Author’s response to reviews

Title: A Method to Reduce Imbalance for Site-Level Randomized Stepped Wedge Implementation Trial Designs

Authors:

Robert Lew (Robert.Lew2@va.gov)
Christopher Miller (christopher.miller8@va.gov)
Bo Kim (bo.kim@va.gov)
Hongsheng Wu (hongshenwu@gmail.com)
Kelly Stolzmann (kelly.stolzmann@va.gov)
Mark Bauer (mark.bauer@va.gov)

Version: 1 Date: 04 Jan 2019

Author’s response to reviews:

Please see Personal Cover Letter for detailed, formatted response. We also paste the specific responses below:

Reviewer #1:

This is a much improved version of the earlier version.

1. I would advise the authors to add some discussion of the unknown tertiary of what these restricted randomizations schemes should be trying to achieve. The authors argue that this type of balance is better than a type which balances the means across intervention and control conditions. However, no proof of this is provided. Type 1 error rates and power could be one way to investigate this - in further work.

The reviewer correctly notes the unintended implication that we prefer one type of balance. We do not, but erred in our zeal to distinguish between the types of balance. The paper opens stating
that very often having too few sites can hinder mean-balance. Perfect mean-balance implies perfect sequential balance. In general, adding sequential balance has little effect on mean-balance, so we propose doing both on page 6 lines108-112 and elaborate on pages 12-13 lines 245-267. We briefly summarize the page 12-13 text:

Some users may prefer mean-balance of time-waves. Preference can be expressed as an explicit loss function, , of the form:

\[ = \text{Loss(imperfect mean-balance)} + \text{Loss(imperfect sequential balance)}. \]

The user-chosen value of the weight favors mean-balance, whereas favors sequential balance. For each choice of we can find designs assigning sites to time-waves that minimize . Designs that nearly minimize for many choices of are robust. In the bed-size example in the Background section an outlier hindered mean-balance, yet many designs had perfect sequential balance regardless of .

2. Minor point: the authors say there is no control here. There is some sort of control and it might be that I am not using the same language as the authors to describe treatment and control conditions. I still suggest that the authors should either show the summary of the characteristics across treatment conditions, or sequences if they prefer.

We regard sites receiving the intervention as cases and sites not receiving it as controls, making case-control status a site characteristic (pages 15-16 lines 331-337). In either case, time-wave is merely a potential covariate in the primary analysis. Sequential balance reduces confounding, time-trend, collinearity and related issues, all potential ill-effects of secular trend in the outcome variable. We considered showing the mean of each characteristic over time, but felt that this might weaken the emphasis on time-trend.

We cannot show characteristics varied over treatment conditions because all sites received the intervention. However, Table 3 indicates how the IMB score varied for each characteristic.
In a study without time-waves all sites would start at the same time. One would find and remove redundant covariates (such as confounders) from the primary analysis model as well as transform variables and insert interaction terms, steps that seldom affect Type I error and power. In contrast, Moulton focuses on mean-balance in case-control comparisons, cautioning against over-constraining design choices, because the few sites chosen from a large pool could misrepresent the population (defined by the pool), introduce bias, and affect Type I error and power. But with a small pool, typically, the chosen sites are most of the pool, as in the BHIP-CCM study.

We have added some text on page 6 lines 108-112 to clarify these issues.

3. The authors adopt the term 'sequential balance' and they say that they adopted this term because of my comments. However, my (implicit) suggestion was to use the term balance on sequence - which seems to be quite different to sequential balance. This leads me to another important point. I am not sure if the authors are balancing on sequence (which they call wave), and this point is still confusing to me. They say they balance on time, but the characteristics are not time varying.

Sorry to have over-interpreted your suggestion. The reviewer is correct. The root cause is that the pragmatic insertion of time-waves in the statistical model allows a time-dependent primary outcome such as hospitalization rate to show a pattern of time-trend with another continuous site characteristic. Issues such as the meaning of confounding require clarification. We address this in text on pages 5-6 lines 99-104 and on page 8 lines 158-162 to indicate the intended interpretation of these terms.

We have extended this concept to categorical characteristics. On pages 10-11 lines 215-220 we propose a formula for expressing time-trend for a categorical variable by evaluating time-trend for each category regarded as a binary variable and then aggregating the categorical scores.

4. In the balance equations T is represented as a linear effect and I wonder the implications of this?
The next paper addresses many issues; the form of the covariates, the form of the imbalance formula, weights assigned to covariates, and constrained optimization. We consider a short-term time pattern of a rise of one unit and then a fall of one unit during the study. The sequential imbalance formula then becomes deviations from the non-linear pattern, but this assumes some advance knowledge of an exact pattern. Assuming a monotone trend, then a link function (such as logit) transforms the data into a linear trend model. But non-monotone patterns require more speculative models often based on post-hoc exploratory analysis such as penalized spline models.

5. The authors give the impression it might be permissible to search for the best balance allocation. However, others have shown that choosing the best balance inflates type 1 errors (see Moulton for a discussion of this). Others in this area therefore select one allocation from a subset of those with good balance.

In a study without time-waves all sites would start at the same time. One would find and remove redundant covariates (such as confounders) from the primary analysis model as well as transform variables and insert interaction terms, steps that seldom affect Type I error and power. In contrast, Moulton focuses on mean-balance in case-control comparisons, cautioning against over-constraining design choices, because the few sites chosen from a large pool could misrepresent the population (defined by the pool), introduce bias, and affect Type I error and power. But with a small pool, typically, the chosen sites are most of the pool, as in the BHIP-CCM study.

6. The authors say there is no information on which to decide how to select those which have an acceptable level of balance and make an arbitrary decision to use the best 32 (I think). But, others have shown that 10% cut point works reasonably well in parallel trials - this might offer a starting point from which to explore the best choice for the cut point value.

We have added your cautionary note on page 13 lines 279-281.

Reviewer #2:

The paper draws attention to the difficulties associated with achieving "mean-balanced" allocation of clusters to the waves/sequences of a stepped-wedge design, and proposes a method
based on eliminating, as far as possible, any linear trends in site characteristics across the time-waves. This is a well-motivated enterprise, but I have some major concerns over its execution.

We thank the reviewer for this summary. In our zeal to show that the types of balance differ we failed to note how they complement each other. Too few sites may hinder mean-balance, but perfect mean-balance implies perfect sequential balance. A large pool of sites promises near mean-balance (as in the cited work of Silber and Rosenberg and the work of Zubizaretta pages 8-9 lines 167-173) and removes the need for sequential balance.

1. Where it can be achieved, mean-balance - such as would be achieved by strict stratification - is surely a more desirable outcome than mere sequential balance since the latter is consistent with identifiable quadratic trends in the site variables. Yet the IMB criterion gives no premium at all to a design that is exactly (or approximately) mean-balanced as opposed to one that is merely sequentially balanced. For example, for a binary characteristic (= 1 or 2), the allocations (1 1 1) (2 2 2) (1 1 1) and (1 2 1) (1 2 1) (1 2 1) both have IMB = 0, yet the latter is perfectly balanced while the former is clearly not. In an extreme case this could mean that the opportunity for a perfect stratification on the characteristic might be missed in favour of a design where all 2-clusters are pre-destined to occupy the middle wave. (This could be avoided by choosing a criterion that is sensitive to both forms of imbalance.)

We agree with the point made with the fortuitous example and made a similar point on page 5 line 99. Patterns such as (1 2 1) (1 2 1) (1 2 1) are optimal for nearly any trend pattern. In our example on page 8 lines 148-159, an outlier hinders mean-balance, but placing the outlier in the middle time-wave controls linear trend indicating the compatibility of these types of balance. With a small pool of sites the data on about 10 factors often contain a few outliers that hinder simultaneous mean-balance.

We have added text on pages 12-13 lines 245-267 dealing with simultaneously balancing both types, briefly summarized below:

Perfect mean-balance over time-waves implies perfect sequential balance, suggesting that, in general, adding sequential balance has little effect on mean-balance. Also, some users may prefer mean-balance. Preference can be expressed as an explicit loss function, , of the form:
The value \( v \) favors mean-balance, whereas \( w \) favors sequential balance. Here, \( v \) is a user-chosen weight. For each choice of \( v \) we wish to find designs assigning sites to time-waves that make small. A robust designs has a small value of \( v \) for many choices of \( v \), as in the bed-size example on page 8 lines 148-159 where an outlier hindered mean-balance, yet many designs had perfect sequential balance regardless of the choice of \( v \).

2. In discussions of balance criteria for parallel studies it is more-or-less automatic that the clusters will have equal probabilities of assignment to any arm. Departures from randomness usually arise only because of imbalance between pairs/triples etc of clusters in the allocation. This cannot be relied on in the present context. The final design will be chosen at random from a limited selection (34, in the example) in which the clusters may not appear equally often in each wave. For example, any 'unusual' cluster is more likely to occupy a middle-wave position. (In a parallel study it would be equally likely to be allocated to any arm.) It would be a good idea to display the prior assignment probabilities for each cluster so that the departure from randomness can be assessed. In general departures from randomness may affect the performance of statistical tests and compromise the face-validity normally associated with randomised studies.

We agree with the reviewer that we cannot simultaneously balance the site assignment in many ways, particularly when the pool of sites is small. The additional text on constrained optimization on pages 12-13 lines 245-267 addresses the issue of tradeoffs. This implicitly includes the choice of limiting selection to 34 allocation patterns, a point addressed in a cautionary note on page 13 lines 279-281.

A formal approach to tradeoffs is the loss function, \( L \), in the previous response, but with a small pool of sites an extended loss function covering too many issues often fails to identify any useful site allocations. The user must choose the few most crucial issues. Thanks to the reviewers, the paper now states that sequential balance can easily be added with little effect on mean-balance.

The next paper describes how one might simultaneously address many issues including the form of the covariates, the form of the imbalance formula, weights assigned to covariates, and constrained optimization. We consider a short-term time pattern of a rise of one unit and then a fall of one unit during the study. The sequential imbalance formula then becomes deviations from the non-linear pattern, but this assumes some advance knowledge of an exact pattern. Assuming a monotone trend, then a link function (such as logit) transforms the data into a linear
trend model. But non-monotone patterns require more speculative models often based on post-hoc exploratory analysis such as penalized spline models.

We believe that our broad computer-intensive approach belongs in a separate paper on the complex tradeoffs needed with a small pool of sites. That paper sets out subjectively chosen thresholds for each balance-like criterion. Sometimes this leads to having no allocations that meet all criteria, forcing perturbation of thresholds until some feasible site allocations are found. Programming in perturbations of each threshold value and running every combination would usually identify feasible allocations, thereby sparing the user the task of guessing which thresholds to adjust.

3. It is proposed (P9) to standardise all continuous variables to unit standard deviation to make them more homogeneous. Then they are automatically replaced by tertiles (incidentally rendering the standardisation redundant). In expressing a continuous variable in categorical form, all information about the ordering of the variable is lost. The proposed approach makes no distinction at all between ordered and unordered categories. For example, for 4 ordered categories

The reviewer is correct. When the study began we had no solution to these problems and chose a conservative approach that discarded information. A solution is to scale up a K-category characteristic by a factor of K/(K-1) to make it more comparable to the continuous factors with unit standard deviation. We have added this detail on page 10 lines 205-206 without a proof, but simulations and a proof in a special case support this assertion.

Minor points

a. P5 line99. It is confusing to say that time and region are completely confounded. In the ordinary sense of the word 'time' refers to the timing of an observation, and is in fact cross-classified with Region in such a stepped-wedge design. Similarly P8, line 151 and line 164.

We have emended the text on pages 5-6 lines 99-102 and page 8 lines 158-162.
Also, we have extended the concept of time-trend to categorical characteristics. On pages 10-11 lines 215-220 we propose a formula for expressing time-trend for a categorical variable by evaluating time-trend for each category regarded as a binary variable and then aggregating the categorical scores.

b. P6 Lines 104-8. The analogy with a BIB design is not immediately compelling, particularly as there are only two treatments in a SWD. What is proposed is a restricted randomisation scheme which, unlike BIB, does not ensure exact compliance with a balancing condition.

We state only that our approach ‘draws’ on BIB designs (page 6 lines 113-118). Once the factor ‘time-wave’ is added, the analytic model contains a major factor that cannot be fully crossed with the factor ‘site’; that is, the two-factor ANOVA model has an incomplete interaction term. This led the authors of the cited references to model the SWD as a BIB model. We did not go so far because time-wave is ordinal and not categorical. Fisher’s easily described BIB designs perfectly balance, whereas our choice of design emerges only after a search. Nevertheless, the design is fixed before any data are gathered. This differs from restricted randomization, which sequentially allocates patients to treatment as the study progresses. Simon’s abstract (Restricted randomization designs in clinical trials, Biometrics. 1979 Jun;35(2):503-12) states:

Abstract

Though therapeutic clinical trials are often categorized as using either "randomization" or "historical controls" as a basis for treatment evaluation, pure random assignment of treatments is rarely employed. Instead various restricted randomization designs are used. The restrictions include the balancing of treatment assignments over time and the stratification of the assignment with regard to covariates that may affect response. Restricted randomization designs for clinical trials differ from those of other experimental areas because patients arrive sequentially and a balanced design cannot be ensured. The major restricted randomization designs and arguments concerning the proper role of stratification are reviewed here. The effect of randomization restrictions on the validity of significance tests is discussed.

b. P11 line 220. The sentence seems to imply that the reverse sequence is a "redundancy". But N=1680 can be obtained only if the reverse sequence is counted separately. Suggest remove sentence beginning "The reverse sequence …"

Done. Thanks.
c. P11 line 231. Whether the SRS method is reliable will surely depend on the size of the problem?

We have corrected this on page 11 lines 279-281.

d. P13 The objection to stratification is that it won't often be feasible. The proposed method is more vulnerable to bias.

This falls into the issue of trade-offs covered in our response to item 2. If a small pool of sites hinders mean-balance and sequential bias reduces the chance that the primary model has ambiguous results, then we have to decide on the tradeoff between this and types of bias that might arise. This is why the next paper will help, not by resolving the reviewer’s issue, but by providing software to assess the feasibility of loss functions that express a variety of preferences.

f. P14 line 293. I don't understand the meaning of "bounds on the differences among the standard deviations across the time-waves". Please clarify.

We have edited the phrase out of this section now found on page 15-16 lines 331-337.

g. P14 line 295. Reference to cases and controls, (intervened/non-intervened) needs clarification. Presumably this is not the "intervention" in the study, but some other characteristic?

See pages 15-16 lines 331-337.

h. P16 line 348. This reads like a throwaway remark that doesn't refer to anything concrete in the paper. What are these long-established principles, and how are they to be applied in parallel designs beyond what is already standard methodology for achieving balance?

We deleted the sentence.