


BRIEF REPORT

Identifying appropriate comparison groups for health system interventions in the COVID-19 era

Samuel T. Savitz^{1,2}  | Jason L. Scott³ | Michael C. Leo³ | Erin M. Keast³ | Lucy A. Savitz³

¹Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, Minnesota, USA

²Division of Health Care Delivery Research, Mayo Clinic, Rochester, Minnesota, USA

³Center for Health Research, Kaiser Permanente Northwest Region, Portland, Oregon, USA

Correspondence

Samuel T. Savitz, Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, 200 1st Street SW, CSHCD—Harwick 2nd Floor, Rochester, MN 55905, USA.

Email: savitz.samuel@mayo.edu

Abstract

Introduction: COVID-19 has created additional challenges for the analysis of non-randomized interventions in health system settings. Our objective is to evaluate these challenges and identify lessons learned from the analysis of a medically tailored meals (MTM) intervention at Kaiser Permanente Northwest (KPNW) that began in April 2020.

Methods: We identified both a historical and concurrent comparison group. The historical comparison group included patients living in the same area as the MTM recipients prior to COVID-19. The concurrent comparison group included patients admitted to contracted non-KPNW hospitals or admitted to a KPNW facility and living outside the service area for the intervention but otherwise eligible. We used two alternative propensity score methods in response to the loss of sample size with exact matching to evaluate the intervention.

Results: We identified 452 patients who received the intervention, 3873 patients in the historical comparison group, and 5333 in the concurrent comparison group. We were able to mostly achieve balance on observable characteristics for the intervention and the two comparison groups.

Conclusions: Lessons learned included: (a) The use of two different comparison groups helped to triangulate results; (b) the meaning of utilization measures changed pre- and post-COVID-19; and (c) that balance on observable characteristics can be achieved, especially when the comparison groups are meaningfully larger than the intervention group. These findings may inform the design for future evaluations of interventions during COVID-19.

KEYWORDS

comparison groups, COVID-19, observational studies, program evaluation

1 | INTRODUCTION

A key challenge for the evaluation of non-randomized health system interventions is the identification of appropriate comparison groups.^{1,2}

Two common choices include historical comparison groups from the same hospital or clinic before the intervention was implemented and concurrent comparison groups receiving care at a different hospital or clinic that was not exposed to the intervention. For both choices, an

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Learning Health Systems* published by Wiley Periodicals LLC on behalf of University of Michigan.

assumption is made that outcomes would be similar in the absence of the intervention after accounting for observable characteristics.³⁻⁵

COVID-19 has greatly increased the challenge of finding appropriate comparison groups. Historical comparison groups pre-pandemic may no longer be comparable due to rapid changes in care delivery like telemedicine⁶ and disruptions in care for patients with non-communicable diseases.⁷ Concurrent comparison groups may not be comparable due to differences in local COVID-19 context including case rate and hospital capacity.^{8,9} These challenges threaten the internal validity for evaluations of interventions during COVID-19 when it is especially important for learning health systems to innovate to address the pressures COVID-19 has brought to staffing,¹⁰ clinician burnout,¹¹ mental health outcomes,^{12,13} and hospital finances.¹⁴

2 | QUESTION OF INTEREST

Our objective is to evaluate approaches for overcoming challenges in identifying a suitable comparison group during COVID-19. We focus on an intervention to provide post-discharge medically tailored meal (MTM) delivery implemented at Kaiser Permanente Northwest (KPNW) to prevent 30-d hospital readmissions and 30-d emergency room (ER) visits following discharge. We evaluate how well the identified groups serve as comparators and present lessons learned for how health systems can identify optimal comparison groups to evaluate interventions.

3 | METHODS

3.1 | Intervention

MTM is a post-discharge, in-home meal delivery program that began April 2020 and was funded internally by KP Community Health. The program was a non-randomized, embedded research study integrated into the hospital discharge process. To be eligible, patients had to be aged 18 y or older, live in one of four counties in the Portland metro area (Multnomah, Clackamas, Clark, or Washington), and have been discharged from a

KPNW hospital with at least one of the qualifying conditions. The conditions were: congestive heart failure, Type II diabetes, chronic obstructive pulmonary disease, chronic kidney disease/end-stage renal disease, and cirrhosis. Patients did not have to have low-income or food insecurity to qualify. Eligible patients were flagged on a daily report and hospital navigators offered MTM to eligible patients as part of the hospital discharge workflow. Meals were delivered to patients who opted in through a partnership with the local Meals on Wheels chapter and provided two meals a day for 4 wk tailored to specific dietary needs. The primary outcomes of the intervention are having readmission within 30 d of discharge and having an ER visit within 30 d of discharge. The readmission or ER visit had to have the referent admission diagnosis or diagnoses in any position

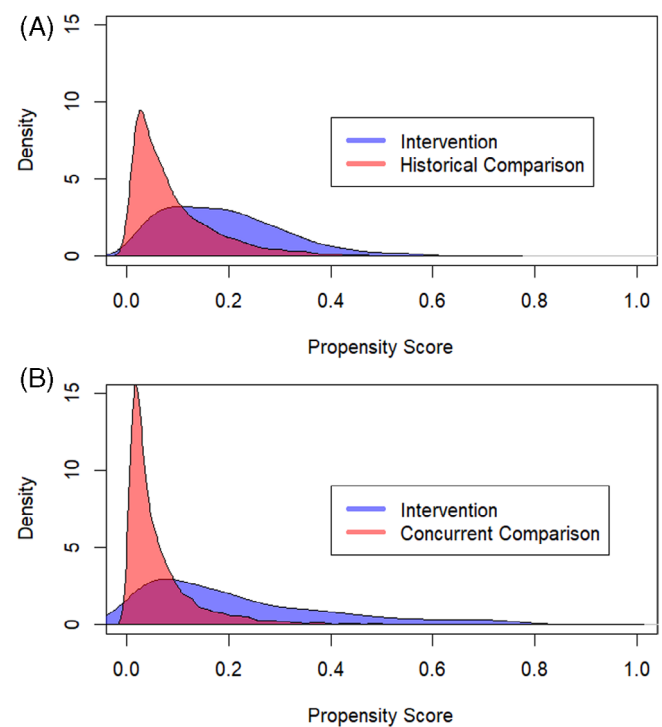


FIGURE 1 (A): Common support for intervention and historical comparison. (B): Common support for intervention and concurrent comparison

TABLE 1 Definition of the treatment and comparison groups

Group	Time period	Facility	Discharge region	Age	Admitting conditions
Intervention	Apr. 2020-Nov. 2021	KPNW Hospital	Portland Metro	≥18	<ul style="list-style-type: none"> • Congestive heart failure (CHF) • Type II Diabetes • Chronic obstructive pulmonary disease (COPD) • Chronic Kidney Disease/End-stage Renal Disease (CKD) • Cirrhosis
Historical Comparison	Apr. 2019-Mar. 2020	KPNW Hospital	Portland Metro		
Concurrent Comparison	Apr. 2020-Nov. 2021	KPNW Hospital	Outside Portland Metro		
		Non-KPNW Hospital	Any	Or ^a	

^aThe Concurrent Comparison group included Kaiser Permanente Northwest (KPNW) patients who were either: (a) discharged from a KPNW hospital to a home outside the Portland metro area; or (b) discharged from a non-KPNW hospital to home regardless of region.

to be counted. Both of these outcomes could be affected by COVID-19 given health system capacity limitations in treating non-COVID patients together with patient fear of exposure during the early months of the pandemic.^{15,16}

For the evaluation, we used the KPNW electronic health record (EHR) and identified a historical comparison group of patients before COVID-19 who would have been eligible and a concurrent comparison group of KPNW patients that were admitted to contracted non-KPNW hospitals or admitted to a KP facility and living outside of the MTM service area but otherwise eligible (see Table 1 for detailed inclusion criteria).

3.2 | Analysis

We evaluated the balance between the intervention and comparison groups using descriptive statistics and standardized differences.¹⁷ We then used three alternative approaches to account for potential observable confounders of the relationship between the intervention and the primary utilization outcomes. First, we performed exact matching on admitting condition and insurance and used 1:1 propensity score matching on the remaining characteristics. These characteristics were: age, sex, area-deprivation index (ADI),¹⁸ Charlson Comorbidity Index,

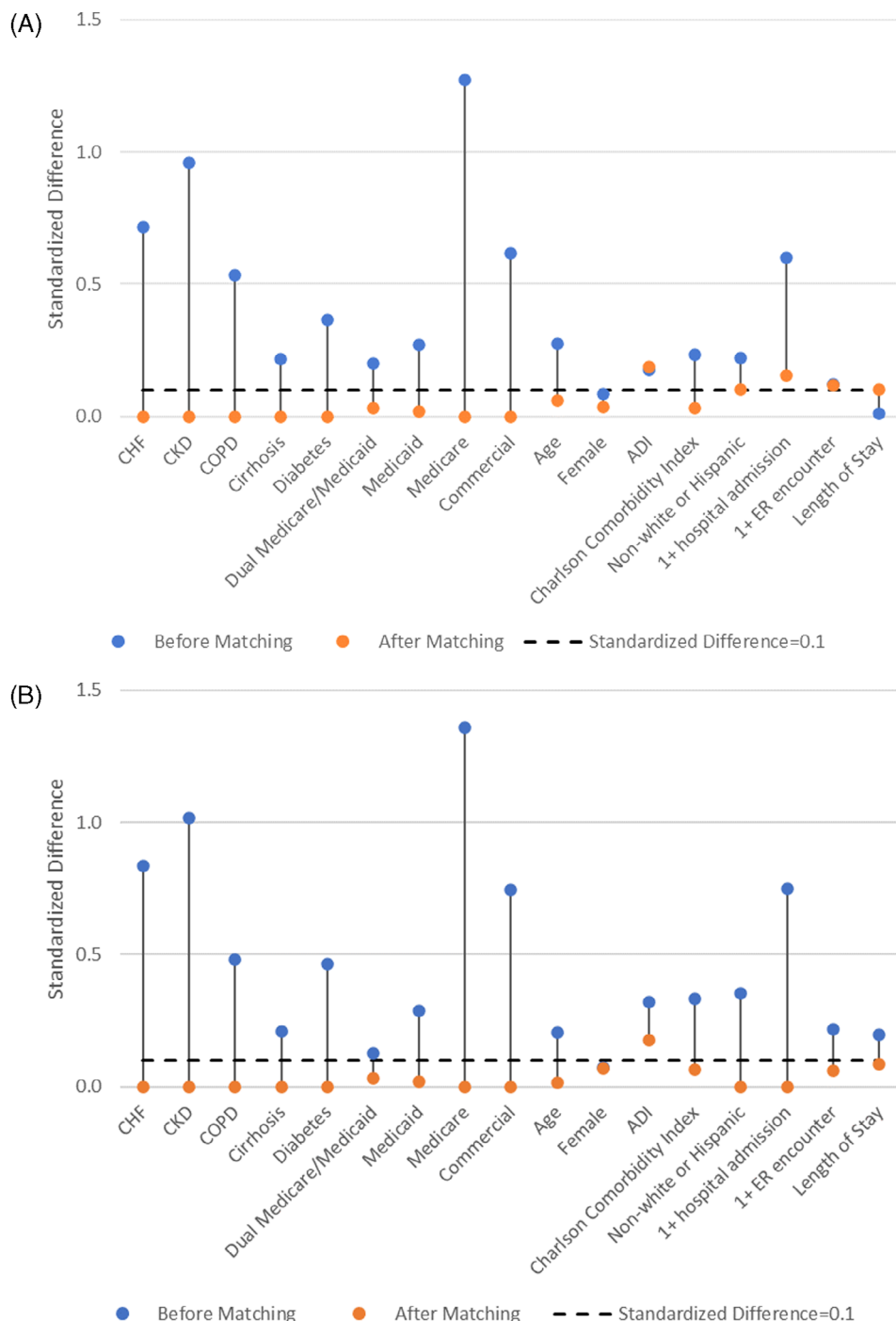


FIGURE 2 (A): Standardized differences after matching for historical comparison group. (B): Standardized differences after matching for concurrent comparison group. CHF, stands for congestive heart failure; CKD, stands for chronic kidney disease; COPD, stands for Chronic obstructive pulmonary disease; ER, stands for emergency room

non-white or Hispanic, one or more hospital admission in the prior 12 mo, one or more ER encounter in the prior 12 mo, and length of stay (LOS).¹⁹ We used optimal matching without replacement and selected observations within calipers of 0.4 SD.

We did not include local COVID-19 infection rates because all patients came from either the Portland Metro area or the surrounding counties and the infection rates were highly correlated. However, local infection rates may be more relevant for analyses that compare patients across different geographic areas. We were unable to control for patient COVID-19 history because there were so few tests available in the early COVID-19 period and many tests were performed at-home and not noted in the EHR.

Second, we performed inverse probability of treatment weighting (IPTW) for the concurrent comparison group using the same characteristics as well as social/financial needs and food security. We kept observations that were in the region of common support.²⁰ We estimated the propensity score using logistic regression for each of the comparison groups. All analyses were performed in SAS 9.4 (SAS Institute, Cary NC).

4 | RESULTS

We observed moderate to large differences between the intervention (N = 452) and the historical (N = 3873) and concurrent (N = 5333)

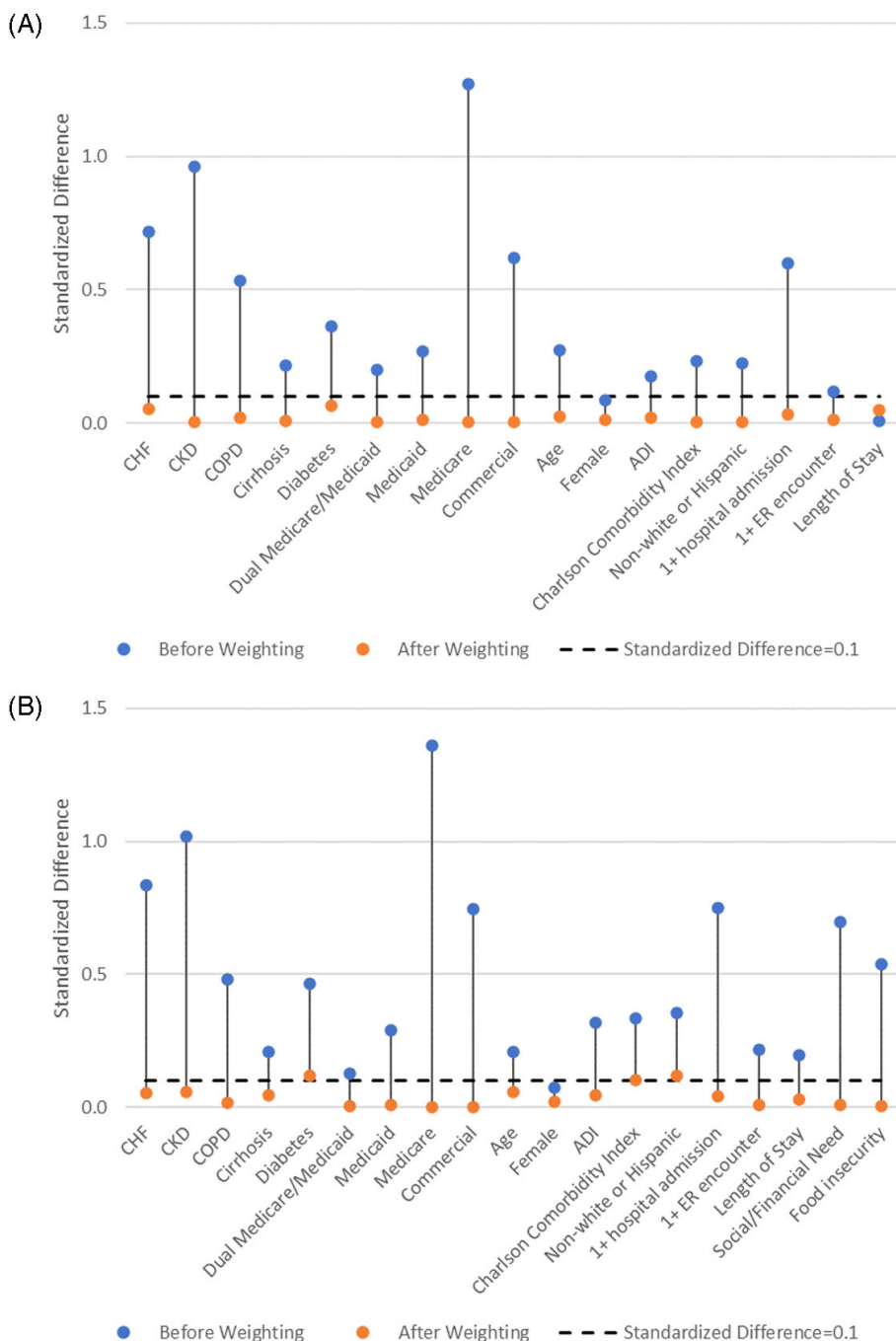


FIGURE 3 (A): Standardized differences after weighting for historical comparison group. (B): Standardized differences after weighting for concurrent comparison group. CHF, stands for congestive heart failure; CKD, stands for chronic kidney disease; COPD, stands for Chronic obstructive pulmonary disease; ER, stands for emergency room

TABLE 2 Descriptive statistics for intervention and comparison cohorts

Variable	Level	Intervention (N = 452) % (N) or mean (SD)	Historical comparison (N = 3873)		Concurrent comparison (N = 5333)	
			% (N) or mean (SD)	Std. Diff.	% (N) or mean (SD)	Std. Diff.
MTM condition (mutually exclusive)	CHF	45.8% (207)	33.4% (1294)	0.72	35.3% (1884)	0.83
	CKD	14.8% (67)	36.2% (1401)	0.96	36.2% (1933)	1.02
	COPD	8.0% (36)	13.8% (533)	0.53	10.7% (573)	0.48
	Cirrhosis	4.0% (18)	3.3% (129)	0.22	2.7% (146)	0.21
	Diabetes	27.4% (124)	13.3% (516)	0.37	14.9% (797)	0.47
MTM condition (not mutually exclusive)	CHF	44.2% (200)	35.1% (1359)	0.19	35.9% (1917)	0.17
	CKD	36.7% (166)	55.4% (2147)	0.38	55.6% (2967)	0.38
	COPD	15.9% (72)	27.1% (1050)	0.26	24.0% (1282)	0.19
	Cirrhosis	5.3% (24)	6.4% (248)	0.05	5.6% (299)	0.01
	Diabetes	37.8% (171)	22.7% (881)	0.35	28.4% (1512)	0.21
Insurance	Dual Medicare/Medicaid	4.2% (19)	3.1% (119)	0.20	1.6% (83)	0.13
	Medicaid	11.9% (54)	6.7% (260)	0.27	6.2% (330)	0.29
	Medicare	58.4% (264)	67.0% (2595)	1.27	66.0% (3522)	1.36
	Commercial	25.4% (115)	23.2% (899)	0.62	26.2% (1398)	0.74
Age	Mean (SD)	64.3 (15.6)	68.4 (14.7)	0.28	67.5 (15.3)	0.21
Sex	Female	45.1% (204)	49.5% (1917)	0.09	48.8% (2605)	0.07
ADI	Mean (SD)	5.3 (2.7)	4.8 (2.6)	0.18	6.1 (2.5)	0.32
ADI	Low Disadvantage	29.4% (133)	37.1% (1437)	0.88	17.6% (938)	0.52
	Moderate Disadvantage	35.2% (159)	33.3% (1290)	0.77	33.9% (1809)	0.85
	High Disadvantage	34.3% (155)	28.1% (1088)	0.67	43.9% (2343)	1.05
	Unknown	1.1% (5)	1.5% (58)	0.16	4.6% (243)	0.35
Charlson	Mean (SD)	5.3 (2.8)	4.6 (2.7)	0.23	4.4 (2.7)	0.33
Charlson Comorbidity Index	0	1.5% (7)	1.8% (68)	0.17	2.4% (128)	0.23
	1–2	16.4% (74)	23.1% (894)	0.68	25.2% (1344)	0.78
	3–5	36.7% (166)	40.9% (1585)	0.91	41.6% (2219)	1.00
	6+	45.4% (205)	34.2% (1326)	0.73	30.8% (1642)	0.74
Non-white or Hispanic		24.8% (112)	16.4% (634)	0.22	12.7% (676)	0.35
1+ hospital admission ^a		57.3% (259)	29.7% (1150)	0.60	24.6% (1310)	0.75
1+ ER encounter ^a		74.8% (338)	79.7% (3086)	0.12	64.5% (3438)	0.22
LOS	Mean (SD)	3.5 (2.6)	3.4 (4.2)	0.01	4.5 (5.6)	0.20
Social/Financial Need ^b		38.9% (176)	NA	NA	13.9% (741)	0.70
Food Insecurity ^b		13.7% (62)	NA	NA	3.2% (171)	0.54

Abbreviations: ADI, stands for Area Deprivation Index; CHF, stands for congestive heart failure; CKD, stands for chronic kidney disease; COPD, stands for Chronic obstructive pulmonary disease; ER, stands for emergency room; NA, stands for not applicable; SD, stands for SD; Std. Diff. stands for Standardized Difference.

^aThe number of admissions or ER encounters in 12 mo prior to index admission.

^bThe social/financial need and food insecurity measures were less commonly assessed for the historical comparison.

comparison groups (Table 2). The intervention group had a higher percentage of visits due to congestive heart failure or diabetes, had a higher Charlson score, was more likely to have a prior hospitalization, and was more likely to have social/financial needs or food insecurity compared to both comparison groups.

There were also differences specific to either comparison group. The intervention group had lower ADI (less disadvantage) and LOS than the concurrent comparison group (ADI 5.3 vs 6.1; LOS 3.5 vs

4.5 d), but not the historical comparison group. Additionally, the intervention group was more likely to have prior ER encounters (74.8%) compared to the historical comparison group (79.7%) but less likely than the concurrent comparison group (64.5%).

Propensity score matching and IPTW achieved adequate balance. We identified adequate common support between the treatment and comparison groups (Figure 1A,B). Propensity score matching resulted in standardized differences below 0.1 for all variables except ADI for

both groups and nonwhite/Hispanic, the utilization measures, and LOS for the historical comparison group (Figure 2A,B and Table S1). IPTW resulted in standardized differences below 0.1 for all variables except the Charlson for the concurrent comparison group (Figure 3A,B and Table S2).

5 | DISCUSSION

In summary, we found that despite large differences in observable characteristics, we were able to successfully reduce these differences for both historical and concurrent comparison groups. Our analysis provides lessons for conducting evaluations in the COVID-19 period.

First, there is value in using both historical and concurrent comparison groups. Both have limitations with respect to COVID-19 as well as general limitations. However, using multiple comparison groups and analytic approaches provides an opportunity to triangulate results. Having similar results across these approaches increases confidence in our findings by demonstrating the effects are robust to differences in historical context or analytical assumptions. Additionally, the use of both groups helps evaluate potential sources of differences. For example, we observed a higher percentage of patients in the intervention group with congestive heart failure and diabetes. This difference may reflect selection into the intervention or changes in the case mix due to COVID-19. Since the distribution of conditions was similar for both comparison groups, these differences are likely due to selection.

Second, the meaning of some measures may change after COVID-19. The intervention group had a higher percentage of prior hospitalizations relative to both comparison groups. While the intervention group also had a higher percentage of prior ER visits relative to the concurrent comparison group, the percentage was lower than for the historical comparison group. This result is likely related to the decline in ER visits observed during COVID-19.¹⁶ As such, utilization measures such as ER visits may measure something different before and after COVID-19. Caution is needed when adjusting for utilization measures or other measures that are recorded during visits that may be affected by COVID-19 utilization patterns.

Third, despite moderate to large differences in some of the initial variables, we were largely able to balance the covariates using two different propensity score approaches. While some of the variables still had standardized differences above 0.1 after using the propensity score methods, the standardized differences were smaller, and we will be able to adjust for these remaining differences in the outcome regressions by including these variables as covariates. These findings suggest that the differences in clinical and demographic characteristics were not so great that we cannot identify comparators. One potential explanation for this finding is that the comparison groups were both much larger than the intervention group. As such, the propensity score approaches were able to focus on the subset of patients in the comparison groups that were more comparable to patients in the intervention group. It would be more difficult to use the propensity score methods if the comparison groups had similar numbers of patients as the intervention

group since there would be fewer comparison group patients that have propensity scores that are similar to intervention group patients.

Our findings on achieving balance should be interpreted in the context of limitations. First, we were unable to evaluate differences in unobservable variables that may have affected outcomes. Unobserved factors that may have affected outcomes include access to care, marital status, caregiver support, COVID-19 status from at-home tests, patient health beliefs and perceived need, and detailed disease severity for the qualifying conditions. While these are important factors affecting health and outcomes, they are not readily or reliably available in the EHR and it is, therefore, possible that imbalances remained for these variables. Second, there may be generalizability issues to non-integrated health systems or geographic areas with different COVID-19 experiences. This study benefitted from having hospitalization and ER visit data on KPNW patients regardless of whether they received care at a KPNW or non-KPNW hospital and from including patients from contiguous counties. Different approaches may be needed for evaluating outcomes when not all hospitalization or ER visit data is captured, or patients are receiving care across different geographic regions.

We did not employ approaches that are often applied but may not be valid in the context of COVID-19. For example, a common approach with historical comparison groups is to perform an interrupted time-series design that evaluates changes in outcomes over time and how the changes relate to when the intervention was implemented.²¹ However, such designs can lead to bias if there are pre-intervention trends such as the changes to healthcare delivery during COVID-19 that began around the same time as the implementation of the intervention. We continue to explore methodological options that improve the strength of findings from this observational study. For example, a difference-in-difference analysis²² will be explored as another approach, which we will report on separately.

5.1 | Conclusions

Careful consideration is needed to identify comparison groups for interventions during COVID-19 and other periods of significant transition. As innovation continues during times of rapid change, evaluating the interventions remains important for learning health systems to establish internal validity. Our comparative analysis suggests that there is value in using both a historical and concurrent comparison group for triangulation of results and utilization measures should be interpreted carefully given changes over periods of rapid transition.

ACKNOWLEDGEMENTS

We wish to acknowledge Kaiser Permanente Community Health for internal research support for the program evaluation of the MTM intervention.

FUNDING INFORMATION

The intervention studied was funded by Kaiser Permanente Community Health. Dr. Savitz' time was supported by the Robert D. and

Patricia E. Kern Center for the Science of Health Care Delivery at Mayo Clinic, Rochester, MN.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

ORCID

Samuel T. Savitz  <https://orcid.org/0000-0003-3190-7740>

REFERENCES

- Stoto M, Oakes M, Stuart E, Brown R, Zurovac J, Priest EL. Analytical methods for a learning health system: 3. *Anal Observat Stud EGEMS (Wash DC)*. 2017;5:30.
- Stoto M, Oakes M, Stuart E, Priest EL, Savitz L. Analytical methods for a learning health system: 2. *Design Observat Stud EGEMS (Wash DC)*. 2017;5:29.
- Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66:688-701.
- Glass TA, Goodman SN, Hernán MA, Samet JM. Causal inference in public health. *Annu Rev Public Health*. 2013;34:61-75.
- Grimes DA, Schulz KF. Bias and causal associations in observational research. *The Lancet*. 2002;359:248-252.
- Patel SY, Mehrotra A, Huskamp HA, Uscher-Pines L, Ganguli I, Barnett ML. Variation in telemedicine use and outpatient care during the COVID-19 pandemic in the United States: study examines variation in total US outpatient visits and telemedicine use across patient demographics, specialties, and conditions during the COVID-19 pandemic. *Health Aff*. 2021;40:349-358.
- Chang AY, Cullen MR, Harrington RA, Barry M. The impact of novel coronavirus COVID-19 on noncommunicable disease patients and health systems: a review. *J Intern Med*. 2021;289:450-462.
- Hong B, Bonczak BJ, Gupta A, Thorpe LE, Kontokosta CE. 2021. Exposure density and neighborhood disparities in COVID-19 infection risk. *Proceedings of the National Academy of Sciences*, 118.
- CDC COVID-19 Response Team. Geographic differences in COVID-19 cases, deaths, and incidence—United States, February 12–April 7, 2020. *Morb Mortal Wkly Rep*. 2020;69:465-471.
- American Association of Critical-Care Nurses. Hear Us Out Campaign Reports Nurses' COVID-19 Reality [Online]. 2021. Accessed May 10, 2022. <https://www.aacn.org/newsroom/hear-us-out-campaign-reports-nurses-covid-19-reality>
- Leo CG, Sabina S, Tumolo MR, et al. Burnout among healthcare workers in the COVID 19 era: a review of the existing literature. *Front Public Health*. 2021;9:750529.
- Pfefferbaum B, North CS. Mental health and the Covid-19 pandemic. *New Engl J Med*. 2020;383:510-512.
- O'Connor RC, Wetherall K, Cleare S, et al. Mental health and well-being during the COVID-19 pandemic: longitudinal analyses of adults in the UK COVID-19 Mental Health & Wellbeing study. *Br J Psychiatry*. 2021;218:326-333.
- Kaufman HAL. *Financial Effects of COVID-19: Hospital Outlook for the Remainder of 2021*. Chicago: American Hospital Association; 2021.
- Bhatt AS, Moscone A, McElrath EE, et al. Fewer hospitalizations for acute cardiovascular conditions during the COVID-19 pandemic. *J Am Coll Cardiol*. 2020;76:280-288.
- Hartnett KP, Kite-Powell A, DeVies J, et al. Impact of the COVID-19 pandemic on emergency department visits—United States, January 1, 2019–may 30, 2020. *Morb Mortal Wkly Rep*. 2020;69:699-704.
- Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28:3083-3107.
- Singh GK. Area deprivation and widening inequalities in US mortality, 1969–1998. *Am J Public Health*. 2003;93:1137-1143.
- Austin PC. An Introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46:399-424.
- Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. *J Econ Surv*. 2008;22:31-72.
- Penfold RB, Zhang F. Use of interrupted time series analysis in evaluating health care quality improvements. *Acad Pediatr*. 2013;13:S38-S44.
- Donald SG, Lang K. Inference with difference-in-differences and other panel data. *Rev Econ Stat*. 2007;89:221-233.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Savitz ST, Scott JL, Leo MC, Keast EM, Savitz LA. Identifying appropriate comparison groups for health system interventions in the COVID-19 era. *Learn Health Sys*. 2022;e10344. doi:10.1002/lrh2.10344