Three outcome designs for pilot trial progression criteria

Duncan T. Wilson[1]

Abstract

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 in advance. 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. By investing at the pilot stage, we aim to identify and correct potential issues prior to the main trial, thereby reducing overall research waste by ensuring it can be conducted successfully.

The recommended method for using pilot estimates to make progression decisions is through so-called *progression criteria*. A simple progression criteria will prescribe a threshold value such that we should progress to the main trial only if the pilot estimate exceeds the threshold, otherwise deciding to not progress. 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

[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

intermediate decision is reached. Three motivations for using a three outcome system in pilot trials can be found in the literature. Firstly, making strict stop/go decisions based on a single threshold has been argued to be in danger of incorrectly leading to the wrong decision by chance alone. By allowing for a buffer zone in between the stop and go regions, we might hope to avoid bad decisions. This argument is paralled in the three outcome literaure, which argue that the designs can improve efficiency and either increase power, or reduce the required sample size.

A second motivation is to allow other information to enter into the progression decision making process in the event of a 'borderline' result on the parameter in question. Again, this was noted as motivation for a three-outcome design, which argued that this is how decisions are typically made even when a usual two-outcome system was nominally used. 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 tests.

A final reason for an intermediate outcome is to provide the flexibility needed to make some adjustment to the intervention or the main 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 'amber' decision often explicitly referred to as 'amend' or 'adjust' (examples). This is quite specific to pilot trials, and has not been considered in the three outcome literature.

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 lacking methodologically, 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'[1][2][3].

In this paper we aim to evaluate progression criteria from a statistical perspective and determine when and how they should be used in pilot trials. We begin in Section 2 by arguing that progression criteria are best viewed as a hypothesis test. In Section ?? we review three-outcome hypothesis test designs, examine their the statistical properties as applied to the progression criteria problem, and consider whether or not they can help with either goal of addressing sampling variability or allowing for adjustments. We illustrate our points via an example in Section ??, before concluding with a discussion in Section ??.

2   Progression criteria as hypothesis tests

Throughout the 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, which extend naturally to the multi-parameter case. Consider a stop/go progression criteria for the adherence rate in a pilot trial with N participants in the intervention arm. Denoting the true (but unknown) adherence rate by ρ, the number of

participants who adhere in the pilot trial, denoted A, will follow a binomial distribution with parameters ρ and n.

One way to make a stop/go decision based on A 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 p the critical value where if $A \geq p$ we reject the null hypothesis and make a 'go' decision, otherwise choosing to 'stop'. We then choose values of n and p which minimise n whilst satisfying the following type I and II error rate constraints:

$$\alpha = P[\hat{\rho} \geq p | \rho = \rho_0] \leq \alpha^* \tag{1}$$
$$\beta = P[\hat{\rho} < p | \rho = \rho_1] \leq \beta^*. \tag{2}$$

Alternatively, we can work backwards and take any given choice for n and p and calculate the resulting error rates. In particular, whenever a pilot trial progression criteria is specified in the form of 'if $A \geq p$ then *go* otherwise *stop*', it is mathematically equivalent to a hypothesis test as described above. The error rates arising from such a procedure are, however, unspecified, and the choice of n and p has not been informed by them.

It may be that p is typically chosen to be the point such that, if $A = p$, we would be indifferent between stopping and going on to the main trial. This would effectively fix the power curve at 0.5 for $\rho = p/n$, which has been described as the only non-arbitrary point on a power curve. The choice of n will then determine power for the rest of the parameter space. It would appear that calculating this power explicitly, either for the whole parameter space of for the specific points ρ_0 and ρ_1, could only help find an appropriate value for n.

3 Three outcome progression criteria

3.1 Three-outcome clinical trial designs

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[4].

Decision-theoretic (non-Bayesian) approach, giving different costs to different types of error. Motivated by constrasting 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[5].

A Class of Phase II Designs with Three Possible Outcomes[6].

A three-outcome design for randomized comparative phase II clinical trials[7].

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"[8].

A review and comparison of early phase trial approaches[9], including the three outcome design of[8].

Although far less prevalent than the standard two-outcome hypothesis test described in Section 2, various designs for trials with three outcomes have been proposed.

three outcome tests have been proposed. Following the same argument, we can show that a three-outcome progression criteria can be viewed as mathematically equivalent to a three-outcome design such as that described in[10]. This design specifies n and two threshold, p_0 and p_1, and determines these based on four operating characteristics:

$$P[\hat{\rho} \geq p_1 | \rho = \rho_0] = \alpha, \tag{3}$$
$$P[\hat{\rho} < p_0 | \rho = \rho_1] = \beta \tag{4}$$
$$P[p_0 \leq \hat{\rho} < p_1 | \rho = \rho_0] = \lambda \tag{5}$$
$$P[p_0 \leq \hat{\rho} < p_1 | \rho = \rho_1] = \delta. \tag{6}$$

By setting upper limits on α, β, η and π, we can identify the optimal trial design n, p_0, p_1.

As noted in Section 1, there are two motivations for using three-outcome progression criteria. In the remainder of this section we discuss if this three-outcome hypothesis testing framework can be used to address either of these goals.

3.2 Addressing sampling variability

For any given choice of nominal α and β, a three outcome design will have a lower required sample size providing the nominal λ and/or δ are 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, as argued in[7]. But this argument rests on a false equivalence, as we will show.

In[10] 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 $\hat{\rho} \geq p_1$; or indirectly, by first obtaining an intermediate result $p_0 \geq \hat{\rho} < p_1$ and then, after due consideration of whatever other information has been obtained in the pilot, 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

$$\eta_0 = P[\text{decide to 'go'} | \rho = \rho_0, p_0 \geq \hat{\rho} < p_1] \tag{7}$$
$$\eta_1 = P[\text{decide to 'stop'} | \rho = \rho_1, p_0 \geq \hat{\rho} < p_1], \tag{8}$$

where e.g. η_0 is the probability of making a 'go' decision following an intermediate result, when the true parameter value is ρ_0. The actual type I error rate, i.e. the probability of making a 'go' decision when $\rho = \rho_0$, is

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

Similarly, the actual type II error rate is

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

Under this reformulation of the three outcome design, we now estimate the probabilities η_0, η_1, set constraints on the actual type I and II error rates $\bar{\alpha}, \bar{\beta}$, and find the optimal design n, p_0, p_1.

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 $\bar{\beta} \leq 0.1$, we plot the required sample size for $0 \leq \eta \leq 0.5$ in Figure **??**.

We see that when $\eta = 0.5$, in which case we can only guess at the correct decision following an intermediate result, the optimal sample size is $n = 53$ and the three outcome design reduces to the two outcome design. At the other extreme, if $\eta = 1$ and we can decide without error, the tree outcome design reduces to a trivial one outcome design where an intermediate result is always obtained, upon which the correct stop/go decision can be reached. We see that for the kinds of sample size reductions given in [10], we require η of around 0.2. That is, we need to have an 80% chance of correctly deciding to stop when we get an intermediate decision and the true parameter value is $\rho = \rho_0$; and similarly for deciding to proceed when $\rho = \rho_1$. In the context of our example, it is hard to think of a reason why our judgements would be so reliable. The appropriate default would seem to be $\eta = 0.5$. Unless we can argue that $\eta << 0.5$, the supposed efficiency of a three outcome design is an illusion, and its use will therefore not achieve the goal of better handling the sampling variability of parameter estimates in pilot trials.

3.3 Incorporating other information

A corrolary of the above analysis is that when $\eta = 0.5$ the optimal deign is a two outcome design, with no amber region. If we want to have an amber region to provide opportunity to use other information in the case of a borderline result, we can go about this in a number of ways.

We could decide on a fixed width, and then design to minimise n s.t. the constraints as above.

We could allow the width to vary, but add another operating characteristic constraint, e.g. conditional on being midway between the null and alternative, we want a certain probability of making an amber decision.

However we go about it, we know from the preceding section that to maintain the actual type I and II errors whilst introducing an amber region, we will have to *increase* the sample size.

3.4 Allowing for adjustments

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 [11]). But generally, the point of running the pilot is to identify issues which were not forseen. See the recruitment issue in [12] - 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.

4   Illustration

5   Discussion

We began by highlighting that 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. 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.

References

1. 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`.

2. 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`.

3. 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`.

4. 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`.

5. 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`.

6. 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`.

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. 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`.

9. 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`.

10. 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`.

11. 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`.

12. 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`.

Appendix