Clinical Trial Methodology
Special Focus Issue: Adaptive Clinical Trials

CLINICAL INVESTIGATION

Multi-arm clinical trials with treatment selection: what can be gained and at what price?

Thomas Jaki
Department of Mathematics & Statistics, Lancaster University, Lancaster, LA1 4YF, UK
Tel.: +44 1524 592318
jaki.thomas@gmail.com

With current success rates of confirmatory studies being only around 50%, new approaches to drug development are paramount. Many trials fail simply because ineffective treatments are identified too late. In this paper, we discuss the utility of multi-arm studies with treatment selection as a potential strategy that can reduce the high attrition rate. We illustrate the large gains in efficiency that are possible based on an example in Alzheimer's disease while outlining the additional challenges that need to be overcome to implement such studies.

Keywords: adaptive design • clinical trial design • multi-arm multi-stage trials • multi-arm trials • treatment selection

The development of medicinal products and health technologies is time consuming and very costly. In the context of pharmaceutical products, for example, it is estimated that the development of a novel item takes 10–15 years and costs several hundred million pounds on average [1]. Among the largest contributors to both time and cost are confirmatory (Phase III) clinical trials that often involve thousands of patients with follow-up period frequently lasting years [2]. In recent years, however, around 50% of confirmatory clinical trials have failed to show a beneficial effect or been rejected at regulatory submission [3] resulting in a large number of participants in these trials being exposed to an ineffective or even harmful treatment while at the same time costing substantial amounts of money. The situation within Phase II studies is even worse with only 18% of these studies progressing a drug candidate into Phase III trials [4]. As a result of these shockingly high failure rates, alternative approaches to drug development are being explored. In this paper, we describe the advantages and additional complications of multi-arm studies that select/drop treatments during the conduct of the study. We begin with a description of the designs followed by relevant additional (practical) aspects that need to be considered before embarking on such a design. We then provide an illustrative example highlighting the efficiency gained by these approaches on the basis of trials in Alzheimer's disease before we finish with some general conclusions.

An overview of different types of multi-arm studies

Multi-arm studies

A multi-arm study is a study which compares several experimental treatments against a common control group. An immediate advantage of such an approach over separate two-arm studies is that only a single control group is used. As a consequence, a patient's chance to receive an experimental treatment is increased which has been argued could help with recruiting patients to such studies [5,6]. Additionally, such studies allow a fair, contemporary comparison of different experimental treatments as the comparisons are made against the same control group and under a single protocol so that relevant features of the study, such as inclusion/exclusion criteria, are the same.

Multi-arm studies with treatment selection

One of the drawbacks of traditional two-armed studies is that they do not allow

FUTURE SCIENCE part of fsg

for early conclusions (for better or for worse) about the treatment. To overcome this, group-sequential designs that allow the study to stop early, either because the evidence is already sufficient to claim superiority of the treatment over control or because it is unlikely to reach such a claim, have been developed [7,8] and are now routinely used in practice.

In the same spirit, multi-arm studies can be made more efficient by adding interim analyses that allow early stopping because the evidence collected is already sufficient to conclude that one or more treatments is superior to control or to stop because none of the experimental treatments looks sufficiently promising. Additionally, interim analysis can be used to select which treatment(s) warrant further experimentation. Typical selection rules used select the best performing or the k-best treatments [9,10], select any treatments that are close to the best performing one [11] or select any treatment that looks promising [12,13]. Figure 1 shows an example of such a design where all promising treatments continue in the study. In this example, two interim analyses and a final analysis are planned. After sufficient patients have been recruited for the first interim analysis, test statistics comparing each experimental treatment to control are found. In the fictitious example, two of the statistics, corresponding to experimental treatments 1 and 3, fall below the lower bound indicating that not further experimentation is warranted on these treatments and consequently these arms are dropped from the study. The remaining two test statistics, corresponding to treatments 2 and 4, are above the lower threshold but not above the upper bound. Therefore, additional information is required on these arms (plus control) to reach a definite conclusion. More patients are consequently recruited to these treatments and control. At the second interim analysis, the test statistic comparing experimental treatment 2 to control exceeds the upper boundary and hence superiority of treatment 2 over control can be claimed and the trial can be stopped.

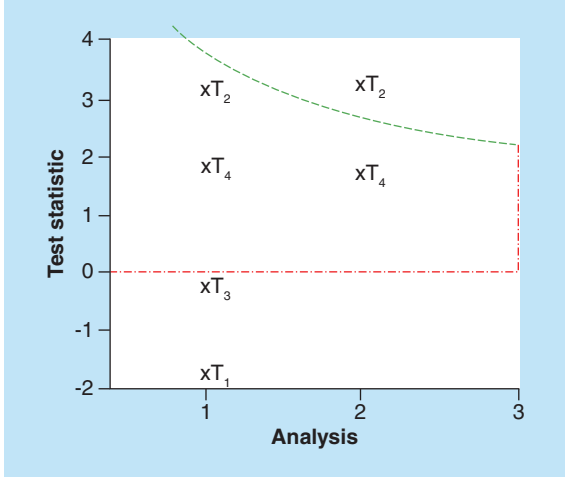

Figure 1. A multi-arm study which selects all promising treatments at interim analysis. Three analyses are planned for a study with four experimental treatments versus control. At the first interim analysis, treatments 1 and 3 are dropped from the study as they are below the futility threshold. At the second interim analysis, the second test statistic exceeds the upper bound so that superiority of treatment 2 over control can be concluded and the study can be stopped.

Because treatments are removed from consideration early and the trial can be stopped before the maximum number of patients are recruited, the required sample size of such studies will typically be smaller than a multi-arm study without treatment selection. It is, however, possible that the realized sample size is in fact larger than a multi-arm study without selection. In particular, in the unlikely case that no treatment arm can be dropped early and no early claim of superiority is possible, the sample size will be larger. This is because allowing for early claim of superiority also gives additional opportunities to wrongly make such a claim. To counteract such mistakes and to ensure that the overall type I error of the procedure is controlled, more stringent critical values than the ones utilized for multi-arm studies are required. The impact of allowing arms to be dropped on the other hand is an increase in type II error if the sample size is kept the same.

To design a multi-arm study with treatment selection, two different statistical approaches can be utilized. The so called 'pre-planned' adaptive designs [9–10,12,14], which are extensions of group-sequential methods, require specification of how treatments will be selected (e.g., select the best treatment or any treatment surpassing a predetermined threshold) while 'fully flexible' adaptive designs [15,16] do not require such pre-specification. The cost for the additional flexibility of the latter approach is a potential loss in efficiency (typically power is lowered by a few% [17]). In either case, we believe that it is paramount that the overall type-I error is controlled [18]. Adding additional treatments would otherwise increase the change of finding an effect (even when all are truly ineffective). As a consequence in a trial without overall type-I-error control, one could simply include numerous doses of the same treatment in a study to be almost certain to show that the treatment has an effect. In current practice, overall type-I-error control is, however, not always adhered to [19].

Besides the added efficiency that will be illustrated below, a further advantage of a multi-arm study with treatment selection, illustrated in Figure 2, is that one can use an interim analysis to mark the end of Phase II and beginning of Phase III thereby allowing a seamless Phase II/III study which removes the white

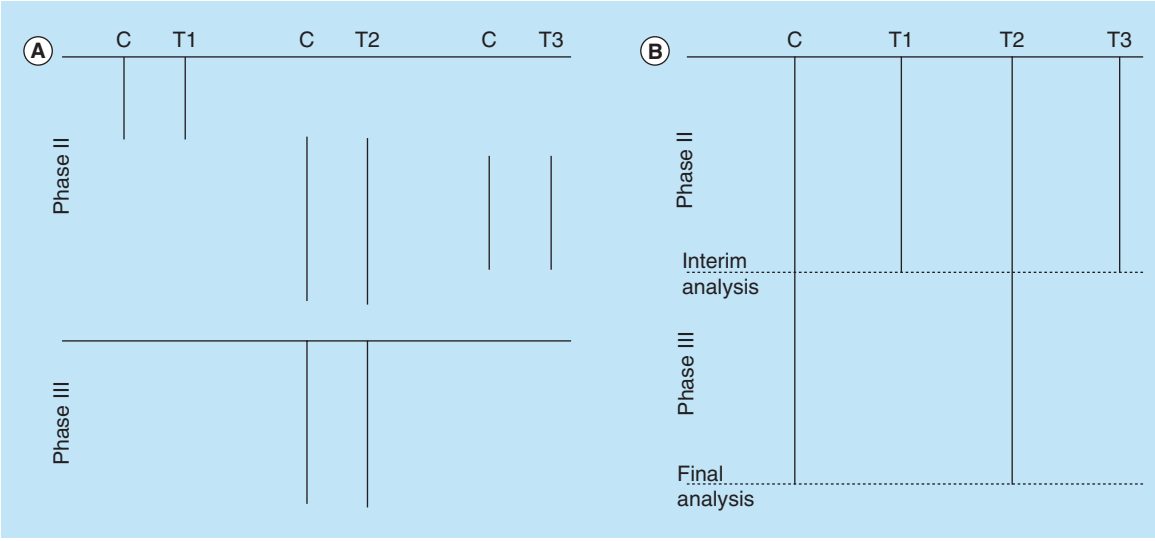

Figure 2. Traditional development versus one based on a multi-arm design. (A) Traditional 'sequential' development process; and **(B)** a multi-arm design with treatment selection. In both, three novel treatments (T1, T2 and T3) are evaluated against control and only treatment 2 is chosen for confirmation in Phase III. In **(A)** each treatment is compared with control in separate trials, while in **(B)** only one control group serves for all treatments. C: Control.

space between the phases. Although this may not always be desirable [20], it can reduce the time of drug development notably.

Adaptive multi-arm studies with treatment selection

When preforming an interim analysis for treatment selection, it is natural to also consider other adaptations. Most commonly sample size reassessment is of interest. Such an adaptation uses the accumulated trial data to verify assumptions made about factors of the study and potentially adjusts the sample size required based on these new estimates. Typically, a factor that is not of primary interest (e.g., variability of the end point) is estimated based on the accumulated trial data and if the estimate deviates from the value assumed on initiation of the study, the new estimate is used to update the sample size required.

Including such additional modifications in the study are straightforward if fully flexible designs are used. For pre-planned adaptive designs, the conditional error approach [21,22] can be used to incorporate such additional adaptations. Using a pre-planned design and making it flexible is, however, only advisable when such adaptations are unplanned as a fully flexible design will typically be more efficient otherwise [23]. For an easy to follow overview of fully flexible adaptive design ideas, see [24].

An example of a multi-arm study with treatment selection

In the above section, we have described the general concept of a multi-arm study with treatment selection and argued that they are an efficient way to investigate several different treatments against a common control group. In this section, we will give a numerical illustration of the gains possible, based on trials in Alzheimer's disease. A review in Alzheimer's disease published in 2010 [25] found that a large number of different treatments are currently being tested. No less than 13 Phase III studies were on going with most using traditional two-arm designs, meaning that equally many control groups are being used. Although not all treatments are targeting the same mechanism of actions and hence are not immediately comparable, there are still three or four treatments being evaluated when only considering treatments that are targeting the same mechanism of action. In this illustration we will compare the sample size requirements of different strategies to evaluate 3 experimental treatments in Alzheimer's disease. The first strategy evaluates all three experimental treatments in three distinct two-arm trials while the second utilizes separate group-sequential designs with triangular stopping boundaries [26]. The third strategy evaluates all three experimental treatments in a single study while a design with multi-arm with treatment selection of all promising treatments [27] is used as the final alternative.

For this hypothetical example, we will use the design parameters used in the recently completed LADDER trial [28]. The primary end point of interest is the change from baseline in the 11-item Alzheimer's Disease Assessment Scale–cognitive subscale [29] at week 24 and we model the outcome as normally distributed. In line with [28], we assume a standard deviation of 6 in the primary outcome and that a 2 point difference

is considered a clinically relevant effect. A one-sided type-I error of 2.5% and power of 90% are used.

Table 1 provides the (maximum) sample sizes of the four different strategies to evaluate three experimental treatments. We consider conducting three separate two-armed trials, three group-sequential trials as well as multi-arm trials with and without selection and use equal allocation of patients to all arms. Calculations were performed using the R package MAMS [30]. The group-sequential design and the multi-arm trial with selection each use one interim analysis conducted at the half-way point of the study. As group-sequential designs and multi-arm designs with treatment selection offer the opportunity to stop early, the expected sample sizes when no treatment is better than control and when exactly one treatment is superior to control are also provided. The (maximum) sample size of using separate trials to evaluate the experimental treatments is larger than the sample size required if a multi-arm trial design (with or without selection) is used. This is despite the fact that no attempt has been made to correct for multiplicity when using separate studies. The sample size of three separate single-stage trials when using a Bonferroni correction to ensure overall type-I-error control – something that the multi-arm designs discussed here do provide automatically – is at 1464 patients, about 50% larger than the multi-arm strategies. When acknowledging that a multi-arm design with treatment selection is expected to drop at least some arms, we find the advantage of the multi-arm design to be even larger. With only around 640 patients expected to be required before a definitive conclusion is reached, a multi-arm strategy is clearly more efficient than conducting separate studies.

Practical considerations for multi-arm studies

Clearly there is a substantial efficiency advantage in using a multi-arm study with treatment selection instead of conducting several separate trials. The above arguments, though statistically correct, are however a little bit over-enthusiastic. This is because some additional considerations and administrative hurdles need to be overcome to benefit from these, potentially large, gains. First there are considerations that come from the desire to evaluate several experimental treatments against a common control and second there are considerations that only apply because interim analyses are used for treatment selection. The latter are by large similar to challenges encountered in two-arm group-sequential designs.

The first challenge introduced by comparing multiple arms is that different trials comparing a single treatment against control are often initiated and conducted by different centres. As a result, they have different inclusion and exclusion criteria, may use different primary and secondary end points and possibly a different comparator treatment. All of these must be standardized for a multi-arm trial that requires negotiations and compromises between investigators. Since a multi-arm trial operates as a single trial under one protocol all treatments in the study need to be available at the same time to ensure contemporary evaluation. Additionally, a multi-arm study implicitly assumes that all experimental treatments start at an equal footing and hence they will only be efficient if there is no reason to believe that one treatment will have a better chance of yielding an improvement over control than any other.

A second challenge is to ensure that no bias in the evaluation is introduced in multi-center multi-arm studies through imbalances between allocations to treatments at different centers/regions. It is therefore paramount that randomization to all arms (including the control arm) is stratified by center or region to ensure that the risk of bias is minimized.

The third challenge concerns the analysis of such studies. At the end of a multi-arm study estimating the effect of the best experimental treatment is often of main interest. Using standard analysis methods for this purpose will result in an over-enthusiastic (upward biased) estimate of the effect [31]. Specialized methods that lead to unbiased estimators [31] or reduce the bias [32] are therefore necessary for analyzing such

Table 1. Sample size requirement for different strategies to evaluate three experimental Alzheimer's treatments against control

Design option	Maximum sample size	Expected sample size	
		All treatments ineffective	One treatment effective
Three separate two-armed trials	1140	1140	1140
Three separate group-sequential trials	1296	788.54	884.88
Multi-arm study without selection	952	952	952
Multi-arm study with treatment selection	1048	642.51	642.88

studies. Similarly, specialized methods to construct confidence intervals are required [33,34].

The first important consideration when allowing interim analyses for treatment selection is that the maximum sample size required will be larger than for a multi-arm study without selection (although the expected number of patients is typically notably smaller). Even though the increase in maximum sample size is typically small, recruiting the maximum number of patients still needs to be possible and investigators need to be prepared to recruit that many patients in the unlikely event that no treatment can be dropped early and no early claim of superiority is possible.

Second, in order to observe (notable) reductions in the sample size required, the end point utilized for treatment selection (typically the primary end point or some short-term surrogate) needs to be available quickly relative to the recruitment rate. The reason for this is that patients will continue to be randomized to each arm while the data to make the interim treatment selection are being collected. In the most extreme case therefore all patients could already be recruited by the time the information from assessed patients is available for making the treatment selection decision.

Third, the organization of interim analyses must be efficient with data monitoring and statistical analysis done to tight deadlines as delays in the selection decision reduces the benefit of such a design as argued above. Additional resource may therefore be required to allow a quick decision making as well as ensuring blinding and trial integrity is maintained. To achieve this, efficient communication between the investigators, data management and statisticians is essential.

A fourth consideration is around communicating the more complex design of a multi-arm study with treatment selection to both patients and investigators. In particular, the informed consent procedure requires careful consideration as patients need to be fully aware of all possibilities. In the STAMPEDE trial, for example, a two-part patient information sheet was used. Information on all arms was provided to all patients while further details on the allocated arms were made available after randomization [35].

The final challenge concerns planning and ensuring treatment supply. The maximum drug supply is uncertain as arms can be stopped prior to the end of the study. Although the same issue is present in group-sequential designs, the additional arms make this challenge more pronounced in multi-arm studies with treatment selection. Accurate planning and in particular precise estimation of recruitment rates – particularly for multi-center studies (e.g., [36]) – are paramount.

Discussion

Multi-arm studies with treatment selection are an efficient means for drug development when several potentially useful treatments are available for testing, and a number of different studies are now being run under this framework in a variety of disease areas [37–39]. In this paper, we have not only highlighted the potential gains that are possible when using such an approach but also discussed the additional complexities such designs bring with them. While we have kept the illustration simple, it should be noted that in-depth evaluations of the design alternatives, usually via simulations, are crucial when deciding which design is best. To support the implementation of these ideas, various software solutions exist. Commercial software such as AddPlan [40] or EAST [41] provide tools to design, simulate and analyze such studies. Additionally, the add-on packages MAMS [30] and asd [42] for the statistical software R [43] are freely available.

Future perspective

Multi-arm designs will become more widely used in the future as an efficient tool to make evidence-based decisions about different licensed treatments. Additionally, their use during the development of novel treatments will increase as familiarity with these ideas rises.

Financial & competing interests disclosure

This report is independent research arising from the author'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 author and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. This work was supported in part by grants from the National Institute for Health Research (NIHR-CDF-2010–03–32) and the Medical Research Council (MR/J004979/1). The author has no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

- Multi-arm designs are an efficient way to evaluate multiple experimental treatments.
- Allowing for selection of treatments enables even bigger gains in efficiency.
- Additional considerations and operational hurdles apply that mean that planning of such studies is more complex and time consuming.

References

Papers of special note have been highlighted as:
• of interest; •• of considerable interest.

1 DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. *J. Health Econ.* 22(2), 151–185 (2003).

2 European federation of pharmaceutical industries and associations. The pharmaceutical industry in figures. www.efpia.org

3 Arrowsmith J. Trial watch: Phase III and submission failures: 2007–2010. *Nat. Rev. Drug Discov.* 10, 87 (2011)

4 Arrowsmith J. Trial watch: Phase II failures: 2008–2010. *Nat. Rev. Drug Discov.* 10, 328–329 (2011).

5 Parmar MKB, Carpenter J, Sydes MR. More multiarm randomised trials of superiority are needed. *Lancet* 384(9940), 283–284 (2014).

6 Halpern SD, Karlawish JHT, Casarett D, Berlin JA, Townsend RR, Asch DA. Hypertensive patients' willingness to participate in placebo-controlled trials: Implication for recruitment efficiency. *Am. Heart J.* 146(6), 985–992 (2003).

7 Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials.* Chapman and Hall, Boca Raton, FL, USA (2000).

• **A comprehensive description of group-sequential methods.**

8 Whitehead J. *The Design and Analysis of Sequential Clinical Trials.* Wiley: Chichester, UK (1997).

9 Stallard N, Todd S. Sequential designs for Phase III clinical trials incorporating treatment selection. *Stat. Med.* 22(5), 689–703 (2003).

10 Stallard N, Friede T. A group-sequential design for clinical trials with treatment selection. *Stat. Med.* 27(29), 6209–6227 (2008).

11 Kelly PJ, Stallard N, Todd S. An adaptive group sequential design for Phase II/III clinical trials that select a single treatment from several. *J. Biopharm. Stat.* 15(4), 641–658 (2005).

12 Magirr D, Jaki T, Whitehead J. A generalised Dunnett test for multi-arm, multi-stage clinical studies with treatment selection. *Biometrika* 99(2), 494–501 (2012).

13 Royston P, Parmar MK, Qian W. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Stat. Med.* 22(14), 2239–2256 (2003).

14 Wason JMS, Jaki T. Optimal design of multi-arm multi-stage trials. *Stat. Med.* 31(30), 4269–4279 (2012).

15 Posch M, Konig F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Stat. Med.* 24(24), 3697–3714 (2005).

16 Bretz F, Konig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. *Stat. Med.* 28(8), 1181–1217 (2009).

17 Friede T, Stallard N. A comparison of methods for adaptive treatment selection. *Biom. J.* 50(5), 767–781 (2008).

• **A comparison of different approach for multi-arm studies.**

18 Wason JMS, Magirr D, Law M, Jaki T. Some recommendations for multi-arm multi-stage trials. *Stat. Methods Med. Res.* (2013) (Epub ahead of print).

•• **An overview of multi-arm trials with treatment selection.**

19 Wason JM, Stecher L, Mander AP. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials* 15, 364 (2014).

20 Cuffe RL, Lawrence D, Stone A, Vandemeulebroecke M. When is a seamless study desirable? Case studies from different pharmaceutical sponsors. *Pharm. Stat.* 13(4), 229–237 (2014).

•• **A great illustration of benefits and pitfalls of adaptive studies.**

21 Müller HH, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. *Stat. Med.* 23(16), 2497–2508 (2004).

22 Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. *Stat. Med.* 27(10), 1612–1625 (2008).

23 Magirr D, Stallard N, Jaki T. Flexible sequential designs for multi-arm clinical trials. *Stat. Med.* 33(19), 3269–3279 (2014).

24 Jennison C, Turnbull BW. Adaptive seamless designs: selection and prospective testing of hypotheses. *J. Biopharm. Stat.* 17(6), 1135–1161 (2007).

25 Mangialasche F, Solomon A, Winblad B, Mecocci P, Kivipelto M. Alzheimer's disease: clinical trials and drug development. *Lancet Neurol.* 9(7), 702–716 (2010).

26 Whitehead J, Stratton I. Group sequential clinical trials with triangular continuation regions. *Biometrics* 39, 227–236 (1983).

27 Jaki T. Designing multi-arm multi-stage clinical studies. In: *Developments in Statistical Evaluation of Clinical Trials.* van Montfort K, Oud J, Ghidey W (Eds.) Springer, Berlin Heidelberg, Germany, 51–69 (2014).

28 Wilkinson D, Windfeld K, Colding-Jørgensen E. Safety and efficacy of idalopirdine, a 5-HT6 receptor antagonist, in patients with moderate Alzheimer's disease (LADDER): a randomised, double-blind, placebo-controlled Phase 2 trial. *Lancet Neurol.* 13(11), 1092–1099 (2014).

29 Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. *Am. J. Psychiat.* 141, 1356–1364 (1984).

30 Jaki T, Magirr D. MAMS: designing multi-arm multi-stage studies, 2014. http://CRAN.R-project.org

31 Bauer P, Koenig F, Brannath W, Posch M. Selection and bias – two hostile brothers. *Stat. Med.* 29(1), 1–13 (2010).

32 Bowden J, Glimm E. Unbiased estimation of selected treatment means in two-stage trials. *Biom. J.* 50(4), 515–527 (2008).

33 Carreras M, Brannath W. Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. *Stat. Med.* 32(10), 1677–1690 (2013).

34 Jaki T, Magirr D. Considerations on covariates and end points in multi-arm multi-stage clinical trials. *Stat. Med.* 32(7), 1150–1163 (2013).

35 Magirr D, Jaki T, Posch M, Klinglmueller F. Simultaneous confidence intervals that are compatible with closed testing in adaptive designs. *Biometrika* 100(4), 985–996 (2013).

36 Sydes MR, Parmar MK, James ND *et al.* Issues in applying multi-arm multi-phase methodology to a clinical trial in prostate cancer: the "MRC STAMPEDE" trial. *Trials* 10(39) doi:10.1186/1745-6215-10-39 (2009).

37 Anisimov V. Predictive modelling of recruitment and drug supply in multicenter clinical trials. *Proc. JSM.* (2009). http://public.ukcrn.org.uk

38 Evaluation of SQ109, high-dose rifampicin, and moxifloxacin in adults with smear-positive pulmonary TB in a MAMS design. http://clinicaltrials.gov

39 Marson AG, Al-Kharusi AM, Alwaidh M *et al.* The SANAD study of effectiveness of carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial. *Lancet* 369(9566), 1000–1015 (2007).

40 AddPlan. www.aptivsolutions.com

41 EAST. www.cytel.com

42 Parsons N, Friede T, Todd S, *et al.* An R package for implementing simulations for seamless Phase II/III clinical trials using early outcomes for treatment selection. *Comput. Stat. Data An.* 56(5), 1150–1160 (2012)

43 R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria (2014). www.R-project.org