Matching in Cluster Randomized Trials Using the Goldilocks Approach

S. Gwynn Sturdevant, Susan Huang, Ken Kleinman

June 20, 2017

Abstract

Matching in cluster-randomized trials (CRTs) is important, but there is no best practice. When baseline data is available, we suggest a strategy that can be used to identify a preferred weighting of pertinent variables to achieve the desired balance between treatment and control groups across numerous potential confounding variables. This strategy involves iteratively: 1)computing the Mahalanobis distance, 2) finding the pairs that minimize the overall within-pair distance; 3) rerandomizing multiple times to generate the potential between-arm imbalance; then 4) reweighting the potential confounder. To aid in the evaluation of step 3, we plot the between-arm differences for each variable with a parallel-coordinates plot. Investigators can compare plots of different weighting scheme to determine the one that best suits their needs. We demonstrate application of the approach with the Swap-Out trial.

1 Introduction

Individually randomized trials (IRTs) with blinding are the "strongest study design available" [Gatsonis and Morton, 2017] to efficacy of a treatment or an intervention. Unfortunately, cost and study design amongst other things mean some interventions can not be randomized on an individual level.

However, for design reasons, some interventions must be delivered to groups of subjects. In these cases, it is at least cost-efficient and possibly scientifically necessary to to gather data on subjects that are correlated with one another. For example, suppose education researchers need to evaluate teacher training for elementary school teachers with respect to it's effect on literacy skills in third graders. Randomizing third graders to treatment or control conditions would not be practical– it would imply training many, many teachers, and ignoring the majority of students as well as the variability of student within class. In addition, some rural schools may have only one teacher, further complicating matters. Instead, researchers would randomize teachers, schools, or counties to the training or to the control condition and evaluate the effect on many students in each classroom. Trials where groups are randomized are called group-randomized or cluster-randomized trials (CRTs). Three reasons for conducting a CRTs are: (i) implementation occurs at the cluster level, (ii) to avoid contamination, and (iii) to measure intervention effects among cluster members who do not receive treatment [Balzer et al., 2012, Hayes, 2009]. CRTs are "the gold standard when allocation of identifiable groups is necessary" [Murray et al., 2004].

One challenge in CRTs is their limited sample size. Most CRTs have less than 30 independent units to randomize, though each unit may have thousands of dependent individuals [Balzer et al., 2012]. In large IRTs, investigators expect randomization to balance confounders across each arm of the trial. The reduced size of CRTs makes imbalance a threat to the causal interpretation of any observed treatment effect. Grouping similar units together, then randomizing, is one solution to this. Scholars debate the sizes of these groups, in particular, matching, which involves grouping 2 units together, vs. stratifying, where many more than 2 units are grouped[DeLong et al.]. This article discusses matching.

Many authors debate matching in CRTs [Balzer et al., 2012, Hayes, 2009, Gatsonis and Morton, 2017, Diehr et al., 1995, Murray, 1998, Imai et al., 2009, DeLong et al., Donner et al., 2007, Klar and Donner, 1997, Donner and Klar, 2000, Martin et al., 1993]. Murray argues that "the choice of matching or stratification [of] factors is critical to the success of the procedure" [Murray, 1998]. Some agree that caution must be used when matching a small number of clusters due to the decrease in power [Donner and Klar, 2000, Klar and Donner, 1997, Balzer et al., 2012, Martin et al., 1993]. Breaking matches, i.e., ignoring the matching during data analysis, addresses this [Diehr et al., 1995], but perhaps only when there are a small number of large

clusters [Donner et al., 2007]. Others argue that matching is effective in a small number of clusters as it "increases the chance of the intervention groups being well-balanced" [Donner et al., 2007]. Imai et al. argue that not matching, in small or large sample, is "equivalent to discarding a considerable fraction of one's data" [Imai et al., 2009]. However, in one trial "matching actually led to a loss in statistical efficiency" [Manun'ebo et al., 1994, Donner and Klar [2000]]. Despite all this debate few authors discuss how to match the clusters [Raab and Butcher, 2001].

Our article is an extension of methods introduced previously [Gatsonis and Morton, 2017]. We suggest a method suitable for *a priori* matching using baseline data. In section 2 we outline our method, section 3 applies it to data for the SWAPOUT trial, and section 4 is a brief discussion.

2 Methods

To approach the complex topic of balancing randomization in CRTs we suggest a new approach. We match the clusters on many variables, using a "weighting" scheme to suggest which variables are most important. Then we perform many randomizations to obtain a distribution of the possible arm assignments that might be obtained when official randomization occurs. Investigators assess these distributions to determine if potential randomizations result in sufficiently balanced treatment assignments. If not, the weighting scheme is adjusted and the process begins again. The details follow.

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

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

where v_{ij} is the i^{th} variable from unit j: each $\overline{V_j}$ contains pertinent variables from unit j. From here, we compute the Mahalanobois distance between two units. This is the generalized n-dimensional distance across the variables; for two units a and b it is calculated as $d(\overline{V_a}, \overline{V_b}) = \sum_{k=1}^{n} \frac{(v_{ak}-v_{bk})^2}{s_k^2}$ where $s_k^2 = \frac{1}{m}\sum_{l=1}^{m}(v_{lk}-\overline{v_{\cdot k}})$ and $\overline{v_{\cdot k}} = \frac{1}{n}\sum_{i=1}^{m} v_{ik}$. Then we find the pairs of clusters that minimize the global Mahalanobis distance across all of the pairs of clusters. This can be done in the R statistical programming environment (CITE!!) using the `nmatch()` function in the `designmatch` [Zubizarreta and Kilcioglu] package. If m is odd, the unmatched unit should be placed in both treatment and control group for exploratory purposes. (SGS– I Don't understand the last sentence. I think it might be better to stipulate even number of clusters here, and suggest what to to with odd number of clusters at the end of this section. –KK) Without loss of generality, we assume m is even for the remainder of this paper and note that to include an odd m either treatment or control groups will include one more set of priority variables and the addition of 1 in the denominator of d_i.

Once the matching is completed and we have pairs $(\overline{C_{11}}, \overline{C_{12}}), (\overline{C_{21}}, \overline{C_{22}}), ..., (\overline{C_{\frac{m}{2}1}}, \overline{C_{\frac{m}{2}2}})$, 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 remainder. 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:

$$d_i = \frac{|\sum_{j=1}^{\frac{m}{2}} C_{ij}^T - \sum_{j=1}^{\frac{m}{2}} C_{ij}^C|}{\frac{m}{2}}$$

for $i = 1, 2, ..., n$. This generates the vector $d^* = (d_1, \ldots, d_n)$ of the average pairwise difference between the arms for each variable. We repeat this process of randomization R times and find d_r^* the vector of average differences between the two arms for the rth re-randomization. To visualize we draw a parallel coordinates plot where the i^{th} axis plots all d_r^* for $r = 1, 2, ..., R$.

The investigators may find the distribution of possible randomizations unacceptable, for example because the mean distance between the arms is too large, or the maximum distance is too large. In that case, we introduce "weights"" $S = (s_1, s_2, ..., s_n)$, which control the strength of matching on each variable. We have

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

which we combine to form

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

If $s_a > s_b$, this has the effect of making clusters "further apart" with respect to variable a, so that the when we re-run the matching algorithm, we will get closer matches for variable a than variable b. After selecting S We again find d_r^* and plot them. The penalty in this process is that closer matches for variable a are likely to imply reduced closeness in another variable, so compromises must be made.

3 Results

To demonstrate the usefulness of this technique we present a brief summary of our randomization process using baseline data from the SWAPOUT trial (Cluster-randomized Non-inferiority Trial Comparing Mupirocin vs. Iodophor for Nasal Decolonization of ICU Patients to Assess Impact on Staphylococcus aureus Clinical Cultures and All-cause Bloodstream Infection During Routine Chlorhexidine Bathing) [Platt]. In this non-inferiority trial, the investigators are studying whether bathing with chlorhexidine gluconate and swabbing with iodophor nasal swabs are an acceptable substitute for bathing with chlorhexidine but swabbing with the antibiotic mupirocin. In the REDUCE trial [Huang et al., 2013] mupirocin nasal swabs and bathing with chlorhexidine reduced methycillin resistant Staphylococcus aureus in Hospital Corporation of America intensive care units (ICU). However, physicians are reluctant to use mupirocin, an antibiotic so broadly, so investigators are assessing "swapping"" it with iodophor, a disinfectant.

Table 1: Abbreviations of variables used to randomize

Primary	Secondary	Tertiary	Quaternary	Quinary
Pt Days	Median LOS	Medicaid	DC SNF	Onc_BMT_Trp
S aur Rate	Comorbidity Score	PCR Blood	Surgery	BMT_Trp
MRSA Rate				
All Blood				
Mup-R				
Hx MRSA				
Mup Adherence				
CHG Adherence				

Data collected from electronic medical records and electronic billing systems were available for matching prior to randomization. We used data from 20 months from 137 hospitals whose administrators had consented to participate. With this data, investigators met to prioritize baseline variables into several categories, as shown in Table 1: primary, secondary, tertiary, quaternary, quinary, and not relevant to randomization. For this trial, the investigators decided that average monthly attributable days (pt_days), Staphylococcus aureus Intensive Care Unit (ICU)-attributable cultures per 1,000 days (S aur rate), MRSA ICU-attributable cultures per 1,000 days (MRSA rate), all pathogen ICU-attributable bacteremia cultures per 1,000 days (All Blood), regional mupirocin resistance estimate (Mup-R), percent of admissions with MRSA diagnosis within a year (Hx MRSA), percent of mupirocin use admission to day 5 (Mup Adherence), and surveyed use of chlorhexidine gluconate (CHG Adherence) were all of primary importance. Of secondary importance were median ICU length of stay (Median LOS), and mean Elixhauser total score (Comorbidity Score). Of tertiary importance were the percentage of ICU medicaid patients (Medicaid), and whether or not a facility uses polymerase chain reactions to identify MRSA in blood (PCR Blood). The next group included percent admissions to skilled nursing facility (DC SNF), and the percent of admissions with Center for Disease Control and Prevention surveillance surgery (Surgery). The final group included whether the ICU had specialty units for oncology, bone marrow transplant, or transplant units (Onc_BMT_Trp), and if the ICU has bone marrow transplant or transplant units (BMT_Trp). More information on each variable is available in Appendix 1.

Prior to randomization, investigators spent time using a web app built using the `Shiny` package in `R` that implements the strategy described in section 2. This enabled the investigators to quickly and easily change the weights applied to each potential matching variable. The web app allows the investigators to set the bounds for each axis as well as the weights. We recommend deciding on tolerable maximum differences between study arms as well as desirable ranges of differences for each variables and using many combinations of strengths of matching until one is found which ensures randomization is likely to satisfy. In the well-known children's fable The Three Bears, Goldilocks tries three bowls of porridge, one is too hot, the other too cold, and the third is just right [STORIES. and HASSALL, 1904]. We recommend a similar procedure applied to strengths of matching, with perhaps more attempts.

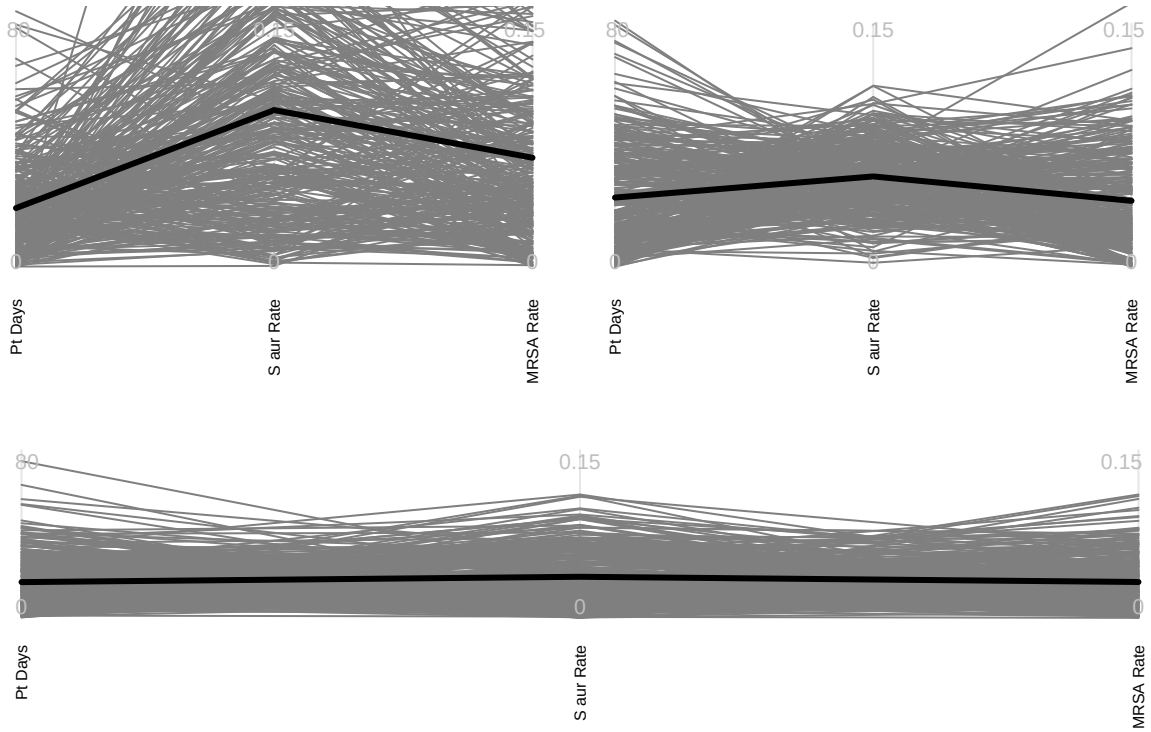
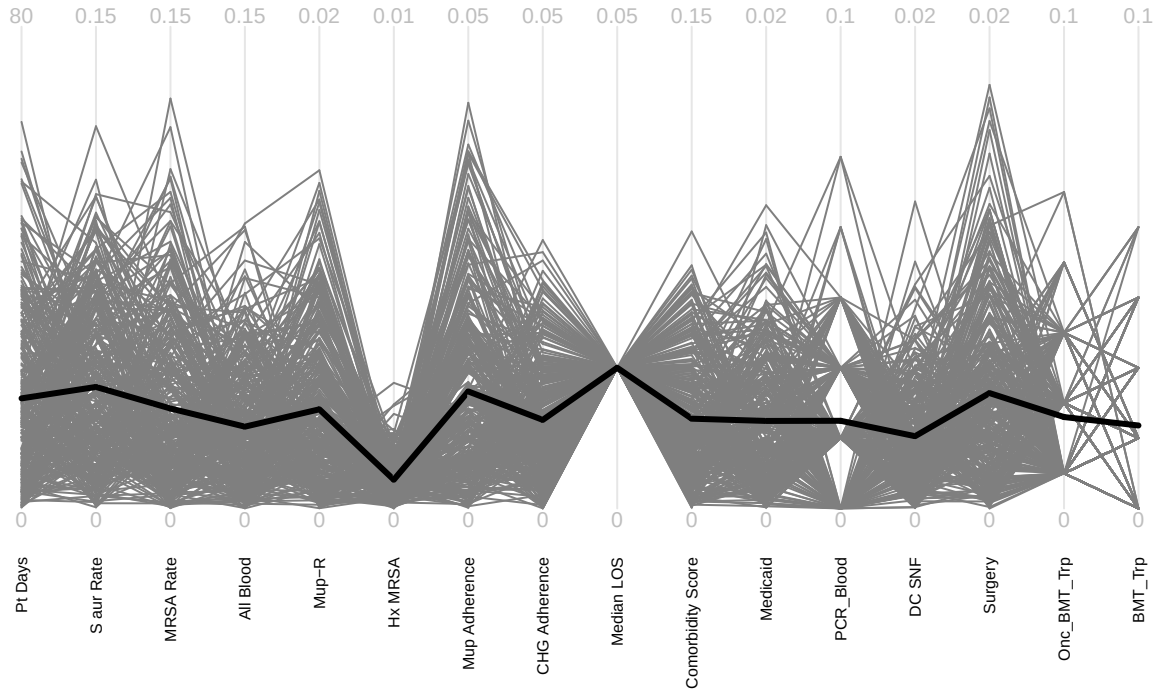


Figure 1 demonstrates this process using three variables: attributable patient days per month, Staphylococcus aureus rate, and MRSA rate. After initial explorations on the web application, investigators agreed that a tolerable maximum mean difference between treatment and control arms for these variables were: 80 attributable patient days per month, 0.15 difference in Staphylococcus aureus infection rates, and 0.15 difference in MRSA rate. The graph on the top left shows the results of 300 (SGS!!?) re-randomizations when all the weights are equal, equivalent to using the raw values of each variables. 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 values in the upper left show that several randomizations exceeded the maximums 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, positive strengths must be added. In the top-right graph a strength 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 low mean difference between the treatment and control arms in that variable. The values on the middle axis are all well below the 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. In particular, our investigators felt that the chance of attaining MRSA rates above 0.15 were too high for these strengths. The bottom plot shows the possible mean balances used in the actual randomization for these three variables, the strengths of matching for each variable were 1, 4, and 2, respectively. In all graphs, the black line indicates the mean value of all points on each axis. We also use this value to help decide whether the matchine is acceptable.

Our investigators used this approach with all 16 variables shown in table 1. After trying many weights they chose one with an agreeable balance between treatment and control arms for the variables of most importance. The results can be seen in Figure 2. For all the variables, none of the randomizations resulted in intolerable between-arm differences, and for most, the mean difference was much closer to 0 than the maximum tolerable. When the trial was randomized we used these strength to match hospitals in the study, then randomized one member of each match to treatment and the other to control.

One variable in Figure 2, median length of stay, has the same value for all the re-randomizations. That is, for this variable, ever 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 on this variable– the vast majority of the hospitals' had the same median length of stay.



4 Discussion

While the Goldilocks approach to randomizing does not ensure balance in the treatment and control arms, it is a tool that provides investigators with a method to explore strengths of matching that impact matching and balance. Investigators that use CRTs and plan to match can use this method prior to randomizing to help ensure balance between treatment and control arms.

In matching for SWAPOUT, we standardized all 16 variables by subtracting the mean from each value, then dividing by the standard deviation. To simplify notation, we reduced the algorithm as this impacts the initial plot, but the process remains the same. In SWAPOUT, the d_{ij} is derived from the nonstandardized data, the standardized data was solely used for matching.

SGS– On the whole, I think we can omit the previous paragraph with a clear conscience. If you like, you can check to make sure that omitting the standardization does not change the results.

If investigators would like to use this method on studies with more than 2 arms, d_i^* can be redefined as, for example, the standard deviation between among the arms.

We plan to publish the `Shiny` web app described above for investigators to use. This application will eventually be an interactive plot that enables users to click on each axis and view where low and high draws of that variable fall for other variables.

4.1 Appendix

Not sure this is needed!! –Ken

A more formal explanation of the variables here, in table format, to be checked with Susan.

Variable	Description
Pt Days	
S aur Rate	
MRSA Rate	
All Blood	
Mup-R	
Hx MRSA	
Mup Adherence	
CHG Adherence	
Median LOS	
Medicaid	
Comorbidity Score	
Medicaid	
PCR Blood	
DC SNF	
Surgery	
Onc_BMT_Trp	
BMT_Trp	

Table 2: Thorough description of baseline variables used in this paper.

References

Laura B Balzer, Maya L Petersen, and Mark J van der Laan. Why match in individually and cluster randomized trials? 2012.

Elizabeth DeLong, Lingling Li, and Andrea Cook. Pair-matching vs stratification in cluster-randomized trials. URL https://www.nihcollaboratory.org/Products/Pairing-vs-stratification_V1.0.pdf.

Paula Diehr, Donald C Martin, Thomas Koepsell, and Allen Cheadle. Breaking the matches in a paired t-test for community interventions when the number of pairs is small. *Statistics in medicine*, 14(13):1491–1504, 1995.

A. Donner and N. Klar. *Design and Analysis of Cluster Randomization Trials in Health Research*. Wiley, 2000. ISBN 9780340691533. URL https://books.google.com/books?id=QJZrQgAACAAJ.

Allan Donner, Monica Taljaard, and Neil Klar. The merits of breaking the matches: a cautionary tale. *Statistics in medicine*, 26(9):2036–2051, 2007.

Constantine Gatsonis and Sally C Morton. Methods in comparative effectiveness research, 2017.

Moulton Hayes. *Cluster Randomised Trials*. Chapman and HallCRC, 2009.

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 decolonization to prevent icu infection. *New England Journal of Medicine*, 368(24):2255–2265, 2013.

Kosuke Imai, Gary King, Clayton Nall, et al. The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Statistical Science*, 24 (1):29–53, 2009.

Neil Klar and Allan Donner. The merits of matching in community intervention trials: a cautionary tale. *Statistics in medicine*, 16(15):1753–1764, 1997.

Manwela N Manun'ebo, Patricia A Haggerty, Muladi Kalen Gaie, Ann Ashworth, and Betty R Kirkwood. Influence of demographic, socioeconomic and. *Journal of tropical medicine and hygiene*, 97:31–38, 1994.

Donald C Martin, Paula Diehr, Edward B Perrin, and Thomas D Koepsell. The effect of matching on the power of randomized community intervention studies. *Statistics in medicine*, 12(3-4):329–338, 1993.

David M Murray, Sherri P Varnell, and Jonathan L Blitstein. Design and analysis of group-randomized trials: a review of recent methodological developments. *American journal of public health*, 94(3):423–432, 2004.

D.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. URL https://books.google.com/books?id=cVLs3m4a9ZoC.

Richard Platt. Mupirocin-iodophor icu decolonization swap out trial. URL https://clinicaltrials.gov/ct2/show/NCT03140423?term=swap+out&rank=1.

Gillian M Raab and Izzy Butcher. Balance in cluster randomized trials. *Statistics in medicine*, 20(3):351–365, 2001.

NURSERY STORIES. and John HASSALL. *The Old Nursery Stories and Rhymes ... Illustrated by John Hassall*. Blackie & Son, London, 1904.

J. R. Zubizarreta and C. Kilcioglu. designmatch: Construction of optimally matched samples for randomized experiments and observational studies that are balanced and representative by design.