Journal Title
XX(X):1–13

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Three outcome designs for pilot trial progression criteria

Duncan T. Wilson[1]

Abstract

The decision of if and how to progress to a definitive trial following a pilot study is often guided by progression criteria. Increasingly, progression criteria with three outcomes (stop, go, and an intermediate) are being used, but there is little methodological work examining how the thresholds or the pilot sample size should be determined. We review existing three-outcome designs and consider if and how they can be used to provide a formal statistical framework for pilot trials. We conclude that, contrary to previous claims, three-outcome designs do not improve efficiency, rather requiring a larger sample size than two-outcome alternatives. We also argue that pre-specified progression criteria are fundamentally ill-suited to pilot trials which lead to modifications of the trial design or the intervention before the main trial.

Keywords

Clinical trial, pilot trial, external pilot, progression criteria, sample size

1 Introduction

When there is some uncertainty about the feasibility of a planned randomised clinical trial (RCT), a small version of the trial, known as an external pilot, can be conducted. The pilot data can be used to estimate various parameters of interest, such as the follow-up rate, and these estimates can then be used to decide if and how to proceed to the main trial. Investing in a pilot trial can identify potential issues at an early stage, make a successful main trial more likely, and reduce research waste.

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

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

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

The CONSORT extension to randomised pilot trials notes that progression decisions can be guided by so-called *progression criteria* [1]. A simple progression criteria will prescribe a single threshold value, such that we should progress to the main trial only if the pilot estimate exceeds the threshold. When progression criteria are specified for several parameters, as is typically the case , these can be combined by progressing to the main trial only if all of the estimates exceed their respective thresholds. In the UK, the NIHR requires that all pilot trials to have these progression criteria pre-specified at the design stage .

Increasingly, more complex 'traffic light' progression criteria are being used. These stipulate two threshold values for a given parameter of interest. If the estimate falls below the lower of these, the decision is to stop; if the estimate falls above the higher threshold, the decision is to proceed immediately to the main trial; and if the estimate falls between the two thresholds, an intermediate decision is reached.

Three motivations for using a three-outcome system in pilot trials can be found in the literature. Firstly, it has been argued that making strict stop/go decisions based on a single threshold may lead to the wrong decision by chance alone: "*estimates of rates in pilot trials may be subject to considerable uncertainty, so that it is best to be cautious about setting definitive thresholds that could be missed simply due to chance variation. In fact it is becoming increasingly common for investigators to use a traffic light system for criteria used to judge feasibility*" [1]. By allowing for a buffer zone in between the 'stop' and 'go' regions, the probability of landing in these critical areas, and therefore of making incorrect decisions, will be reduced.

A second motivation is to allow other information to inform the progression decision in the event of a 'borderline' result. This might be attractive in the context of a pilot trial, where several aspects (quantitative and qualitative) are often being studied and we would like to be able to take these results into account rather than everything being decided on the result of a handful of estimates and their thresholds. Moreover, it could be argued this is how decisions are typically made even when a usual two-outcome system is nominally used.

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, we might consider moving from a postal follow-up strategy to one based on contacting the participants over the phone. This is perhaps the most prevalent reasoning given for three outcomes in pilot trials, with the intermediate 'amber' decision often explicitly referred to as 'amend' or 'adjust' (examples).

Although prevalent, there is little methodological guidance to help researchers decide what threshold values to use in their progression criteria. The related question of pilot trial sample size is also methodologically undeveloped, with work in this area almost exclusively focussing on pilots whose primary function is to estimate the primary outcome variance to inform the main trial sample size calculation. These methods nevertheless are used when this is not the primary aim of the pilot, often in the form of simple 'rules-of-thumb' [3–5]. There are a

number of papers, however, which propose designs for trials with three outcomes. Although typically proposed for phase II trials of cancer treatments, these three-outcome designs were motivated by the same factors given above, and so may provide a useful framework for the pilot trial setting.

In this paper we consider if, and how, three-outcome designs can be used to determine optimal pilot progression criteria and sample size. We begin in Section 2 by arguing that progression criteria are mathematically equivalent to hypothesis tests and best viewed as such. In Section 3 we review three-outcome trial designs. We examine their statistical properties as applied to the progression criteria problem, and consider whether or not they can help with any of the three motivating goals (addressing sampling variability, incorporating other information, or allowing for adjustments), in Section 4. We illustrate our arguments via an example in Section 5, before concluding with a discussion in Section 6.

2 Progression criteria as hypothesis tests

Throughout this article we will focus on the simple case of a single parameter being assessed in the pilot trial: the adherence rate in the intervention arm. This will be sufficient to motivate and illustrate all our arguments. In particular, consider a two-outcome 'stop/go' progression criterion for the adherence rate in the intervention arm of a pilot trial. Denoting the true (but unknown) adherence rate by ρ, the number of participants who adhere in the pilot trial will follow a binomial distribution with parameters ρ and the intervention-arm sample size, denoted n. This data is then used to calculate the estimated adherence rate, denoted $\hat{\rho}$.

One way to make a stop/go decision based on $\hat{\rho}$ is through a hypothesis test, which can be constructed 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. We denote by x the critical value where if $r\hat{h}o \geq x$ we reject the null hypothesis and make a 'go' decision, otherwise choosing to 'stop'. We then choose values of n and x which minimise n whilst satisfying the type I and II error rate constraints

$$\alpha = P[\hat{\rho} > x \mid \rho = \rho_0] \leq \alpha^* \tag{1}$$

$$\beta = P[\hat{\rho} \leq x \mid \rho = \rho_1] \leq \beta^*. \tag{2}$$

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 'if $\hat{\rho} \geq x$ then *go* otherwise *stop*', it is mathematically equivalent to a hypothesis test. For example, setting $n = 15$ and $x = 10/15$, we can plot the probability of making a 'go' decision as a function of the unknown parameter ρ. This power curve is illustrated in Figure 1.

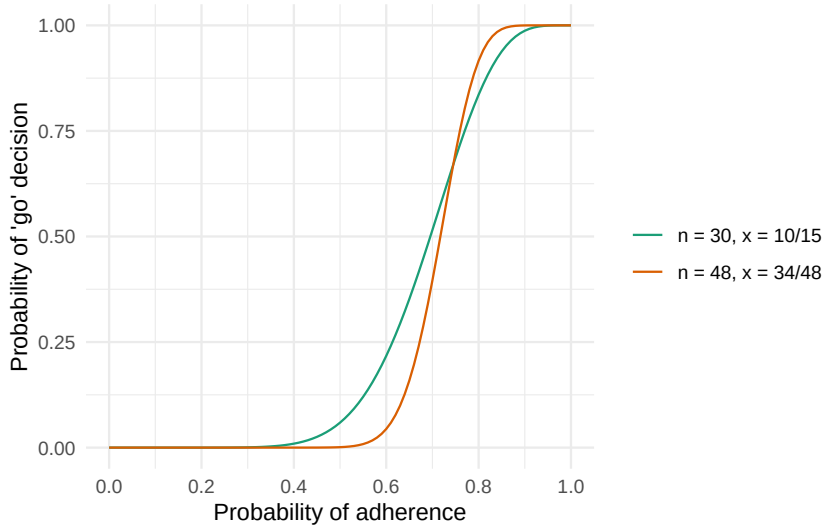

Figure 1. Power curves.

If we suppose the null and alternatives hypotheses of interest are $\rho_0 = 0.6, \rho_1 = 0.8$ respectively, choosing $n = 15, x = 10/15$ gives error rates of $\alpha = 0.22$ and $1 - \beta = 0.84$. Alternatively, we can constrain the error rates to, for example, $\alpha^* = 0.05$ and $\beta^* = 0.1$, in which case the smallest possible sample size is $n = 48$ and the progression threshold is $x = 34/48$. The power curve for this design is plotted in Figure 1 for comparison.

3 Three-outcome clinical trial designs

The two-outcome hypothesis test framework described in Section 2 can be extended to three outcomes by using two critical values, x_0 and x_1, where we 'stop' if $\hat{\rho} \leq x_0$ and 'go' if $\hat{\rho} > x_1$. When the estimate falls $x_0 < \hat{\rho} \leq x_1$ we obtain an intermediate result, with a prescribed action that will depend on the specific context.

In order to guide the choice of the thresholds x_0, x_1 and the sample size n, some operating characteristics have been proposed. One proposal[6] defines

$$\begin{aligned} P[\hat{\rho} > x_1 | \rho = \rho_0] &= \alpha, \\ P[\hat{\rho} \leq x_0 | \rho = \rho_1] &= \beta \\ P[x_0 < \hat{\rho} \leq x_1 | \rho = \rho_0] &= \lambda \\ P[x_0 < \hat{\rho} \leq x_1 | \rho = \rho_1] &= \delta. \end{aligned} \tag{3}$$

Thus, α is the probability of making a 'go' decision under the null hypothesis $\rho = \rho_0$, i.e. a type I error rate. Similarly, β represents the usual type II error rate. The remaining operating characteristics, λ and δ, are the probabilities

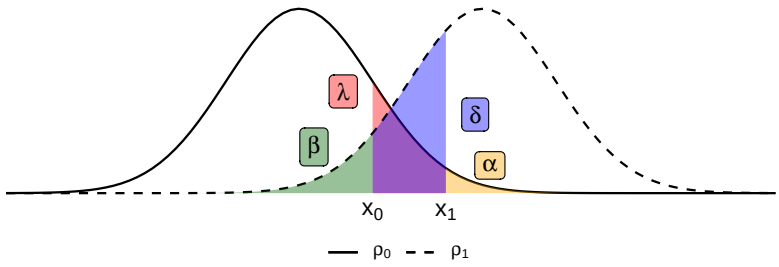

Figure 2. Operating characteristics for Sargent *et al.*[6] three-outcome design. The curves represents the sampling distribution of the estimate under different values of the parameter ρ.

of obtaining an intermediate result under the null and alternative hypotheses respectively. These four operating characteristics are illustrated in Figure 2. The proposal is to set constraints on these four operating characteristics and choose n, x_0, x_1 to minimise n whilst satisfying these constraints.

An alternative three-outcome design[8] uses the operating characteristics

$$\begin{aligned} P[\hat{\rho} > x_0 | \rho = \rho_0] &= \alpha, \\ P[\hat{\rho} \leq x_\cdot | \rho = \rho_1] &= \beta \\ P[\hat{\rho} \leq x_0 | \rho = \rho_m] &= \gamma_L \\ P[x_1 < \hat{\rho} | \rho = \rho_m] &= \gamma_U. \end{aligned} \tag{4}$$

Note that the operating characteristics α and β are different to those in (3), now being the probability of *not stopping* under the null and *not going* under the alternative. The operating characteristics γ_L, γ_U are the probabilities of making 'stop' and 'go' decisions when the parameter takes a specific value ρ_m between the null and alternative hypotheses, with the suggestion to set $\rho_m = (\rho_1 - \rho_0)/2$. Figure 3 illustrates the operating characteristics for this design.

4 Three-outcome designs for progression criteria

As noted in Section 1, adding a third outcome to pilot trial progression criteria has been motivated on grounds of i) improved (statistical) efficiency; ii) the need to incorporate other information into progression decisions; and iii) the ability to make modifications to the intervention or the trial design before commencing the main trial. In this section we discuss if the three-outcome design frameworks reviewed in Section 3 can be used to address these goals.

4.1 Statistical efficiency

To explore the question of sampling variability, we will apply the framework of Sargent *et al.*[6] which, it is claimed, will reduce the required sample size by allowing a region of uncertainty.

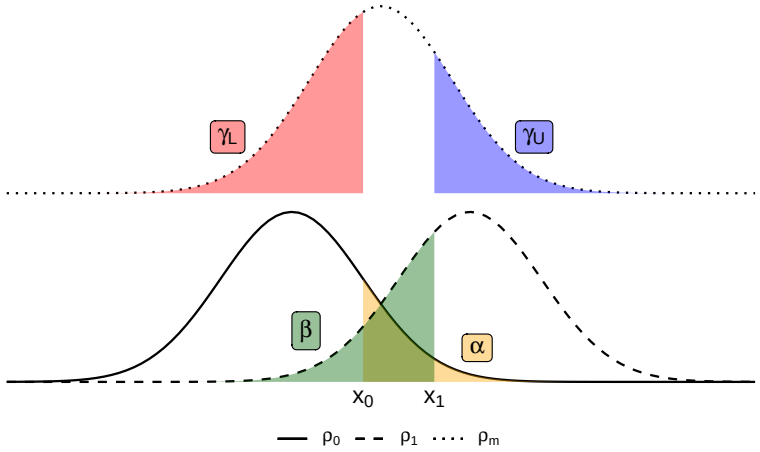

Figure 3. Operating characteristics for Storer's[8] three-outcome design. The curves represents the sampling distribution of the estimate under different values of the parameter ρ.

For any given choice of nominal α and β, as defined in (3), an optimal three-outcome design will indeed have a lower sample size than a two-outcome design, 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 $\alpha \leq 0.05$ and $\beta \leq 0.1$. In contrast, by allowing $\lambda \leq 0.1$ and $\delta \leq 0.1$ in a three-outcome design, we can obtain $\alpha \leq 0.05$ and $\beta \leq 0.1$ with only $n = 42$. This would suggest three outcome designs are indeed more efficient[6,7]. But this argument rests on a false equivalence, as we will show.

In[6] the authors state that "We can interpret α the usual manner, i.e., the maximum probability of making an erroneous decision by rejecting the null hypothesis when in fact it is true". To "reject the null hypothesis" here is to arrive at a 'go' decision and proceed to the main trial. This decision, however, can be arrived at in two ways: directly, by obtaining $\hat{\rho} > x_1$; or indirectly, by first obtaining an intermediate result $x_0 < \hat{\rho} < x_1$ and then deciding to proceed. Ultimately, we must always decide to either 'stop' or 'go' to the main trial following an intermediate result in the pilot. Define

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

$$\eta_1 = P[\text{decide to 'stop'} \mid \rho = \rho_1, x_0 < \hat{\rho} \leq x_1]. \tag{6}$$

For example, η_0 is the probability of making a 'go' decision following an intermediate result and when the true parameter value is ρ_0. The correct probability of making a 'go' decision when $\rho = \rho_0$ (i.e., the type I error rate), then, is not α but

$$\bar{\alpha} = \alpha + \eta_0 \lambda.$$

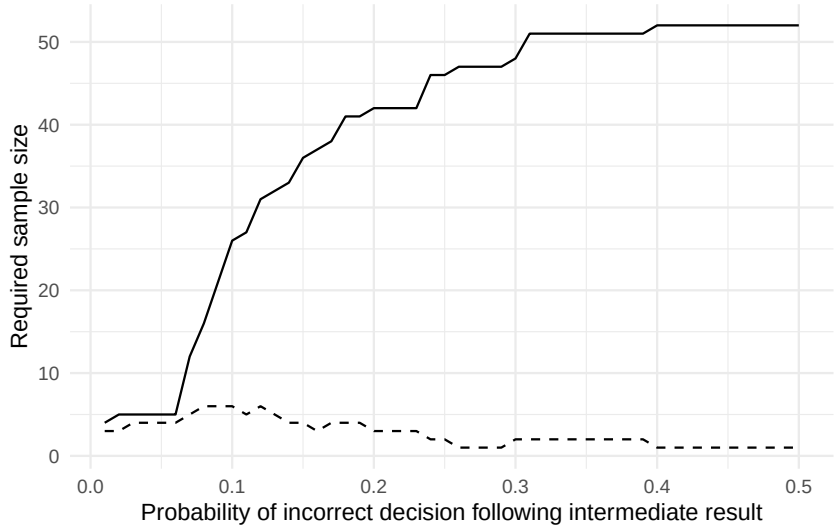

Figure 4. Minimum required sample size for a three-outcome design as a function of η (solid line), along with the corresponding size of the intermediate zone $x_1 - x_0$ (dashed line).

Similarly, the correct type II error rate is

$$\bar{\beta} = \beta + \eta_1 \delta.$$

Under this reformulation, an optimal three-outcome design can be found by first estimating the probabilities η_0, η_1, then setting constraints on the actual type I and II error rates $\bar{\alpha}, \bar{\beta}$, 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 $\eta_0 = \eta_1 = \eta$; that is, the probability of eventually making the wrong decision following an intermediate result is the same for $\rho = \rho_0$ as for $\rho = \rho_1$. Returning to our example where $\rho_0 = 0.5, \rho_1 = 0.7, \bar{\alpha} \leq 0.05$ and $\bar{\beta} \leq 0.1$, recall that the optimal two-outcome design had a sample size of $n = 53$. We plot the required sample size for $0 \leq \eta \leq 0.5$ in Figure 4, which also illustrates the size of the intermediate zone $x_1 - x_0$ (dashed line).

When $\eta = 0.5$, in which case we can only guess at the correct decision following an intermediate result, the optimal sample size is $n = 52$. We might have expected the three- and two-outcome designs to coincide at this point, and indeed this is the case if we employ a normal approximation for $\hat{\rho}$. When using exact binomial probabilities as we have done here, the optimal three-outcome design has $x_0 = 31, x_1 = 32$, in comparison to the $x = 32$ of the two-outcome design. This suggests the true optimal x for a two-outcome design lies in the interval $(31, 32)$. We have not considered $\eta > 0.5$ here, since this represents a decision making ability worse than making a random choice and so the optimal design remains the usual two-outcome design.

When $\eta = 1$, in which case we can make the correct decision with certainty whenever an intermediate result is obtained, the optimal three-outcome design is the trivial $n = 2, x_0 = 0, x_1 = 1$. Figure 4 shows that, for a 20% reduction from the $n = 53$ two-sample design down to $n = 42$, we would require $\gamma = 0.2$. That is, we must be confident that 80% of the times we get an intermediate result, but with a true $\rho = \rho_i (i = 0, 1)$, we will make the correct progression decision. In the context of our simple example, it is hard to think of a reason why our judgements would be so reliable. In particular, the estimate $\hat{\rho}$ is a sufficient statistic for the parameter ρ, and so we cannot hpe to obtain any other information relevant to this particular judgement. The appropriate default would then seem to be $\eta = 0.5^*$, in which case the optimal three-outcome design will reduce to a two-outcome design. We can conclude that three-outcome progression criteria are not capable of improving statistical efficiency in pilot trials.

4.2 Incorporating other information

Although three-outcome designs cannot improve statistical efficiency, they may be attractive as a way to allow other information to formally enter into the progression decision making process. This is the view which motivated Storer's three-outcome proposal[8] as described in Section 3, and so we will adopt their framework here. As in Section 4.1, we use the corrected type I and II error rates $\bar{\alpha}, \bar{\beta}$, but here we assume $\gamma = 0.5$. We then encourage the design to have an appropriate intermediate zone by constraining the operating characteristics γ_L, γ_U defined in (3), limiting the chance of making a conclusive 'stop' or 'go' decision when $rho = \rho_m$ and we would ideally like to bring other factors into play. With $\rho_0 = 0.5, \rho_1 - 0.7, \bar{\alpha} \leq 0.05$ and $\bar{\beta} \leq 0.1$ as before, we set $\rho_m = (\rho_1 - \rho_0)/2 = 0.6$ and find optimal designs for a range of γ^*, where $\gamma_L, \gamma_U \leq \gamma^*$. The sample size of these designs is plotted in Figure 5.

When we set the constraint at $\gamma \leq 0.5$, no intermediate zone is required and so the optimal design, which minimises the required sample size, is the usual two-outcome design. As we decrease the nominal level on this constraint, we require a larger probability of obtaining an inconclusive result conditional on $\rho = \rho_m$. This leads to an increasing width of the intermediate zone alongside increasing sample size. The required increse in sample size beyond the two-outcome design can be substantial. For example, to ensure at least a 20% chance of obtaining an intermediate result when $\rho = \rho_m$, we must increase the sample size from $n = 53$ to $n = 115^\dagger$. Thus, providing the associated increase in sample size is considered worhtwhile, three-outcome designs could be used in pilot trials to ensure other information can be considered in the event of a 'borderline' value in the parameter of interest, with some desired probability.

---

*In fact, it is plausible that $\eta > 0.5$ if the final stop/go decision is influenced by other endpoints which are negatively correlated with the endpoint of interest. As noted already, the optimal design in such situations is the same as for $\eta = 0.5$.

†This is lower than that required if we use the (incorrect) α and β as given in (3), since $\alpha \geq \bar{\alpha}$ and $\beta \geq \bar{\beta}$.

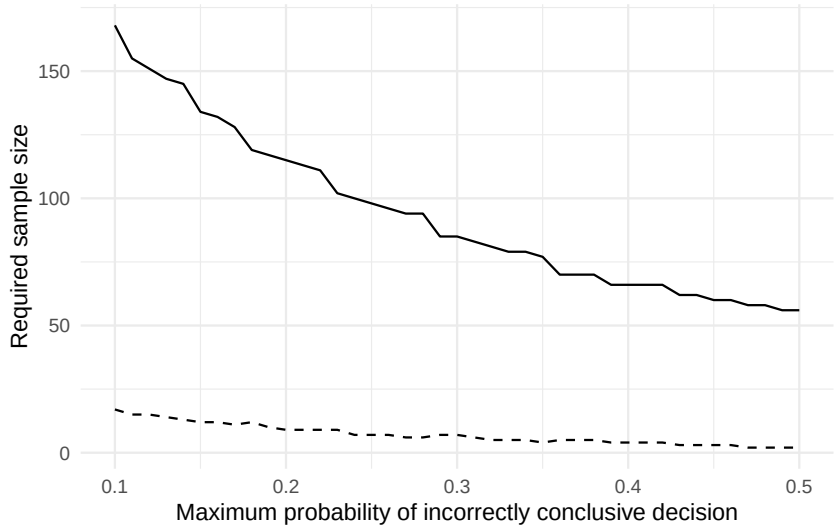

Figure 5. Minimum required sample size for a three-outcome design as a function of γ (solid line), along with the corresponding size of the intermediate zone $x_1 - x_0$ (dashed line).

4.3 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).

An alternative rationale for allowing three outcomes in a pilot trial is to enable some kind of adjustments to the trial design or intervention to be made upon obtaining a 'borderline' result in the pilot. Under this setup, an intermediate result leads to an eventual 'go' decision, but only after making some modification(s). Under the assumption that such modifications are always successful, the type I and II error rates are as above.

We assume that a parameter in the amber region is salvageable, and will be salvaged if we get an amber decision. A red parameter is unsalvagable, regardless of the outcome. So, we could set up OCs: prob of infeasible trial = prob of a or g under R; prob of discarding intervention = prob of r under A or G, and of g under A; and prob of needless modification = prob of a under R or G. These are as in WP2.

This only makes sense if we know what modifications will look like. In some cases this could feasibly apply. For example, we could consider improving recruitment by increasing the number of centres (although even here the size of the effect might be uncertain - see the Bayesian approach in[9]). But generally, the point of running the pilot is to identify issues which were not forseen. See the recruitment issue in[10] - a staffing problem, which could be remidied and

with a huge impact on recruitment rate, but where we would be guided into a stop decision becasue we did not allow for such a large modification effect. In the other direction, we might assume a large modification will be possible but not find any, landing us with a modify and go decision when we don't think this is sensible. In either case, we would ignore the pre-specified thresholds and use our judgement. But if this is the plan, why are we pre-specifying thresholds in the first place? If we are not certain about relying on our judgements, the best course of action may be to run another pilot with the changes made to check they are adequate. An adaptive design might make this process more efficient; or a factorial design trying lots of approaches at once and identifying the best to use in the main trial.

5 Illustration

6 Discussion

We began by highlighting that two-outcome pilot progression criteria are mathematically equivalent to a hypothesis test, but not designed with respect to the corresponding error rates. This might be a sensible reaction to the hegemony of this particular method, which has many shortcomings and expectations regarding arbitrary error nominal error rates. Nevertheless, if we are working in a frequentist framework then the power curve as a function of ρ and n must be useful information, even if we use it more sensibly than via rigid type I and II error rate constraints. Note in particular that if progression criteria thresholds indicate the *parameter* value which we would consider on the border of stop/go, then this will lead to a test with 50% power at this inflection point - and it has been argued elsewhere that this is sensible. The choice of sample size can then be made conditional on this, using some other criteria, which could be power at a another specific point.

Pilot trial progression criteria are typically viewed as flexible guides to decision making, rather than cast-iron rules. This makes complete sense, and applies equally to the usual primary analysis tests where the p-value is not the last word in decision making. So, although we are not suggesting the rules are strictly followed, they can still be useful as a guide to decision making and choice of sample size.

An alternative approach to SSD is to focus on precision, rather than testing. But this appears to be much harder to reason about - how much precision do we want? Perhaps this is something which will emerge from experience?

Intermediate outcomes in pilots are never (?) used in an adaptive sense, i.e. where intermediate means collect more data. In contrast, this is the most common type of three outcome design in the drug trial literature, where many two-stage designs exist. An adaptive approach would help with objective 1 (efficiency), and could be used with a terminal intermediate outcome (in contrast to two stage designs which will have stop/go at the end to force a decision), using the above analysis to work out how much more sample size is needed to provide this flexibility. And indeed this aligns with[6], who proposed one and two stage

versions of their design. In the context of allowing modifications, we have the same problem of not being able to anticipate what they are or their effect. If we know what they are at least, an adaptive MAMS or factorial type approach evaluating several strategies and picking the winner - might be more efficient if data can be pulled for other parameters (e.g. if we have several recruitment strategies, but one follow-up, all arms can inform the follow-up estimate).

Multiple testing procedures for phase II trials, which we can think of as a three outcome design at each stage where the middle outcome is "go to next stage and test again". So, quite different to our problem because we want to change the intervention and then possibly start over again [11]. Similarly, [12] is just a two stage design, similar to [13].

Motivated by the fact that we can't have both a low sample size and low alpha and betas. Suggests a three outcome model where the decisions are stop / slow development / accelerated development, allowing two extra types of errors to be allowed and controlled. Seems to be equivalent to a two-stage design since the middle outcome is to 'do another study' [14], but maybe more pertinent to our external pilot situation.

Decision-theoretic (non-Bayesian) approach, giving different costs to different types of error. Motivated by contrasting one and two sided tests, looking for a method which can tell us if there is a significant direction in either direction or if there is no difference [15].

A review and comparison of early phase trial approaches [16], including the three outcome design of [14].

A Bayesian adaptive approach, anticipating the effect of recruitment adjustments [9].

Acknowledgements

Acknowledgements.

Declaration of conflicting interests

The Authors declare that there is no conflict of interest.

Funding

This work was supported by the Medical Research Council [grant number MR/N015444/1].

References

1. Eldridge SM, Chan CL, Campbell MJ et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *BMJ* 2016; : i5239DOI: 10.1136/bmj.i5239. URL `http://dx.doi.org/10.1136/bmj.i5239`.
2. S G and AP D. Defining and identifying the effect of treatment on the treated. In Illari P, Russo F and Williamson J (eds.) *Causality in the Sciences*. Oxford Univ. Pre, 2011. pp. 728 – 49. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.370.9724&rep=rep1&type=pdf`.

3. Browne RH. On the use of a pilot sample for sample size determination. *Statistics in Medicine* 1995; 14(17): 1933–1940. DOI:10.1002/sim.4780141709. URL http://dx.doi.org/10.1002/sim.4780141709.
4. Teare M, Dimairo M, Shephard N et al. Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials* 2014; 15(1): 264. DOI:10.1186/1745-6215-15-264. URL http://www.trialsjournal.com/content/15/1/264.
5. Whitehead AL, Julious SA, Cooper CL et al. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Statistical Methods in Medical Research* 2015; DOI:10.1177/0962280215588241. URL http://smm.sagepub.com/content/early/2015/06/27/0962280215588241.abstract. http://smm.sagepub.com/content/early/2015/06/27/0962280215588241.full.pdf+html.
6. Sargent DJ, Chan V and Goldberg RM. A three-outcome design for phase II clinical trials. *Controlled Clinical Trials* 2001; 22(2): 117 – 125. DOI:http://dx.doi.org/10.1016/S0197-2456(00)00115-X. URL http://www.sciencedirect.com/science/article/pii/S019724560000115X.
7. Hong S and Wang Y. A three-outcome design for randomized comparative phase II clinical trials. *Statist Med* 2007; 26(19): 3525–3534. DOI:10.1002/sim.2824. URL http://dx.doi.org/10.1002/sim.2824.
8. Storer BE. A class of phase ii designs with three possible outcomes. *Biometrics* 1992; 48(1): 55–60. URL http://www.jstor.org/stable/2532738.
9. Hampson LV, Williamson PR, Wilby MJ et al. A framework for prospectively defining progression rules for internal pilot studies monitoring recruitment. *Statistical Methods in Medical Research* 2017; 0(0): 0962280217708906. DOI:10.1177/0962280217708906. URL http://dx.doi.org/10.1177/0962280217708906. PMID: 28589752, http://dx.doi.org/10.1177/0962280217708906.
10. Avery KNL, Williamson PR, Gamble C et al. Informing efficient randomised controlled trials: exploration of challenges in developing progression criteria for internal pilot studies. *BMJ Open* 2017; 7(2): e013537. DOI:10.1136/bmjopen-2016-013537. URL https://doi.org/10.1136%2Fbmjopen-2016-013537.
11. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982; 38(1): 143–151. URL http://www.jstor.org/stable/2530297.
12. Shuster J. Optimal two-stage designs for single arm phase ii cancer trials. *Journal of Biopharmaceutical Statistics* 2002; 12(1): 39–51.
13. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989; 10(1): 1 – 10. DOI:http://dx.doi.org/10.1016/0197-2456(89)90015-9. URL http://www.sciencedirect.com/science/article/pii/0197245689900159.
14. Brown MJ, Chuang-Stein C and Kirby S. Designing studies to find early signals of efficacy. *Journal of Biopharmaceutical Statistics* 2012; 22(6): 1097–1108. DOI:10.1080/10543406.2011.570466. URL https://doi.org/10.1080/10543406.2011.570466. PMID: 23075010, https://doi.org/10.1080/10543406.2011.570466.
15. Emerson JD and Tritchler D. The three-decision problem in medical decision making. *Statistics in Medicine* 1987; 6(2): 101–112. DOI:10.1002/

sim.4780060202. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780060202. https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780060202.

16. Kirby S and Chuang-Stein C. A comparison of five approaches to decision-making for a first clinical trial of efficacy. *Pharmaceutical Statistics* 2016; 16(1): 37–44. DOI:10.1002/pst.1775. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.1775. https://onlinelibrary.wiley.com/doi/pdf/10.1002/pst.1775.

Appendix