Clinical Infectious Diseases

INVITED ARTICLE

IDSA Infectious Diseases Society of America | hivma hiv medicine association | OXFORD

INNOVATIONS IN DESIGN, EDUCATION AND ANALYSIS (IDEA): Victor De Gruttola and Scott R. Evans, Section Editors

Resist the Temptation of Response-Adaptive Randomization

Michael Proschan[1] **and Scott Evans**[2]

[1]Mathematical Statistician, Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Rockville, Maryland, USA, and [2]Department of Biostatistics and Bioinformatics; Director, Biostatistics Center, Milken Institute School of Public Health, George Washington University, Washington, DC, USA

Response-adaptive randomization (RAR) has recently gained popularity in clinical trials. The intent is noble: minimize the number of participants randomized to inferior treatments and increase the amount of information about better treatments. Unfortunately, RAR causes many problems, including (1) bias from temporal trends, (2) inefficiency in treatment effect estimation, (3) volatility in sample-size distributions that can cause a nontrivial proportion of trials to assign more patients to an inferior arm, (4) difficulty of validly analyzing results, and (5) the potential for selection bias and other issues inherent to being unblinded to ongoing results. The problems of RAR are most acute in the very setting for which RAR has been proposed, namely long-duration "platform" trials and infectious disease settings where temporal trends are ubiquitous. Response-adaptive randomization can eliminate the benefits that randomization, the most powerful tool in clinical trials, provides. Use of RAR is discouraged.

Keywords. response-adaptive randomization; temporal trend; platform trials; frequentist approach; Bayesian approach.

Clinical trials provide the highest level of evidence for the safety and effectiveness of new interventions. At their cornerstone is randomization, which separates clinical trials from all other forms of medical studies by ensuring the expectation of between-group balance with respect to all factors known or unknown, measured or unmeasured, except for treatment assignment. But not all randomization is created equal. For instance, imagine a clinical trial of 20 patients randomized using treatment:control allocation ratios of 9:1 for the first 10 patients and 1:9 for the second 10 patients. That would be very foolish because nearly all treatment patients are in the first half and all control patients are in the second half of the trial. Any observed difference between treatment and control could easily be explained by a temporal trend. The only way to separate the treatment effect from a temporal trend is to compute the treatment effect estimate separately in the 2 halves of the study, average them, and use a stratified variance. Given that 1 group has 9 times the sample size of the other within each half, the variance of that treatment effect estimator is proportional to 1/9 + 1/1 instead of 1/5 + 1/5 with 1:1 randomization; the variance for 9:1 randomization is approximately 2.8 times the variance for 1:1 randomization within each half. In other words, the only way to be confident that the treatment effect estimate is not biased by temporal trends is to use an extremely inefficient estimate that requires 2.8 times as many patients to maintain the same power.

Now consider a trial with the primary outcome of mortality, and suppose that the mortality rate on the standard treatment is high. For instance, mortality on the standard treatment in the trial of extracorporeal membrane oxygenation (ECMO) in infants with primary pulmonary hypertension was expected to be 80% [1]. In the PREVAIL II trial in Ebola virus disease in Liberia, Sierra Leone, and Guinea in western Africa, mortality in the control arm was close to 40% [2]. Middle East respiratory syndrome (MERS) is another infectious disease with a high mortality rate. In these settings, it is tempting to change randomization probabilities on the basis of ongoing trial results, such that future patients are more likely to receive the treatment doing better. This is called response-adaptive randomization (RAR). RAR requires the primary outcome status to be known relatively quickly after enrollment. Proponents tout RAR as being more ethical than conventional randomization because fewer patients receive the inferior treatment [3]. But RAR can produce periods during which most of the assignments are to only 1 arm. The resulting treatment effect estimate can be seriously biased by temporal trends unless one uses an inefficient treatment effect estimate analogous to that depicted in the preceding paragraph. Ironically, the increased sample size required to achieve the desired power actually results in more, not fewer, patients receiving the inferior treatment compared with conventional randomization [4]. Proponents of RAR suggest that it may be especially appealing in multiarmed trials [3], but in an authors' response, Korn and Friedlin [5] present similarly discouraging simulation results for a 4-armed trial.

Received 21 November 2019; editorial decision 21 March 2020; accepted 26 March 2020; published online March 28, 2020.

Correspondence: M. Proschan, Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Room 4C30, MSC 9820, 5601 Fishers Lane, Rockville, MD 20852 (proscham@niaid.nih.gov).

Clinical Infectious Diseases® 2020;71(11):3002–4

DOI: 10.1093/cid/ciaa334

In their simulations, the proportion of responders was a little higher with RAR, although the total sample size and number of nonresponders were larger. Another issue is that RAR results in highly variable per-arm sample sizes, resulting in a nontrivial probability of a larger sample size with the inferior arm than with the superior arm [6]. As noted in reference [6], this negative consequence of RAR has not been appreciated because only the expected sample sizes are typically reported, which obscures the large variability in actual sample sizes per arm.

As pointed out in reference [7], temporal trends seem especially likely in 2 settings: (1) trials of long duration, such as platform trials in which treatments may continually be added over many years, and (2) trials in infectious diseases such as MERS, Ebola virus, and coronavirus. Some conditions leading to changing event rates over time in trials in infectious diseases include resistance and changes in viral species; evolving standard of care, seasonal effects; and the introduction of a vaccine that may not only reduce disease incidence but lessen disease severity [7, 8]. Given the likelihood of temporal trends, RAR should be avoided in long-duration trials and trials in infectious diseases.

A BAD START

Response-adaptive randomization had an inauspicious debut in the aforementioned ECMO trial. A very unstable urn randomization scheme was used. Initially, 1 new (N) treatment ball and 1 standard (S) treatment ball were in the urn. A death on N or survival on S would cause a new S ball to be added, whereas a death on S or survival on N would cause a new N ball to be added. The first infant received N and survived, the second received S and died, and the next 10 received N and survived. Randomization stopped and the new treatment was ultimately declared superior. Although the conclusion proved correct in a subsequent large randomized trial in the United Kingdom [9], the original ECMO trial was very controversial. The lone infant assigned to S was objectively sicker than other infants, and may have died irrespective of treatment. Proponents and opponents of RAR agree that the randomization scheme employed in the ECMO trial was too volatile. Current thinking about RAR is that it should be preceded by a "burn in" period of conventional, eg, permuted block, randomization. We agree with this conclusion, but we would "burn in" for the entire trial! There are other adaptations of RAR that similarly ameliorate its disadvantages. For instance, one could make use of randomization probabilities intermediate between conventional randomization and the purest currently proposed form of RAR, which randomizes according to the posterior probability that an arm is best. Is it not preferable to eliminate these disadvantages by abandoning RAR than to lessen disadvantages by modifying RAR?

Incidentally, the fact that the lone infant receiving standard therapy was sicker than the infants treated with ECMO is not unexpected. Large between-arm imbalances in a prognostic baseline characteristic when there are temporal trends is a serious risk in trials using RAR. In I-SPY 2, a phase 2 trial comparing standard neoadjuvant chemotherapy plus neratinib with control in patients with high-risk stage II or III breast cancer, 65 of 115 neratinib patients (57%) and 22 of 78 control patients (28%) were human epidermal growth factor receptor 2-positive [10]. Imbalances in some baseline covariates is common in clinical trials, but imbalances of that size are very uncommon with more conventional randomization methods.

OTHER PROBLEMS WITH RESPONSE-ADAPTIVE RANDOMIZATION

Potential bias from temporal trends is not the only drawback of RAR. Response-adaptive randomization creates analysis problems for the so-called frequentist approach. There is a schism in the statistical community. The frequentist approach results in the conclusion that treatment is effective on the basis of how unlikely the observed or better results would be if the treatment truly had no effect. The Bayesian approach, on the other hand, quantifies prior belief about the treatment effect, and then updates that to a posterior belief after observing data. Both methods of analysis have proven useful, and we take no position that one method is universally better than the other. The Bayesian approach allows seamless analysis of results of a trial that uses RAR. The frequentist approach faces great difficulties in the setting of RAR, as illustrated in the attempts to analyze the ECMO trial (see reference [11] and its commentary). A major problem is that most analysis methods, such as *t* tests, tests of proportions, linear models, etc, treat sample sizes in different arms as fixed constants. That is ill advised in a trial using RAR because sample sizes themselves contain important information about whether treatment works. The reason so many infants in the original ECMO trial were assigned to the new treatment is that infants in that arm were doing well. Conditioning on per-arm sample sizes, as we would do for virtually any other form of randomization, conditions away evidence of a treatment benefit. Wei [12] showed that the *P* value in the original ECMO trial is approximately 0.62 or 0.051 depending on whether we condition on per-arm sample sizes or not. Use of RAR eliminates the great majority of standard analysis methods and ensures that a substantial fraction of clinical trial statisticians will not view results with the same confidence that a trial with conventional randomization would engender.

Another disadvantage of RAR is that strings of treatment assignments predominantly to 1 arm unblind clinical trial staff to results of the trial. Maintaining treatment blinding is critical to protecting the integrity of a clinical trial. An investigator who knows that the next treatment assignment is likely to be to a given arm might inadvertently introduce selection bias by vetoing potential patients until one who is perceived as ideally

suited for that intervention arrives. This could result in systematic differences between patients assigned to treatment and control. Consequently, any observed difference between arms could be caused by these systematic patient differences rather than a true effect of treatment.

CONCLUSIONS

Response-adaptive randomization has the potential to nullify many of the advantages of randomization. Disadvantages of RAR include bias in the presence of temporal trends unless one uses an inefficient treatment effect estimate that can result in more, not fewer, patients receiving the inferior treatment (less ethical); analysis issues that could make a substantial proportion of readers question whether trial results are convincing; and problems inherent in loss of blinding. Proponents of RAR point out that bias is less of a problem with less-volatile assignment probabilities and a burn-in period, and that bias can be addressed through modeling. For some people, the perceived ethical advantages of RAR outweigh the above disadvantages, although one can never be confident that all relevant variables have been measured or accounted for correctly in modeling. A properly randomized clinical trial should yield valid inferences without the need to appeal to models to correct for bias. Response-adaptive randomization has noble intent, but introduces serious problems that jeopardize the integrity of a clinical trial.

Notes

Acknowledgments. The authors thank Dean Follmann and Toshi Hamasaki for reviewing this manuscript.

Potential conflicts of interest. S. E. reports personal fees from Takeda/Millennium, Pfizer, Roche, Novartis, Achaogen, Huntington's Study Group, Analgesic, Anesthetic, and Addiction Clinical Trial Translations, Innovations, Opportunities, and Networks (ACTTION), Genentech, Amgen, GlaxoSmithKline (GSK), the American Statistical Association, the Food and Drug Administration, Osaka University, National Cerebral and Cardiovascular Center of Japan, the National Institutes of Health (NIH), the Society for Clinical Trials, Statistical Communications in Infectious Diseases (DeGruyter), AstraZeneca, Teva, Austrian Breast & Colorectal Cancer Study Group/Breast International Group and the Alliance Foundation Trials, Zeiss, Dexcom, the American Society for Microbiology, Taylor and Francis, Claret Medical, Vir, Arrevus, Five Prime, Shire, Alexion, Gilead, Spark, Clinical Trials Transformation Initiative, Nuvelution, Tracon, Deming Conference, Antimicrobial Resistance and Stewardship Conference, World Antimicrobial Congress, WAVE, Advantagene, Braeburn, Cardinal Health, Lipocine, Microbiotix, and Stryker; and grants from the National Institute of Allergy and Infectious Diseases/NIH, outside the submitted work. M. P. reports no potential conflicts. Both authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Bartlett RH, Roloff DW, Cornell RG, Andrews AF, Dillon PW, Zwischenberger JB. Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. Pediatrics **1985**; 76:479–87.
2. PREVAIL II Writing Group. A randomized, controlled trial of ZMapp for Ebola virus infection. N Engl J Med **2016**; 375:1448–56.
3. Berry DA. Adaptive clinical trials: the promise and the caution. J Clin Oncol **2011**; 29:606–9.
4. Korn EL, Friedlin B. Outcome-adaptive randomization: is it useful? J Clin Oncol **2011**; 29:771–6.
5. Korn EL, Friedlin B. Reply to Y. Yuan et al. J Clin Oncol **2011**; 29:e393.
6. Thall P, Fox P, Wathen J. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. Ann Oncol **2015**; 26:1621–8.
7. Huskins WC, Fowler VG, Evans S. Adaptive designs for clinical trials: application to healthcare epidemiology research. Clin Infect Dis **2018**; 66:1140–6.
8. Fleming TR, Ellenberg SS. Evaluating interventions for Ebola: the need for randomized trials. Clin Trials **2016**; 13:6–9.
9. UK Collaborative ECMO Trial Group. UK collaborative randomised trial of neonatal extracorporeal membrane oxygenation. Lancet **1996**; 348:75–82.
10. Park JW, Liu MC, Yee D, et al; I-SPY 2 Investigators. Adaptive randomization of neratinib in early breast cancer. N Engl J Med **2016**; 375:11–22.
11. Begg CB. On inferences from Wei's biased coin design in clinical trials. Biometrika **1990**; 77:467–73.
12. Wei LJ. Exact two-sample permutation tests based on the randomized play-the-winner rule. Biometrika **1988**; 75:603–6.