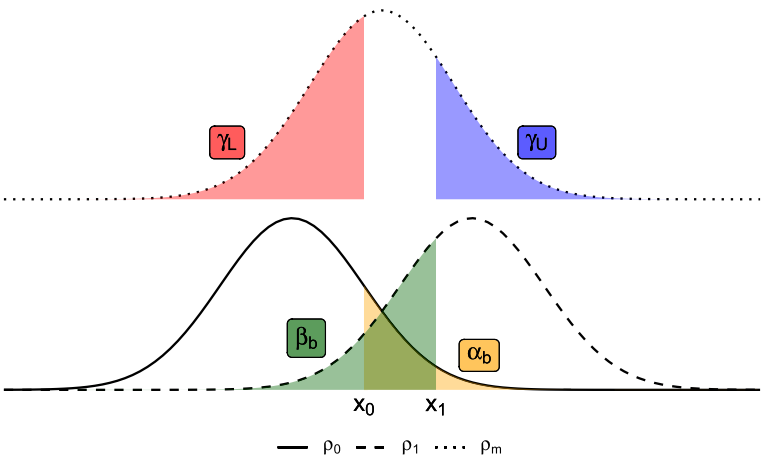

Fig. 2 Graphical illustration of the operating characteristics for Storer's three-outcome design [18]. The curves represents the sampling distribution of the estimate under the null hypothesis $\rho = \rho_0$ (solid line), the alternative hypothesis $\rho = \rho_1$ (dashed line), and a mid point $\rho = \rho_m = (\rho_1 + \rho_0)/2$ (dotted line)

$$\eta_1 = \Pr[\text{decide to } stop \mid \rho = \rho_1, x_0 < \hat{\rho} \leq x_1]. \quad (6)$$

For example, η_0 is the probability of making a *go* decision following a *pause* outcome and when the true parameter value is ρ_0. The probability of making a *go* decision when $\rho = \rho_0$, then, is not α_a but

$$\alpha = \alpha_a + \eta_0\lambda. \quad (7)$$

Similarly, the type II error rate is

$$\beta = \beta_a + \eta_1\delta. \quad (8)$$

Previous authors have suggested these operating characteristics in the context of multi-armed screening trials [19, 20]. For simplicity we will assume that $\eta_0 = \eta_1 = \eta$; that is, the probability of eventually making the wrong decision following an intermediate result is the same when $\rho = \rho_0$ as when $\rho = \rho_1$. Under this reformulation an optimal three-outcome design can be found by first estimating the probability η, setting constraints on the type I and II error rates α, β, and finally searching for the values of n, x_0, x_1 which minimise n whilst satisfying the constraints. Note that we no longer need to set constraints on the operating characteristics λ and δ, as these probabilities of obtaining an intermediate decision under the null and alternative hypothesis will be automatically limited by the constraints on α and β (as defined in Eqs. 7 and 8) respectively.

A similar argument applies when considering Storer's method, where we can replace the operating characteristics α_b, β_b with α and β. As the operating characteristics γ_L, γ_U are designed to encourage an intermediate outcome under ρ_m, rather than limit it as in the method of Sargent et al., we keep these in our reformulation. Thus, an optimal three-outcome design under the reformulated Storer method can be found by estimating the probability η, setting constraints on α, β, γ_L and γ_U, and finally searching for the values of n, x_0, x_1 which minimise n whilst satisfying the constraints.

For simplicity, we will assume that the cost of an incorrect *stop* or *go* decision when $\rho = \rho_m$ are the same, and replace the two error rates γ_U, γ_L by the single error rate

$$\gamma = \gamma_L + \gamma_U,$$

the probability of making an incorrect conclusive decision of either type. Note that the reformulated method of Sargent et al. is a special case of this method when we set the trivial constraint $\gamma \leq 1$, and so we have a single unified framework for designing and analysing three-outcome pilot studies which don't allow for adjustments to the intervention or trial design prior to the definitive trial.

Allowing for adjustments following a *pause* outcome

We now further generalise the three-outcome testing framework to allow for adjustments to be made following a *pause* outcome. Denote the effect of this adjustment by τ, such that the parameter in the main trial will equal $\rho' = \rho$ if no adjustment is made and $\rho' = \rho + \tau$ if it is. We will assume that the adjustment effect is known up to an interval $\tau \in [\tau_{min}, \tau_{max}]$, and that $\tau_{min} \geq 0$. We then refine our definitions of the error rates α and β as

- α: the probability of proceeding to the main trial when $\rho' \leq \rho_0$
- β: the probability of not proceeding to the main trial when $\rho' \geq \rho_1$;

Because we can make a *go* decision in two ways, α is now the maximum probability of proceeding either directly or following a *pause* outcome (in which case the adjustment is made) when this will lead to $\rho' \leq \rho_0$. As before, we assume there is a constant probability of mistakenly deciding to proceed following a *pause* outcome when in fact $\rho + \tau \leq \rho_0$, denoted by η. That is,

$$\alpha = \max\left[\max_{\rho \leq \rho_0} \Pr(x_1 < \hat{\rho}),\ \max_{\rho+\tau \leq \rho_0} \eta \Pr(x_0 < \hat{\rho} \leq x_1) + \Pr(x_1 < \hat{\rho})\right].$$

The first term is maximised at $\rho = \rho_0$. The second term can be written as

$$\begin{aligned}\eta \Pr(x_0 < \hat{\rho} \leq x_1) + \Pr(x_1 < \hat{\rho}) &= \eta \Pr(\hat{\rho} \leq x_1) - \eta \Pr(\hat{\rho} \leq x_0) + 1 - \Pr(\hat{\rho} \leq x_1) \\ &= 1 + (\eta - 1)\Pr(\hat{\rho} \leq x_1) - \eta \Pr(\hat{\rho} \leq x_0),\end{aligned}$$

and so is maximised at $\rho = \rho_0 - \tau_{min}$, giving

$$\begin{aligned}\alpha = \max\big[&\Pr(x_1 < \hat{\rho} \mid \rho = \rho_0),\ \eta \Pr(x_0 < \hat{\rho} \leq x_1 \mid \rho = \rho_0 - \tau_{min}) \\ &+ \Pr(x_1 < \hat{\rho} \mid \rho = \rho_0 - \tau_{min})\big].\end{aligned} \quad (9)$$

An incorrect *stop* decision may again occur two ways - directly, or following a *pause* outcome. The error rate β can therefore be written as

$$\beta = \max\left[\max_{\rho > \rho_1} \Pr(\hat{\rho} \leq x_0),\ \max_{\rho+\tau > \rho_1} \Pr(\hat{\rho} \leq x_0) + \eta \Pr(x_0 < \hat{\rho} \leq x_1)\right]. \quad (10)$$

The first term is maximised at $\rho = \rho_1$, while the second term can be written as

$$\eta \Pr(\hat{\rho} \leq x_1) + (1-\eta)\Pr(\hat{\rho} \leq x_0),$$

which is maximised at $\rho = \rho_1 - \tau_{max}$. Since $\tau_{max} > 0$, then $\Pr(\hat{\rho} \leq x_0 | \rho = \rho_1) \leq \Pr(\hat{\rho} \leq x_0 \mid \rho = \rho_1 - \tau_{max})$ and so Eq. 10 can be simplified to

$$\beta = \Pr(\hat{\rho} \leq x_0 \mid \rho = \rho_1 - \tau_{max}) + \eta \Pr(x_0 < \hat{\rho} \leq x_1 \mid \rho = \rho_1 - \tau_{max}). \tag{11}$$

Note that the general error rates of Eqs. 9 and 11 are valid in the special case where no adjustments are going to be made following a *pause* decision (i.e. when $\tau_{min} = \tau_{max} = 0$). As such, we have a general formulation of error rates in three-outcome designs regardless of the possibility of adjustment.

Implementation

To determine appropriate values for n, x_0 and x_1 which will satisfy nominal error rate constraints α^*, β^* and γ^*, we first consider n and x_1 to be fixed and such that $\Pr(\hat{\rho} > x_1 \mid \rho = \rho_0) \leq \alpha^*$. Then we set the second term in Eq. 9 equal to α^* and rearrange to get

$$\Pr(\hat{\rho} \leq x_0) = \frac{1}{\eta} + \frac{\eta - 1}{\eta}\Pr(\hat{\rho} \leq x_1) - \frac{\alpha^*}{\eta}. \tag{12}$$

Using the inverse of $\hat{\rho}$'s distribution function, we can then find the value of x_0 which gives us $\alpha = \alpha^*$ (or for the case of a binary outcome, the x_0 which maximises α whilst ensuring $\alpha \leq \alpha^*$).

To choose x_1 we continue to fix n and then use a numerical search to find the largest x_1 such that $x_0 \leq x_1$ (with x_0 determined using Eq. 12) and $\beta \leq \beta^*$. Finally, we can choose n by finding the smallest value such that the third constraint $\gamma \leq \gamma^*$ is satisfied when the corresponding x_1 and x_0 are chosen using the above procedure.

The `R` package `tout` uses this approach to determine optimal three-outcome designs for the case of univariate progression criteria based on a binary or continuous endpoint in single arm, single stage trials. See the supplementary material and package documentation for full details and illustrations.

Evaluation

As noted in the introduction, adding a third outcome to pilot trial progression criteria has been motivated on grounds of i) statistical efficiency; ii) the need to incorporate other information or stakeholders into progression decisions; and iii) the need to make modifications to the intervention or the trial design before commencing the main trial. In this section we examine to what extent the three-outcome design framework described in the previous section can be used to meet these goals.

Statistical efficiency

Sargent et al. show that three-outcome designs based on the operating characteristics given in Table 1 will have a lower sample size than corresponding two-outcome designs whilst constraining the error rates α_a, β_a to the same levels, providing λ and/or δ are allowed to be greater than 0. For example, take $\rho_0 = 0.5$ and $\rho_1 = 0.7$. A two-outcome design will require $n = 53$ to ensure $\alpha_a \leq 0.05$ and $\beta_a \leq 0.1$. In contrast, by allowing $\lambda \leq 0.1$ and $\delta \leq 0.1$ in a three-outcome design, we can obtain $\alpha_a \leq 0.05$ and $\beta_a \leq 0.1$ with only $n = 42$, suggesting three-outcome designs are indeed more efficient [11, 21]. However, this apparent advantage breaks down when using our reformulation. To illustrate this we found optimal sample sizes for this example problem over the range $0 \leq \eta \leq 0.5$ (where η denotes the probability of making an incorrect progression decision following a *pause* outcome), using the constraints $\alpha \leq 0.05, \beta \leq 0.2, \gamma \leq 1$. We have not considered $\eta > 0.5$ as this represents a decision-making ability worse than random, in which case the optimal design remains the usual two-outcome design. The optimal sample sizes are plotted in Fig. 3, where the discrete nature of sample size leads to step functions.

When $\eta = 0.5$, in which case we can only guess at the correct decision following a *pause* outcome, the optimal sample size is $n = 52$. Figure 3 shows that a low value of η is required to achieve a meaningful reduction in sample size. For example, for a 20% reduction from the $n = 52$ two-outcome design down to $n = 41$, we would require $\eta = 0.2$. That is, we must be confident that following a *pause* outcome, but with a true $\rho = \rho_i (i = 0, 1)$, we will make the correct progression decision with a probability of 0.8. In the context of our simple one-parameter example, the estimate $\hat{\rho}$ is a sufficient statistic for ρ and so we cannot obtain any other information relevant to this particular judgement. This would lead to $\eta = 0.5$, in which case the optimal three-outcome design will reduce to a two-outcome design.

We may expect $\eta < 0.5$ if measures of another outcome in the trial, correlated with the outcome of interest, are going to inform the progression decision following a *pause* outcome. For example, patient adherence and retention may be correlated. If a *pause* outcome was observed when assessing adherence but retention was seen to be high, we might infer the true adherence rate to be larger than the estimate. The extent of this will depend, however, on the strength of the correlation, which may be hard to judge at the design stage.

To explore the implications of incorrectly assuming $\eta < 0.5$, we took each of the optimal designs found over the range $0 \leq \eta \leq 0.5$ and calculated their type I and II error rates under a true value of $\eta = 0.5$. These

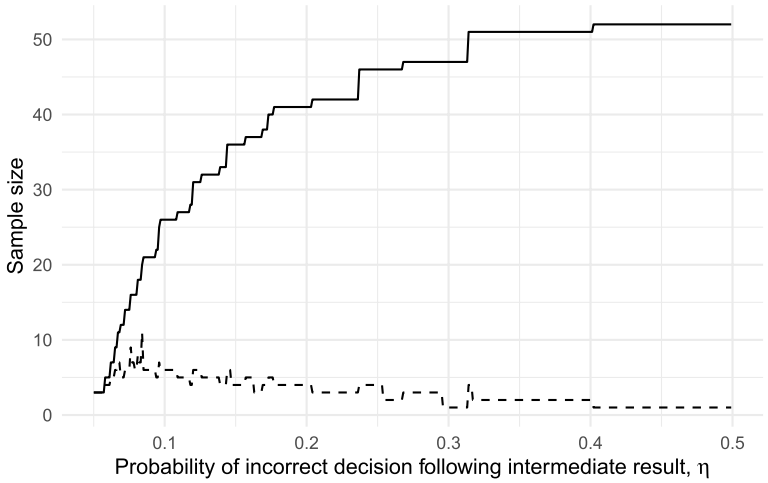

Fig. 3 Minimum required sample size for a three-outcome design as a function of η (solid line), along with the corresponding size of the intermediate zone $x_1 - x_0$ (dashed line)

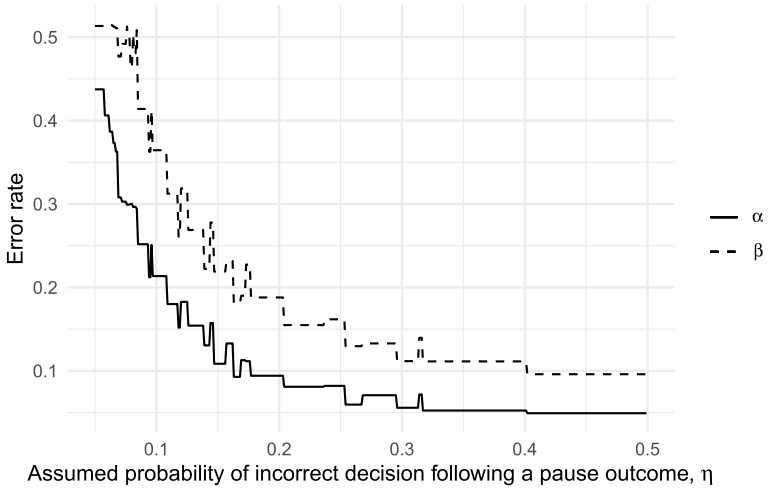

Fig. 4 Type I (solid line) and II (dashed line) error rates of three-outcome designs, each locally optimal for an assumed η, when in fact $\eta = 0.5$

error rates are plotted in Fig. 4. We find that error rates will be substantially inflated whenever η was incorrectly assumed to be small enough to lead to a meaningful reduction in sample size. For example, if we incorrectly assume $\eta = 0.2$ when in fact $\eta = 0.5$ the 'optimal' design will lead to actual type I and II error rates of 0.094 and 0.188, rather than the nominal 0.05 and 0.1.

Given the challenges of estimating η and the implications of doing it badly, we follow previous suggestions [19, 20] that a default assumption of $\eta = 0.5$ is appropriate. Under this conservative assumption, three-outcome progression criteria will not improve statistical efficiency in pilot trials beyond two-outcome alternatives.

Incorporating other information

Although three-outcome progression criteria are not more efficient than their two-outcome alternatives, we may nevertheless wish to use them to allow other information to inform the progression decision rather then being ignored completely. For example, in addition to requiring sufficient adherence we may also want to see good recruitment and retention in the pilot trial. In the event of a *pause* outcome when assessing adherence, we could then decide to proceed to the main trial only if the estimated recruitment and retention rates are

large enough. A *pause* outcome could also provide an opportunity for discussion amongst the various stakeholders (such as the trial team, steering committee, funder, and patients) to arrive at a collective decision on progression.

To facilitate this we can encourage the design to have an appropriate intermediate zone $|x_1 - x_0|$ by constraining the operating characteristic $\gamma = \gamma_L + \gamma_U$ defined in Table 1, thus limiting the chance of making a conclusive *stop* or *go* decision when $\rho = \rho_m$. With $\rho_0 = 0.5, \rho_1 = 0.7, \alpha \leq 0.05$ and $\beta \leq 0.1$ as before, we set $\rho_m = (\rho_1 - \rho_0)/2 = 0.6$ and found optimal designs for a range of γ^* while assuming throughout that $\eta = 0.5$. The sample sizes of these designs are plotted in Fig. 5.

When we set $\gamma^* = 1$, no intermediate zone is required and so the optimal design is the usual two-outcome design. As we decrease the nominal level on this constraint we permit an ever smaller probability of obtaining a conclusive *stop* or *go* outcome when $\rho = \rho_m$. This leads to an increasing width of intermediate zone $|x_1 - x_0|$, alongside an increasing sample size. The required increase in sample size beyond the two-outcome design can be substantial. For example, to ensure a maximum 40% chance of obtaining a conclusive result when $\rho = \rho_m$, we must increase the sample size from $n = 52$ to $n = 98$. Providing such increases in sample size are considered worthwhile, we conclude that three-outcome designs can be used in pilot trials to allow other information and stakeholders to feed into progression decisions.

Allowing for adjustments

A final rationale for an intermediate outcome in pilot trials is to enable some modifications to be made prior to commencing the main trial. These could be adjustments to the trial design (e.g. to improve recruitment) or to the intervention itself (e.g. to improve adherence). The intermediate *pause* outcome now leads to the decision to either *stop* or to make these modifications and then *go* to the main trial. Recall that the effect of such an adjustment is denoted by τ.

Known adjustment effect

Assume that the effect of adjustment τ is known a priori. Considering the same problem as before $(\rho_0 = 0.5, \rho_1 = 0.7, \alpha^* = 0.05, \beta^* = 0.1, \eta = 0.5)$ we found optimal designs for a range of known adjustment effects spanning $\tau \in [0, 0.125]$. The required sample size of these designs is illustrated in Fig. 6.

When adjustments have no effect ($\tau = 0$), the optimal three-outcome design reduces to the usual *stop/go* two-outcome design with $n = 52$. As τ increases the required sample size increases with it exponentially. For example, for $\tau = 0.125$ we require $n = 275$. This can be explained by looking back at our error rate definitions in Eqs. 9 and 11, which show that α constrains the term $\Pr(x_1 < \hat{\rho} \mid \rho = \rho_0)$ and thus places a lower limit on x_1; meanwhile, β constrains the term $\Pr(\hat{\rho} < x_0 \mid \rho = \rho_1 - \tau_{max})$ and thus forces x_0 to be lowered as τ_{max} increases, leading to a larger intermediate zone and correspondingly worse error rates.

Partially known adjustment effect

We now consider the case where the adjustment effect τ is known only up to an interval $\tau \in [\tau_{min}, \tau_{max}]$. We considered a range of values for τ_{min} from 0 up to 0.1 alongside a range of interval widths $\tau_{max} - \tau_{min}$ from 0 to 0.05. The resulting sample sizes are plotted in Fig. 7.

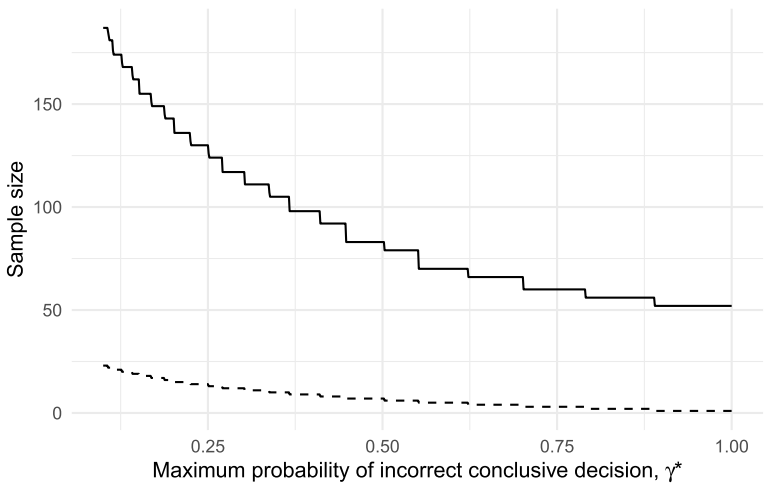

Fig. 5 Minimum required sample size for a three-outcome design as a function of γ^* (solid line), along with the corresponding size of the intermediate zone $x_1 - x_0$ (dashed line)

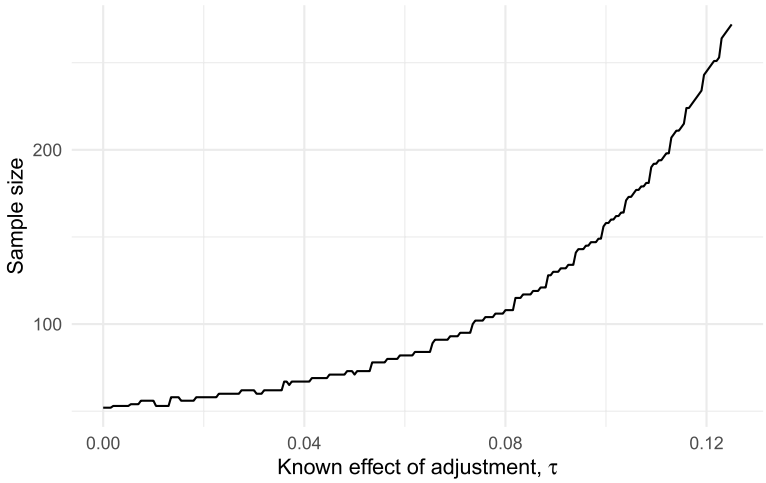

Fig. 6 Minimum required sample size for a three outcome design as a function of the known adjustment effect τ

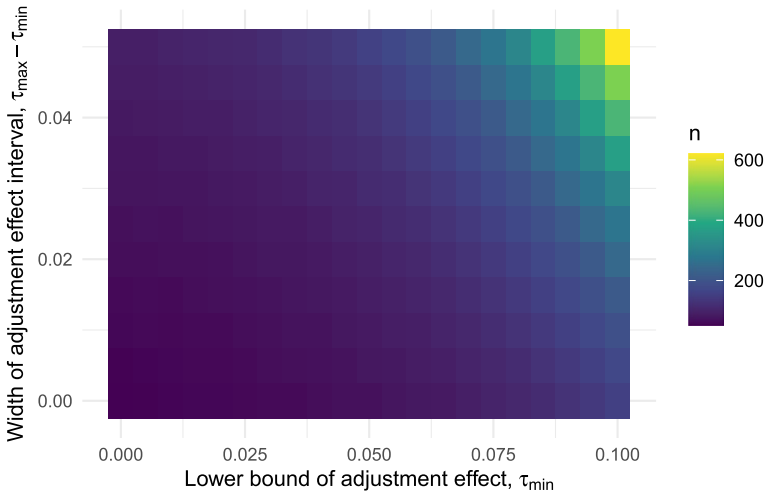

Fig. 7 Minimum required sample size for a three outcome design as a function of the lower limit of the adjustment effect τ_{min} and the width of the effect interval, $\tau_{max} - \tau_{min}$

We see that increasing the width of the adjustment effect interval leads to an increased sample size, and the rate at which this happens increases with the lower interval limit τ_{min}. For example, moving from $[\tau_{min}, \tau_{max}] = [0, 0]$ to $[0,\ 0.05]$ leads to a change in sample size from $n = 52$ to $n = 93$; while moving from $[\tau_{min}, \tau_{max}] = [0.1, 0.1]$ to $[0.1,\ 0.15]$ takes us from $n = 158$ to $n = 620$. Figure 7 suggests that the main driver of sample size is the upper limit τ_{max}, but changing τ_{min} while keeping this fixed can still lead to considerable changes in n. From example, $[\tau_{min}, \tau_{max}] = [0.05, 0.05]$ (that is, a known $\tau = 0.05$) requires $n = 71$ while $[\tau_{min}, \tau_{max}] = [0.01, 0.05]$ requires $n = 87$.

The case $\eta < 0.5$

In the preceding subsections we have assumed $\eta = 0.5$. In the context of allowing adjustments we might argue that $\eta < 0.5$ is plausible because, although we may not be able to learn anything about ρ beyond what is provided by $\hat{\rho}$, we might learn something about the adjustment effect τ during the pilot. If we can assume this will reduce η, this will lead to a corresponding reduction in sample size in a manner similar to the no-adjustment case discussed previously (see supplementary material for more details). We must bear in mind, though, that learning about τ can only take us so far. Even in the extreme case where we can learn τ exactly, the residual uncertainty about ρ will place a lower limit on our

ability to make the correct decision following a *pause* outcome.

We conclude that three-outcome progression criteria can be used to allow for adjustments, but note that this ability will come at the cost of an increased sample size. This increase can be especially large if the effect of adjustment may be substantial, or if there is considerable uncertainty about how large it is.

Discussion

We have shown how the three-outcome progression criteria commonly used in pilot trials can be viewed as three-outcome hypothesis tests, and described how related clinical trial designs from the phase II setting can be used (with some reformulation) to optimise these criteria and the pilot sample size. This allowed for a formal comparison to be made between three- and two-outcome designs for pilot trials, with the latter as a special case of the former. We have shown that three-outcome designs do not improve efficiency in comparison to two-outcome alternatives, but that three-outcome designs can be used to allow for a more realistic decision-making process involving multiple sources of information and multiple stakeholders in the event of a borderline result. We have also shown that three-outcome progression criteria can facilitate making adjustments to the intervention or trial design following the pilot when the effect of such an adjustment is known in advance. We found that there is a price to pay for these benefits, with three-outcome pilot trials needing a (sometimes considerably) larger sample size than two-outcome alternatives to obtain the same operating characteristics. This suggests that the small sample sizes typically seen in pilot trials [22] may be inadequate for their goals.

We have quantified the impact of using three-outcome progression criteria through the resulting required sample size while keeping operating characteristics constrained. An alternative would be to fix the pilot sample size and examine the impact on operating characteristics. For example, taking $\rho_0 = 0.5$ and $\rho_1 = 0.7$ as before and fixing $n = 30$, a two-outcome design with threshold $x_0 = x_1 = 17$ will give us operating characteristics of $\alpha = 0.18, \beta = 0.08, \gamma = 1$. Moving to a three-outcome design will allow us to reduce γ, but only at the expense of an increased α and β. For example, the design with $x_0 = 15$ and $x_1 = 20$ gives $\alpha = 0.22, \beta = 0.21, \gamma = 0.35$. Although this type I error rate is larger than conventional constraints, previous authors have argued that these may be relaxed in the setting of phase II [23, 24] and pilot [25] trials.

In order to apply the proposed method in practice we need to specify null and alternative hypotheses for the parameter of interest and put constraints on three error rates. Specifying hypotheses for feasibility parameters like adherence rates may be challenging when compared to the more typical setting of assessing efficacy. Generally speaking there will be no default choice for the null of a feasibility parameter, as opposed to the default efficacy null of no difference to standard care. Moreover, the concept of Minimal Clinically Important Difference, which often guides the choice of target difference in efficacy, will not apply to feasibility. It may help to begin by determining the midpoint $\rho_m = (\rho_0 + \rho_1)/2$ which we would consider a borderline value and where we would ideally like to obtain a *pause* outcome, then determine the width of the interval $|\rho_1 - \rho_0|$, and finally set constraints on error rates α and β based on the relative impact of these errors when defined with respect to ρ_0 and ρ_1. Having determined these values, one can find optimal designs for a range of constraints on the third operating characteristic γ as shown in Fig. 5. Alternatively, the choice of ρ_0 and ρ_1 could be driven by the corresponding impact on any subsequent trial as measured though, for example, its power [7].

In addition to defining hypotheses, the proposed method also requires values for η_0 and η_1, the probabilities of making an incorrect progression decision *with respect to the parameter* ρ following a *pause* outcome under the null and alternative hypothesis respectively. We have argued that a default of $\eta_0 = \eta_1 = 0.5$ may be justified as a conservative assumption. Alternatively, we may anticipate a bias towards progressing from the pilot to the definitive trial. This could be modelled by setting $\eta_0 > 0.5, \eta_1 < 0.5$ whilst constraining $\eta_0 + \eta_1 = 1$, although we found this to have little impact on the optimal design in our running example.

Our results have highlighted the difficulties of allowing adjustments to be made following a pilot trial when using pre-specified progression criteria, even when the effect of the adjustment is known (or partially known) in advance. This may be an unrealistic assumption, since a primary goal of many pilot trials is to identify *unforeseen* problems and solutions to these. In this context, pre-specifying an upper threshold x_1 can help identify cases which are feasible enough, without modification, to proceed to the main trial. In contrast, the lower threshold x_0 appears somewhat arbitrary and may force inappropriate decisions. For example, if x_0 is set too high, we may be led to a *stop* decision (i.e. $\hat{\rho} \leq x_0$) despite believing, based on what was seen in the pilot, that a certain modification would lead to an adjusted adherence rate greater than ρ_1. When a new and unanticipated adjustment is proposed following the pilot trial, a second (and potentially internal) pilot may be needed to establish that it works as expected. Brown et al. suggested a similar strategy in the context of phase II drug trials, and it is in line with guidance

on the development and evaluation of complex interventions [26] which emphasises the iterative nature of the process. Alternatively, if the cost of adjustment is low then a two-outcome design could be used where we assume the modification will always be made following a *go* outcome. Note that the power of this design will depend on the unknown effect of adjustment τ, and a conservative assumption of $\tau = \tau_{max}$ would lead to large sample size requirements.

The underlying statistical framework considered in this paper is frequentist, focused on pre-specified decision rules chosen based on their long-run operating characteristics. An alternative is to design and analyse the pilot trial under a Bayesian framework [27–29]. This could improve efficiency by allowing external information or expert knowledge to be incorporated into decision-making, and would enable a more flexible approach to analysis which could formally account for anticipated effects of adjustments based on what was seen in the pilot. Willan and Thabane [30] do not consider the question of optimising pilot sample size, but show through an example how a Bayesian analysis of pilot data can help quantify uncertainty around feasibility parameters by producing posterior distributions which can be used to design the main trial. A Bayesian approach would also help address the aforementioned difficulties in specifying parameters (including the the probability of making correct decisions following a *pause* outcome, and the anticipated effect of adjustment) by allowing uncertainty regarding these to be expressed through prior distributions.

Our work has been motivated by external pilot trials assessing the feasibility of a subsequent study, but may be equally relevant to other settings such as phase II drug trials where the parameter of interest is a measure of efficacy. For example, the three-outcome framework could be used when making post trial adjustments to improve efficacy by changing eligibility criteria in an attempt to focus on a subgroup of patients. Although our findings should also broadly apply to internal pilot trials, care may be needed when the error rates of the final analysis may be affected by a formal internal pilot analysis of a correlated endpoint (for example, adherence). Finally, we expect our conclusions to carry over from the univariate setting considered here to the more general multivariate setting, where several progression criteria are applied simultaneously [6], although it has been shown that such multivariate tests can be counter-intuitively inefficient [7].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02351-x.

Additional file 1.

Acknowledgements
We would like to thank Martyn Lewis for his helpful comments on an earlier draft of this manuscript.

Authors' contributions
DTW conceived the project, programmed the software application and drafted the manuscript. EH and SB contributed to the design of the work and the interpretation of the results, and critically reviewed and revised the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by the Medical Research Council under Grant MR/N015444/1 to DTW.

Availability of data and materials
Supplementary material, including all `R` code used to generate the results in this manuscript and the associated `R` package `tout`, can be found at `https://github.com/DTWilson/tout`.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 17 May 2024 Accepted: 24 September 2024
Published online: 02 October 2024

References

1. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios L, et al. A tutorial on pilot studies: the what, why and how. BMC Med Res Methodol. 2010;10(1):1. https://doi.org/10.1186/1471-2288-10-1.
2. Arain M, Campbell M, Cooper C, Lancaster G. What is a pilot or feasibility study? A review of current practice and editorial policy. BMC Med Res Methodol. 2010;10(1):67. https://doi.org/10.1186/1471-2288-10-67.
3. Eldridge SM, Lancaster GA, Campbell MJ, Thabane L, Hopewell S, Coleman CL, 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.
4. Morgan B, Hejdenberg J, Hinrichs-Krapels S, Armstrong D. Do feasibility studies contribute to, or avoid, waste in research? PLoS ONE. 2018;13(4):1–8. https://doi.org/10.1371/journal.pone.0195951.
5. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ. 2016;355:i5239. https://doi.org/10.1136/bmj.i5239.
6. Lewis M, Bromley K, Sutton CJ, McCray G, Myers HL, Lancaster GA. Determining sample size for progression criteria for pragmatic pilot RCTs: the hypothesis test strikes back! Pilot Feasibility Stud. 2021;7(1):1–14.
7. Wilson DT, Brown J, Farrin AJ, Walwyn REA. A hypothesis test of feasibility for external pilot trials assessing recruitment, follow-up, and adherence rates. Stat Med. 2021;40(21):4714–4473. https://doi.org/10.1002/sim.9091.
8. National Institute for Health Research. Research for Patient Benefit (RfPB) Programme Guidance on Applying for Feasibility Studies. Version 4.0 - November 2023. https://www.nihr.ac.uk/guidance-applying-feasibility-studies. Accessed 26 Sept 2024.
9. Mbuagbaw L, Kosa SD, Lawson DO, Stalteri R, Olaiya OR, Alotaibi A, et al. The reporting of progression criteria in protocols of pilot trials designed to assess the feasibility of main trials is insufficient: a meta-epidemiological study. Pilot Feasibility Stud. 2019;5(1):120. https://doi.org/10.1186/s40814-019-0500-z.

10. Avery KNL, Williamson PR, Gamble C, Francischetto EO, Metcalfe C, Davidson P, 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.
11. Sargent DJ, Chan V, Goldberg RM. A Three-Outcome Design for Phase II Clinical Trials. Control Clin Trials. 2001;22(2):117–25. https://doi.org/10.1016/S0197-2456(00)00115-X.
12. Herbert E, Julious SA, Goodacre S. Progression criteria in trials with an internal pilot: an audit of publicly funded randomised controlled trials. Trials. 2019;20(1):493. https://doi.org/10.1186/s13063-019-3578-y.
13. Browne RH. On the use of a pilot sample for sample size determination. Stat Med. 1995;14(17):1933–40. https://doi.org/10.1002/sim.4780141709.
14. 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.
15. 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):1057–73. https://doi.org/10.1177/0962280215588241.
16. Kirby S, Chuang-Stein C. A comparison of five approaches to decision-making for a first clinical trial of efficacy. Pharm Stat. 2016;16(1):37–44. https://doi.org/10.1002/pst.1775.
17. A'Hern RP. Sample size tables for exact single-stage phase II designs. Stat Med. 2001;20(6):859–66. https://doi.org/10.1002/sim.721.
18. Storer BE. A Class of Phase II Designs with Three Possible Outcomes. Biometrics. 1992;48(1):55–60.
19. Sargent DJ, Goldberg RM. A flexible design for multiple armed screening trials. Stat Med. 2001;20(7):1051–60. https://doi.org/10.1002/sim.704.
20. Dehbi HM, Hackshaw A. Sample size calculation in randomised phase II selection trials using a margin of practical equivalence. Trials. 2020;21:1–7.
21. Hong S, Wang Y. A three-outcome design for randomized comparative phase II clinical trials. Stat Med. 2007;26(19):3525–34. https://doi.org/10.1002/sim.2824.
22. Billingham S, Whitehead A, Julious S. 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 Med Res Methodol. 2013;13(1):104. https://doi.org/10.1186/1471-2288-13-104.
23. Rubinstein LV, Korn EL, Freidlin B, Hunsberger S, Ivy SP, Smith MA. Design Issues of Randomized Phase II Trials and a Proposal for Phase II Screening Trials. J Clin Oncol. 2005;23(28):7199–206. https://doi.org/10.1200/JCO.2005.01.149.
24. Hutson AD, Wilding GE. An examination of the relative impact of type I and type II error rates in phase II drug screening trial queues. Pharm Stat. 2012;11(2):157–62. https://doi.org/10.1002/pst.529.
25. Lee E, Whitehead A, Jacques R, Julious S. The statistical interpretation of pilot trials: should significance thresholds be reconsidered? BMC Med Res Methodol. 2014;14(1):41. https://doi.org/10.1186/1471-2288-14-41.
26. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ. 2021;374:n2061. https://doi.org/10.1136/bmj.n2061.
27. 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. 2018;27(12):3612–27. https://doi.org/10.1177/0962280217708906.
28. Wilson DT, Wason JMS, Brown J, Farrin AJ, Walwyn REA. Bayesian design and analysis of external pilot trials for complex interventions. Stat Med. 2021;40(12):2877–92. https://doi.org/10.1002/sim.8941.
29. Lv D, Grayling M, Zhang X, Zhao Q, Zheng H. A Bayesian approach to pilot-pivotal trials for bioequivalence assessment. BMC Med Res Methodol. 2023;23(1):301. https://doi.org/10.1186/s12874-023-02120-2.
30. Willan AR, Thabane L. Bayesian methods for pilot studies. Clin Trials. 2020;17(4):414–9. https://doi.org/10.1177/1740774520914306.

Publisher's Note