Design of a multi-arm randomized clinical trial with no control arm

Amalia Magaret^{*1,2}, Derek C. Angus³, Neill K.J. Adhikari⁴, Patrick Banura⁵, Niranjana Kissoon⁶, James V. Lawler^{5,7}, Shevin T. Jacob^{5,8,9}

Author information

Departments of Laboratory Medicine¹, Biostatistics², Medicine⁸, and the International Respiratory and Severe Illness Center (INTERSECT)⁹, University of Washington, Seattle, WA, US

Department of Critical Care Medicine³, University of Pittsburgh School of Medicine, University of Pittsburgh, Pittsburgh, PA, US

Department of Critical Care Medicine and Sunnybrook Research Institute⁴, Sunnybrook Health Sciences Centre and Interdepartmental Division of Critical Care, University of Toronto, Toronto, Ontario, Canada

Austere environments Consortium for Enhanced Sepsis Outcomes⁵, Frederick, MD, US

Departments of Pediatrics and Emergency Medicine⁶, British Columbia's Children's Hospital and The University of British Columbia, Vancouver, BC, CA

United States Naval Medical Research Center⁷, Frederick, MD, US

*Corresponding author

Amalia Magaret

Virology Research Clinic

Harborview Medical Center

325 9th Ave, Box #359928

Seattle, WA 98104

amag@uw.edu

ph: 206-520-4316

fax: 206-520-4371

Keywords: clinical trial, sepsis, multi-arm, randomized, study design

Abstract

Background: Clinical trial designs that include multiple treatments are currently limited to those that perform pairwise comparisons of each investigational treatment to a single control. However, there are settings, such as the recent Ebola outbreak, in which no treatment has been demonstrated to be effective; and therefore, no standard of care exists which would serve as an appropriate control.

Methods/Design: For illustrative purposes, we focused on the care of patients presenting in austere settings with critically ill ‘sepsis-like’ syndromes. Our approach involves a novel algorithm for comparing mortality among arms without requiring a single fixed control. The algorithm allows poorly-performing arms to be dropped during interim analyses. Consequently, the study may be completed earlier than planned. We used simulation to determine operating characteristics for the trial and to estimate the required sample size.

Results: We present a potential study design targeting a minimal effect size of a 23% relative reduction in mortality between any pair of arms. Using estimated power and spurious significance rates from the simulated scenarios, we show that such a trial would require 2550 participants. Over a range of scenarios, our study has 80 to 99% power to select the optimal treatment. Using a fixed control design, if the control arm is least efficacious, 640 subjects would be enrolled into the least efficacious arm, while our algorithm would enroll between 170 and 430. This simulation method can be easily extended to other settings or other binary outcomes.

Conclusion: Early dropping of arms is efficient and ethical when conducting clinical trials with multiple arms.

Background

Clinical setting

Sepsis is defined as acute life-threatening organ dysfunction secondary to infection.

Sepsis affects tens of millions of people each year globally, with a substantially higher burden and mortality risk in resource poor settings [1-4]. Although typically due to bacterial infections, many systemic manifestations of sepsis are also seen in several viral infections, including Ebola. Sepsis management guidelines for resource poor settings adapted from Surviving Sepsis Campaign (SSC) guidelines have been proposed by World Health Organization (WHO) and global intensive care organizations for deployment in both interpandemic and pandemic situations [5-7]. These guidelines are primarily consensus-based, however, and lack a basis in robust clinical trial evidence. Importantly, recent large sepsis trials conducted in the US, Australia and England (i.e., PRoCESS [8], ARISE [9], ProMISe [10]) showed no survival benefit with protocolized resuscitation when compared to usual care. Resuscitation strategies for sepsis may be context-specific, with lack of generalizability of randomized trials from high-income settings [11]. Thus, there is great interest in determining optimal resuscitation strategies in resource poor settings.

Study design choices

With no current standard of care proven effective, we desired a multi-arm clinical trial design able to consider all treatments as potentially optimal, rather than to compare each arm to a fixed control. Comparing each arm to a fixed control has two drawbacks: 1) it does not allow comparison between experimental arms and 2) it requires a control arm to be retained throughout the study. We sought to develop an efficient, ethical study design that could select an optimal treatment, while also allowing for the

possibility of any arm to be dropped for poor performance. We included the consideration of interim analyses so that treatment differences larger than the minimal important difference could be detected early. Poorly performing arms could be stopped at these interim analyses or the entire trial might be halted if a clear winner were apparent. Inherent in interim analyses was the potential for inflation in the type I error rate. To maximize efficiency and control type I error in such trials, we considered several design approaches, including sequentially designed trials, Bayesian adaptively designed trials, and simulation-based approaches.

Sequentially designed trials control type I error by requiring more stringent cutoffs for significance at each interim analysis, thereby controlling the trial-wide spurious significance rate [12]. The methods used to compute these amended cutoffs are well-established for a variety of settings involving two-arm trials, which can be expanded to multi-arm trials so long as there is a common control arm (not necessarily a placebo) against which all others are compared [13-15].

Bayesian design is purportedly more efficient than sequential design for a multi-arm setting, without requiring a pre-specified control arm. Its primary gains in efficiency are claimed through adaptive randomization and the potential to drop arms mid-study [16, 17]. Adaptive randomization is the process whereby allocation proportions are updated at interim analyses toward favoring the currently most effective treatment(s). However, any increased efficiency of this approach may come at the cost of reduced interpretability. This situation arises when changing allocation proportions among arms make treatment benefit impossible to distinguish from changing baseline risk levels over the period of accrual, also known as “drift” [18]. For large effectiveness trials

with longer term accrual periods, where drift is plausible and the study findings are intended to be decisive, confounding of the treatment effect with an unobservable drift is highly undesirable. For findings potentially influenced by drift, inference made with adaptive randomization should be confirmed with subsequent trials [19]. Post-hoc adjustments for stage of randomization can ameliorate the drift-induced bias, but at the cost of reduced power [18, 20].

Design of multi-arm multi-stage (MAMS) clinical trials, proposed by Royston et al., involves a simulation-based approach that efficiently reduces the number of arms in which the primary clinical endpoint will be tested by weeding out poorly performing arms in an earlier stage, often using an intermediate endpoint [21]. These trials require a single, common control arm that is not dropped during the course of the trial and a validated, intermediate endpoint. This method evaluates type I error during the design phase of the trial through simulation, using random number generators to repeatedly create and assess hypothetical data under a variety of plausible study conditions. In comparing the ability of the trial to detect true and spurious treatment differences in survival under these varying conditions, simulations determine the number of participants needed to achieve desired operating characteristics.

We wanted to conduct a trial with the potential to drop arms at pre-determined intervals without requiring the specification of a single control arm. Since both standard sequential design and MAMS required a fixed control arm and Bayesian adaptive randomization is more appropriate for exploratory rather than confirmatory trials [19], none of the study design contexts described above provided an appropriate precedent for this trial. Therefore, we developed and conducted simulations to suit our

particular conditions, permitting the dropping of any arm, with the goal of selecting the most effective treatment arm or arms by trial termination. The operating characteristics of our trial were achieved by repeatedly simulating trial conditions under as realistic as possible assumptions and observing the detection rate of real (power) and spurious (type I error) treatment differences. As such, our design used frequentist stopping rules (p-values).

Methods

We described a clinical trial designed to discern the optimal treatment regimen for severe sepsis treatment in resource poor settings. We anticipated four arms with varying treatments, including local best available care and protocolized strategies, and anticipate a primary outcome of mortality at 28 days. These methods, however, would apply to any multi-arm study with a binary endpoint, and simulations could easily be extended to include more than 4 arms.

Simulated conditions

The common assumptions for each simulated scenario included 24-30 months of accrual and equal randomization to all active treatment arms. We imposed a 23% relative reduction in mortality between arms as our minimal clinically important difference and designed the sample size to achieve 80% power to detect that difference. This relative reduction resulted in different assumed absolute reductions in risk depending on the starting point for comparison. For example, a 23% relative reduction translated to a 7% absolute reduction from a 31% mortality to 24% (see figure 1 for other absolute reductions).

We evaluated scenarios having no treatment differences and those having successive reductions in mortality for some to all arms. Scenarios A-J describe the variety of plausible conditions for a sepsis trial in a resource-poor setting (such as sub-Saharan Africa), with most assuming a highest observed mortality of 40% (figure 1) [22]. Often we assumed at least three distinct mortality rates between the four treatments. We would have more power than that shown to detect treatment benefit if relative reductions in mortality between arms were larger than 23%.

To make our simulations as realistic as possible, we imposed facility-specific differences in baseline mortality among 5 anticipated study sites and allowed mortality to vary uniformly within 5% of the assumed treatment-specific rate. For each scenario, three interim analyses and one final analysis were simulated; the process was repeated 4000 times in order to achieve $\pm 1\%$ precision in power and type I error rates. Through trial and error, three interim analyses were selected in order to provide reasonably high probabilities of early stopping. More frequent interim analyses provided no substantial benefit in power while increasing the spurious significance rate.

Treatment arm comparison algorithm

We needed a means of comparing all arms to one another efficiently, with the objective of determining the most effective arm. Therefore, we developed a procedure to minimize the number of comparisons with the goal of dropping any and all arms shown to perform significantly worse than the best performing arm. In contrast to a procedure that would perform all pair-wise comparisons, this process began by ranking all active arms by observed 28-day mortality at the time of analysis. Limiting the number of tests performed increased power and reduced the spurious significance rate. The algorithm

was designed without a designated control arm. At each interim analysis, the order of testing was as follows.

- Step 1. Was there an arm with significantly *lower* mortality than the next lowest?
 - a. If yes, drop all remaining arms and stop for efficacy.
 - b. If no, continue to step 2.
- Step 2. Were there arms with significantly *higher* mortality than the best arm?
 - a. If yes, drop those arms.
 - b. Either way, continue to step 3, at least two arms will remain.
- Step 3. Did remaining best and worst performing arms have similar mortality?
 - a. If yes, stop for futility.
 - b. If no, continue enrolling.

Whenever arms were dropped, subsequently enrolled participants were allocated evenly between remaining active treatment arms. The efficacy cutoff was used in steps 1 and 2, and the futility cutoff in step 3. Cutoffs were determined through simulation such that regardless of the actual differences between arms, type I error rate was limited and we had at least 80% power to determine the most efficacious arm or arms. During the trial, this order of hypothesis testing should be repeated exactly as simulated in order to assure achievement of the targeted operating characteristics (see Appendix A for details of statistical testing and Appendix B for associated statistical code in R). Both relative and absolute mortality reduction could be reported at study end.

Definition of type I error in trials with no control group

The null scenario ‘N’ with all arms having an equal mortality allowed us to test the spurious significance rate and to choose a cutoff p-value for efficacy (α_E). With a single control arm, the type I error can be consistently defined as the probability of preferring any arm when no arms perform better than the control. In this context of no pre-specified control arm, spurious significance has two potential definitions: 1) as above,

the probability of preferring any single arm as optimal when all arms have the same mortality; and 2) the probability of falsely declaring any arms to be significantly worse than the best performing arm when the null hypothesis is true. We opted to evaluate both definitions and provided both overall as well as arm-specific spurious significance rates.

Additional measures recorded

In each simulation, we recorded the timing of dropping of each arm and the timing of early stopping for either efficacy or for futility. From these results, we computed the proportion of simulations in which the trial was stopped early for either reason and the proportion of trials in which the arm with highest mortality was dropped at each interim analysis. This latter proportion informs the probability of being randomized to the least effective intervention.

We additionally recorded the number of participants completed overall and on each arm at the time of early termination. This allowed us to calculate potential reductions in the planned sample size. Expected sample sizes for each scenario were then computed as the average of observed sample size over many randomly generated trials. In practice, the study would make decisions at pre-planned interim analyses only, when $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{3}{4}$ of the total outcomes have been observed. Therefore, as the study would accrue approximately $2550/4 \approx 640$ persons between each interim analysis, the study would terminate after either ~ 640 , ~ 1280 , ~ 1920 , or ~ 2550 persons enrolled. An expected enrollment number of 1800, however, might come from a 25% chance of stopping with 1280 persons enrolled, a 69% chance of stopping with 1920, and a 6% chance of fully enrolling.

Comparison to fixed control design

We also computed the power to conduct this trial comparing each arm to a fixed control, a scenario that might be closest to a standard sequentially designed trial. While in all previous simulations, no treatment designation was associated with a particular mortality rate, here we designated 40% as the mortality under the control arm and compared each other arm to the observed mortality in that specified arm, performing pairwise comparisons of all active arms versus the control at each interim analysis. We recomputed the operating characteristics under this alternate algorithm. These simulations performed were otherwise similar to those described above, including site variability and allowing arms to drop once either the futility or efficacy cutoff was achieved. Subsequent participants were allocated equally to remaining arms as before, and efficacy and futility cutoff were determined to achieve 80% power and control overall type I error.

Results

Power

With 2550 participants, we found at least 80% power to identify the arm with the lowest mortality rate, given relative mortality reductions of 23% relative or greater, with most tested scenarios having $\geq 90\%$ power (figure 2a). With a 10% loss to follow-up for the mortality outcome, this study would require 2800 enrolled persons.

Early Stopping

It was of interest to determine the probability of completing the trial early under the different scenarios. All scenarios had at least 60% probability of stopping by the time of the third interim analysis, or when ~1920 participants have accrued (figure 2b), with some scenarios (E, G, H and N) having over a 90% probability of stopping by the third

interim analysis. In addition, all scenarios with a single highest mortality arm (B, C, D, F and G) had at least 79% power to drop that arm at the first interim analysis (figure 2c) and at least a 99% probability of dropping that arm by the second interim analysis (figure 2d). Scenarios with two equally high mortality arms had at least a 44% chance of dropping both those arms by the second interim analysis and a $\geq 80\%$ chance of dropping them by the third interim analysis (Supplementary Figure 1).

Planned versus effective sample size

Dropping poorly performing arms at interim analyses allowed the enrolled sample size to be lower than planned under many scenarios. Figure 3 shows the average number enrolled over the 4000 simulations plotted against the planned number to enroll, for planned sample sizes between 1000 and 3000. Only when we assumed two arms with equally lowest mortality (scenarios C, G and I) did we anticipate the full 2550 enrolled. Scenarios with more than one arm having lowest mortality were more likely to continue accruing to the final analysis to confirm that these treatments have a less than 23% relative mortality difference. The remaining scenarios had expected sample size substantially lower than planned, particularly those that assume two arms share the highest mortality (scenarios A, E and H). With those two highest mortality arms likely to be dropped early, all remaining participants were enrolled into the active arms, decreasing the time needed to complete the trial. The null scenario N was also expected to stop early for futility with ~1950 enrolled compared to the maximum potential of 2550.

Type I error

Spurious significance rates were well-controlled by the study stopping rules. The cutoff for futility, α_F , was determined to be 0.50, which minimized the probability of halting efficacious treatment scenarios, while also allowing null trials a substantial probability of stopping early. The cutoff for efficacy (α_E) was determined to be 0.055 for all analysis time points. Cutoffs for efficacy that varied over time, i.e. lower for early interim analysis and subsequently higher, were assessed but provided little additional protection against spurious significance at the expense of efficiency. The chosen cutoffs provided the following type I error rates: the probability of putting either of the equally low mortality arms forth as best when there are 2 arms having that same optimal (lowest) mortality was controlled at 5% overall (with one exception at 7%), and 2-3% for each specific arm. For the full null scenario there was only a 0.4% probability of selecting any arm as best. When the null hypothesis was true (all arms are equal), the probability that all four arms would be demonstrated equivalent was 86%. This left a 14% chance of not detecting equivalence, composed of a 13% chance and a 1% chance, respectively, of falsely determining one and two arms have higher mortality than the remaining. This last probability, 14%, is different from a spurious significance rate because no single winner is declared; rather, one to two arms were falsely declared to have a higher mortality than the remaining two to three equivalently optimal arms. We felt comfortable with this probability being higher than the traditional 5% type I error, since it involved falsely declaring one or two arms as ineffective rather than falsely declaring a single arm to be optimal. Finally, when any two arms exist (say, A and B) such that the mortality on A was at least 23% lower relative to B, the probability of selecting B as

optimal was less than .01%. In summary, in all scenarios we were extremely unlikely to choose the wrong arm as optimal.

Comparison to fixed control design

We also simulated an alternative algorithm of comparing each arm to a pre-selected control arm, which was designed under simulation to have 40% mortality. Due to the larger number of comparisons made, the cutoff for efficacy (α_E) was changed from 0.055 to 0.040 to maintain an overall 5% type I error, where here type I error is the probability that any arm is claimed to have lower or higher mortality than the control arm when, in actuality, all arms perform equally. Two main differences emerge. Firstly, since the control arm was maintained throughout the trial and served as the only comparator, we could not evaluate the probability of selecting the arm with the lowest mortality relative to all others. We can only show that (perhaps several) treatments are better than the control, but not choose between them. Secondly, we can show that for most scenarios and planned sample sizes, our algorithm enrolls fewer persons into the least efficacious arm, minimizing risk to participants (Figure 4). The exception is scenario H, where the higher enrollment for our algorithm is due to the large and erroneous tendency (10-40% of the time) of the fixed control algorithm to stop early for futility, enrolling 13-24% fewer participants overall. The early stopping in that case is due to the fixed algorithm evaluating futility the same way it evaluates efficacy, which is pairwise, instead of using our algorithm which only stops for futility when all remaining arms are determined equal by the futility threshold.

Discussion

We used simulation to directly assess the operating conditions of a hypothetical trial comparing resuscitation strategies for sepsis in a resource-poor setting. This trial did not have a fixed control arm and thus did not follow a standard design. Under a variety of reasonable assumptions, we demonstrated sufficient power to detect treatment differences of clinical interest.

We were able to compare all arms to one another and to identify the overall best performing arm with the lowest mortality. This objective of identifying the best arm is often not possible in designs which require only pairwise comparisons to a pre-determined control and do not allow comparisons between experimental arms. Additionally, in nearly all scenarios, the number of persons enrolled into the least efficacious arm was smaller using our algorithm versus that determined using the fixed control design. Our design was particularly superior to the fixed control design in scenarios where the designated control arm had the highest mortality. In that case, while the fixed control design would retain the control arm throughout accrual, our algorithm enrolled as few as 26% of the maximum number of participants in that arm. Lives were therefore saved and research resources conserved over previous designs. Furthermore, by using interim efficacy analyses, our design may result not just in dropping of arms but in early stopping of our trial to determine efficacy or futility. Thus, our algorithm appears optimal by several criteria.

“Drift”, or changing baseline risk of the outcome, is likely in our study setting of interest, sepsis, in which disease etiology may change. However, the occurrence of drift will not bias inference when using this study design. Because participants continue to be

allocated evenly to remaining arms following any dropping of ineffective arms, the arms will remain comparable to one another with respect to baseline risk factors over all stages of the study.

Our study design was targeted for a trial of four sepsis treatments without a currently proven standard of care and a primary outcome of 28-day mortality. However, this simulation method could be applied to other settings with varying numbers of arms and other binary outcomes and event rates.

Competing interests

None declared.

Authors contributions

DCA, NKJA, PB, NK, JVL, and STJ conceived the study design. AM is a statistician who designed the algorithm, performed the simulations, and summarized the results. All authors contributed to and approved the final manuscript.

Acknowledgments

The design of this study was supported by a Cooperative Agreement (N62645-PA-13-1) between the Naval Medical Research Center and the Henry M. Jackson Foundation for the Advancement of Military Medicine. The authors wish to thank Roger Lewis, Dennis Amundson, Allen Cheng, Arjen Dondorp and Andrew Kambugu for their critical review of this design during our many discussions.

Copyright statement

JVL is a military service member and this work was prepared as part of his official duties. Title 17 U.S.C. §105 provides that “Copyright protection under this title is not available for any work of the United States Government.” Title 17 U.S.C. §101 defines a U.S. Government work as a work prepared by a military service member or employee of the U.S. Government as part of that person’s official duties.

Disclaimer

The views expressed in this article are those of the author and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U. S. Government.

Figure legends

Figure 1. Each row (labeled A-J) presents a single simulated scenario showing 28-day mortality in each of four treatment arms. These scenarios were constructed to include 23% relative reductions between each arm and successively lower mortality arms (mortality of 18%, 24%, 31%, 40%, 52% and 67%). The size of the dot indicates the number of arms with that mortality.

(Fig1.hypothetical_effect_sizes.jpg here)

Figure 2. Power for scenarios A-J and N=the null scenario to detect treatment differences

Figure 2a. Power to drop all but lowest mortality treatment(s) (vertical axis) by the end of study for each scenario, plotted against hypothetical numbers enrolled (horizontal axis). Dashed horizontal lines show 0.05 and 0.80 power, while dashed vertical line shows planned sample size of 2550.

(Fig2a.manu2powerH2055.jpg here)

Figure 2b. Power to drop all but lowest mortality treatment(s) by the 3rd (last) interim analysis.

(Fig2b.manu2earlyeff055.jpg here)

Figure 2c. Power to drop highest mortality arm(s) at the first interim analysis.

(Fig2c.manu2pdropworst.first.int055.jpg here)

Figure 2d. Power to drop highest mortality arm(s) at the second interim analysis.

(Fig2d.manu2pdropworst.sec.int055.jpg here)

Figure 3. Expected overall sample size under scenarios A-J and N, the sum of all arms. Vertical axis shows number enrolled by the end of study (EOS).

(Fig3.manu2enroll055.jpg here)

Figure 4. Number enrolled into the least efficacious arm, all scenarios, comparing our algorithm (y-axis) to an algorithm in which each arm is compared to a fixed control arm (x-axis).

(Fig4.compareworst_scatter.jpg here)

References

1. Healthcare Cost and Utilization Project. . 2013. <http://www.hcup-us.ahrq.gov/reports/statbriefs/sb122.pdf>.
2. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. *Critical care medicine*. 2001;29(7):1303-10.
3. Kumar G, Kumar N, Taneja A, Kaleekal T, Tarima S, McGinley E et al. Nationwide trends of severe sepsis in the 21st century (2000-2007). *Chest*. 2011;140(5):1223-31. doi:10.1378/chest.11-0352.
4. Wenzel RP. Treating sepsis. *N Engl J Med*. 2002;347(13):966-7. doi:10.1056/NEJMp020096.
5. Organization WH. Integrated Management of Adolescent and Adult Illness District Clinician Manual. 2011.
6. Jacob ST, Lim M, Banura P, Bhagwanjee S, Bion J, Cheng AC et al. Integrating sepsis management recommendations into clinical care guidelines for district hospitals in resource-limited settings: the necessity to augment new guidelines with future research. *BMC medicine*. 2013;11:107. doi:10.1186/1741-7015-11-107.
7. Dunser MW, Festic E, Dondorp A, Kisson N, Ganbat T, Kwizera A et al. Recommendations for sepsis management in resource-limited settings. *Intensive care medicine*. 2012;38(4):557-74. doi:10.1007/s00134-012-2468-5.
8. Yealy DM, Kellum JA, Huang DT, Barnato AE, Weissfeld LA, Pike F et al. A randomized trial of protocol-based care for early septic shock. *N Engl J Med*. 2014;370(18):1683-93. doi:10.1056/NEJMoa1401602.
9. Peake SL, Delaney A, Bailey M, Bellomo R, Cameron PA, Cooper DJ et al. Goal-directed resuscitation for patients with early septic shock. *N Engl J Med*. 2014;371(16):1496-506. doi:10.1056/NEJMoa1404380.
10. Mouncey PR, Osborn TM, Power GS, Harrison DA, Sadique MZ, Grieve RD et al. Trial of early, goal-directed resuscitation for septic shock. *N Engl J Med*. 2015;372(14):1301-11. doi:10.1056/NEJMoa1500896.
11. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380(9859):2095-128. doi:10.1016/S0140-6736(12)61728-0.
12. Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. Boca Raton, FL: Chapman & Hall / CRC; 2000.
13. Barthel FM, Babiker A, Royston P, Parmar MK. Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over. *Stat Med*. 2006;25(15):2521-42. doi:10.1002/sim.2517.
14. Freidlin B, Korn EL, Gray R, Martin A. Multi-arm clinical trials of new agents: some design considerations. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 2008;14(14):4368-71. doi:10.1158/1078-0432.CCR-08-0325.
15. Wason JM, Jaki T. Optimal design of multi-arm multi-stage trials. *Stat Med*. 2012;31(30):4269-79. doi:10.1002/sim.5513.

16. Berry DA. Bayesian statistics and the efficiency and ethics of clinical trials. *Stat Sci.* 2004;19(1):175-87.
17. Berry DA. Adaptive Clinical Trials: The Promise and the Caution. *J Clin Oncol.* 2011;29(6):606-9.
18. Emerson SS, Fleming TR. Adaptive methods: telling "the rest of the story". *J Biopharm Stat.* 2010;20(6):1150-65. doi:10.1080/10543406.2010.514457.
19. FDA. Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics. <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm> 2010.
20. Simon R, Simon NR. Using randomization tests to preserve type I error with response adaptive and covariate adaptive randomization. *Stat Probabil Lett.* 2011;81(7):767-72.
21. Royston P, Parmar MK, Qian W. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Stat Med.* 2003;22(14):2239-56. doi:10.1002/sim.1430.
22. Jacob ST, Banura P, Baeten JM, Moore CC, Meya D, Nakiyingi L et al. The impact of early monitored management on survival in hospitalized adult Ugandan patients with severe sepsis: a prospective intervention study*. *Critical care medicine.* 2012;40(7):2050-8. doi:10.1097/CCM.0b013e31824e65d7.

Figure

Figure

Figure

Figure

Figure

Figure

Figure

Figure