# A review of statistical designs for improving the efficiency of phase II studies in oncology

James M. S. Wason[1] and Thomas Jaki[2]

[1]MRC Biostatistics Unit, Cambridge, United Kingdom

[2]Department of Mathematics and Statistics, Lancaster University, United Kingdom

## Abstract

Phase II oncology trials are carried out to assess whether an experimental anti-cancer treatment shows sufficient signs of effectiveness to justify being tested in a phase III trial. Traditionally such trials are conducted as single arm studies using a binary response rate as the primary endpoint. In this article we review and contrast alternative approaches for such studies. Each approach uses only data that are necessary for the traditional analysis. We consider two broad classes of methods: ones that aim to improve the efficiency using novel design ideas, such as multi-stage and multi-arm multi-stage designs; and ones that aim to improve the analysis, by making better use of the richness of the data that is ignored in the traditional analysis. The former class of methods provides considerable gains in efficiency, but also increases the administrative and logistical issues in running the trial. The second class consists of viable alternatives to the standard analysis that come with little additional requirements and provide considerable gains in efficiency.

**Keywords:** augmented methods; continuous; multi-stage; Phase II; RECIST; oncology.

# 1. Introduction

Phase II oncology trials are carried out to assess whether an experimental anti-cancer treatment shows sufficient signs of effectiveness to justify being tested in a phase III trial. Currently the success rate of phase III oncology trials is found to be unacceptably low (1;2) and a major reason is due to inadequate early phase trials.

In most tumour types, the gold-standard endpoint of time-to-death (which is often used in phase III trials) is not feasible for use in phase II trials. This is due to substantial progress in increasing survival rates for most tumour types, meaning the time taken to observe time-to-death is too long. Instead, endpoints that are relatively quickly observed and that are thought to be informative for time-to-death are used. It is clear that, whichever endpoint is used instead of the gold-standard endpoint of time-to-death, needs to be clinically relevant and the collected information needs to be expected to provide a reliable assessment of whether or not to proceed to phase III. For solid-tumours, the focus of this paper, the first attempt to standardise the assessment of anticancer agents was the World Health Organisation (WHO) criteria (3). The WHO criteria categorised the change in number and size of measurable tumour lesions. As technology for scanning tumour lesions advanced, several modifications to the WHO criteria were introduced. This led to different organisations using inconsistent evaluations of anti-tumour effect. The RECIST criteria, originally proposed in 2000 (4) and updated in 2009 (5), was an attempt to provide a consistent mechanism for evaluating anti-cancer treatments. RECIST categorises the change in the size and number of tumour lesions into four levels: complete response (CR), partial response (PR), stable disease (SD) and progressive disease (PD) to which death is frequently added as a further category. RECIST has had a wide uptake and currently endpoints that use RECIST categories are used in the vast majority of solid-tumour Phase II trials.

Endpoints based on RECIST that are regularly used in early oncology trials are: 1) the response rate (RR), which is a binary outcome with the treatment classed as successful if the patient experiences a CR or PR and as a failure otherwise; 2) the disease control rate (DCR), which modifies RR to include SD as a success; 3) progression-free-survival (PFS), which is the time until a patient experiences PD or is dead. RR and DCR can be further divided into best observed response and response at a fixed time. For best observed response, patients are assessed at regular intervals until they progress, and the best observed RECIST outcome before progression is classified as a success or failure using RR or DCR. The fixed time analysis assesses RR or DCR at a fixed time after randomisation, for example after the treatment finishes.

Many statistical methods have been proposed for improving early oncology trials. These improvements may be to increase the power of the trial, to improve the efficiency of the trial (i.e. reduce the number of patients required on average), or to improve the ability of the trial to predict whether the treatment will be successful at phase III. In this paper we aim to provide a review of some important proposed methods, and to provide recommendations about their use. While our focus of this paper is Phase II studies, many aspects discussed here overlap with pilot studies, which have been defined as a “small study for helping to design a

further confirmatory study" (6), and hence some of the methods will also be applicable to them.

## 2. A standard design and its multi-stage extension

A standard trial design for early cancer studies is to use a single arm design with response rate as the primary endpoint. This design tests

$$H_0: p = p_0 \qquad \text{vs} \qquad H_1: p > p_0 \tag{1}$$

where $p$ is the response rate on the treatment under investigation and $p_0$ is a fixed constant that represents the response rate on the standard treatment or standard of care. The value of $p_0$ may be chosen from clinical judgement or from historical control data. Generally a total number of $n$ patients would be recruited so that the power of the trial to reject $H_0$ is equal to $1-\beta$ when the true response rate on the experimental treatment is $p=p_1$ with $p_1$ being a response rate that indicates the treatment is promising and warrants further trials. The null hypothesis is rejected in favour of the alternative if the total number of successes exceeds a critical value $r$ which is chosen to give a one-sided significance level of $\alpha$. The sample size and critical value are typically determined either by employing a normal approximation or using exact binomial probabilities (see e.g. Machin (7)).

A frequently used modification of this standard design is to utilize (group-) sequential methods that allow interim analyses to be carried out after groups of patients have been assessed. Often, a two-stage design is used for early cancer studies which, as its name implies, splits the trial into two stages. In the first stage a total of $n_1$ patients are recruited and assessed. If the number of responders in the first stage is below a pre-specified value, $f$, the trial stops for futility, and $H_0$ is not rejected. If the number of responders exceeds an upper bound, $e_1$, the trial stops for efficacy and $H_0$ is rejected. If the trial continues to the second stage, $n_2$ additional patients are recruited and assessed. At the end of the second stage, $H_0$ is rejected if the total number of responders is above a threshold $e_2$.

The advantage of a two-stage design is that the expected sample size (ESS) is often lower than $n$, the number of patients required in the one-stage design. This is good for the average cost and length of the trial and also is more ethical as a truly ineffective drug will be given to fewer patients, on average. However a drawback is that the maximum sample size (MSS), equal to $n_1+n_2$, is generally larger than $n$ would be for a single stage design with the same type I error rate and power. That is, if the two-stage trial continues to a second stage, the sample size used will usually be greater than would have required for a trial without an interim analysis. We should note that ESS (but not MSS) depends on the true response rate, $p$. Although it is not necessary to allow both stopping for futility and efficacy in the first stage, doing so provides a broader range of values of $p$ that result in ESS being lower than $n$.

Simon (8) proposed two different two-stage designs that allow early stopping for futility at an interim analysis. The first, called the optimal design, is the two-stage design with required type I error rate and power that minimises the ESS when $p=p_0$. The second, called the minimax design, minimises the maximum sample size (MSS) i.e. ($n_1+n_2$). Since MSS is an

integer value, several designs will have this property: the minimax design is the one amongst these with the lowest ESS when $p=p_0$. Both the optimal and minimax designs have ESS values that are considerably lower than the sample size that would be required for a one-stage design without interim analyses.

Simon's two-stage design has been used frequently in practice (9;10) and the original paper cited over 2000 times. It has also been extended in many ways. Firstly, the original design only allows stopping for futility, effectively setting the value of $e_1$ to $n_1$. Schuster (11) and Mander and Thompson (12) consider optimal designs that allow early stopping for efficacy. In the case of early stopping for futility and efficacy, the ESS generally starts low for very low values of $p$ (as early stopping for futility is likely), increases to a maximum, and then decreases again for large values of $p$ (where early stopping for efficacy is likely). Mander and Thompson consider the design that has the lowest ESS when $p=p_1$, and Shuster considers the design that minimises the maximum ESS (i.e. the peak of the ESS curve). Designs that allow early stopping for efficacy will increase the efficiency of the trial considerably when the drug is truly effective. They also have the drawback that estimates of the success probability for a successful treatment will be based on fewer patients, which may be undesirable for the planning of a new trial.

A second useful modification is generalising optimal designs to consider more than one optimality criteria. Designs that optimise a single criterion often perform poorly on other criteria that might be of interest. For example, Simon's optimal design typically has a large MSS. Jung *et al.* (13) propose using admissible designs. For a specified set of optimality criteria (Jung *et al.* consider the ESS under $p = p_0$ and the MSS), an admissible design is one which is optimal for a weighted sum of the two criteria. That is, it minimises:

$$\pi ESS(p = p_0) + (1-\pi)\mathrm{MSS}, \qquad (2)$$

for some value of $\pi \in (0,1]$. There will in general be a finite set of admissible designs, all of which balance the criteria in different ways. An interesting observation made by Jung *et al.* is that an admissible design for $\pi$ close to 1 will have an ESS close to the ESS of the optimal design, but a notably lower MSS than the optimal design. Admissible designs can be applied with more than two criteria - designs that consider the ESS at $p=p_0$, the ESS at $p=p_1$ and the MSS were considered by Mander et al. (14). Admissible designs can also consider criteria other than sample sizes – for example, Bowden and Wason (15) considered the expected sample size and subsequent estimation performance, in terms of bias and mean squared-error, following a two-stage design.

A third modification is to consider more than two-stages. Ensign *et al.* (16) and Chen (17) considered extending Simon's optimal design to three stages. Chen and Shan (18) considered optimal and minimax three-stage designs that also allow early stopping for efficacy. Generally adding stages in a multi-stage design provides diminishing returns - going from a two-stage design to a three-stage design reduces the ESS by a small amount compared to the reduction from going from a one-stage to a two-stage.

Although Simon's original paper (8) considered a single-arm trial with a binary endpoint, group-sequential design theory (for a broad and thorough overview, see Jennison and Turnbull (19)) can more generally be applied to two-arm trials comparing an experimental arm to a control arm. It can also be applied to a wide variety of endpoints, including normally distributed and time-to-event. Normally distributed endpoints are rare in early oncology trials (with the exception of considering the change in tumour size directly, which we consider further in section 3); time-to-event endpoints are common however, with progression-free survival and time-to-progression two frequently used endpoints in randomised phase II trials. In this case, multi-stage designs can still be applied to improve efficiency – Schaid *et al.* (20) describe the required methodology when the log-rank test is used as the test statistic. Other modifications described above are also more generally applicable, with minor modifications, to randomised phase II trials (for example, see (21) for use of admissible designs in a randomised two-arm setting) and other primary endpoints, such as progression-free survival, as well.

Recent methodological work has focused on development of multi-arm multi-stage (MAMS) designs. These allow more than one experimental arm to be compared against a common control arm. At interim analyses, experimental arms can be dropped for futility if they have performed poorly (22-24) . Efficacy stopping can also be allowed if it is of interest tofind only a single effective experimental treatment. For a recent review on different classes of MAMS designs, and how they might be useful in phase II trials, see Wason (25) while (26) provides some general recommendations on these designs.

## 3. Alternative endpoints

### 3.1 Using a categorical endpoint

Using a binary endpoint for early phase cancer trials has become an established standard, yet it is well known that dichotomizing a categorical variable leads to a loss of information and is hence inefficient (27;28). As a consequence a natural alternative is to utilize the categorical information directly in the design of the early oncology studies. Such a trial could consider the log-odds ratio, defined as

$$\theta_k = \log\left(\frac{Q_{Ek}(1-Q_{0k})}{Q_{0k}(1-Q_{Ek})}\right)$$

where $Q_{jk} = p_{j1} + \ldots + p_{jk}$ and $p_{jk}$ is the is the probability that a patient on treatment $j$ has an outcome in category $C_k$. It is clear from the definition of the parameter of interest, however, that both information on the experimental treatment and a comparator, indexed by zero here, is required. As a consequence response values for the comparator would need to be specified and assumed for single arm studies frequently utilized in early cancer studies which may be difficult to achieve. For comparative studies, however, such an approach will be efficient and details on how to obtain sample sizes have been studied (29) and extended to multiple experimental treatments (30). In either case the solution is based on efficient score statistics and hence approximate and proportional odds (27) are assumed. For the two-arm comparative setting, the critical value, $c$, and the corresponding sample size have been derived as

$$c = z_{1-\alpha}\sqrt{V} \quad \text{and} \quad n = \frac{3(R+1)^2(z_{1-\alpha}+z_{1-\beta})^2}{R\theta^2(1-\sum_{i=1}^{k}\bar{p}_i^3)} \qquad (2)$$

for a trial randomizing in an R:1 ratio in favour of control. In equation (2), θ represents the clinically relevant effect and $\bar{p}_i$ is the average proportion across arms in response category *i*.

3.2. Using tumour size

Another natural alternative to using binary response based on RECIST is based on the recognition that the RECIST categories are largely based on the change in the size of the lesions. In their simplest form (31;32) the change in (possibly transformed) tumour size is modelled as being normally distributed allowing standard methods to be used. This brings the advantage that the continuous variable underlying the RECIST categories is utilized directly and hence the well known loss in efficiency resulting from categorizing a continuous variable (33) is avoided. A short-coming of using tumour size directly is that other relevant information used in RECIST, such as death or new lesions, are not used. Two different approaches have been proposed to overcome this issue.

The first, by Karrison *et al.* (32), uses the continuous tumour shrinkage (on the log tumour ratio scale) as a primary endpoint. It is suggested that patients who die or progress due to new lesions are included in the analysis, but their outcome is set to the worst observed log tumour size ratio amongst other patients. Similarly, as complete responses have an undefined log tumour ratio, their outcome is set to the best observed outcome amongst other patients. As this procedure is likely to cause deviations from normality, it is recommended that a non-parametric test is used.

Jaki *et al.* (34) propose to overcome this issue by combining two separate test statistics that can be related to the same endpoint and describe their solution for a comparative study. In particular they construct the design based on survival to the end of the study and a continuous tumour size measurement for all patients that are alive at the end of follow-up. Jaki *et al.* suggests that one begins by considering the binary variable that captures whether a patient has survived until the tumour size measurement post treatment is scheduled (often several months after initiation of the treatment). Using a generalized linear model the treatment effect, $\theta_S$, on the binary survival variable for patients who died can be straightforwardly estimated. Similarly, the treatment effect on the change in tumour size, $\theta_{CTS}$, can be estimated using a linear model based on all patients that have survived long enough, provided that the endpoint is (approximately) normally distributed. Note that this construction allows for a natural way to include covariates in the analysis. Using a score statistic framework, the corresponding estimators will be approximately normal with the mean corresponding to the underlying parameter times the information level and variance equal to the information level so that standard tests for each of the parameters separately can be constructed.

To avoid having to consider two separate hypothesis tests that may contradict each other, Jaki *et al.* proceed by combining the two test statistics into a single statistic testing one parameter namely the log-hazard ratio for overall survival, $\theta_{OS}$. To do so they recognize, that the binary

survival parameter in the generalized linear model is exactly as the log-hazard ratio provided that a complementary log-log link is used (35). Additionally this relationship is still approximately true if a logit link is used. To derive a similar relationship between overall survival and change in tumour size, they use a historical dataset on the control treatment to estimate this relationship using a Cox proportional hazards model.

Using these ingredients, the distributions of the score statistics, denoted by S, can be written as

$$S_s \sim N(\theta_{OS}, V_S) \quad \text{and} \quad S_{CTS} \sim N\left(\frac{\theta_{OS}}{-\beta_{CTS}}, V_{CTS}\right)$$

where $V_s$ and $V_{CTS}$ are the Fisher information for the binary survival and change in tumour size, respectively, and $\beta_{CTS}$ is the coefficient of the Cox proportional hazards model associated with change in tumour size. Combining the two test statistics as $S_{\mathrm{S}} + \frac{1}{-\beta_{\mathrm{cts}}} S_{\mathrm{cts}}$ then allows construction of a single test for whether the log hazard ratio is 0 (i.e. whether there is a treatment effect on overall survival) based on change in tumour size information. To design the study, standard results for score statistics can be used to find the critical value, $u$, and the sample size as

$$u = z_{1-\alpha}\sqrt{V_{\mathrm{S}} + \frac{1}{\beta_{\mathrm{cts}}{}^2} V_{\mathrm{CTS}}} \quad \text{and} \quad n = \frac{(R+1)^2}{R}\left(\frac{z_{1-\alpha}+z_{1-\beta}}{\theta_{\mathrm{OS}}}\right)^2 \frac{\bar{p}\beta_{\mathrm{CTS}}{}^2\sigma^2}{q^2(1-\bar{p})\beta_{\mathrm{CTS}}{}^2\sigma^2+\bar{p}}$$

were $\theta_{\mathrm{OS}}$ is the interesting treatment effect in terms of overall survival, $\sigma^2$ the variance of the change in tumour size measurements, $\bar{p}$, is the average proportion of patients dying before the post treatment tumour size measurement can be taken and $q = -\log(1-\bar{p})$.

The original publication discusses using a binary endpoint capturing survival until a post-treatment measurement of the tumour size can be obtained. It is, however, straightforward to use other measures used in RECIST besides survival as well. For example the binary variable could capture if additional lesions have developed.

### 3.3. Improving analyses of existing endpoints

Despite many alternative endpoints existing that provide potential benefits over traditional early phase endpoints based on RECIST, they have only been used in a limited number of trials. For better or for worse, RR, DCR, TTP and PFS have gained widespread acceptability with clinicians conducting trials. Instead of attempting to shift practice, an alternative strategy is to attempt to improve analyses that use existing endpoints.

With this in mind, Wason and Seaman (36) proposed the augmented binary method. This assumes the endpoint of interest is a binary outcome, such as RR or DCR, which is measured at a fixed time (in terms of number of assessments rather than calendar time). In this case, a treatment is successful for a patient if the tumour shrinkage is above a certain threshold, and there is no treatment failure for other reasons (such as new lesions).

Tumour size measurements are assumed to be taken at baseline, some intermediate time, and at the end of treatment. These are denoted for patient *i* respectively as $(z_{0i}, z_{1i}, z_{2i})$. Indicators for whether failure occurred for a reason other than unacceptable tumour growth (henceforth called non-growth failures) are denoted as: $D_{i1}$ = 1 if patient *i* has a non-growth failure before the interim measurement, and $D_{2i}$ = 1 if patient *i* has a non-growth failure between the interim measurement and the end of treatment. Typically, the log tumour-size ratio:

$$\mathrm{y}=(y_{1i}, y_{2i}) = \left( \log\left(\frac{z_{1i}}{z_{0i}}\right), \log\left(\frac{z_{2i}}{z_{0i}}\right) \right) \tag{2}$$

is used due to it being well approximated by the normal distribution (31).

The *y* and *D* variables determine the composite outcome. $S_i$ denotes the composite success indicator for patient *i*. If RR is used, then $S_i$ is equal to 1 if $D_{i1}=0$, $D_{2i}=0$ and $y_{2i} < \log(0.7)$. That is, it is 1 if patient *i* has a tumour shrinkage of more than 30% at the end of treatment (which is the minimum requirement for PR in RECIST) and no non-shrinkage failure. It is missing if the patient drops out of the trial for a reason other than one of the failure criteria. More generally, a different threshold can be used if an endpoint other than RR is used.

To estimate the probability of success, i.e. $\mathrm{P}(S_i = 1)$, models for *y* and $D=(D_{i1}, D_{2i})$ are fitted. The model for *y* is:

$$(y_{i1}, y_{i2})^T \mid z_{i0} \sim N((\mu_{1i}, \mu_{2i})^T, \Sigma), \tag{2}$$

where:

$$\begin{aligned} \mu_{1i} &= \alpha + \gamma z_{i0} \\ \mu_{2i} &= \beta + \gamma z_{i0} \end{aligned}. \tag{2}$$

Additional covariates can also be included in the tumour-shrinkage. For *D*, two separate logistic regressions are used:

$$\begin{aligned} &\mathrm{Logit}(\mathrm{P}(D_{i1} = 1) \mid Z_{i0}) = \alpha_{D1} + \gamma_{D1} z_{i0} \\ &\mathrm{Logit}(\mathrm{P}(D_{i2} = 1 \mid D_{i1} = 0, Z_{i0}, Z_{i1})) = \alpha_{D2} + \gamma_{D2} z_{i1}. \end{aligned} \tag{2}$$

Models (2) and (2) can be combined to give the probability of success for patient *i*, with baseline tumour size $z_{0i}$, which is given by:

$$\int_{-\infty}^{\log(0.7)} \int_{-\infty}^{\infty} \mathrm{P}(D_{1i} = 0 \mid z_{0i}, \theta) \mathrm{P}(D_{2i} = 0 \mid D_{1i} = 0, z_{0i}, y_{1i}, \theta) f_{Y_1, Y_2}(y_{1i}, y_{2i}; \theta) dy_{1i} dy_{2i}, \tag{2}$$

where $f_{Y_1, Y_2}(y_{1i}, y_{2i}; \theta)$ is the pdf of the bivariate normal distribution from equation (2).

By repeating this for all patients in the dataset, an estimate of the overall probability of success is reached. To get a measure of uncertainty for the estimated probability of success the delta method is applied. For each arm in the trial, an estimated probability of success and associated 95% CI is found. For a single-arm trial, this CI can be compared against the null

probability of success that was pre-specified; for a two-arm trial, the difference in success probably and 95% CI for the difference can be straightforwardly found.

The augmented binary method can improve the precision of the estimated success probability considerably. The mean width of the CI using the method is equivalent to the mean width of the CI from the traditional approach with a 35% higher sample size. The efficiency gain does depend on the dichotomisation threshold used – if the number of treatment successes or failures in the trial is very low then the method does not do well. So if few partial or complete responses were to be expected, then one should use the augmented binary method on the DCR endpoint instead of the RR endpoint.

The main current limitation of the augmented binary method is that it can only be applied to endpoints measured at a fixed timepoint. Thus it cannot be used on analyses using best observed response or PFS. It should be possible to extend the method to work on these endpoints, and this is an active area of research. A second limitation is that there are currently no analytical formulae for the sample size required for a trial using the augmented binary method. Wason and Seaman (36) suggested that the sample size calculation should assume the traditional analysis is to be used, and then the augmented binary method used to gain power. A simulation-based approach could alternatively be used if a specified power is required.

## 4. Comparison of methods

A variety of alternative designs have been proposed to improve Phase II cancer trials. From our description above it is clear that different approaches have advantages and disadvantages. In the conclusion of this work we provide comparisons of the different ideas under a broad range of criteria. A summary comparison of different designs is provided in Table 1 and some more details provided below.

### Uptake of methods

Despite the availability of numerous alternative approaches, single arm trials based on a binary endpoint are by a large margin the most common. Whenever comparative trials are used they also tend to be based on binary response or PFS. There is some suggestion that evaluation of newer treatments, such as cytostatic agents, use comparative studies more frequently. Approaches using change in tumour size have been discussed for some time but their uptake to date is still limited while the augmented method, being a recent development, has not been applied to our knowledge.

### Simplicity

In terms of simplicity two different aspects need to be considered. Is the approach intuitive and easy to explain to a non-statistician and how difficult is the method in terms of the underlying statistics. Once again the traditional approaches based on binary response fare well on both accounts. The underlying statistical ideas are standard and clinical experts are happy to embrace a design based on such an endpoint – possibly helped by the fact that this

design has been used for many years. Similarly simple to explain to non-experts are designs based on continuous change in tumour size. This is helped by the wide use of RECIST and the familiarity with its derivation and experience with continuous data by many scientists. Using just the size of the tumour as in (32) is also straightforward statistically as standard methods for continuous data can be used. Once one allows for patients possibly not surviving long enough in order to obtain a tumour size measurement the statistical complexity increases drastically.

The augmented approach in (36) is somewhat more complex to explain to non-experts as the underlying method is fairly statistically advanced. The design based on a categorical response mainly suffers from the difficulty to explain the proportional odds assumption to non-experts. The statistics underneath the approach is, however, relatively straightforward and should not pose any challenge to implement.

Efficiency

One of the most important considerations in any statistical design is to make best use of the available data. In the context of designing early oncology study we consider the number of subjects required to attain a particular power. It is clear, that single-arm studies have an unfair advantage over two-arm trials as no data on the control group is collected. Yet, Jaki *et al.* (34) show convincingly that their approach for a comparative trial based on continuous tumour size yields comparable sample sizes to a single arm trial based on a binary response. This can be explained by the well known fact that dichotomizing a continuous variable leads to a substantial loss on efficiency. Consequently the popular approach based on binary response is less efficient that using a categorical variable which is again less efficient than using a continuous response. The augmented binary method certainly improves efficiency over the binary analysis and has been shown to outperform Karrison's continuous method (32) in some, but not all, instances (36). A more detailed comparison against the more complex continuous method is yet to be undertaken, however.

Single arm, two-armed and beyond

All of the methods discussed in this work are applicable to two-arm comparisons. The continuous approach discussed in (34), however has not been developed for single arm studies. Similarly the design using categorical outcomes has not been developed except a simplified version that allows for partial and complete responses to be considered separately (37). For the approaches currently proposed for comparative trials, some modifications to the methods could be used to allow for single armed versions. As these would require some additional assumptions (e.g. proportions in each category for the control arm) that would be difficult to obtain, these approaches are unlikely to be useful for single arm studies in practice.

Going beyond two arms, methods have been developed for multi-arm trials for binary responses (38;39), ordinal response (30;40) and simple continuous data (22;23). The more complex continuous method and the augmented approach have not been extended to this setting as of yet.

Additional assumptions and information

Underlying to all the discussed approaches is the desire to use information that is required in order to obtain the RECIST criteria. Despite that fundamental commonality the different approaches differ in the assumptions or information they require. It is, for example, clear that the RECIST categories themselves plus information on survival is sufficient to design a trial using a categorical endpoint. Similarly the simplification used by a binary response variable to group categories also does not require any additional data. In order to use a continuous change in tumour size endpoint and the augmented binary approach, however, the raw data on the size of the tumours will be required. In general this will not pose a particular challenge as these data are necessary in order to obtain the RECIST classification. For the method of Jaki *et al.* (34), however, a historical dataset on the control treatment which includes the raw tumour size measurements will be necessary in addition in order to relate the survival endpoint to the change in tumour size endpoint.

Although not requiring any additional data, the categorical analysis and the augmented approach do make some additional assumptions. For categorical responses, the proportional odds assumption (27) is typically made while an assumption about missing at random (MAR – see for example, (41)) is used for the augmented binary method. In both cases sensitivity analyses have shown that the methods are not highly sensitive to the respective assumptions, yet it is clear that some thought about the appropriateness of them is necessary.

Interim analyses

Finally, we have not discussed the utility of multi-stage designs any further nor included them in our comparison table. This is because using interim analysis is, in principle, possible for all different endpoints discussed. Our view is that allowing for early stopping should routinely be considered, independent of the endpoint/method used for the trial otherwise. This is because, if done well, allowing for an interim analysis can substantially reduce the realized sample size which is then directly reflected in the expected sample size. The cost for doing so is usually a small increase in the maximum sample size. Nevertheless there are a few points to consider when considering including interim analyses in a trial.

Firstly and most importantly, these designs will only result in a benefit if the recruitment is slow in comparison to the time it takes to observe the primary endpoint. In situations where most or even all patients can be recruited before enough information is collected on the primary endpoint, allowing for multiple analyses will not result in any worthwhile benefit (except perhaps in terms of time taken). Additionally overrunning, that is having recruited patients to the study who are not part of the data that have been used to the decision to stop. This reduces the efficiency gain from multi-stage approaches, especially if recruitment is quick in comparison to the delay between recruitment and assessment.

Secondly it is important to ensure that at the time of the first analysis, the totality of the data (and not just the information on the primary endpoint) is convincing. It may therefore be appropriate to allow stopping for lack of benefit earlier than stopping for efficacy as safety data are not mature.

Thirdly, interim analyses do, besides requiring specific statistical methods, add an additional element of complexity to the logistics of a study through additional data cleaning, analysis and so on. Moreover, the additional reduction in expected sample size decreases with the number of analyses. As a consequence a maximum of two interim analyses is usually sufficient for early cancer studies.

The final point to make is that sequential methods lead to standard maximum likelihood estimators of the treatment effect being biased due to the option to stop early (42). Although this bias tends to be small, special estimators that overcome these issues have been developed (43).

## 5. Conclusions

There are a large number of recent methodological developments that have aimed to improve Phase II cancer trials. Many provide an improvement in efficiency over traditional designs and analyses used. We believe that several different novel approaches are available that should be considered as alternatives to single arm trials using binary endpoints to ensure the maximum use of information is made. Additionally we believe that multi-stage designs should routinely be considered. Finally it is worth highlighting, that the popularity of single armed designs based on response stems from investigating cytotoxic compounds. Newer treatments, such as cytostatic agents, are becoming more commonly evaluated through comparative trials, possibly using endpoints such as progression-free survival. These designs can, however, still be improved further by some of the ideas outlined in this review. Multi-stage designs, for example, will be efficient in these designs and should be routinely considered.

## Acknowledgments

Both authors have made equal contributions to this manuscript. This report is independent research arising in part from Dr. Jaki's Career Development Fellowship (NIHR-CDF-2010-03-32) supported by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. JW is funded by the Medical Research Council [grant number G0800860] and the NIHR Cambridge Biomedical Research Centre.

Reference List

(1) Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nature reviews Drug discovery 2004;3(8):711-6.

(2) DiMasi JA, Grabowski HG. Economics of new oncology drug development. Journal of Clinical Oncology 2007;25(2):209-16.

(3) Miller AB, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. Cancer 1981;47(1):207-14.

(4) Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, et al. New guidelines to evaluate the response to treatment in solid tumors. Journal of the National Cancer Institute 2000;92(3):205-16.

(5) Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). European journal of cancer 2009;45(2):228-47.

(6) Arnold DM, Burns KE, Adhikari NK, Kho ME, Meade MO, Cook DJ. The design and interpretation of pilot trials in clinical research in critical care. Critical care medicine 2009;37(1):S69-S74.

(7) Machin D, Campbell MJ, Tan SB, Tan SH. Sample size tables for clinical studies. John Wiley & Sons; 2011.

(8) Simon R. Optimal two-stage designs for phase II clinical trials. Controlled clinical trials 1989;10(1):1-10.

(9) Lee JJ, Feng L. Randomized phase II designs in cancer clinical trials: current status and future directions. Journal of Clinical Oncology 2005;23(19):4450-7.

(10) Jaki T. Uptake of novel statistical methods for early-phase clinical studies in the UK public sector. Clinical Trials 2013;10(2):344-6.

(11) Shuster J. Optimal two-stage designs for single arm phase II cancer trials. Journal of biopharmaceutical statistics 2002;12(1):39-51.

(12) Mander AP, Thompson SG. Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials. Contemporary Clinical Trials 2010;31(6):572-8.

(13) Jung S-H, Lee T, Kim K, George SL. Admissible two-stage designs for phase II cancer clinical trials. Statist Med 2004;23(4):561-9.

(14) Mander AP, Wason JMS, Sweeting MJ, Thompson SG. Admissible two-stage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. Pharmaceutical Statistics 2012 Mar 1;11(2):91-6.

(15) Bowden J, Wason J. Identifying combined design and analysis procedures in two-stage trials with a binary end point. Statist Med 2012;31(29):3874-84.

(16) Ensign LG, Gehan EA, Kamen DS, Thall PF. An optimal three-stage design for phase II clinical trials. Statist Med 1994;13(17):1727-36.

(17) Chen TT. Optimal three-stage designs for phase II cancer clinical trials. Statist Med 1997;16(23):2701-11.

(18) Chen K, Shan M. Optimal and minimax three-stage designs for phase II oncology clinical trials. Contemporary Clinical Trials 2008;29(1):32-41.

(19) Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. CRC Press; 1999.

(20) Schaid DJ, Wieand SAM, Therneau TM. Optimal two-stage screening designs for survival comparisons. Biometrika 1990;77(3):507-13.

(21) Wason J, Mander AP, Thompson SG. Optimal multistage designs for randomised clinical trials with continuous outcomes. Statist Med 2012;31(4):301-12.

(22) Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. Statist Med 2003;22(5):689-703.

(23) Magirr D, Jaki T, Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. Biometrika 2012;99(2):494-501.

(24) Wason J, Jaki T. Optimal design of multi-arm multi-stage trials. Statist Med 2012;31(30):4269-79.

(25) Wason JM. Reducing the average number of patients needed in a phase II trial through novel design. Clinical Research and Regulatory Affairs 2013;30(4):47-54.

(26) Wason J, Magirr D, Law M, Jaki T. Some recommendations for multi-arm multi-stage trials. Statistical methods in medical research 2012.

(27) McCullagh P. Regression models for ordinal data. Journal of the royal statistical society Series B (Methodological) 1980;109-42.

(28) Agresti A. Analysis of ordinal categorical data. 656 ed. John Wiley & Sons; 2010.

(29) Whitehead J. Sample size calculations for ordered categorical data. Statist Med 1993;12(24):2257-71.

(30) Whitehead J, Jaki T. One- and two-stage design proposals for a phase II trial comparing three active treatments with control using an ordered categorical endpoint. Statist Med 2009;28(5):828-47.

(31) Lavin PT. An alternative model for the evaluation of antitumor activity. Cancer clinical trials 1980;4(4):451-7.

(32) Karrison TG, Maitland ML, Stadler WM, Ratain MJ. Design of phase II cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non-small-cell lung cancer. Journal of the National Cancer Institute 2007;99(19):1455-61.

(33) Altman DG, Royston P. Statistics notes: the cost of dichotomising continuous variables. BMJ: British Medical Journal 2006;332(7549):1080.

(34) Jaki T, Andre V, Su TL, Whitehead J. Designing exploratory cancer trials using change in tumour size as primary endpoint. Statist Med 2013;32(15):2544-54.

(35) Collett D. Modelling Survival Data in Medical Research. Boca Raton, Fla : Chapman & Hall/CRC 2003.

(36) Wason J, Seaman SR. Using continuous data on tumour measurements to improve inference in phase II cancer studies. Statist Med 2013;32(26):4639-50.

(37) Panageas KS, Smith A, G+Ânen M, Chapman PB. An optimal two-stage phase II design utilizing complete and partial response information separately. Controlled clinical trials 2002;23(4):367-79.

(38) Thall PF, Simon R, Ellenberg SS. Two-stage selection and testing designs for comparative clinical trials. Biometrika 1988;75(2):303-10.

(39) Bratton DJ, Phillips PP, Parmar MK. A multi-arm multi-stage clinical trial design for binary outcomes with application to tuberculosis. BMC medical research methodology 2013;13(1):139.

(40) Jaki T, Magirr D. Considerations on covariates and endpoints in multi-arm multi-stage clinical trials selecting all promising treatments. Statist Med 2013;32(7):1150-63.

(41) Molenberghs G, Kenward M. Missing data in clinical studies. 61 ed. John Wiley & Sons; 2007.

(42) Bauer P, Koenig F, Brannath W, Posch M. Selection and bias - two hostile brothers. Statist Med 2010;29(1):1-13.

(43) Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. Biometrika 1990;77(4):875-92.

Table 1: Characteristics of different designs for early oncology trials.

| | | | **Continuous** | | |
|---|---|---|---|---|---|
| | **Binary** | **Categorical** | **Karrison** | **Jaki et al** | **Augmented** |
| 1. Current uptake | High | Low | Moderate | Low | Low |
| | | | | | |
| 2. Simplicity | | | | | |
| 2a Ease of non-technical explanation of method | High | Low | High | High | Moderate |
| 2b Statistical complexity | Low | Moderate | Low | High | High |
| | | | | | |
| 3. Efficiency | Low | Moderate | High | High | High |
| | | | | | |
| 4. Number of arms | | | | | |
| 4a Single arm | Yes | No | No | No | Yes |
| 4b Comparative 2-arm trial | Yes | Yes | Yes | Yes | Yes |
| | | | | | |
| 5. Additional requirements or assumptions beyond the standard design? | NA | Proportional odds | NA | Historical data | Missing at random |
| | | | | | |