EUROPEAN JOURNAL OF CANCER 47 (2011) 983–989

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

Reducing sample sizes in two-stage phase II cancer trials by using continuous tumour shrinkage end-points

James M.S. Wason [a,*], *Adrian P. Mander* [a], *Tim G. Eisen* [b]

[a] *MRC Biostatistics Unit, Hub for Trials Methodology Research, Cambridge, United Kingdom*
[b] *Cambridge Cancer Trials Centre, Cambridge, United Kingdom*

ARTICLE INFO

Article history:
Received 14 September 2010
Received in revised form 3 December 2010
Accepted 7 December 2010
Available online 14 January 2011

Keywords:
Phase II cancer clinical trials
Sample sizes
Trial design
RECIST

ABSTRACT

Reducing the number of patients required for a clinical trial is important for shortening development time. Phase II cancer trials assess the tumour-shrinking effect of a novel compound through a binary end-point formed from the percentage change in total lesion diameter. We compare single-arm two-stage designs which use the binary end-point to those which directly use the continuous end-point.

Using the continuous end-point results in lower expected and maximum sample sizes. For larger trials the reduction is around 37%. This assumes that the dichotomisation point of the continuous end-point is chosen to give the best sample size, with the trial design using the binary end-point performing even worse otherwise. We consider a previous trial designed using a Simon two-stage design and show that if the continuous end-point had been used, the expected and maximum sample sizes of the trial would be reduced by around 50%.

Using the continuous end-point in a two-stage cancer trial results in large sample size reductions. The methods discussed in this paper work best when the number of complete responses is low, as is true in several types of cancer. We discuss what could be done if this is not the case.

1. Introduction

Two of the greatest barriers for improved cancer treatment are the expense and time taken by clinical trials. Methods that maintain statistical power whilst reducing the required sample size are of great practical interest.

In phase II cancer trials for which the primary end-point is the tumour-shrinking effect of a treatment, the underlying measurement is continuous. It is commonplace to use the RECIST criteria[1] to dichotomise this measurement and form a binary end-point representing whether the treatment was successful or not.

It is well known in the statistical literature that dichotomising a continuous end-point loses information. This leads to the need for larger studies due to loss of statistical power. The degree of loss has been well studied in certain situations,[2] but in many areas of medical statistics it is still common to dichotomise.[3]

The idea of using the continuous end-point to power the trial is not new and was originally discussed by Lavin.[4] Karrison et al.[5] discuss using a continuous end-point in a randomised, single-stage design. The authors conclude that using the continuous end-point with a control group would reduce the high rates of failure found in phase III trials, as discussed by Tang et al.[6]

* *Corresponding author:* Address: MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, United Kingdom. Tel.: +44 01223 760717; fax: +44 01223 330365.
E-mail address: james.wason@mrc-bsu.cam.ac.uk (J.M.S. Wason).
0959-8049/$ - see front matter
doi:10.1016/j.ejca.2010.12.007

The Simon two-stage design[7] which allows for early termination of the trial due to lack of efficacy is commonly used in cancer trials.[8] Two-stage designs which use the continuous end-point should require even lower sample sizes. Many such designs, for both continuous and dichotomous end-points, have been proposed and discussed in the literature.[9–15]

In this paper we compare the sample size requirements of two-stage cancer trial designs using the binary and continuous end-points, respectively. We show that the required sample size for the Simon two-stage design depends on which point the continuous distribution is dichotomised at; that for smaller trials this factor makes a larger relative difference in sample sizes; and that the dichotomisation point implied by RECIST will rarely be the most efficient one to use. Lastly we discuss a recently completed trial which was designed using a Simon two-stage design, and show that a continuous two-stage design would have resulted in a 50% saving in sample size required.

2. Materials and methods

We assume that the percentage decrease in the sum of lesion diameters is normally distributed with mean δ and variance σ^2. Positive values represent shrinkages in the tumour size.

In practice, a binary end-point is formed, equal to 1 if the tumour shrinkage is above a threshold, d, and 0 otherwise. d is the dichotomisation threshold. If treatment success is defined as partial or complete responses, d is 30% by the RECIST criteria. We define p as the response probability, equivalent to the probability of the tumour shrinkage being above d.

In a phase II cancer trial the hypothesis $H_0{:}p \leqslant p_0$ is tested, with the trial powered to reject the null with probability $1-\beta$ when $p = p_1$, the clinically relevant difference (CRD), and below α when H_0 is true. α and $1-\beta$ are the significance level and power, respectively. We define a feasible design as one with correct significance level and power.

When using a two-stage design, n_1 patients are recruited in the first stage. If the number of treatment successes is above r_1, further n_2 patients are recruited. The expected sample size (ESS) is the number of patients recruited on average if the trial was repeated many times, equal to $n_1 + (1 - \text{PET})\, n_2$ where PET is the probability of early termination of the trial. ESS depends on the trial design and also the true response probability of the treatment. A second quantity of interest is the maximum sample size (MSS), $n_1 + n_2$.

For p_0, p_1, α, β, the optimal Simon two-stage design[7] has the lowest ESS when $p = p_0$ amongst all feasible designs. The minimax design has the lowest MSS amongst all feasible designs; since several such designs may exist, the one with the lowest ESS when $p = p_0$ is chosen.

In the continuous case, the hypothesis tested is: $H_0{:}\delta \leqslant \delta_0$. The trial is designed with significance level $\leqslant\alpha$ and power at $\delta = \delta_1$ is $\geqslant 1-\beta$. Here, δ_1 is the CRD, analogous to p_1 for the binary case.

The continuous two-stage design we use consists of recruiting n_1 patients and carrying out a t-test, giving test statistic T_1. The trial stops if T_1 is less than f, a futility threshold. Otherwise, further n_2 patients are recruited, with T_2 being the t-test of their treatment responses. H_0 is rejected if a weighted average of T_1 and T_2 is above e. More details of the design are provided in the online supplementary material.[16] The optimal and minimax designs are defined similarly as for the binary end-point.

Given (p_0, p_1, σ, d), δ_0 and δ_1 can be found straightforwardly, as shown in the supplementary material. This means that given a value of σ, the trial using the continuous end-point can be easily designed from the response probabilities.

3. Results

3.1. Comparison of optimal binary and continuous designs

We used the simon2stage command described in Mander and Thompson[17] to find optimal Simon two-stage designs, and a programme in C (available upon request) to find optimal continuous designs. For this section we set α, the significance level, to be 5%, and, the power, $1-\beta$, to be 90%. We show results for $\delta_1 = 10$, i.e. a 10% mean tumour shrinkage, with $\delta_0 = 0$, and $\sigma \in (10,20,40)$. The values of σ used represent the range of sample sizes used in phase II cancer trials: $\sigma = 10$ results in 11 patients required for a single-stage trial with the continuous end-point; $\sigma = 40$ results in 139 patients required.

Fig. 1 shows how the ESS of the optimal binary and continuous designs varies for different values of d, the dichotomisation threshold.

The lines representing the ESS of the optimal binary designs are not smooth as d changes because all parameters in the Simon design are integers. This means small changes in d do not necessarily mean small changes in ESS. As σ increases, the binary lines become smoother, since small changes in d result in relatively small changes to the design parameters.

The impact of d is extremely large for the smaller trials that use the binary end-point, and still makes a considerable relative difference when $\sigma = 40$. As d increases beyond 15, the ESS increases further. Since defining partial and complete responders as treatment successes implies $d = 30$, using RECIST to dichotomise leads to higher ESS than if the dichotomisation threshold was chosen differently. The best d to use is difficult to determine exactly, but halfway between δ_0 and δ_1 gives good results for this set of trial parameters.

Using the continuous tumour end-point always provides a lower ESS. Table 1 shows the magnitude of reduction compared with using the best optimal binary design and the optimal binary design using RECIST to dichotomise. For the best binary design, the reduction increases with σ. As σ increases beyond 40, the reduction converges to around a 37% reduction in ESS. This shows that even if the dichotomisation point is chosen carefully, substantial reductions in sample size are still possible by using the continuous end-point. For binary designs using RECIST, far bigger reductions are possible, especially for smaller trials.

Table 2 shows the design parameters, together with ESS, for the optimal continuous design and the optimal binary design when $d = 5$. The table shows that using the continuous end-point results in a lower MSS as well as ESS.

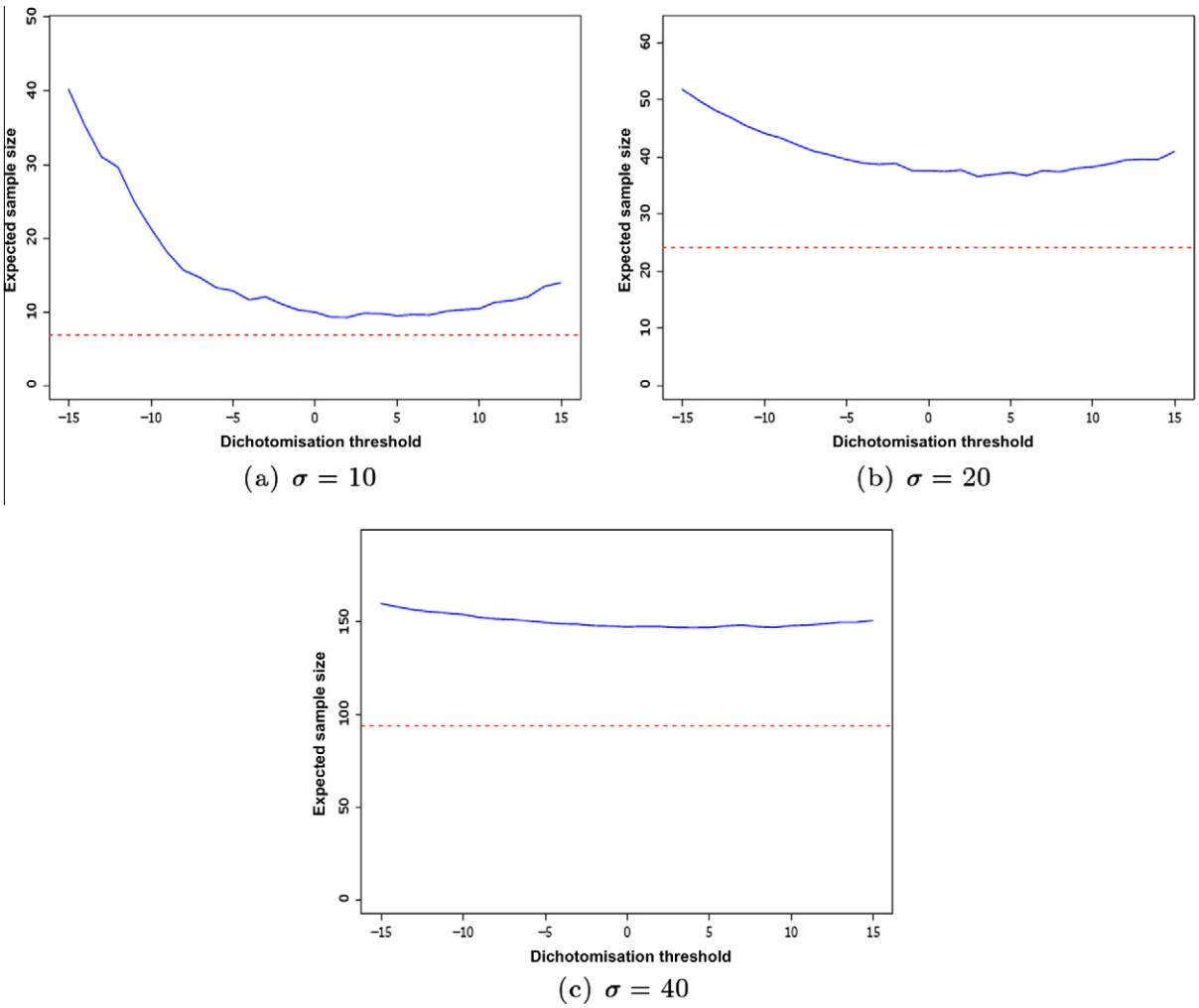

Fig. 1 – Plot of expected sample sizes of optimal binary (solid) and continuous (dashed) designs under null hypothesis as σ, the standard deviation of tumour shrinkage, and d, the dichotomisation threshold, vary.

Table 1 – Percentage reduction in expected and maximum sample sizes of optimal and minimax designs from using continuous end-point, $\delta_1 = 10$.

σ	Reduction in optimal design expected sample size (%) compared to		Reduction in minimax maximum sample size (%) compared to	
	Best binary design	RECIST binary design	Best binary design	RECIST binary design
10	25.8	94.4	26.7	93.5
20	34.3	64.8	31.5	59.8
40	36.0	45.0	35.6	44.2

Table 2 – Table of binary and continuous optimal two-stage designs for $\delta_1 = 10$.

σ	Optimal binary design				ESS at H_0	Optimal continuous design				ESS at H_0
	n_1	n_2	r_1	r		n_1	n_2	f	e	
10	8	7	3	7	9.47	4	9	0.55	1.94	6.79
20	25	45	11	34	37.25	15	26	0.40	1.64	24.03
40	88	172	41	129	146.96	62	103	0.50	1.55	93.87

In this section we only used $(\delta_0, \delta_1) = (0,10)$. Using the continuous design, the properties of the design only depend on $(\delta_1 - \delta_0)$. The properties of the binary design, given a fixed d, depend on the actual values of δ_0 and δ_1, especially so when σ is small. The binary design will perform best when the interval (δ_0, δ_1) contains d.

3.2. *Comparison of minimax binary and continuous designs*

Using the same parameters as previously, we found and compared the binary and continuous minimax designs. Such designs minimise $n_1 + n_2$, the MSS. Since there may be several designs with minimum MSS, the minimax design is the one which minimises the ESS when $p = p_0$. Minimax designs may be preferable when the potential number of patients to recruit is lower.

Fig. 2 shows the MSS of the binary and continuous minimax designs as d changes. The pattern is similar to before: as σ increases, choosing d makes less of a relative difference, and the reduction in sample size by using the continuous design compared to the best binary design increases.

Table 1 shows how the reduction in MSS of the minimax continuous design compared to that of the best minimax binary design changes as σ increases. Again, the reduction increases as σ increases. Although not clear from the table, as σ increases, the reduction converges to around 37%, the same limit as for the ESS of the optimal designs. Substantially higher reductions can be made if the dichotomisation threshold is not chosen well, however.

Table 3 shows the parameters of the minimax designs as σ increases. The minimax binary design given is actually the one for $d = 5$. Interestingly, as the size of the trial increases, most of the sample size is allocated to the first stage for both end-points.

As for the optimal designs, using the continuous endpoint provides a large decrease in MSS. For smaller trials, choosing a dichotomisation threshold makes a larger relative difference, with RECIST only performing well for a small subset of trials.

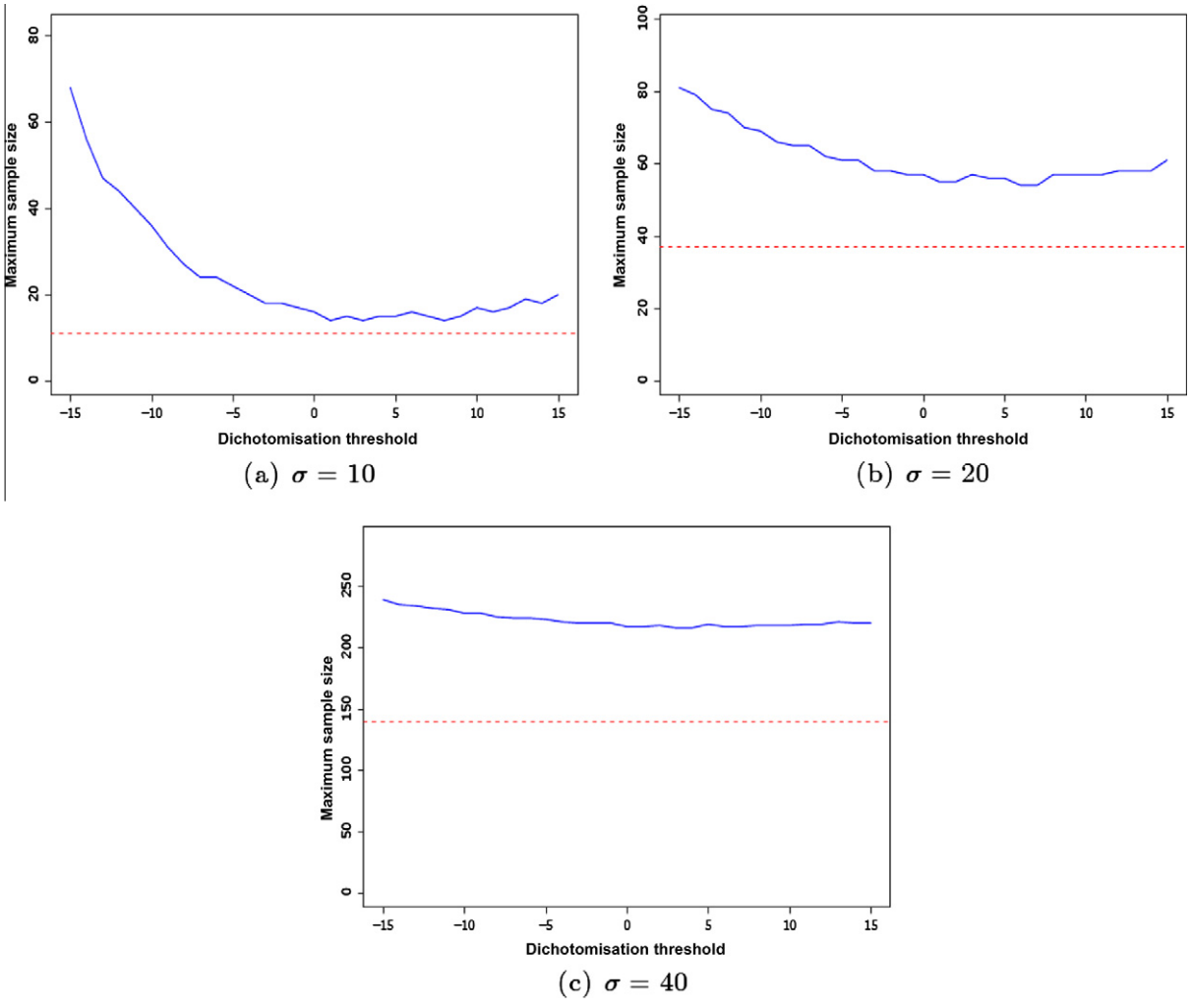

Fig. 2 – Plot of maximum sample sizes of minimax binary (solid) and continuous (dashed) designs under null hypothesis as σ, the standard deviation of tumour shrinkage, and d, the dichotomisation threshold, vary.

Table 3 – Table of binary and continuous minimax two-stage designs for $\delta_1 = 10$.

σ	Minimax binary design				Minimax continuous design			
	n_1	n_2	r_1	r	n_1	n_2	f	e
10	8	7	3	7	6	5	0.7	1.94
20	29	27	12	2	20	17	0.4	1.70
40	175	44	85	110	92	47	0	1.67

3.3. *Case study*

To illustrate the advantages of using the continuous tumour end-point, we have used a completed but, at present, unpublished trial. The underlying continuous measure, maximum shrinkage in total lesion diameter, was observed, but the trial was designed as an optimal Simon two-stage design which used the RECIST criteria to form a binary outcome (a partial or complete response was classed as a success, other categories as failures).

The current standard therapy had previously been shown to have tumour response rate in the range 10–20%, and so the value of p_0 was assumed to be 15%, with $p_1 = 30\%$. The study was designed to have 5% significance level and 90% power. The optimal Simon two-stage design has first and second stage sample sizes of 30 and 52, respectively, and an ESS at $p = p_0$ of 45.1.

In order to use the continuous end-point with the methods discussed in this paper, it must be normally distributed. Statistical tests such as the Kolmogorov–Smirnov test did not detect any significant difference between this distribution and a normal distribution. It seems that in this case, the percentage change in sum of lesion diameters is a valid end-point to use. It is also more comparable than other possible continuous end-points, since it is the one underlying the RECIST criteria.

From the actual trial data, the estimated value of σ is 36.4. In practice this would have been estimated from previous trials or preclinical studies. Together with p_0 and p_1, σ can be used to find δ_0 and δ_1, the mean tumour shrinkages under the null hypothesis and the CRD (see supplementary material Eq. (2)). In this case, $\delta_0 = -7.6\%$, $\delta_1 = 11.1\%$.

These values of δ_0, δ_1, and σ result in the optimal continuous design having a first and second stage sample size of 15 and 24, respectively. The ESS under the null is 22.9. This is a huge saving in terms of number of patients needed – over 50% for the MSS and just under 50% for the ESS.

The large savings indicate that for this trial the RECIST criteria did not provide the best point to dichotomise the distribution. A 30% reduction is well above what is expected on average even under the alternative hypothesis. Results from Section 3.1 indicated that an optimal dichotomisation threshold is between the values under the null and alternative hypotheses.

4. Discussion

In this paper we have compared the properties of optimal and minimax two-stage trials which use binary and continuous tumour response outcomes. Cancer trials record continuous measurements, but use dichotomised end-points for the purposes of design and analysis. This is statistically disadvantageous, and leads to trials which require larger numbers of patients to detect the same treatment response.

A considerable proportion of phase II cancer trials use the Simon two-stage design to reduce expected patient numbers.[8] We show herein that even further reductions could be made with a continuous two-stage design.

In practice, RECIST is often used to dichotomise, with tumour shrinkages of 30% or more being treatment successes. We have shown that the best point to dichotomise, from a power point of view, is around halfway between the assumed shrinkage under the null and clinically relevant differences. Thus, RECIST will perform well when this is true. If 30% tumour shrinkage is more than expected on average, using the dichotomised end-point is even worse. However, since the 30% threshold was originally proposed to minimise the impact of measurement error,[18] an interesting piece of future work would be how to choose the dichotomisation threshold which takes both factors into account.

As the size of the trial increases, the gain from using the continuous end-point increases to an upper limit. This limit appears to correspond to needing around 37% fewer patients to detect the same treatment effect when the dichotomisation threshold is chosen optimally.

Although this paper has focused on reductions in sample size, this is not necessarily the best use of the power advantages from using the continuous end-point. Instead the sample size could be maintained and the effect size that the trial is powered to detect lowered. As well as ensuring a lower proportion of significant results are false-positives, it would allow better information to be gathered about secondary end-points such as survival which could be useful for planning a subsequent trial.

In this paper, we only provided results for designs with significance level 5% and 90% power. Similar results hold for other commonly used combinations. Equivalent results for 80% power are provided in the online supplementary material.[16]

To illustrate the theoretical results, we examined a trial designed using an optimal Simon two-stage design. Redesigning using the continuous end-point saves almost 50% in ESS, and over 50% in MSS. The reason for the larger reduction is that the RECIST criteria are not the best choice of dichotomisation threshold for this trial. Few of the patients were expected to achieve a response, even if the drug worked well.

A current difficulty to designing a trial using the continuous end-point is that little prior information is available. For example, an estimate for the standard deviation is required to determine the sample size. Currently the only published source of information about the standard deviation is in

waterfall plots. As an initial step, trials could use the continuous end-point as a secondary outcome and publish relevant information. Over time, trials would become easier to design using the continuous end-point as a primary end-point.

Using the methods discussed relies on the percentage shrinkage in tumour to be normally distributed. Although the t-test is robust to deviations from normality[19] a problem occurs when complete responses are expected. Complete responses are rare in many cancer areas, including non-small-cell lung cancer,[20] metastatic melanoma[21] and renal cancer.[22] The methods should work well when the proportion of complete responses is low. Karrison et al.[5] suggest that when the proportion is higher non-parametric tests could be used. This is a sensible suggestion, but a parametric test which allows for complete responses may be more efficient. If the observed tumour shrinkage is not normally distributed for another reason, a transformation such as that of Box and Cox[23] can be used.

Karrison et al. also suggest that if the continuous end-point is used, a control group is required. Although there are many advantages of including a control group, we do not agree that it is more necessary with the continuous end-point than the binary one. One of the reasons stated for requiring a control group is that there is less information about the distribution of the tumour shrinkage end-point, for example the standard deviation. This is true, but just as much information is needed for designing a two-arm trial as for a one arm trial. In fact, this is an argument for using a design where the standard deviation is reestimated at the interim analysis, such as that of Whitehead et al.[15] Arguments such as the poor translation of results to phase III are correct, but are not unique to trials with the continuous end-points.

However, significant results from randomised trials provide considerably better evidence than from single-arm trials. A small one-arm study would be best followed up with a larger, randomised study before deciding whether to start a large and expensive phase III trial. The methods used in this paper are straightforwardly generalisable to randomised phase II designs, with similar results.

Another possible source of error that we did not consider in this paper was the measurement error. Measuring lesion diameter is not exact, with Erasmus et al. [24] finding an average difference of 12% between reported readings of two observers. We have evaluated the effect of measurement error on the continuous and binary outcome. Measurement error affects the power of the continuous design, and both the significance level and power of the binary design (unpublished data). The fact that the continuous design controls the significance level in the presence of measurement error is an advantage.

There are several possible extensions to this work which would provide utility for designing clinical trials. One is to address the aforementioned problem of complete responses.

Secondly, more than two stages may be considered. Such designs will reduce the ESS further, but have a higher MSS. For smaller trials, they will not provide much advantage above two-stage designs.

Thirdly, there is work needed in reducing bias in the case of patients dropping out of the trial. If a patient drops out early due to toxicity, it is not obvious how to include their continuous response. For the binary end-point it is straightforward enough to include them as a treatment failure. One possibility is to impute the missing tumour shrinkage value, but more investigation is needed. The impact of this problem could be reduced if the conduct of phase II trials was changed, with living patients being followed up after drop out for tumour size measurements.

Overall, we have shown that using the continuous end-point underlying the RECIST criteria allows large reductions in required sample size, and is not significantly more difficult to design than the Simon two-stage design. We expect that RECIST will remain widely used in trials, but we hope that trials using continuous outcomes will become more common. This would reduce trial costs, minimise the number of patients exposed to non-effective treatments, and allow more trials to take place in areas of cancer with fewer patients available to be recruited.

Conflict of interest statement

None declared.

Acknowledgements

This work was supported by the UK Medical Research Council (grant codes G08008600 and U.1052.00.014), the Cambridge Biomedical Research Centre, and Cancer Research UK. We thank Dr. Michael Sweeting for his helpful comments on the manuscript. We also thank the anonymous reviewer for their useful review.

REFERENCES

1. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur J Cancer* 2009;**45**:228–47.
2. Farewell VT, Tom BDM, Royston P. The impact of dichotomization on the efficiency of testing for an interaction effect in exponential family models. *JASA* 2004;**99**:822–31.
3. DeCoster J, Iselin AR, Gallucci M. A conceptual and empirical examination of justifications for dichotomization. *Psychol Methods* 2009;**14**:349–66.
4. Lavin PT. An alternative model for the evaluation of antitumor activity. *Cancer Clin Trials* 1981;**4**:451–7.
5. Karrison TG, Maitland ML, Stadler WM, Ratain MJ. Design of phase II cancer trials using a continuous endpoint of change in tumour size: application to a study of sorafenib and erlotinib in non-small-cell lung cancer. *JNCI* 2007;**99**:1455–61.
6. Tang H, Foster NR, Grothey A, et al. Comparison of error rates in single-arm versus randomized phase II cancer clinical trials. *J Clin Oncol* 2010;**28**:1936–41.
7. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clin Trials* 1989;**10**:1–10.
8. Lee J, Feng L. Randomized phase II designs in cancer clinical trials: current status and future directions. *J Clin Oncol* 2005;**23**:4450–7.
9. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977;**64**:191–9.

10. O'Brien PC, Flemming TR. A multiple-testing procedure for clinical trials. *Biometrics* 1979;**35**:549–56.
11. Emerson SS, Flemming TR. Symmetric group sequential designs. *Biometrics* 1989;**45**:905–23.
12. Eales JD, Jennison C. An improved method for deriving optimal one-sided group sequential tests. *Biometrika* 1992;**79**:13–24.
13. Whitehead J, Stratton I. Group sequential clinical trials with triangular continuation regions. *Biometrics* 1983;**39**:227–36.
14. Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. Chapman and Hall; 2000.
15. Whitehead J, Valdes-Marquez E, Lissmats A. A simple two-stage design for quantitative responses with application to a study in diabetic neuropathic pain. *Pharm Stat* 2009;**8**:125–35.
16. Supplementary material for "Reducing sample sizes in two-stage phase II cancer trials by using continuous tumour shrinkage endpoints". www.mrc-bsu.cam.ac.uk/Publications/addmaterial.
17. Mander AP, Thompson SG. Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials. *Contemporary Clin Trials* 2010;**31**:572–8.
18. Miller AB, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. *Cancer* 1981;**47**:207–14.
19. Cressie NAC, Sheffield LJ, Whitford HJ. Use of the one sample *t*-test in the real world. *J Chron Dis* 1984;**37**:107–14.
20. Azim HA, Ganti AK. Targeted therapy in advanced non-small cell lung cancer (NSCLC): where do we stand? *Cancer Treat Rev* 2006;**32**:630–6.
21. Lui P, Cashin R, Machado M, et al. Treatments for metastatic melanoma: synthesis of evidence from randomized trials. *Cancer Treat Rev* 2007;**33**:665–80.
22. Nelson EC, Evans CP, Lara PN. Renal cell carcinoma: current status and emerging therapies. *Cancer Treat Rev* 2007;**33**:299–313.
23. Box GEP, Cox DR. An analysis of transformations. *J R Stat Soc Ser B* 1964;**26**:211–52.
24. Erasmus JJ, Gladish GW, Broemeling L, et al. Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumour response. *J Clin Oncology* 2003;**21**:2574–82.