Published in final edited form as:

Contemp Clin Trials. 2016 January ; 46: 48–51. doi:10.1016/j.cct.2015.11.010.

A Small n Sequential Multiple Assignment Randomized Trial Design For Use in Rare Disease Research

Roy N. Tamura^a, Jeffrey P. Krischer^a, Christian Pagnoux^b, Robert Micheletti^c, Peter C. Grayson^d, Yeh-Fong Chen^{e,*}, and Peter A. Merkel^f

^aHealth Informatics Institute, University of South Florida, Tampa, FL

^bDepartment of Rheumatology, University of Toronto, Toronto, CA

^cDepartments of Dermatology and Medicine, University of Pennsylvania, Philadelphia, PA

^dSystemic Autoimmunity Branch, National Institute of Arthritis and Musculoskeletal Diseases, Bethesda, MD

^eDivision of Biometrics, US Food and Drug Administration, Silver Spring, MD

^fDivison of Rheumatology, University of Pennsylvania, Philadelphia, PA

Abstract

Background—Clinical trials in rare diseases are difficult to conduct due to the limited number of patients available with each disorder. We developed a Phase 2 trial which is a small n sequential multiple assignment randomized trial (snSMART) design to test several treatments for a rare disease for which no standard therapy exists.

Purpose—This paper illustrate the design, sample size estimation and operating characteristics of an snSMART.

Methods—We investigate the performance of a class of weighted Z statistics via computer simulations.

Results—We demonstrate the increase in power over traditional single stage designs, and indicate how the power changes as a function of the weight given to each stage.

Conclusion—The snSMART design is promising in a rare diseases setting where several alternative treatments are under consideration and small sample sizes are necessary.

Corresponding author: Roy N. Tamura, Health Informatics Institute, University of South Florida, 3650 Spectrum Boulevard, Tampa, FL 33612. Tel.: +1 813 396 9548, roy.tamura@epi.usf.edu. jeffrey.krischer@epi.usf.edu (J.P. Krischer), Cpagnoux@mtsina.on.ca (C. Pagnoux), Robert.Micheletti@uphs.upenn.edu (R. Micheletti), peter.grayson@nih.gov (P. C. Grayson), YehFong.Chen@fda.hhs.gov (Y. Chen), Peter.Merkel@uphs.upenn.edu (P. A. Merkel)

*This paper reflects the views of the authors and should not be construed to represent the views or policies of the U. S. Food and Drug Administration.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

re-randomization; binary data; weighted Z statistic

Background

Conducting clinical research in rare diseases poses numerous challenges due to the small number of patients with any given disorder. Griggs et al. [1] make a number of recommendations for clinical trials in rare diseases, including testing more than one therapy within each patient. Most rare diseases have no approved therapy [2] and several drug therapies are often used even though the efficacy data for any of the drugs is limited. This was the case for a recent situation faced by the Vasculitis Clinical Research Consortium and this paper describes our approach to the design and analysis of a clinical trial for patients with a rare form of vasculitis. The design is a small n sequential multiple assignment randomized trial (snSMART).

Sequential multiple assignment randomized trials (SMART) in the literature [3,4,5,6,7] are most often conducted in large patient populations with well established therapeutic options. In these situations, the primary objective of the trials is to determine optimal sequences and combination of sequences of the therapies. In contrast, oftentimes in rare diseases, drugs are used without strong evidence of efficacy. Given this difference and also given the limited population size, our objectives were to first determine the best treatment in the initial stage in terms of response rate and second, to make inference around that chosen treatment in comparison to the other treatments. We use the re-randomization feature of the SMART to borrow information over the different stages of the trial to make conclusions for individual treatments. Re-randomization of non-responders in SMART is also attractive to patients as it allows patients to try a second therapy.

The scarcity of patients with vasculitis makes a placebo controlled trial of one or more drugs difficult. In oncology, Simon et al. [8] argues against independent Phase 2 trials of different drugs because differences in patient selection and other factors makes the comparison of results across trials difficult; the same argument holds for most diseases. In our form of vasculitis, a crossover trial is problematic for two reasons: i) the disease activity waxes and wanes as opposed to that seen in a stable chronic condition; and ii) patients and physicians are reluctant to have a subject crossover from a drug that is effective and commercially available. We sought to compare the efficacy of three drugs and the primary goals of the trial were to determine the drug which demonstrates the best efficacy and to compare that drug with the next best drug.

A goal of randomized early phase trials of active drugs is to ensure with high probability that the chosen treatment is truly the one with the highest underlying response rate. Simon et al. [8] show how the sample size can be determined such that with high probability, the study would correctly select the drug with the highest response rate when the difference in response rates between that drug and the remaining drugs is assumed to be d . If the response rate for an ineffective treatment is 0.25 and an effective agent of interest has a response rate of 0.50, which we assumed as the base case for our trial, then a sample size of 27 per arm

assures with 95% probability that the correct drug will be chosen. This sample size is markedly smaller than the sample size to detect a statistically significant difference between the two correctly determined drugs with response rates of 0.25 and 0.50. In order to achieve 95% power with a two-tailed alpha level of 0.05, the study would require enrollment of 95 subjects per arm.

Trial Description

In psychiatric diseases, the sequential parallel design [9] is used to augment the placebo controlled trial with a second stage in which placebo non-responders are re-randomized to drug or placebo. The purpose of the sequential parallel design is to pool the placebo versus drug information from placebo non-responders in the second stage with the placebo versus drug information from the first stage. Adopting a similar approach, we propose a trial design comparing efficacy of three medications in which non-responders of each drug are re-randomized to one of the other two drugs. The resulting design is shown in Figure 1. In Stage 1, patients are randomized equally to one of the three drugs. Patients who respond at the end of Stage 1 are continued on the drug. Patients who do not respond to drug in Stage 1 are re-randomized equally to receive one of the other two drugs. Response is again assessed at the end of Stage 2. The transition from Stage 1 to Stage 2 happens immediately for all patients and the analysis is conducted when all patients complete the trial. The definition of response is established at the design stage of the trial and documented within the protocol. At the end of the trial, we identify the drug with the best response rate in Stage 1 and then compare this drug with the best of the two remaining drugs. Re-randomizing patients in the trial increases power in the inference between the observed best treatment versus the best of the other treatments. For example, if the two best treatment groups in Stage 1 are the Drugs A and B, then we compare these two groups using both the data in Stage 1 and also the data from Stage 2 when those two groups are randomized to subjects who failed to respond to Drug C. The idea of first selecting groups and then randomizing additional subjects to those groups is the characteristic feature of two stage selection and testing designs [10]; the difference in our trial is that the two stages of the trial are being implemented simultaneously as opposed to in parallel for the traditional two stage selection and testing designs. A second difference is that the second stage of interest for the comparison of Drugs A and B consists of subjects who have first failed to respond on Drug C.

This design can easily be extended to more than 3 treatment groups. In this case, the number of Stage 2 re-randomizations increases and at least theoretically, the number of stages can also increase. As the number of groups increases, the necessary total sample size also increases which may force some limitations on the number of treatment groups for a study in a rare disease. The definition of response could also be extended to jointly consider both efficacy and toxicity.

Simulation Study of Test Statistics

If we consider stage/previous drug as a stratification factor, the analysis of binary data in treatment comparisons usually involves pooling of the inference across strata. In our analysis, the two treatment groups from Stage 1 are chosen as the two best treatment groups.

Kim [11] has shown that for normal data, under the null hypothesis when the best and next best treatment groups are from the same normal distribution, the t-statistic of these two groups has Type 1 error less than or equal to the nominal level. For binary data, a similar argument shows that the score test from Stage 1 has appropriate Type 1 error.

Mehotra and Railkar [12] investigated several weighted average methods for stratified binary data where the score test statistics from each strata are weighted. The most commonly used weights are based on the sample size harmonic mean (SSIZE) or based on estimated variance (INVAR) from the strata. The appendix gives the formulas for the INVAR and SSIZE weights. When the two treatment groups in Stage 1 are chosen based on the data, it is unclear whether these weights are still appropriate. Hence, we also investigate several fixed weights in addition to SSIZE and INVAR weights. The weights considered here are similar to those used in the analysis of a sequential parallel design [5] in that the weight is for the standardized difference for a stage and is made to optimize the power of the test statistic. They differ in the weights utilized in the SMART literature [13] which are used to estimate the effect of a particular sequence of interventions. In the SMART literature, the weights are for an individual response and are a function of the observed response rates and the randomization probabilities.

Without loss of generality, define the three groups of Stage 1 such that the true response rates, $\pi_1 \leq \pi_2 \leq \pi_3$. Let the observed response rates in Stage 1 be p_1, p_2 , and p_3 and let k and l be the indices defined by:

$$p[k] = \max p_i \quad \text{and} \quad p[l] = \max_{j \neq k} p_j.$$

The indices k and l are the identifiers of the best and next best treatment groups in terms of response rate in Stage 1. Let π_4 and π_5 be the second stage response rates for the best two drugs in Stage 1 in the subpopulation of non-responders to the third group. The joint null hypothesis that we test is $H_0: \pi_2 = \pi_3$ and $\pi_4 = \pi_5$. The difference in response rates between groups k and l in the two stages are defined as d_1 and d_2 , and the statistics we consider are:

$$Z_w = (w(d_1) + (1 - w)d_2) / \sqrt{w^2 \text{Var}_0(d_1) + (1 - w)^2 \text{Var}_0(d_2)}$$

where $0 \leq w \leq 1$ and Var_0 represents the usual null variance estimate of the difference in responses.

In order to characterize the operating characteristics of the statistics, we use the generalized power concept [10] for two stage designs. In this definition, a significant result occurs when the best treatment group is correctly identified and when the difference between this treatment group and the next best treatment group is statistically significant. A Type 1 error occurs when the best two treatment groups are in fact equal ($\pi_2 = \pi_3$ and $\pi_4 = \pi_5$) but the test statistic Z_w is in the rejection region. In our simulation study, we looked at two-tailed alpha levels of 0.05 and 0.10 and sample size per arm, n , ranging from 20 to 40. These small sample sizes are typical in trials of rare diseases.

For each set of response rates, sample size, and alpha level, Bernoulli random variables were generated for Stage 1 data; and from the observed number of non-responders in each group, Bernoulli random variables were again generated for Stage 2 data. The test statistics were calculated for the resulting dataset. All simulations assumed an equal allocation for Stage 1. A total of 10,000 realizations for each sample size were generated to estimate level and power. All of the weighted statistics had Type 1 error less than or equal to the nominal level when $\pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5$. When all response rates equaled 0.25, and the nominal Type 1 error rate was 0.05, the empirical Type 1 error rates for Z_w ranged from 0.017 - 0.028 for all w and n (Table 1). At the same response rates values, with nominal Type 1 error of 0.10, the empirical Type 1 error rates ranged from .047 - .067. Additional simulations showed that under equal allocation to the three treatment groups and when $\pi_2 = \pi_3$ and $\pi_4 = \pi_5$, the Type 1 error uniformly increases as $\pi_1 \rightarrow 0$.

Table 1 shows the estimated generalized power of Z_w using various weights for a variety of assumed response rates. For these assumed response rates, the power rapidly decreases as w decreases from 0.50. Note that $w=1$ corresponds to the power of the standard single stage design. We also examined the Wald statistic from a logistic regression with treatment and strata as independent factors; the resulting estimated powers were lower than the other test statistics. The simulations were conducted in SAS 9.4 in a Linux SAS Grid and a copy of the code is included as Supplemental Information. As indicated in Thall, et al. (10), the generalized power is minimized when $\pi_1 = \pi_2$ so the majority of scenarios assumed this. We also examined situations where the response rates in Stage 2 were lower than in Stage 1 since it is possible that a large percentage of patients fail to respond to any treatment and also a moderate percentage of patients respond to any treatment. Table 1 indicates that for $n=30$ and for $\alpha = .05$, the power generally ranges from 0.30 to 0.60 and for $\alpha = .10$, the power ranges from 0.45 to 0.75. When the response rates are 0.25 for an ineffective treatment and 0.50 for an effective treatment, the power increases from 0.56 to 0.68 as the Type 1 error increases from 0.05 to 0.10. In this non-registration trial, we felt that the trade-off of an increase of 0.05 for Type 1 error for an increase of 0.12 for power was acceptable and a priori chose a 0.10 Type 1 error.

In planning a study, one must first make assumptions regarding the true underlying response rates in Stage 1 and Stage 2. For a specified total sample size, one can then use these values and the expectation of the sample sizes in Stage 2 to calculate approximate weights from either the SSIZE or INVAR formulas. The power of Z_w in the neighborhood around these approximate weights can be estimated using the SAS program in the Supplemental Information to determine an appropriate weight for a study. The weight needs to be pre-specified prior to the start of the study. The optimal weights from Table 1 are lower than the weights for SSIZE or INVAR using this approach which suggests that the optimal weight slightly down weight Stage 1 data. For any rare disease, the sample size has to consider both the power and enrollment feasibility.

Discussion

The proposed snSMART design expands the design options for comparative effectiveness trials and increases power over single stage designs, an especially important advantage for

the study of rare diseases There are two ways that we believe the statistical analysis could be further enhanced. When the worst treatment group has response rate greater than zero, the weighted Z statistics are conservative in terms of Type 1 error. If the null variance of d_1 could be more accurately estimated, there should be an increase in power. A second improvement would be to incorporate the information from all patients in Stage 2. If the two best treatment groups in Stage 1 were Drug A and Drug B, one might be able to use the information from the re-randomization of failures to those two group. Using information from the A-C and B-C comparison to draw inference for A-B comparison is called an indirect comparison [14]. The difficulty in formulating the indirect comparison is because the patient populations for the A-C comparison (B failures) and the B-C comparison (A failures) are potentially different.

In some rare diseases, there may be a standard therapy which could serve as a control group. In these situations, the primary objective would be to find the therapy which is the best over the control group. In this situation, the design would likely be different as the control group would be included in all of the Stage 2 re-randomizations, even for subjects who initially do not respond to the control group. A similar weighted Z statistic could be constructed from all strata which directly compare the best arm versus control arm; however, the variance of the statistic would be more complicated because some subjects would be included in more than one stratum.

Although the sample size in an snSMART will be small; there may be situations in which the sequence of treatments is also of interest. As examples, it may happen that the response rate of a drug in the second stage of an snSMART is much larger than the response rate of the same drug in the initial stage, or perhaps the response rate of all drugs is larger than expected. The appropriate role of estimation of the effectiveness of sequences of treatments for an snSMART is an open and interesting research question.

The first trial using this design will be for a trial in patients sharing a rare form of vasculitis. We believe that the snSMART is an effective and practical design strategy in the rare disease setting and hope that other researchers will find the design useful.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding:

The Vasculitis Clinical Research Consortium has received support from the NIAMS (U54AR057319), the National Center for Research Resources (U54RR019497), the Office of Rare Disease Research and the national Center for Advancing Translational Science. The Health Informatics Institute has received support from the National Center for Advancing Translational Sciences (9U01TR001263).

Appendix

Using the same notation as the manuscript, define n_k and n_l as the allocated sample sizes for the two best treatment groups for Stage 1 and define n_4 and n_5 as the corresponding sample sizes for these groups in Stage 2. Then, SSIZE weight is defined by:

$$w(SSIZE) = (n_k^{-1} + n_l^{-1})^{-1} / \left[(n_k^{-1} + n_l^{-1})^{-1} + (n_4^{-1} + n_5^{-1})^{-1} \right].$$

Similarly, if $\text{Var}_0(d_1)$ and $\text{Var}_0(d_2)$ are the estimated pooled variances of the treatment differences in Stage 1 and Stage 2 respectively for the two best treatment groups, then INVAR weight is defined by:

$$w(INVAR) = (\text{Var}_0(d_1))^{-1} / \left[(\text{Var}_0(d_1))^{-1} + (\text{Var}_0(d_2))^{-1} \right].$$

References

1. Griggs RC, Batshaw M, Dunkle M, Gopal-Srivastava R, Kaye E, Krischer J, Nguyen T, Paulus K, Merkel PA. Clinical research for rare disease: opportunities, challenges, and solutions. *Molecular Genetics and Metabolism*. 2009; 96:20–26. [PubMed: 19013090]
2. iRDirC-EMA. Preparatory document for joint workshop on small population clinical trials challenges in the field of rare disease. European Medicines Agency. Aug 19.2015
3. Murphy SA. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*. 2005; 24:1455–1481. [PubMed: 15586395]
4. Lavori PW, Dawson R. Introduction to dynamic treatment strategies and sequential multiple assignment randomization. *Clinical Trials*. 2014; 11:393–399. [PubMed: 24784487]
5. Thall PF, Millikan RE, Sug HG. Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine*. 2000; 19:1011–1028. [PubMed: 10790677]
6. Jones, H. Reinforcement-Based Treatment for Pregnant Drug Abuses (HOME II). National Institute of Health; Bethesda, MD: <http://clinicaltrials.gov/ct2/show/NCT01177982>
7. Oslin, D. Managing alcoholism in people who do not respond to naltrexone (EXTEND). National Institute of Health; Bethesda, MD: <http://clinicaltrials.gov/ct2/show/NCT0011503>
8. Simon R, Wittes RE, Ellenberg SS. Randomized Phase II Trials. *Cancer Treatment Reports*. 1985; 12:1375–1381. [PubMed: 4075313]
9. Fava M, Evins AE, Dorer DJ, Schoenfeld DA. The problem of the placebo response in clinical trials for psychiatric disorders; culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics*. 2003; 72:115–127. [PubMed: 12707478]
10. Thall PF, Simon R, Ellenberg SS. Two-stage selection and testing for comparative clinical trials. *Biometrika*. 1988; 75:303–310.
11. Kim WC. On detecting the best treatment. *Journal of the Korean Statistical Society*. 1988; 17:82–92.
12. Mehotra DV, Railkar R. Minimum risk weights for comparing treatments in stratified binomial trials. *Statistics in Medicine*. 2000; 19:811–825. [PubMed: 10734285]
13. Nahum-Shani I, Almirall D, Fabiano GA, Yu J, Qian M, Pelham W, Gnagy B, Waxmonsky JG, Murphy SA. Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychological Methods*. 2012:457–477. [PubMed: 23025433]
14. Signorovitch JE, Wu EQ, Yu AP, Gerrits CM, Kantor E, Bao Y, Gupta SR, Mulani PM. Comparative effectiveness without head-to-head trials. *Pharmacoeconomics*. 2010; 28:935–945. [PubMed: 20831302]

Figure 1.
Diagram of proposed sequential multiple assigned randomized trial.

Table 1

Estimated Type 1 error (first set of π_i) and power (remaining sets of π_i) for different weights, two-tailed alpha levels, and n per arm. Each line of the table represents results of 10,000 independent realizations and SE of all entries are less than 0.005.

Response Rates													
Stage 1		Stage 2			Alpha	n	.60	.65	.70	.75	1.0	INVAR	SSIZE
π_1	π_2	π_3	π_4	π_5									
0.25	0.25	0.25	0.25	0.25	.05	20	.028	.024	.020	.017	.012	.022	.020
						30	.029	.025	.022	.018	.012	.023	.021
						40	.027	.024	.022	.018	.011	.021	.020
0.25	0.25	0.25	0.25	.10	20	.067	.060	.054	.050	.035	.055	.052	
					30	.063	.058	.054	.048	.039	.054	.053	
					40	.066	.061	.053	.047	.036	.055	.052	
0.25	0.25	0.50	0.50	.05	20	.374	.373	.374	.359	.203	.366	.368	
					30	.546	.556	.554	.539	.355	.549	.548	
					40	.686	.698	.700	.692	.500	.696	.698	
0.25	0.25	0.15	0.40	.10	20	.509	.515	.513	.501	.310	.508	.510	
					30	.678	.684	.687	.680	.485	.680	.683	
					40	.783	.798	.802	.797	.618	.798	.797	
0.25	0.25	0.50	0.15	.05	20	.411	.407	.401	.378	.201	.404	.396	
					30	.591	.592	.578	.558	.357	.582	.567	
					40	.734	.738	.732	.713	.498	.728	.725	
0.30	0.30	0.30	0.50	.10	20	.534	.536	.530	.512	.306	.528	.529	
					30	.715	.716	.711	.696	.476	.707	.705	
					40	.820	.826	.821	.806	.614	.820	.814	
0.30	0.30	0.50	0.30	.05	20	.235	.231	.227	.214	.108	.220	.218	
					30	.339	.340	.338	.326	.205	.333	.329	
					40	.435	.447	.447	.446	.299	.437	.440	
0.30	0.30	0.50	0.20	.10	20	.332	.336	.333	.320	.186	.325	.325	
					30	.461	.466	.467	.461	.301	.463	.465	
					40	.569	.583	.581	.573	.399	.575	.572	
0.30	0.30	0.50	0.20	0.40	.05	20	.244	.240	.235	.218	.112	.229	.226

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Response Rates					Weight (w)								
Stage 1		Stage 2			Alpha	n	.60	.65	.70	.75	1.0	INVAR	SSIZE
π_1	π_2	π_3	π_4	π_5									
						30	.360	.356	.344	.327	.205	.346	.332
						40	.486	.490	.486	.473	.304	.481	.480
					.10	20	.347	.345	.338	.326	.184	.337	.334
						30	.494	.495	.488	.477	.301	.485	.484
						40	.592	.596	.591	.579	.395	.590	.581
0.20	0.30	0.50	0.30	0.50	.05	20	.304	.306	.307	.298	.183	.303	.304
						30	.431	.442	.445	.439	.311	.444	.445
						40	.550	.565	.571	.565	.425	.567	.570
				.10		20	.416	.426	.429	.423	.283	.422	.426
						30	.552	.564	.567	.569	.418	.565	.567
						40	.668	.689	.693	.693	.500	.692	.691