Does Balancing Site Characteristics Result in Balanced Population Characteristics in a Cluster-Randomized Controlled Trial?

Kelly Stolzmann (✉ kelly.stolzmann@va.gov)

VA Boston Healthcare System <https://orcid.org/0000-0002-3491-9803>

Robert Lew

VA Boston Healthcare System

Christopher J Miller

VA Boston Healthcare System

Bo Kim

VA Boston Healthcare System

Hongsheng Wu

VA Boston Healthcare System

Mark S Bauer

VA Boston Healthcare System

Research

Keywords: cluster randomized controlled trial, stepped-wedge design, nested control trial, collaborative care, mental health

Posted Date: March 8th, 2021

DOI: <https://doi.org/10.21203/rs.2.20639/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Health Services and Outcomes Research Methodology on February 14th, 2022. See the published version at <https://doi.org/10.1007/s10742-022-00271-1>.

Abstract

Background

Intervention trials with nested designs seek to balance sites randomized regarding key site characteristics. Among the goals of such site-level balancing is to accrue patient-level equivalence among treatment arms. We investigated patient-level equivalence in a cluster randomized controlled trial, which balanced study waves on site-level characteristics.

Methods

The Behavioral Health Interdisciplinary Program - Collaborative Chronic Care Model project utilized a stepped wedge design to stagger implementation of an evidence-based team-oriented mental health patient management system at 9 Veteran Affairs Medical Centers. Study sites were balanced on eight site-level characteristics over time (3 balanced waves [consecutive time periods] with 3 sites per wave) to minimize trend. Sites were balanced on selected site-level characteristics but not on patient-level variables. We explored internal differences in patient demographics across the three study waves. Eligible patients had at least two visits to a participating mental health clinic in the prior year and did not have a diagnosis of dementia (n=5,596).

Results

We found modest but statistically significant inter-site differences in age, marital status, ethnicity, service-related disability, mental health hospitalizations, and selected diagnoses by study wave. Although many of the differences in patient demographics by study wave were statistically significant, only a few results were practically meaningful as measured by effect size. A bipolar diagnosis (49.0%, 21.0%, 17.0% in waves 1-3, respectively; Cramer's $V=.3124$) and Hispanic ethnicity (2.9%, 29.6%, 2.0% in waves 1-3, respectively; Cramer's $V=.3949$) resulted in differences that were considered a 'moderate' effect size. The number of patient characteristics that were both statistically and meaningfully different by study wave among all possible site assignments was comparable to the 34 most balanced site assignments identified in our balancing algorithm.

Conclusions

Using a balancing algorithm to reduce imbalance among site characteristics across time periods did not appear to negatively affect the balance of patient characteristics across sites over time. A site-level balancing algorithm that includes characteristics with a direct relationship to relevant patient-level factors may improve the overall balance across key elements of the study, and aide in the interpretation of results.

Trial registration: Hybrid Collaborative Care Randomized Program Evaluation (BHIP-CCM)

NCT02543840 (September 7, 2015)

Background

Cluster randomized designs are used when individual patients are members of a larger unit, and randomization is done at the level of the unit, rather than the level of the individual patient.(1). This design is commonly used in health services or implementation research when it is necessary, or most relevant, to randomize at the region, site, or clinic level, but measure outcomes at the level of members of those units, e.g., patients. Site-level intervention trials with cluster-randomized designs hope to balance sites randomized with respect to a large number of key characteristics most pertinent to patient care.(2) A parallel, but often unstated, goal of such site-level balancing is to achieve patient-level equivalence among sites and treatment arms. Even if site characteristics are balanced there is no guarantee that this will translate into similar patient populations.

Numerous multivariate matching methods from various disciplines have been used to match patients to simultaneously balance groups on many variables.(3–7) These methods produce best matches in terms of a defined “distance”; they are in effect close but inexact matches. Rosenbaum and Rubin demonstrated that certain distances are highly effective in producing balanced clusters, even when exact matches are not used.(8) Although matching algorithms in a controlled trial design can reduce bias due to the observed variables and increase the precision of analytical adjustments, these matching methods will not likely remove all biases presented by unobserved covariates.(9) Additionally, cluster-randomized designs are unable to match at the patient level, as characteristics on study patients are not available when the study is designed.

We have demonstrated that a novel balancing algorithm can, as hypothesized, successfully reduce imbalance across clusters in terms of site-level characteristics.(7) Therefore we subsequently investigated the degree to which successful balancing at the site level resulted in similar patient populations across treatment groups. We utilized the site-balancing algorithm to ensure similar groupings of sites across three waves of a stepped wedge(2) implementation trial to establish evidence-based team care in general mental health clinics in the Department of Veterans Affairs (VA).(10) Therefore, after balancing sites across study waves, using site characteristics, we explored whether this site-level balancing resulted in balanced patient-level demographic and clinical characteristics across study waves. We are aware of no health services trial data that analyze whether site-level balancing results in patient-level equivalence among treatment arms.

Methods

The Behavioral Health Interdisciplinary Program - Collaborative Chronic Care Model (BHIP-CCM) cluster randomized controlled trial (ClinicalTrials.gov, NCT02543840) utilized a stepped wedge design (SWD) to stagger implementation of an evidence-based team care program in general mental health clinics at nine VA medical centers.(10, 11) Briefly, all sites received implementation support to establish mental health teams structured according to the evidence-based collaborative care model;(12, 13) however, sites were

randomized with regard to the timing of the receipt of implementation support, which was stepped out in three waves of three sites each.(10)

Types of balance in controlled trials

Prior to implementation, the nine BHIP VA study sites were balanced on eight site-level characteristics over the three waves, with three sites assigned to each wave. Our method (7) allocates sites to study waves so as to minimize a pre-specified imbalance score. There are different types of balance in cluster randomized controlled trials. The most familiar, mean balance, tries to make the mean values of each of many site characteristics the same across every time-wave of a SWD. This is difficult to implement with a small pool of potential sites because an extreme outlier ruins mean balance. Another type of balancing, sequential balance, tries to minimize the time trend of continuous factors over the time-waves of a SWD. For example, a study hoping to reduce infection rates should avoid a design with all small hospitals in the first wave and all large hospitals in the last wave, as hospital size would be confounded with time (i.e. did infection rates fall over time or did large hospitals do better at preventing infection than small hospitals?) Perfect mean balance implies perfect sequential balance, but the latter is easier to achieve. The BHIP-CCM study sequentially balanced sites, but the same central question of obtaining balanced patient-level characteristics would be relevant had the study mean-balanced sites.

For the initial site balancing we drew on the concepts behind the method of *fine balance* proposed by Silber, Rosenbaum, and Ross for medical outcome analyses that profile and rank hospital performance. (14) With an abundance of patient data, their method can simultaneously balance up to 60 factors. We also considered stable balancing weights proposed by Zubizarreta et al.(15) These methods complement the propensity score methods put forth by Rosenbaum and Rubin.(8) However, these methods require a large pool of potential controls. In contrast, we only had a pool of nine sites.

Balancing sites in the BHIP-CCM study

The BHIP-CCM research team, which consisted of healthcare system operational leaders and health services and implementation researchers, began with a list of over 20 site characteristics that were winnowed down to a set of 8 factors (after removing redundant or correlated characteristics) which were then used to balance sites across the three waves: urban/rural; site complexity (a 5-level index including relevant characteristics such as array of available services, inpatient care intensity, size of training programs, etc.); region; proportion of care received via telephone calls (an index of use of non-traditional treatment methods); percentage of primary care patients seen in integrated primary care / mental health teams (an index of system redesign experience); number of BHIP-CCM teams established at the site prior to the study; VA employee psychological safety score (a critical component both of team-building and of system redesign efforts); and number of patients seen in general mental health clinics in the prior year (an indicator of size and flow through the mental health programs).

Detailed information about the site balancing algorithm that was used for the current study has been published previously.(7) To measure the aggregate imbalance over many characteristics requires creating

an imbalance score with a weight and an imbalance term for each characteristic. All characteristics were categorized into tertiles. We generated $n = 1,680$ distinct site assignments from 20,000 permutations of the nine sites assigning the first three sites to wave 1, the next three sites to wave 2, and the last three sites to wave 3. After computing the imbalance score for all 1,680 assignments, we restricted to those 34 with the lowest overall imbalance scores, and randomly chose one permutation. Table 1 shows the imbalance scores among all 1,680 site assignments and the 34 site assignments with minimal scores. The primary set of factor representations used equal weights and re-expressed all characteristics as categorical variables.

Table 1

Imbalance scores among all 1,680 site assignments and to the 34 site assignments with minimal scores

Site characteristics (categorical)	All site assignments (n = 1,680) Mean (sd)	Site assignments with minimal scores (n = 34) Mean (sd)
VA Employee Psychological Safety Score	0.40 (0.42)	0.12 (0.11)
Number of BHIP teams	0.38 (0.40)	0.15 (0.13)
Number of patients seen in general mental health clinics in the prior year	0.36 (0.34)	0.09 (0.07)
Primary Care - Mental Health integration	0.40 (0.42)	0.10 (0.07)
Proportion of care via telephone	0.38 (0.40)	0.13 (0.11)
VA geographic region	0.40 (0.42)	0.09 (0.07)
Urban / Rural	0.40 (0.42)	0.18 (0.15)
Site Complexity	0.38 (0.40)	0.13 (0.11)
Overall Imbalance Score	3.10 (0.91)	0.99 (0.32)

BHIP-CCM study patient population

Having reduced imbalance among the site characteristics over time, for the current manuscript, we then explored internal differences in patient demographics, hospitalizations, and diagnoses by facilities in each study wave ($n = 3$ sites in $k = 3$ study waves). The eligible study population consisted of Veterans who had at least two visits in the prior year (with at least one visit within the past three months) to a general mental health clinic at one of nine Veteran Affairs Medical Centers (VAMCs) ($n = 5,596$).^(10, 11, 16) A subset of up to 500 individuals from each VAMC were selected to complete a telephone-administered survey battery at baseline, 6, and 12 months with a total of 1,050 Veterans who completed an interview ($n = 1,436$ total interviews). Women were oversampled to increase gender balance. Two weeks prior to calls, potential participants received mailed study information and opt-out instructions if they chose not to be interviewed.

Statistical analyses

We calculated the differences in demographic and clinical variables across our three study waves via chi-squared tests for categorical variables and analysis of variance for continuous variables. The study waves were sequentially balanced on site-level variables (as discussed above) whereas we tested for the more stringent mean balance in patient-level variables across these waves. Effect size was obtained from Cramer's V test for categorical variables and by calculating eta squared (i.e. the between-groups sum of squares divided by the total sum of squares) for continuous variables.(17, 18) For Cramer's V, an effect size of 0.1 is considered a 'small' effect size, 0.3 a 'medium' effect size, and 0.5 a 'large' effect size.(19) For eta squared, an effect size of 0.01 was considered a 'small' effect size, 0.06 represented a 'medium' effect size and 0.14 a 'large' effect size.(19) Statistical significance was assessed at the $p < .05$ and $p < .0001$ level. We also explored the number of patient-level variables with a medium or large effect size difference among all 1,680 site assignments versus the 34 site assignments with minimal scores, identified in our site balancing algorithm.(7) All data analysis was generated using SAS software, Version 9.4. All study procedures were approved by the VA Central Institutional Review Board.

Results

With a large sample size, we found modest but statistically significant inter-site differences in age, marital status, ethnicity, service-related disability, mental health hospitalizations, and selected diagnoses by study wave (i.e. with the site assignments determined by our balancing algorithm; Table 2). Mean age by waves 1–3 was, respectively: 53.2 years, 51.2 years and 52.4 years ($p < .0001$). Sites in wave 2 had a much higher percentage of Hispanic/Latino (29.6%) than study sites in wave 1 (2.9%) or wave 3 (2.0%). The percent married was similar in waves 1 (43.4%) and 3 (42.0%), but 53.9% in wave 2 ($p < .0001$). A bipolar disorder diagnosis in the year prior was 49.0% in Wave 1, 21.0% in Wave 2, and 17.0% in Wave 3 (Chi-Square $p < .0001$). There were no statistically significant differences across study waves by gender ($p = .31$). Race ($p = .0088$), substance use ($p = .0107$), and personality disorder ($p = .0038$) were all significant at the $p < .05$ level but did not reach the $p < .0001$ level (Table 2).

Table 2
Differences in patient demographics by study wave

	Wave 1 (n = 1993)	Wave 2 (n = 2060)	Wave 3 (n = 1543)	p value	Effect Size
Age (mean (± sd))	53.2 (14.3)	51.2 (14.7)	52.4 (14.3)	< .0001	0.0035
Race (%)				0.0088	0.0359
White	79.1	81.7	77.3		
Black	18.6	15.4	19.4		
Other	2.4	2.9	3.4		
Hispanic/Latino (%)	2.9	29.6	2.0	< .0001	0.3949 ^b
Female (%)	14.9	16.7	15.6	0.3074	0.0205
Service Connected	73.1	88.0	69.3	< .0001	0.1951 ^a
Marital Status (%)				< .0001	0.0863
Married	43.4	53.9	42.0		
Never Married	16.2	15.6	21.6		
Divorced/Widowed/Separated	40.4	30.6	36.4		
Mental Health Hospitalization (%)	6.1	2.0	5.4	< .0001	0.0897
Diagnoses (%)					
PTSD	48.9	59.2	38.8	< .0001	0.1628 ^a
Substance Use	21.5	17.7	19.6	0.0107	0.0403
Anxiety Disorder	33.2	27.0	35.2	< .0001	0.0750
Bipolar Disorder	49.0	21.0	17.0	< .0001	0.3124 ^b
Depression	50.3	35.2	48.7	< .0001	0.1415 ^a
Pain	51.3	44.9	39.0	< .0001	0.0981
Personality Disorder	4.2	4.9	6.6	0.0038	0.0446
Schizophrenia	6.5	4.9	10.8	< .0001	0.0941
^a small effect size; ^b medium effect size					

Although many of these differences in patient characteristics by study wave were statistically significant, only a few results were practically meaningful as measured by effect size (Table 2). The difference in

depression diagnosis across study waves (50.3%, 35.2% and 48.7% in waves 1–3, respectively) was associated with a small effect size (Cramer's $V = .1415$). A small effect size (Cramer's $V = .1620$) was also found in PTSD diagnoses across waves (48.9%, 59.2%, and 38.8% in waves 1–3, respectively) and among the percent with a service-related disability (Cramer's $V = .1951$; 73.1%, 88.0%, and 69.3% in waves 1–3, respectively). Only the difference across study waves among bipolar diagnoses (49.0%, 21.0%, 17.0% in waves 1–3, respectively; Cramer's $V = .3124$) and Hispanic ethnicity (2.9%, 29.6%, 2.0% in waves 1–3, respectively; Cramer's $V = .3949$) resulted in differences that were considered a 'moderate' effect size. Among all 1,680 site allocation schemes, the number of patient characteristics that were both statistically and meaningfully different (of a medium effect size, no large effect sizes were found) among waves was 2.7% (+/- .8%). Among the 34 site allocations that were identified as the most balanced according to our site-level algorithm(7), the number of patient characteristics that were both statistically and meaningfully different among waves was 1.2% (+/- 3.7%).

Discussion

Balancing site-level characteristics over time should improve the precision of results in a multi-site stepped wedge design intervention. Patient demographics among study sites may differ, despite a balanced design to mute potential imbalance at the site level. The results of the current study demonstrate that although allocating sites to study waves can reduce sequential imbalance among site-level characteristics, patient characteristics among study waves may still differ. We are aware of no health services trial data that analyze whether site-level balancing results in patient-level equivalence among treatment arms.

Among study sites balanced with respect to eight site-level characteristics using a sequential balancing algorithm(7), the three waves of sites differed in multiple patient demographic variables. Although statistically significant, many of the differences in patient demographics across study waves were of small to moderate effect. Patient characteristics vary greatly by region, and by site-specific specialization, making imbalance more likely. We found a higher percentage of Veterans who self-identified as Hispanic/Latino among our wave 2 study sites. Although sites overall were balanced by region two of the three study sites in wave 2 are located in Texas. Differences in diagnoses across study wave were also found, most notably among bipolar disorder. Regional variation in provider diagnostic practices and availability of site-level services may explain some of these differences across VA medical centers.(20)

There are limitations to our study. First, data were collected from multiple sites of a large, publicly-funded healthcare system, and results may differ in populations treated in other healthcare systems. Although this study was conducted within the VA, external validity limitations of VA-based studies are diminishing as health care organizations move toward integrated care models.(21–23) Second, the data on VA sites originates from regular reports required by the VA administrative and financial management, but not explicitly designed for research, and these national data might not contain some of the factors pertinent to the research question. However, the VA employs one of the largest electronic health record systems in

the nation, which is extensive in scope and captures a multitude of patient demographic and diagnostic characteristics that were included in this study.(24)

Cluster randomized controlled trials that seek to balance characteristics at the site level should also explore cross-group patient-level demographics to understand the extent to which these characteristics are balanced across study groups. In this way, the interpretation of findings can appropriately consider patient-level imbalance. Algorithms to balance site-level variables should include, to the extent available, factors that reflect relevant patient-level characteristics. For example, a study on the effect of patient employment status on health outcomes may want to include 'proportion employed' among the site-level factors to promote balance. If the number of patients in the study is very large, then covariate adjustment may adequately correct for any overall imbalance at the site level.

Conclusions

Cluster randomized trials may risk imbalance across study arms even if the total number of patients enrolled in the study is large. Using a balancing algorithm(7) to reduce imbalance among site characteristics across time periods did not appear to negatively affect the balance of patient characteristics across sites over time, further bolstering the potential utility of our site-level balancing algorithm. In the future, site-level balancing algorithms that include characteristics with a direct relationship to relevant patient-level factors may further improve the overall balance across key elements of the study, and aid in the interpretation of results.

Abbreviations

BHIP-CCM

Behavioral Health Interdisciplinary Program - Chronic Care Model

SWD

Stepped wedge design

VA

Veterans Affairs

VAMC

Veteran Affairs Medical Center

Declarations

Ethics approval and consent to participate:

The overall study was reviewed and approved by the VA Central IRB, Project Number 15-08. The study was approved for a HIPAA waiver for informed consent by the VA Central IRB.

Consent for publication:

Not applicable

Availability of data and materials:

The datasets generated and/or analyzed during the current study contained protected VA data; as such, these datasets are not publicly available; however, a limited dataset is available from the

corresponding author on reasonable request.

Competing interests:

The authors declare that they have no competing interests.

Funding:

QUERI 15-289. The funding body did not participate in the design of the study, collection,

analysis or interpretation of data, or in writing the manuscript.

Authors' contributions:

RL, CM and BK made substantial contributions to the conception and design of the work. RL and HU made substantial contributions on the acquisition, analysis, and interpretation of the data. MB oversaw all aspects of the conception and design of the study and was a major contributor in writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements:

Not applicable

References

1. Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. 1st ed. New York: Oxford University Press Inc.; 2000.
2. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. BMC Med Res Methodol. 2006;6:54.
3. Silber JH, Rosenbaum PR, Trudeau ME, Even-Shoshan O, Chen W, Zhang X, et al. Multivariate matching and bias reduction in the surgical outcomes study. Med Care. 2001;39(10):1048–64.
4. Blackford JU. Propensity scores: method for matching on multiple variables in down syndrome research. Intellect Dev Disabil. 2009;47(5):348–57.
5. Zhan C, Miller MR. Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. JAMA. 2003;290(14):1868–74.

6. Karkouti K, Beattie WS, Dattilo KM, McCluskey SA, Ghannam M, Hamdy A, et al. A propensity score case-control comparison of aprotinin and tranexamic acid in high-transfusion-risk cardiac surgery. *Transfusion*. 2006;46(3):327–38.
7. Lew RA, Miller CJ, Kim B, Wu H, Stolzmann K, Bauer MS. A method to reduce imbalance for site-level randomized stepped wedge implementation trial designs. *Implement Sci*. 2019;14(1):46.
8. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. *Biometrics*. 1985;41(1):103–16.
9. Ming K, Rosenbaum PR. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*. 2000;56(1):118–24.
10. Bauer MS, Miller C, Kim B, Lew R, Weaver K, Coldwell C, et al. Partnering with health system operations leadership to develop a controlled implementation trial. *Implement Sci*. 2016;11:22.
11. Bauer MS, Miller CJ, Kim B, Lew R, Stolzmann K, Sullivan J, et al. Effectiveness of Implementing a Collaborative Chronic Care Model for Clinician Teams on Patient Outcomes and Health Status in Mental Health: A Randomized Clinical Trial. *JAMA Netw Open*. 2019;2(3):e190230.
12. Miller CJ, Grogan-Kaylor A, Perron BE, Kilbourne AM, Woltmann E, Bauer MS. Collaborative chronic care models for mental health conditions: cumulative meta-analysis and metaregression to guide future research and implementation. *Med Care*. 2013;51(10):922–30.
13. Woltmann E, Grogan-Kaylor A, Perron B, Georges H, Kilbourne AM, Bauer MS. Comparative effectiveness of collaborative chronic care models for mental health conditions across primary, specialty, and behavioral health care settings: systematic review and meta-analysis. *Am J Psychiatry*. 2012;169(8):790–804.
14. Silber JH, Rosenbaum PR, Ross RN, Ludwig JM, Wang W, Niknam BA, et al. A hospital-specific template for benchmarking its cost and quality. *Health Serv Res*. 2014;49(5):1475–97.
15. Zubizarreta JR, Reinke CE, Kelz RR, Silber JH, Rosenbaum PR. Matching for Several Sparse Nominal Variables in a Case-Control Study of Readmission Following Surgery. *Am Stat*. 2011;65(4):229–38.
16. US Department of Veterans Affairs.: 172VA10P2: VHA Corporate Data Warehouse – VA. 79 FR 4377; Accessed August 15, 2018.
17. Effect Size: Statistics Solutions: Advancement Through Clarity; 2013 [Available from: <http://www.statisticssolutions.com/academic-solutions/resources/directory-of-statistical-analyses/effect-size/>].
18. Levine TRH. C.R. Eta squared, partial eta squared, and misreporting of effect size in communication research. *Hum Commun Res*. 2002;28:612–25.
19. Cohen J. *Statistical power and analysis for the behavioral sciences*. 2nd ed. Hillsdale: Lawrence Erlbaum Associates, Inc.; 1988.
20. Wennberg JE. Dealing with medical practice variations: a proposal for action. *Health Aff (Millwood)*. 1984;3(2):6–32.
21. Luft HS. Becoming accountable-opportunities and obstacles for ACOs. *N Engl J Med*. 2010;363(15):1389–91.

22. Kilbourne AM, Neumann MS, Pincus HA, Bauer MS, Stall R. Implementing evidence-based interventions in health care: application of the replicating effective programs framework. *Implement Sci.* 2007;2:42.
23. Shortell SM, Wu FM, Lewis VA, Colla CH, Fisher ES. A taxonomy of accountable care organizations for policy and practice. *Health Serv Res.* 2014;49(6):1883–99.
24. Fihn SD, Francis J, Clancy C, Nielson C, Nelson K, Rumsfeld J, et al. Insights from advanced analytics at the Veterans Health Administration. *Health Aff (Millwood).* 2014;33(7):1203–11.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [STROBEchecklistobservationalcohort.docx](#)