Three outcome designs for pilot trial progression criteria

Journal Title
XX(X):??–??

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

Duncan T. Wilson[1]

Abstract

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 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,

[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]*

an intermediate decision is reached. Including an intermediate outcome has been motivated in part by the imprecision of the pilot estimate and the desire to avoid arriving at strict 'stop/go' decisions by chance. Thus, an intermediate decision could be to pause, consider the bigger picture illustrated by the pilot trial, and make a progression decision informed by other factors. Another 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.

Although prevalent, there is little methodological guidance to help researchers decide what threshold values to use in their progression criteria and what sample size is required. In this paper we will evaluate progression criteria from a statistical perspective and determine when and how they should be used in pilot trials. 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. We will first argue in that progression criteria are best viewed as a hypothesis test. We then consider the statistical properties of a three-outcome hypothesis test, and argue that it will often not achieve the goals of addressing sampling variability or allowing for adjustments.

2 Progression criteria as hypothesis tests

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 A will follow a binomial distribution with parameters ρ and n. For simplicity, we will approximate the sampling distribution of the estimate $\hat{\rho} = A/n$ with a normal distribution with mean ρ and variance $\rho(1-\rho)/n$.

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 arrive at a 'stop' decision with high probability, denoted $1-\alpha$. Similarly, we identify ρ_1 such that if $\rho \geq \rho_1$ we would like to arrive at a 'go' decision with high probability $1-\beta$. We then choose values of p and n such that

$$P[\hat{\rho} \geq p | \rho = \rho_0] = \alpha, \tag{1}$$

$$P[\hat{\rho} < p | \rho = \rho_1] = \beta. \tag{2}$$

These probabilities are the type I and II error rate of the test, respectively.

Alternatively, we can work backwards and take any given choice for n and p and calculate the resulting error rates. Thus, the standard approach of choosing n based on a simple rule-of-thumb and choosing p *somehow* can be viewed as mathematically equivalent to a formal hypothesis test, but with error rates which are unspecified due to the null and alternative hypotheses ρ_0, ρ_1 being unspecified. We would therefore argue that an explicit hypothesis testing view of stop/go progression criteria, where these error rates are calculated and used to inform the choice of p and n, is sensible.

It may be that p is typically chosen to be the point such that, if $\rho = p$, we would be indifferent between stopping and going on to the main trial. This would effectively fix the power curve at 0.5 for $\rho = p$, where this 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

Although far less prevalent than the standard two-outcome hypothesis test described in Section **??**, 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 **?** . 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 **??**, 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.

3.1 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 **?** . But this argument rests on a false equivalence, as we will show.

In **?** 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 **?** , 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.

3.2 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 **?**). But generally, the point of running the pilot is to identify issues which were not forseen. See the recruitment issue in **?** - 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

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].

Appendix