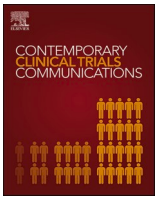
Contemporary Clinical Trials Communications 22 (2021) 100746

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc



Matching in cluster randomized trials using the Goldilocks Approach

S. Gwynn Sturdevant [a,*], Susan S. Huang [b], Richard Platt [c], Ken Kleinman [d]

[a] Laboratory for Innovation Science at Harvard, 175 N. Harvard Street, Suite 1350, Boston, MA, 02134, USA
[b] University of California, Irvine, 101 The City Drive South, City Tower, Suite 400, Mail Code: 4081, Orange, CA, 92868, USA
[c] Harvard Pilgrim Health Care Institute, 401 Park Drive, Suite 401 East, Boston, MA, 02215, USA
[d] Department of Biostatistics and Epidemiology, 715 North Pleasant Street, University of Massachusetts, Amherst, MA, 01003, USA

ARTICLE INFO

Keywords:
Matching
Randomized trials
Randomization
Baseline covariates

ABSTRACT

In group or cluster-randomized trials (GRTs), matching is a technique that can be used to improve covariate balance. When baseline data are available, we suggest a strategy that can be used to achieve the desired balance between treatment and control groups across numerous potential confounding variables. This strategy minimizes the overall within-pair Mahalanobis distance; and involves iteratively: 1) making pairs that minimize the distance between pairs of clusters with respect to potentially confounding variables; 2) visually assessing the potential effects of these pairs and resulting possible randomizations; and 3) reweighting variables of selecting weights to make pairs of clusters. In step 2, we plot the between-arm differences with a parallel-coordinates plot. Investigators can compare plots of different weighting schemes to determine the one that best suits their needs prior to the actual, final, randomization. We demonstrate application of the approach with the Mupirocin-Iodophor Swap Out trial. A webapp is provided.

1. Introduction

Individually randomized trials with blinding are the most rigorous way of determining whether a causal relation exists between an intervention and an outcome (e.g. Ref. [1]). However, for scientific and practical design reasons some interventions must be delivered to groups of subjects. Trials where groups are randomized are called group-randomized or cluster-randomized trials (GRTs). Three reasons for conducting a GRT are: (i) because implementation occurs at the cluster level, (ii) to avoid treatment contamination between subjects who are in contact with one another, and (iii) to measure intervention effects among cluster members who do not themselves receive treatment [2,3]. GRTs are "the gold standard when allocation of identifiable groups is necessary" [4].

One challenge in GRTs is that there is typically a small number of clusters. Many GRTs have fewer than 30 independent clusters to randomize, and most have fewer than 200. Thus, even though each cluster may have thousands of individuals [2], there may well be concern about confounding. In contrast, in large individually randomized trials investigators expect randomization to balance potential confounders across each arm of the trial. The smaller number of randomizable cluster in GRTs makes imbalance a threat to the causal interpretation of any observed treatment effect.

Several approaches to this problem have been proposed, including minimization [5], constrained randomization [6,7], and matching or stratification (see, e.g. Ref. [8]). Briefly, minimization can be seen as a sequential assignment of each randomized cluster to each arm such that the imbalance after the addition of that cluster is minimized. It is better suited to studies in which clusters are accrued as they are randomized. In cases where many clusters are assembled before randomization begins, it is dependent on the initial cluster and can be nearly deterministic.

Covariate constrained randomization effectively enumerates all possible treatment assignments and eliminates those that do not meet with desired features of balance. Usually schemes that have less than some maximum value of covariate difference are selected, and then one is chosen at random. For each group to have equal probability of assignment to each arm of the trial, half of the selected schemes should have it in one arm, the other half in the other. Although this is not impossible, it is unlikely. To some trialists, any deviation from an equal probability of assignment to each arm will be unacceptable; in any case it is unclear how to make principled decisions about how much inequity in arm assignment probability is allowable.

Extensive simulations compared analyses of constrained randomization, simple randomization, and the truth for both binary and

* Corresponding author.
E-mail address: nzgwynn@gmail.com (S.G. Sturdevant).

https://doi.org/10.1016/j.conctc.2021.100746
Received 3 August 2020; Received in revised form 20 November 2020; Accepted 9 February 2021
Available online 5 May 2021

continuous, normally distributed outcomes [9,10]. For continuous outcomes, they demonstrate that adjustments for covariates at the analysis stage are important even after design based adjustments. An adjusted F-test must be used, and permutation tests must account for the balanced scheme, otherwise constrained randomization improves power while maintaining type I error rates. For binary outcomes, prior knowledge should drive careful selection of covariates used in constrained randomization to maximize power and maintain type I error rates.

Other research shows that constrained randomization has smaller total sum of squares distance than simple randomization, minimization, matching, and stratification when all clusters are known in advance [11].

In stratified randomization similar clusters are grouped together prior to randomization, and randomization takes place within these smaller groups. There is debate about the optimal sizes of these groups. In particular, there is disagreement about the merits of matching, which involves grouping 2 clusters together, vs. stratification, where more than 2 clusters are grouped [12].

If there are a small number of groups in a trial, stratification is most useful when there are only a few covariates to balance. Otherwise, strata of size 4 are said to have all the advantages of matching with none of the drawbacks [13].

Many authors address the value of matching in GRTs in both the design stage and in the analysis [2,3,8,12,14–20]. Murray argues that "the choice of matching or stratification [of] factors is critical to the success of the procedure" [8]. Others suggest that caution must be used when matching a small number of clusters due to the decrease in power [2,18–20]. Breaking the matches, i.e., ignoring the matching during data analysis, addresses this [15], but perhaps only when there is a small number of large clusters [17]. Breaking the matches may also increase the type I error rate for analyses that are not the intervention effect [17]. Further drawbacks include difficulties in estimating the intracluster correlation coefficient, an inability to test for homogenity of odds ratio, and predictions that are restricted to cluster-level baseline risk factors [17]. Another complication involves removal of a cluster due to protocol violations [21].

Imai et al. develop an estimator that gives accurate standard errors when matched pairs are used; ignoring the matching gives slightly conservative standard errors [16]. However, in one trial "matching actually led to a loss in statistical efficiency" [19,22]. Despite this ongoing debate, few authors discuss how to match the clusters [7].

This article describes an extension of methods discussed previously [14]. We suggest a method suitable for *a priori* matching using baseline data. In section 2, we outline our method. In section 3, we show how it was applied in a large cluster-randomized trial, the Mupirocin-Iodophor Swap Out trial [23]. In section 4 we discuss the implications of our approach.

2. Methods

We suggest an approach to the complex topic of balancing randomization in GRTs. We match the clusters on many variables, using a "weighting" scheme to suggest which variables are most important. Then we perform many practice or "false" randomizations to obtain a distribution of the possible average arm differences that might be obtained when actual randomization occurs. Investigators assess these distributions to determine if potential randomizations would result in sufficiently balanced treatment assignments. If not, the weighting scheme is adjusted and the process begins again. The details follow. Our approach is the same of that proposed by Greevy and colleagues [24], of which we were unaware until writing this manuscript. In our approach, we facilitate weight selection through a novel visual approach for assessing the potential randomization quality for a given set of weights.

The initial step involves prioritizing variables (1, 2, …, n) from clusters (1, 2, …, m) to be randomized. We have

$$\begin{aligned} V_1 &= (v_{11},\ v_{12}, \ldots, v_{1n}) \\ V_2 &= (v_{21},\ v_{22}, \ldots, v_{2n}) \\ \vdots \quad &= \quad \vdots \\ V_m &= (v_{m1},\ v_{m2}, \ldots, v_{mn}) \end{aligned}$$

where v_{ij} is the j^{th} variable from cluster i: each V_i contains pertinent variables from cluster i. From here, we compute the Mahalanobis distance between two clusters. This is the generalized n-dimensional distance across the variables; for two clusters a and b it is calculated as $d(V_a, V_b) = \sum_{k=1}^{n} \frac{(v_{ak} - v_{bk})^2}{s_k^2}$ where $s_k^2 = \frac{1}{m}\sum_{l=1}^{m}(v_{lk} - v_{\cdot k})^2$ and $v_{\cdot k} = \frac{1}{n}\sum_{i=1}^{m} v_{ik}$.

Then we find the way of pairing the clusters that minimizes the global Mahalanobis distance across all of the possible pairs of clusters. This is a short way of describing a lengthy process: we pair cluster 1 with cluster 2 and cluster 3 with cluster 4, and so forth. Then we calculate the Mahalanobis distance between each of these pairs, and sum it. Then we pair cluster 1 with cluster 3 and cluster 2 with cluster 4, and we continue until we have the summed Mahalanobis distance for all of the possible ways to pair the clusters. The set with the minimum sum is the best way to match the clusters. This process can be done in the R statistical programming environment [25] using the `nmatch` function in the `designmatch` package [26].

Once the matching is completed, we have pairs $(C_{11}, C_{12}), (C_{21}, C_{22}), \ldots, \left(C_{\frac{m}{2}1}, C_{\frac{m}{2}2}\right)$, where C_{ij} is the jth cluster in the ith pair. The first match in each pair will be randomized to either treatment or control, the second to the other arm. If cluster C_{11} is randomized to treatment, we denote this as C_{11}^T, and this implies C_{12}^C, where the superscript indicates either treatment (T) or control (C). Next, we find the per variable difference between the two groups, averaged across the clusters in the trial:

$$d_j = \frac{\left|\sum_{i=1}^{\frac{m}{2}} C_{ij}^T - \sum_{i=1}^{\frac{m}{2}} C_{ij}^C\right|}{\frac{m}{2}}$$

for $j = 1, 2, \ldots, n$. This generates the vector $D = (d_1, \ldots, d_n)$ of the average pairwise difference between the arms for each variable. When the trial is complete, these differences are likely to be reported as evidence of the balance achieved in the randomization.

We repeat this process of randomization R times and find D_r, the vector of average differences between the two arms for the rth practice randomization. For study designs with more than 2 arms, D_r can be redefined as, for example, the standard deviation between the arms. To visualize we draw a parallel coordinates plot where the j^{th} axis plots the difference between study arms for variable j. On the plot we include D_r for all practice randomizations $r = 1, 2, \ldots, R$, as shown in the Figures below.

Upon review of the plot, we may find that the balance between the arms is unacceptable for some variables. For example, the mean or maximum distance between the arms may be too large. To accommodate this possibility, we introduce "weights" $S = (s_1, s_2, \ldots, s_n)$, which control the strength of matching on each variable. We have

$$v_{ij}^{*} = \prod_{i=1}^{m} v_{ij} s_j$$

which we combine to form

$$\begin{aligned} V_1^* &= (v_{11}^*, v_{12}^*, \ldots, v_{1n}^*) \\ V_2^* &= (v_{21}^*, v_{22}^*, \ldots, v_{2n}^*) \\ \vdots \quad &= \quad \vdots \\ V_m^* &= (v_{m1}^*, v_{m2}^*, \ldots, v_{mn}^*). \end{aligned}$$

If $s_j > s_j^*$, we are multiplying variable j by a larger value than variable

j^*, and this has the effect of increasing the distance between clusters for variable j, relative to variable j^*. Then, counter-intuitively, when we re-run the matching algorithm, we will get closer matches for variable j than variable j^*, because the Mahalanobis distance minimization will minimize this larger distance on variable j. Similarly, as the weight s_v for some variable v approaches 0, the distance between any two clusters with respect to variable v becomes very small, relative to the other variables. If $s_v = 0$, v is effectively not included in the matching at all – all clusters are perfectly matched on that variable during the matching process, and any two clusters make an equally good match on that variable. After selecting the weights S and matching on V^*, we again repeatedly find the vector of between-arm differences for each variable D_r and plot it.

The cost of a high weight for variable j in this process is that closer matches for variable j may result in reduced closeness in another variable. If so, compromises must be made. Investigators can perform iterative selections of the weights S and arrive at a set of weights S that generates a distribution of randomizations that best reflect the most desired and tolerable differences in specific characteristics between arms.

3. Results

To demonstrate the usefulness of this technique we present a brief summary of our randomization process using baseline data from the Mupirocin-Iodophor Swap Out trial (www.clinicaltrials.gov, NCT03140423) [23]. This trial follows the REDUCE MRSA trial [27] in which universal use of mupirocin nasal swabs and daily bathing with chlorhexidine was shown to markedly reduce methicillin resistant *Staphylococcus aureus* (MRSA) clinical cultures and all-cause bloodstream infection in adult intensive care units (ICU) of hospitals belonging to HCA Healthcare (HCA). One concern about the mupirocin regimen is that *S. aureus* resistance to mupirocin is relatively common in some communities and so the agent would be ineffective for many patients. Another is that routine use of mupirocin, an antibiotic, may provide selective pressure for resistant strains, thus rendering mupirocin less effective for all uses. It would thus be desirable to be able to use a substitute nasal component of the decolonizing regimen for which resistance is less likely to be present or to develop as a result of treatment. The Swap Out trial is a cluster-randomized non-inferiority trial, comparing the antibiotic mupirocin (the current standard of care) to the antiseptic iodophor for nasal decolonization of ICU patients to assess impact on *Staphylococcus aureus* clinical cultures and all-cause bloodstream infection during routine chlorhexidine bathing.

Baseline data collected from HCA's centralized data warehouse were available for matching prior to randomization. We used data from 20 months from 137 participating hospitals. Investigators prioritized 16 baseline variables into several categories. For this trial, the investigators put the highest priority on baseline values of the primary outcome measures, *Staphylococcus aureus* ICU-attributable clinical cultures per 1000 days, MRSA ICU-attributable cultures per 1000 days, and all pathogen ICU-attributable bloodstream infections per 1000 days, as well as average monthly attributable days, regional mupirocin resistance estimates, percent of ICU admissions with a prior history of MRSA, current usage of mupirocin (percent of mupirocin use in the first 5 days of ICU admission), and current usage of chlorhexidine (percent adherence to daily chlorhexidine gluconate for bathing). Of secondary importance were median ICU length of stay, and mean Elixhauser total score [28]. Of tertiary importance were the percentage of ICU Medicaid patients, and whether or not a facility uses polymerase chain reactions to identify MRSA in blood. The next group included percent of admissions involving a skilled nursing facility, and the percent of surgical admissions. The final group included whether the ICU had specialty units for oncology, bone marrow transplant, or transplant units, and if the ICU has bone marrow transplant or transplant units.

Prior to randomization, investigators used an interactive web-based application, built using the `Shiny` package in R, which implements the strategy described in section 2. The application accepted an Excel spreadsheet as input. This enabled the investigators to quickly and easily change the weights applied to each potential matching variable. The application allowed the investigators to set the desirable maximum between-arm differences for each variable as well as the relative weights. We input tolerable maximum differences between study arms as well as desirable ranges of differences for each variable and compared many sets of variable weights until we found one that was suitable.

To begin, we show a version of this process using just three of the 16 variables; the actual randomization preparation is described below. Fig. 1 demonstrates how preparation for randomization would proceed using 1) attributable patient days per month, 2) *Staphylococcus aureus* rate, and 3) MRSA rate. To read a parallel coordinates plot, trace a single gray line from "Pt Days" to "S aur rate" to "MRSA rate"; this shows the between-arm differences obtained from a single randomization. The investigators agreed that the tolerable maximum absolute mean difference between treatment and control arms for these variables were: 80 attributable patient days per month, 0.15 difference in *Staphylococcus aureus* infection rates, and 0.15 difference in MRSA rates. These define the top of our axis lines in each graph. The black line indicates the mean value of all points on each axis. We can also use this value to help decide whether the matching was acceptable. To be completely clear, this process begins in the knowledge that none of the particular practice arm assignments that resulted in these D values will be used in the actual trial: these are hypothetical randomizations that might be applied to the hospitals. In contrast, the pairs established with these weights are set by the minimizing process and are fixed.

The graph on the left is a parallel coordinates plot displaying the results of 300 randomizations when all the weights are equal, equivalent to using the raw values of each variable. The number of possible randomizations for a given matching is $2^{\frac{m}{2}}$ so more than 300 may need to be assessed for an accurate representation. The values in the plot show that several randomizations exceeded the desired maximum between-arm difference in the second and third axis: there is a reasonable chance that if randomization occurred with this weighting, the *Staphylococcus aureus* and MRSA rates would be imbalanced between the treatment and control arms. To rectify this, we should increase the weights s_r for those variables. In the center graph a weight of 8 has been applied to the *Staphylococcus aureus* rate. In this graph, the matching of hospitals is strongly adjusted so that hospitals with similar *Staphylococcus aureus* rates are paired. This results in smaller mean difference between the treatment and control arms for that variable. The values on the middle axis are all well below the desired maximum value: if randomization occurred using these strengths we are likely to get suitable balance in this variable. Unfortunately, there is a penalty. Hospitals with similar *Staphylococcus aureus* rates do not have similar attributable patient days per month and MRSA rates, which results in a few of these values exceeding the maximum tolerable difference between arms. In particular, the chance of a trial randomization with a difference in MRSA rates greater than 0.15 is too high with these weights. The right plot shows the randomizations when the matching weights for each variable were 1, 4, and 2, respectively. This plot shows all 300 randomizations comfortably below the predetermined maximum mean arm differences.

In the actual study, we used this approach with all 16 variables listed above. After trying many weights we chose a set of weights that balanced the covariates between the two arms, as seen in Fig. 2. Weights are recorded in the figure legend. For all the variables, none of these randomizations resulted in intolerable between-arm differences, and for most, the mean difference was much closer to 0 than the maximum tolerable. When it was time to assign the hospitals to their interventions, we used these weights to match hospitals in the study into pairs, then formally randomized one member of each match to treatment and the other to control. Note that some weights were 0; these variables were not used in the matching, but the figure still helps to visualize the between-

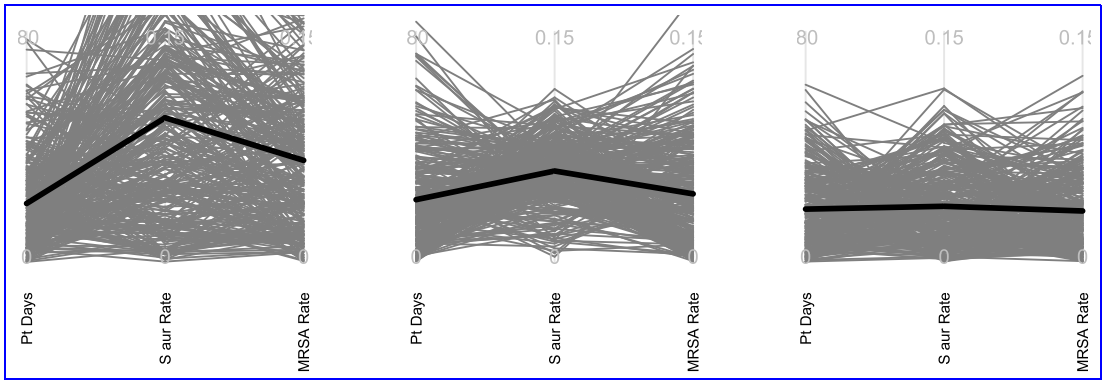


Fig. 1. Possible randomizations for 3 different sets of weights for three attributes: average monthly attributable days (Pt days), *Staphylococcus aureus* ICU-attributable cultures per 1000 days (S aur rate), MRSA ICU-attributable cultures per 1000 days (MRSA rate). Each light gray line represents a single randomization and the black line is the mean difference between arms. The left image has no weighting and two axes exceed maximum values. The center image is matched well on the middle axis, but the first and third have some randomization draws that would exceed the desired maximum values for the mean difference between the groups. The right image reaches a happy medium.

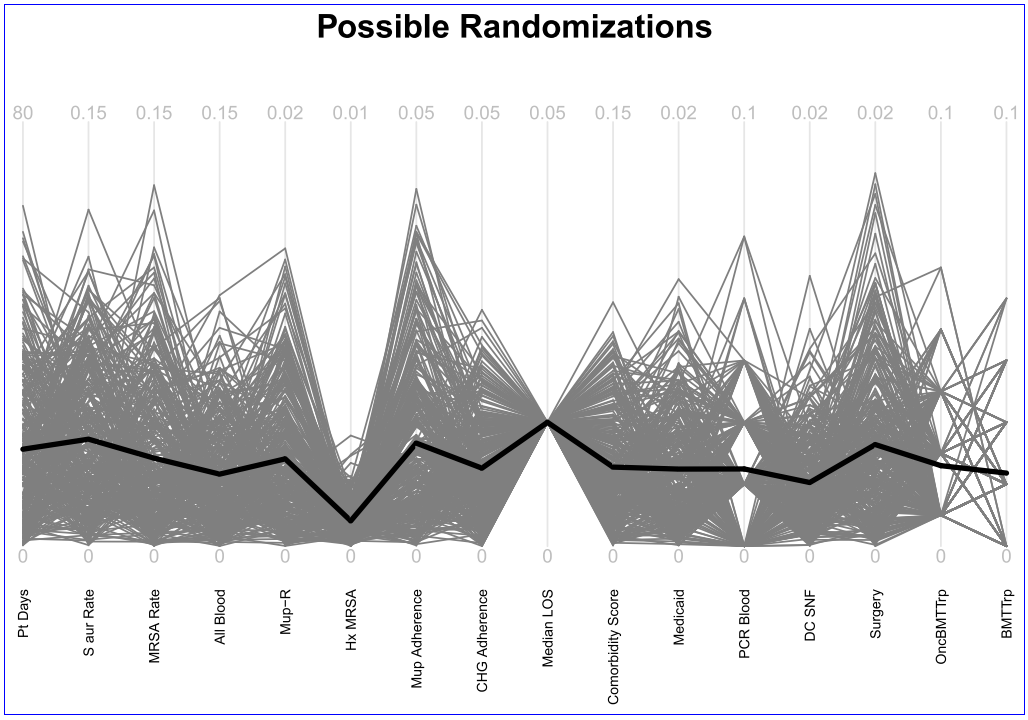


Fig. 2. Weighting scheme used in the Mupirocin-Iodophor Swap Out Trial. The variables are: patient days (Pt days, weight = 1), *Staphylococcus aureus* ICU-attributable cultures per 1000 days (S aur rate, weight = 4), MRSA ICU-attributable cultures per 1000 days (MRSA rate, weight = 2), all pathogen ICU-attributable bloodstream infections per 1000 days (All Blood, weight = 4), regional mupirocin resistance estimates (Mup R, weight = 2), percent of ICU admissions with a prior history of MRSA (Hx MRSA, weight = 1), baseline usage of mupirocin (percent of mupirocin use in the first 5 days of ICU admission (Mup Adherence, weight = 1), current usage of chlorhexidine (percent adherence to daily chlorhexidine gluconate for bathing (CHG Adherence, weight = 1), median ICU length of stay (Median LOS, weight = 3), mean Elixhauser total score (Comorbidity Score, weight = 1), percent ICU patients insured by Medicaid (Medicaid, weight = 0), whether or not a facility uses polymerase chain reactions to identify MRSA in blood (PCR Blood, weight = 0), percent admissions involving a skilled nursing facility (DC SNF), percent surgical admissions (Surgery, weight = 1), whether the ICU had specialty units for oncology, bone marrow transplant, or transplant units (OncBMTTrp, weight = 2), if the ICU has bone marrow transplant or transplant units (BMTTrp, weight = 0). Note that Median LOS has the same value for all the re-randomizations. That is, for this variable, every assignment of treatment and control within the pairs results in the same mean difference in median length of stay between the control and treatment arms. This is likely due to the very small variability of this variable. The vast majority of the hospitals had the same median length of stay.

arm differences obtained in the planning randomizations.

4. Discussion

In this article, we discuss using an iterative process to 1) make pairs that minimize the Mahalanobis distance between pairs of clusters with respect to potentially confounding variables; 2) visually assessing the potential effects these pairs and the resulting randomization; and 3) reweighting variables by selecting weights to make pairs of clusters. This process is similar to that proposed by Greevy and colleagues [24]. The main differences are i) that we use a visualization method, the parallel coordinates plot, to help investigators assess the effects of different weighting schemes and that ii) we emphasize and clarify that weighting must be an iterative and collaborative process. We also show a study where the method was applied, as opposed to a hypothetical example. In addition to the ongoing Swap Out trial shown in the Results section [23], we also used the method in a recently completed and published trial [27, 29].

For general use, we recommend deciding on tolerable maximum differences between study arms *a priori* and testing many combinations

of variable weights (S) until one is found which ensures that the eventual randomization is likely to satisfy. We call this the Goldilocks Approach, after the well-known fable, The Three Bears, in which Goldilocks tries three bowls of porridge – one is too hot, another too cold, and the third is just right [30]. More than three attempts to find a suitable combination of variable weights may be needed.

Another advantage of the Goldilocks Approach is that many covariates can be accounted for in this method, and many more explored. We also note that each cluster has equal probability of being assigned to treatment or control, something that constrained randomization forgoes.

It may bear reinforcement at this point that the many randomizations performed in the Goldilocks Approach do not constitute a search for the study randomization and treatment assignment with acceptable covariate balance. That description better suits the constrained randomization approach described previously. In contrast, the treatment assignments used in Goldilocks Approach are purely hypothetical. We should think of them as addressing the question: "If we were to match with these weights, what sort of covariate balance would we be likely to obtain in our actual randomization?" After we have found the set of weights that are just right, we formally randomize to assign the members of each matched set to a study arm. We expect a covariate balance that is similar to the ones seen in the parallel coordinates plot, but it is unlikely to be identical to any of the ones seen.

While it is often possible to obtain satisfactory balance on many covariates at the same time using the Goldilocks approach, there are limits, of course. For example, we can effectively require perfect matches on categorical variables by using large weights for them. If some categories have few members, the matches on the remaining variables are unlikely to be very close. For example if we place a large weight on suburban vs. urban hospital location, and have only 8 urban hospitals, we will be unlikely to find good matches on the other characteristics among those 8 hospitals.

The web-based application described above can be found at bit.ly/GoldilocksApp, and an instructional video explaining the use is here bit.ly/GoldilocksVid. We invite the community to use these resources, which are still under development.

While the Goldilocks approach to trial randomization cannot ensure balance between the treatment and control arms, it allows us as investigators to explore different weighting schemes. Choosing weights and assessing their likely impact means that the effects of matching and balance for relevant potential confounders can be observed and compared. Investigators who conduct GRTs and plan to match can use this method prior to randomizing to help ensure balance between treatment and control arms.

As our reviewers noted, we must also recommend caution when matching in both the design phase and analysis phase of research. Matching has consequences. It can result in reduced power and difficulties in calculating the intracluster correlation coefficient along with the multitude of faults mentioned in the introduction. Take care.

While the Goldilocks approach to trial randomization cannot ensure balance between the treatment and control arms, it allows us as investigators to explore different weighting schemes. Choosing weights and assessing their likely impact means that the effects of matching and balance for relevant potential confounders can be observed and compared. Investigators who conduct GRTs and plan to match can use this method prior to randomizing to help ensure balance between treatment and control arms.

Declaration of competing interest

All authors have no conflicts of interest to declare.

Acknowledgement

This project was funded by the National Institutes of Health Common Fund and administered by the National Institute of Allergy and Infectious Diseases (UH2/UH3 AT007769). The findings and conclusions expressed in this article are those of the authors and do not necessarily represent the official position of the National Institutes of Health or the CDC.

References

[1] B. Sibbald, M. Roland, Understanding controlled trials. Why are randomised controlled trials important? BMJ 316 (7126) (Jan 1998) 201.
[2] Laura B. Balzer, Maya L. Petersen, Mark J. van der Laan, Why Match in Individually and Cluster Randomized Trials? U.C. Berkeley Division of Biostatistics Working Paper Series, vol. 294, 2012. http://biostats.bepress.com/ucbbiostat/paper294.
[3] Richard J. Hayes, Lawrence H. Moulton, *Cluster Randomised Trials*. Chapman and HallCRC, 2009.
[4] David M. Murray, Sherri P. Varnell, Jonathan L. Blitstein, Design and analysis of group-randomized trials: a review of recent methodological developments, Am. J. Publ. Health 94 (3) (2004) 423–432.
[5] Neil W. Scott, Gladys C. McPherson, Craig R. Ramsay, Marion K. Campbell, The method of minimization for allocation to clinical trials: a review, Contr. Clin. Trials 23 (6) (2002) 662–674. ISSN 0197-2456, https://doi.org/10.1016/S0197-2456(02)00242-8, http://www.sciencedirect.com/science/article/pii/S0197245602002428.
[6] Lawrence H. Moulton, Covariate-based constrained randomization of group-randomized trials, Clin. Trials 1 (3) (2004) 297–305, https://doi.org/10.1191/1740774504cn024oa, doi: 10.1191/1740774504cn024oa. URL, PMID: 16279255.
[7] Gillian M. Raab, Izzy Butcher, Balance in cluster randomized trials, Stat. Med. 20 (3) (2001) 351–365.
[8] David M. Murray, Design and analysis of group-randomized trials. Number v. 29; v. 1998, in: Design and Analysis of Group-Randomized Trials, Oxford University Press, 1998. ISBN 9780195120363.
[9] Fan Li, Yuliya Lokhnygina, David M. Murray, Patrick J. Heagerty, Elizabeth R. DeLong, An evaluation of constrained randomization for the design and analysis of group-randomized trials, Stat. Med. 35 (10) (2016) 1565–1579, https://doi.org/10.1002/sim.6813. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6813.
[10] Fan Li, Elizabeth L. Turner, Patrick J. Heagerty, David M. Murray, William M. Vollmer, Elizabeth R. DeLong, An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes, Stat. Med. 36 (24) (2017) 3791–3806, doi: 10.1002/sim.7410. URL, https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7410.
[11] Esther de Hoop, Steven Teerenstra, G. Betsie, I. van Gaal, Mirjam Moerbeek, George F. Borm, The "best balance" allocation led to optimal balance in cluster-controlled trials, J. Clin. Epidemiol. 65 (2) (2012) 132–137, https://doi.org/10.1016/j.jclinepi.2011.05.006. ISSN 0895-4356, http://www.sciencedirect.com/science/article/pii/S0895435611001594.
[12] Elizabeth DeLong, Lingling Li, Andrea Cook, Pair-matching vs stratification in cluster-randomized trials. https://www.nihcollaboratory.org/Products/Pairing-vs-stratification_V1.0.pdf, 2017.
[13] Elizabeth Turner, Li Fan, John Gallis, Melanie Prague, David Murray, Review of recent methodological developments in group-randomized trials: Part 1—design, Am. J. Publ. Health 107 (e1–e9) (2017), https://doi.org/10.2105/AJPH.2017.303706, 04.
[14] Ken Kleinman, Cluster-randomized trials, in: Constantine Gatsonis, Sally C. Morton (Eds.), Methods in Comparative Effectiveness Research, CRC Press, 2017.
[15] Paula Diehr, Donald C. Martin, Thomas Koepsell, Cheadle Allen, Breaking the matches in a paired t-test for community interventions when the number of pairs is small, Stat. Med. 14 (13) (1995) 1491–1504.
[16] Kosuke Imai, Gary King, Clayton Nall, The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation, Stat. Sci. 24 (1) (2009) 29–53.
[17] Allan Donner, Monica Taljaard, Neil Klar, The merits of breaking the matches: a cautionary tale, Stat. Med. 26 (9) (2007) 2036–2051.
[18] Neil Klar, Allan Donner, The merits of matching in community intervention trials: a cautionary tale, Stat. Med. 16 (15) (1997) 1753–1764.
[19] Allan Donner, Neil Klar, Design and Analysis of Cluster Randomization Trials in Health Research, Wiley, 2000. ISBN 9780340691533. URL, https://books.google.com/books?id=QJZrQgAACAAJ.
[20] Donald C. Martin, Paula Diehr, Edward B. Perrin, Thomas D. Koepsell, The effect of matching on the power of randomized community intervention studies, Stat. Med. 12 (3–4) (1993) 329–338.
[21] A.V. Bartlett, S.J. Englender, B.A. Jarvis, L. Ludwig, J.F. Carlson, J.P. Topping, Controlled trial of giardia lamblia: control strategies in day care centers, 1001–6, 08, Am. J. Publ. Health 81 (1991), 0.2105/ajph.81.8.1001.
[22] Manwela N. Manun'ebo, Patricia A. Haggerty, Muladi Kalen Gaie, Ann Ashworth, Betty R. Kirkwood, Influence of demographic, socioeconomic and, J. Trop. Med. Hyg. 97 (1994) 31–38.
[23] Richard Platt, Mupirocin-iodophor icu decolonization swap out trial. https://clinicaltrials.gov/ct2/show/NCT03140423, 2017.
[24] Robert A. Greevy Jr., Carlos G. Grijalva, Christianne L. Roumie, Cole Beck, Adriana M. Hung, Harvey J. Murff, Xulei Liu, Marie R. Griffin, Reweighted mahalanobis distance matching for cluster-randomized trials with missing data, Pharmacoepidemiol. Drug Saf. 21 (S2) (2012) 148–154, https://doi.org/10.1002/pds.3260. https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.3260.

[25] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2016. https://www.R-project.org/.
[26] J.R. Zubizarreta, C. Kilcioglu, Designmatch: Construction of Optimally Matched Samples for Randomized Experiments and Observational Studies that Are Balanced and Representative by Design, 2017.
[27] Susan S. Huang, Edward Septimus, Ken Kleinman, Julia Moody, Jason Hickok, Taliser R. Avery, Julie Lankiewicz, Adrijana Gombosev, Leah Terpstra, Fallon Hartford, et al., Targeted versus universal decolo- nization to prevent icu infection, N. Engl. J. Med. 368 (24) (2013) 2255–2265.
[28] A. Elixhauser, C. Steiner, D.R. Harris, R.M. Coffey, Comorbidity measures for use with administrative data, Med. Care 36 (1) (1998) 8–27.
[29] Susan S. Huang, Edward Septimus, Ken Kleinman, Julia Moody, Jason Hickok, Lauren Heim, Adrijana Gombosev, Taliser R. Avery, Katherine Haffenreffer, Lauren Shimelman, Mary K. Hayden, Robert A. Weinstein, Caren Spencer-Smith, Rebecca E. Kaganov, Michael V. Murphy, Forehand Tyler, Julie Lankiewicz, Micaela H. Coady, Lena Portillo, Jalpa Sarup-Patel, John A. Jernigan, Jonathan B. Perlin, Richard Platt, Chlorhexidine versus routine bathing to prevent multidrug-resistant organisms and all-cause bloodstream infections in general medical and surgical units (abate infection trial): a cluster-randomised trial, Lancet 393 (10177) (2019) 1205–1215, https://doi.org/10.1016/S0140-6736(18)32593-5. ISSN 0140-6736, http://www.sciencedirect.com/science/article/pii/S0140673618325935.
[30] John Hassall, The Old Nursery Stories and Rhymes, Blackie & Son, London, 1904.