© Health Research and Educational Trust

DOI: 10.1111/1475-6773.12832

METHODS ARTICLE

Key Design Considerations When Calculating Cost Savings for Population Health Management Programs in an Observational Setting

Shannon M. E. Murphy , Douglas E. Hough, Martha L. Sylvia, Linda J. Dunbar, and Kevin D. Frick

Objective. To illustrate the impact of key quasi-experimental design elements on cost savings measurement for population health management (PHM) programs.

Data Sources. Population health management program records and Medicaid claims and enrollment data from December 2011 through March 2016.

Study Design. The study uses a difference-in-difference design to compare changes in cost and utilization outcomes between program participants and propensity score-matched nonparticipants. Comparisons of measured savings are made based on (1) stable versus dynamic population enrollment and (2) all eligible versus enrolled-only participant definitions. Options for the operationalization of time are also discussed.

Data Collection/Extraction Methods. Individual-level Medicaid administrative and claims data and PHM program records are used to match study groups on baseline risk factors and assess changes in costs and utilization.

Principal Findings. Savings estimates are statistically similar but smaller in magnitude when eliminating variability based on duration of population enrollment and when evaluating program impact on the entire target population. Measurement in calendar time, when possible, simplifies interpretability.

Conclusion. Program evaluation design elements, including population stability and participant definitions, can influence the estimated magnitude of program savings for the payer and should be considered carefully. Time specifications can also affect interpretability and usefulness.

Key Words. Population health management, program evaluation, attrition, intention-to-treat, cost savings

U.S. federal health care reform is centered on providing high-quality, affordable health care while reducing spending (Key features of the Affordable Care Act, 2014). Developing innovative ways to address the needs of high-risk

populations is key for improving health and care quality. However, an equally critical component is the ability to demonstrate, using rigorous evaluation techniques, that these innovations achieve the desired impact on outcomes and costs. Quantifying savings is particularly important for making decisions about funding and sustainability (Wilson 2003; McGlynn and McClellan 2017).

The fundamental question underlying any program evaluation is, “Does the program work?” That is, does the intervention achieve the desired aims such as improve outcomes and/or reduce care spending? Although defining the question is easy, obtaining the answer can be a daunting task. Randomized controlled trials are generally considered the gold standard for measuring the causal impact of an intervention, but such designs are often infeasible due to ethical or contractual issues (Fetterolf, Wennberg, and Devries 2004; Bonell et al. 2011; Handley, Schillinger, and Shiboski 2011). Thus, researchers face the challenge of devising a quasi-experimental framework that minimizes common validity threats. More rigorous design improves our ability to differentiate program impact from alternative explanations of change and moves us closer to measuring the causal impact of the intervention in an observational setting (Shadish, Cook, and Campbell 2002).

The purpose of this study was to describe a rigorous method for evaluating population health management (PHM) programs with an emphasis on payer cost and utilization outcomes. The framework synthesizes some best practices for maximizing internal validity to help estimate the causal impact of the intervention. The design is tailored to align with the intervention’s population health focus. Lastly, but perhaps most important, the paper demonstrates how key design elements influence the measurement of cost savings (i.e., population stability, participant definition, and time operationalization) and provides rationale for each element to help researchers assess the framework’s appropriateness and applicability to other programs and settings. Applying the framework should help researchers provide sound evaluation results to aid decisions related to program funding and sustainability.

Address correspondence to Shannon M. E. Murphy, M.A., Johns Hopkins HealthCare LLC, 6704 Curtis Court, Glen Burnie, MD 21060; e-mail: SMurphy@jhhc.com. Douglas E. Hough, Ph.D., is with the Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD. Martha L. Sylvia, Ph.D., M.B.A., R.N., is with the Medical University of South Carolina, Charleston, SC. Linda J. Dunbar, Ph.D., is with the Johns Hopkins HealthCare LLC, Glen Burnie, MD. Kevin D. Frick, Ph.D., is with the Johns Hopkins University Carey Business School, Baltimore, MD.

EVALUATION DESIGN CONSIDERATIONS

The difference-in-difference design is frequently applied to minimize validity threats (Baker et al. 2011; Lin et al. 2012; Grossmeier et al. 2013; Hawkins et al. 2015; Cuellar et al. 2016). This method estimates the pre-post difference for the intervention group and compares it to the pre-post difference for a similar group not exposed to the intervention. Thus, the comparison group provides a counterfactual estimate for the expected difference in outcomes had the participants not been exposed to the intervention. Changes in outcomes influenced by history, maturation, and regression to the mean should affect both groups in similar ways, and the difference between groups should identify the intervention effect (Shadish, Cook, and Campbell 2002).

In the difference-in-difference design, the ability of the comparison group to mitigate validity threats depends on the quality of the comparison group (Shadish, Cook, and Campbell 2002). Ideally, the comparison group should be as similar as possible on baseline characteristics, both observed and unobserved (Heckman 1979). Propensity scores are a common way to control for baseline differences when estimating intervention effects. Propensity scores quantify the probability that a study participant received the intervention based on observed baseline characteristics. Using propensity scores to match intervention to comparison participants helps ensure that the distribution of baseline characteristics is similar between groups, thus mitigating observed confounding factors (Stuart 2010; Stuart and Ialongo 2010; Ryan, Burgess, and Dimick 2015).

Researchers need to determine the best way to apply difference-in-difference designs with propensity score matching within a population health framework. With the current emphasis on population health strategies to address the needs of entire populations, the evaluation should also employ a similar framework for measuring intervention impact while maintaining the highest possible level of rigor. Three design considerations for achieving these goals require defining the following concepts: (1) the population, (2) the participant, and (3) the timing of the intervention.

The “population” is a fundamental component of the evaluation design. Although the general definition of the population is clear and stable (e.g., individuals in a geographic area, members within a health plan), individuals within a population are often changing (e.g., through birth, death, relocation, and plan enrollment/disenrollment). Researchers must be able to distinguish changes *among* the population from changes *within* the population. For

instance, when viewing changing trends in health outcomes, it can be difficult to determine whether the changes result from intervention-induced improvements for individuals within the population or from changing case mix as participants enter and exit the population. One way to resolve this issue is to identify a stable cohort (or cohorts) of individuals and measure differences in their outcomes over time. Another related, and commonly used, strategy is to require a minimum duration of enrollment in the population (Baker et al. 2011; Cuellar et al. 2016). Illustrating the importance of this factor, Cuellar et al. (2016) conducted a patient-centered medical home (PCMH) evaluation and found improved cost and utilization outcomes, but the magnitude of benefit was smaller for those continuously attributed to a PCMH site over the multiyear study period, as compared to a larger group requiring a minimum of only one-quarter. Requiring a minimum duration in the population minimizes the impact of attrition on the observed results. Although both strategies may reduce the sample size and generalizability of the results, their benefits to internal validity when trying to measure causal impact likely outweigh the costs.

The “participant” definition is also central to program evaluation design. For example, does participation start at the point of identification, at enrollment, or after a certain “dose” of the intervention has been received? All are valid participant definitions, but they affect validity and generalizability in different ways. Thus, selecting from among the participant options depends on the estimate of interest (Heckman 1992; Heckman, Smith, and Taber 1998). Evaluating program impact at the point of identification aligns most closely with the goals of population health. Similar to an intention-to-treat (ITT) approach, it measures the impact of the program on the entire target population and not just the subset who chose to enroll (Banerjee and Duflo 2009). Additionally, the ITT model addresses many issues associated with selection bias (Lin et al. 2012). Selection bias can be difficult to avoid in observational settings because, even if the comparison group has similar observed characteristics, differences in unobserved characteristics that influence group selection (i.e., whether the individual chooses to enroll in the program) may remain. The ITT model mitigates these differences because it includes the entire target population, including those who opt to enroll and those who refuse (Wilson 2003; Lin et al. 2012), thus avoiding the sample specification error, and associated biases, as described by Heckman and colleagues (Heckman 1979; Heckman et al. 1998). A downside of the ITT design is that it may be more difficult to detect an intervention effect if the effect is small and/or if a large portion of the target population does not participate. The ITT approach emphasizes not only the effect size for those who received the intervention, but also efforts to

maximize intervention uptake. Both are essential for a successful population health strategy as there would be less utility of an effective program if only a small portion of eligible participants receives the intervention. Thus, the ability to measure the full impact of the intervention, or the average treatment effect, is particularly ideal for answering policy-level questions such as return-on-investment (Wilson 2003). Another factor influencing the use of an ITT design is the requirement that the comparison participants be drawn from a pool of similar members who are not offered the intervention (Heckman et al. 1998). Although this approach is ideal for mitigating selection bias, it may not always be feasible if a suitable comparison population cannot be found.

The counterpart of the ITT model is the per-protocol (PP) model, which evaluates the impact on those who received the intervention as intended (the average treatment effect on the treated) (Brody 2011). The strengths of one approach are generally the weaknesses of the other. Under the PP model, selection bias remains a threat to internal validity and the results do not consider the nonenrolled portion for which the intervention has no impact. On the other hand, the PP model measures the effect size for those who received at least some portion of the intervention, similar in some ways to a dose analysis. Considering the strengths and weaknesses of each approach, an ideal evaluation design may employ both, if feasible. Although the ITT model is the preferred and more rigorous approach, the combination would help distinguish issues of effect size versus uptake, if they exist. The latter approach would also quantify the potential impact of the intervention if recruitment efforts are increased.

A third important design consideration, closely related to the definitions of the population and participant, is the operationalization of time. When specifying the pre- and post-periods, two common approaches are to (1) use calendar months to apply the same start date for all study participants or (2) allow participants to have their own start date based on the date of identification and/or enrollment. Each has strengths and weaknesses. For cost outcomes, preserving calendar time may be preferred for three primary reasons. The first, and perhaps most important, is that plan-level changes in costs (i.e., rate changes) likely impact intervention and comparison groups at specific points in time, and therefore, it is important to use the same pre- and post-periods for both groups to reduce cost-specific historical threats. The second reason is so that the display of results can more closely mirror the cost trending carried out by finance groups, which is typically displayed by calendar month, quarter, or year. The increased ease of understanding achieved by this approach, and its comparability with other population-level financial trends,

better facilitates funding decisions. A third benefit of calendar time is the ease of specifying study periods for comparison group members. If study periods are allowed to vary for intervention participants, it is difficult to estimate an equivalent start date for comparison group members that does not introduce some form of bias.

METHODS

The program evaluation framework is demonstrated using the Medicaid results from the Johns Hopkins Community Health Partnership (J-CHiP) community-based PHM program (Berkowitz et al. 2016; Hsiao et al. unpublished data). Supported by a Health Care Innovation Award (HCIA) from the Center for Medicare and Medicaid Innovation (CMMI), the PHM program component of J-CHiP was designed to improve health and reduce spending for high-risk Medicaid and Medicare populations in East Baltimore. The program aimed to achieve these goals by improving care coordination and addressing the clinical and social determinants of health. Care coordination teams were formed with site-embedded care managers and behavioral health specialists, community health workers, and neighborhood navigators.

The J-CHiP program was implemented at most primary care sites throughout East Baltimore between December 2012 and June 2013, focusing initially on Priority Partners Managed Care Organization (PPMCO) Medicaid members. The full detail of the J-CHiP initiatives has been described in the literature (Berkowitz et al. 2016; Hsiao et al. unpublished data).

Participants

Using an ITT approach, the evaluation includes chronically ill adult (18 +) Medicaid members eligible for the J-CHiP PHM program, regardless of their level of program engagement. The target population includes high-risk members with one or more chronic conditions who receive primary care at a J-CHiP site. High-risk individuals are identified by predictive modeling or physician referral. From among the entire J-CHiP population, a cohort is identified that includes all eligible participants at a specific time. This time point began once the intervention was rolled out to the majority of sites and when there was a sufficient sample size to estimate the intervention effect. In addition, the time point is early enough to ensure at least 1 year of postimplementation data. Based on

the implementation timing, and to allow each site at least 1 month for intervention ramp-up, the cohort was identified in August 2013. Enrollment had nearly reached its goal of 1,000 participants.

To measure program impact for those who received the intervention as intended, a second cohort is identified. This cohort is a subset of J-CHiP participants who worked with a care manager (CM) for at least 3 months. In addition to the outreach services offered to the full cohort, the CM cohort also received CM assistance and maintained a minimum participation level (3 months). Analysis of this cohort helps measure the differential impact of J-CHiP for those who received more intensive intervention services as compared to the full cohort under the ITT design.

To reduce the influence of confounding factors, other exclusion criteria are applied, as described in the Appendix. A critical confounding factor is length of health plan enrollment given the more transient nature of the Medicaid population. To limit the influence of plan enrollment length (attrition bias) on the results, the primary analysis is limited to members enrolled in Medicaid for at least 1 year in both study periods (specifically the entire pre-period and the first 12 months of the post-period) to ensure that changes over time reflect participant-level changes and not changes in case mix due to plan enrollment and/or duration. This group is referred to as the “stable” population. However, to quantify the impact of this design element, the results are replicated with a larger “dynamic” population where only 1 month of Medicaid enrollment is required in each period.

After applying the inclusion and exclusion criteria, the comparison group is identified from a pool of similar Medicaid members who received primary care from a nonintervention site but would have otherwise been eligible. Using 1-to-1 nearest neighbor matching without replacement, each J-CHiP participant is matched to the chronically ill member most similar in terms of the following characteristics, as quantified by propensity scores: age, gender, race, residence in Baltimore City, residence in the J-CHiP target area, baseline health risk indicators (i.e., condition prevalence, predictive modeling score for the probability of an inpatient admission), and pre-period utilization (i.e., resource utilization level, cost, and any ED visits, admissions, and readmissions). For the “dynamic” population cohort that does not require continuous plan enrollment, the matching process also includes a binary indicator for 2-year plan enrollment.

Data

The evaluation uses multiple data sources spanning the period from December 2011 through March 2016. Medicaid (PPMCO) administrative records are used to determine plan enrollment characteristics (i.e., timing, length, and primary care site), as well as demographic information such as age, gender, race, and residence location. Johns Hopkins ACG[®] system data provide baseline indicators of health risk including condition prevalence, health resource utilization level, and risk of future hospitalization (Weiner et al. 1991; Weir, Aweh, and Clark 2008; About the ACG system, 2018). Hospitalization risk is characterized further by an internally developed predictive modeling score based on ACGs, claims, laboratory results, and EMR data (Berkowitz et al. 2016; Hsiao et al. unpublished data). Lastly, medical and pharmacy paid claims data are used to define cost and utilization metrics.

The final evaluation dataset contains 2–4 yearly records per member that correspond to the year before and up to 3 years after the J-CHiP program was implemented. The primary predictor variables include indicators for group (J-CHiP vs. Comparison) and study period (pre vs. post). The outcomes include total paid costs (i.e., inpatient and outpatient)¹ and count of inpatient admissions per member per year (PMPY). The covariates used for both matching and risk adjustment include age, gender, race, residence location, baseline health risk indicators, and pre-period utilization. Refer back to the *Participants* section for additional details on the covariates.

Evaluation Periods

As previously noted, the J-CHiP program was implemented in phases by site. The evaluation cohort is identified at a time that allowed the J-CHiP sites to refine the implementation process prior to measuring impact while also ensuring a sufficient sample and time for follow-up. The study periods are defined corresponding to the implementation schedule and the timing of cohort identification: pre (December 2011–November 2012), ramp-up (December 2012–July 2013), and post (August 2013–March 2016). The interim period between the pre- and post-periods, referred to as the “ramp-up” period, represents a period when the individual sites were rolling out the intervention, finalizing the implementation process, and enrolling participants. Data from this ramp-up period are excluded; analyses compare the pre- and post-periods.

Analysis

The J-CHiP program is evaluated using a difference-in-difference design. Using this approach, cost savings are realized if the improvement in outcomes is significantly better for J-CHiP participants than comparable nonparticipants.

The outcomes are modeled using the generalized estimating equations (GEE) technique with a working independence correlation structure to account for repeated yearly observations (Liang and Zeger 1986). Costs are analyzed in dollars using the sum of medical and pharmacy paid costs.² The utilization outcome is analyzed as count of inpatient admissions in the observation year. Pre-post and between-group (“difference-in-difference”) differences are measured using Poisson and linear distributions for the utilization and cost outcomes, respectively. Statistical significance is measured using a bootstrapped 95 percent confidence interval. Bootstrapping of matched pairs (with 1,000 repetitions) is used to estimate the standard errors to adjust for the possible misspecification of the correlation structure (Bertrand, Duflo, and Mullainathan 2004) and/or underlying distribution, particularly for costs that are assumed to be normally distributed (Jiang and Zhou 2004). The primary predictors include group, study period, and their interaction. All analyses are risk-adjusted using the same (or similar) factors included in the matching process. The dynamic models are also adjusted by plan enrollment length. For each outcome, the analysis is conducted four times to compare the differences based on the following evaluation design elements:

- Population Definition: Stable versus Dynamic
 - Stable: static cohort of program participants who were continuously enrolled in Medicaid for a minimum of 12 months in both the pre- and post-periods
 - Dynamic: cohort of program participants with varying lengths of plan enrollment in both of the study periods, with a minimum of 1 month
- Participant Definition: ITT versus PP
 - ITT: includes everyone eligible for the program, regardless of their degree of participation, referred to as the “Full” participant cohort
 - PP: a subset of the ITT population that received a minimum level of the intervention (i.e., received CM for at least 3 months), referred to as the “CM” participant cohort

RESULTS

Participant Characteristics

The final sample includes 1,710 study participants with 50 percent from each intervention group (Table 1). The majority of J-CHiP participants from the

dynamic population also met the criteria for the stable population (82 percent for the full participant cohort and 85 percent for the CM cohort). Additionally, most of the J-CHiP-eligible participants worked with a care manager for at least 3 months and therefore are also included in the CM cohort analysis (79 percent for the stable population and 75 percent for dynamic). As illustrated in Figures 1a and b, the matching process was generally successful in achieving balance³ between the J-CHiP and comparison groups on observed baseline characteristics. The remaining imbalance is primarily related to the proportion residing in the target area, as well as the plan enrollment length for the dynamic population. To further address any lingering baseline group differences, the pre-post approach allows each group to serve as its own control and the analysis also includes covariate adjustment for all of the factors included in Figure 1, including residence in the target area and number of months enrolled in Medicaid in the post-period.⁴

Impact of Evaluation Design Elements on Program Results

Results by Population: Stable versus Dynamic. As indicated in Table 2, the magnitude of the results differs when the study population is allowed to vary by observation length (i.e., months enrolled in the health plan), although the overall lack of statistical significance is unchanged. For costs, the savings rate is larger for the dynamic population than the stable population (−\$2,751 vs. −\$1,171), as exemplified by the full participant results. Upon review of the pre-post changes by group, PMPY costs increased by a larger amount for the J-CHiP group using the dynamic population, as compared to the stable population (from +\$2,536 to +\$3,396), but the comparison group increased by an even larger amount (+\$3,706 to +\$6,147).

The impact of using a dynamic population is similar for inpatient utilization (Table 2). As exemplified by the full participant results, the degree of program savings is larger for the dynamic population (ratio of incident rate ratios

Table 1: Sample Size by Population and Participant Definitions

	<i>J-CHiP</i>	<i>Comparison</i>	<i>Total</i>
Stable population			
Full participants	699	699	1,398
CM participants	550	550	1,100
Dynamic population			
Full participants	855	855	1,710
CM participants	645	645	1,290

Figure 1: (a) Baseline Group Differences before and after Matching: Stable Population. (b) Baseline Group Differences before and after Matching: Dynamic Population [Color figure can be viewed at wileyonlinelibrary.com]

[IRRs] = 0.91) than the stable population (ratio of IRRs = 1.01). As was found with costs, the difference appears to be linked to worsening outcomes for the comparison group [i.e., pre-post IRR worse/higher under the dynamic population (IRR = 1.01) than the stable population (IRR = 0.86)] more so than improved outcomes for program participants. For the full participant group, J-CHiP participants also showed an increase in the pre-post IRR under the dynamic population (i.e., from IRR = 0.87 to 0.92), but to a smaller degree than the comparison group.

Results by Participant Definition: Full versus CM. Table 2 also illustrates the program savings variation under an ITT approach versus one that focuses on the subset of the full participant population who received a minimum dose of the intervention. The results for these two participant cohort definitions are listed under the “Full” and “CM” participant sections, respectively. Using the stable population, both outcomes demonstrate a larger magnitude of program

Table 2: Results by Population and Participant Definitions

<i>Total Costs (PMPY) By Participant Cohort</i>	<i>J-CHiP Pre-Post Difference: Mean (CI)</i>	<i>Comp Pre-Post Difference: Mean (CI)</i>	<i>Difference-in- Difference: Mean (CI)</i>
Stable population			
Full	\$2,536 (\$−65, \$5,386)	\$3,706 (\$1,314, \$6,276)*	\$−1,171 (\$−4,968, \$2,145)
CM	\$2,058 (\$−931, \$5,319)	\$5,109 (\$2,184, \$8,260)*	\$−3,051 (\$−7,517, \$994)
Dynamic population			
Full	\$3,396 (\$699, \$6,099)*	\$6,147 (\$3,418, \$8,770)*	\$−2,751 (\$−6,433, \$1,240)
CM	\$2,372 (\$−414, \$5,405)	\$5,702 (\$2,684, \$8,834)*	\$−3,331 (\$−7,385, \$1,081)

<i>Admissions (PMPY) By Participant Cohort</i>	<i>J-CHiP Pre-Post Difference: IRR (CI)</i>	<i>Comp Pre-Post Difference: IRR (CI)</i>	<i>Difference-in-Difference: IRR (CI)</i>
Stable population			
Full participants	0.87 (0.76, 0.98)*	0.86 (0.75, 0.99)*	1.01 (0.83, 1.22)
CM participants	0.85 (0.74, 0.97)*	0.93 (0.80, 1.07)	0.91 (0.75, 1.12)
Dynamic population			
Full participants	0.92 (0.81, 1.04)	1.01 (0.87, 1.17)	0.91 (0.75, 1.11)
CM participants	0.85 (0.74, 0.98)*	0.97 (0.85, 1.13)	0.88 (0.72, 1.05)

*Indicates significant difference at $p < .05$ (based on bootstrapped 95% confidence interval or CI).

impact for the CM group than the full participant group (−\$3,051 vs. −\$1,171 for costs and 0.91 vs. 1.01 for admissions).

These findings also hold for the dynamic population, except that the differences between CM and full participant cohorts are smaller. For example, the difference in total cost savings between the full and CM cohorts is \$1,880 for the stable cohort (−\$3,051 for CM vs. −\$1,171 for full), and \$580 for the dynamic cohort (−\$3,331 for CM vs. −\$2,751 for full).

DISCUSSION

The quasi-experimental evaluation framework described here is designed to minimize common threats to internal validity (i.e., sampling selection bias, self-selection bias, attrition bias, history, maturation, and regression to the mean) when estimating the causal impact of PHM programs. The paper demonstrates the impact of key evaluation design elements on measurements of program savings.

First, it is essential to include a comparison group or employ another rigorous method for calculating expected (or counterfactual) outcomes for program participants if they had not been exposed to the intervention. Relying on pre-post results alone can be insufficient and misleading for understanding program impact. In the current analysis, the J-CHiP group showed significant (or nearly significant) decreases in utilization outcomes from pre to post. However, after viewing the comparison group's similar reduction, the difference-in-difference results indicate that at least part of this improvement should be attributed to factors other than program participation.

Second, the designation of the "population" and "participant" cohorts can affect the calculation of cost savings. In the current analysis, across all population and participant definitions, the results are in similar directions and consistent in terms of statistical significance, but the magnitude of the estimates differed. When comparing a stable versus dynamic population, the cost savings rate was lower for the stable population after removing the variability due to plan enrollment length, consistent with Cuellar et al. (2016). Furthermore, the stable cohort detected larger differences between the full and CM participant cohorts than the dynamic cohort. The differences between the stable and dynamic populations appear to stem from negative changes to the comparison group estimates, while the participant group remained stable. Overall, the differences by population definition suggest that ignoring confounding produced by plan attrition may overestimate program impact, particularly in terms of costs, and may underestimate the benefits of increased program intensity as provided by CMs.

Third, the "participant" definition impacts validity and generalizability. For J-CHiP, the results were generally better when limiting the intervention group to those who received a minimum dose of the program, consistent with the findings of Baker et al. (2011). This specification helps program managers assess intervention impact for those who are treated. It can also highlight more or less effective program components for the purpose of quality improvement. However, the results based on this design are less helpful for administrators making program funding decisions (Wilson 2003). If the enrollment rate is low, and/or if the effect size is small, then using only the treated participants may overestimate the savings rate. For funding decisions, it is more helpful to know the impact across the entire population, especially considering the outreach funds invested for those who do not choose to enroll. Another benefit of the ITT design is that it mitigates the selection bias threat that may still be present when only the treated population is used.

Lastly, although the focus of this study is on the bias associated with cohort definitions, the operationalization of time is an important design element that warrants further research. The method described here evaluated outcomes based on calendar time, which has a number of potential benefits, including simplification of the comparison group selection process, mitigation of historical threats related to systematic changes in rate setting, ease of understanding, and comparability to existing finance trend reporting. This approach can also be adapted using groupings of participants based on timing of enrollment to evaluate program impact when start dates span longer periods or with phased implementation. This adaptation would also be useful for further separating secular trends from the treatment effect. Furthermore, although time is modeled in years for this example, using smaller units of time (e.g., months, quarters) may detect trend variations that could be useful for program monitoring and quality improvement. Smaller time units would also allow more model flexibility as may be preferred under the dynamic population approach. Further research is needed to test these hypotheses and to demonstrate the impact of alternative time choices (e.g., time since enrollment) on the validity and utility of each approach.

One of the main limitations of the approach is excluding a portion of the program participants (e.g., those who did not have stable health plan enrollment) to reduce the influence of confounding factors. This may reduce both generalizability and power. However, if carefully considered and appropriately applied, the benefits to internal validity when trying to measure causal impact should outweigh the disadvantages. For example, to address issues of generalizability for this PHM program evaluation, the analysis was conducted using both the stable and dynamic populations. Although generally similar, the results from the stable population are preferred due to the increased internal validity achieved by removing plan enrollment length as a confounder. Additionally, power and generalizability could both be improved through periodic replications of the evaluation on subsequent PHM participant cohorts.

These design factors should have wide applicability, either separately or in combination. For example, many of them can be applied to evaluations of accountable care organizations (ACOs). When assessing impact at the population level, cohorts of attributed beneficiaries can be identified (e.g., based on year of initial attribution to the ACO) and used to calculate changes in outcomes in the year(s) before and after attribution. Use of cohorts enables the differentiation of case-mix changes from within-person changes. Identifying cohorts based on year of attribution aligns measurement periods based on calendar time and enables comparability to other

financial trends. Use of multiple cohorts allows for replication of pre-post measurement with newer cohorts and maximizes sample size. A challenge for evaluating ACOs will be identifying comparison groups. At the ACO level, claims data may not be available for a comparison group of nonattributed beneficiaries. When evaluating ACO components, such as a PHM program, it may be difficult to employ the ITT approach if the target population includes the entire ACO. If feasible, it may be ideal to consider a staggered implementation approach (e.g., stepped wedge design) to allow for a suitable comparison pool using an ITT design (McGlynn and McClellan 2017).

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: The authors wish to acknowledge the late Dr. Fred Brancati for his instrumental role in the launch of the J-CHiP program and the development of the evaluation design described in this article. Others who made valuable contributions to the acquisition and/or analysis of the data include Sarah Kachur, Xuan Huang, Yanyan Lu, Chorong Biann, and Tom Richards.

Disclosures: The project described was supported by Grant number 1C1CMS331053 from the Department of Health and Human Services, Centers for Medicare & Medicaid Services. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the U.S. Department of Health and Human Services or any of its agencies. The research presented here was conducted by the awardee. Findings might or might not be consistent with or confirmed by the findings of the independent evaluation contractor.

Disclaimer: None.

NOTES

1. All claims paid by PPMCO are included to estimate cost savings to the payer. Due to the mental/behavioral health carve-out in Maryland, claims for these services are excluded.
2. Costs are analyzed using a normal distribution to produce estimates that do not require retransformation and can be easily used for return-on-investment calculations. As costs typically violate the normality assumption, two sensitivity analyses are conducted: (1) replication of the analysis using a gamma distribution and a log

link; and (2) trimming costs to the 99th percentile to assess the impact of outliers. The results from both analyses are similar to the main analysis.

3. As suggested in Stuart (2010), postmatching balance is assessed as follows: (1) standardized differences after matching are <0.25 for each predictor, squared predictor, and two-way interaction, (2) variance ratio of the propensity scores between groups is between 0.5 and 2, and (3) ratio of residual variances between groups for each predictor is between 0.5 and 2.
4. To prevent multicollinearity, the baseline measure for prior admissions is excluded from the admissions model.

REFERENCES

- About the ACG system. 2018. [accessed on January 17, 2018] Available at <https://www.hopkinsacg.org/advantage>
- Baker, L. C., S. J. Johnson, D. Macaulay, and H. Birnbaum. 2011. "Integrated Telehealth and Care Management Program for Medicare Beneficiaries with Chronic Disease Linked to Savings." *Health Affairs (Project Hope)* 30 (9): 1689–97.
- Banerjee, A. V., and E. Duflo. 2009. "The Experimental Approach to Developmental Economics." *Annual Review of Economics* 1: 151–78.
- Berkowitz, S. A., P. Brown, D. J. Brotman, A. Deutschendorf, L. Dunbar, A. Everett, A. Everett, D. Hickman, E. Howell, L. Purnell, C. Sylvester, R. Zollinger, M. Bellantoni, S. C. Durso, C. Lyketsos, P. Rothman, and J-CHiP Program. 2016. Case study: Johns Hopkins Community Health Partnership: A Model for Transformation. *Healthcare (Amsterdam, Netherlands)* 4: 264–70.
- Bertrand, M., E. Duflo, and S. Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119 (1): 249–75.
- Bonell, C. P., J. Hargreaves, S. Cousens, D. Ross, R. Hayes, M. Petticrew, and B. R. Kirkwood. 2011. "Alternatives to Randomisation in the Evaluation of Public Health Interventions: Design Challenges and Solutions." *Journal of Epidemiology and Community Health* 65 (7): 582–7.
- Brody, T. 2011. "Chapter 8: Intent to Treat Analysis vs. Per Protocol Analysis." In *Clinical Trials: Study Design, Endpoints, and Biomarkers, Drug Safety, and FDA and ICH 6 Guidelines*, pp. 143–64. Burlington: Elsevier Science.
- Cuellar, A., L. A. Helmchen, G. Gimm, J. Want, S. Burla, B. J. Kells, and L. M. Nichols. 2016. "The CareFirst Patient-Centered Medical Home Program: Cost and Utilization Effects in Its First Three Years." *Journal of General Internal Medicine* 31 (11): 1382–8.
- Fetterolf, D., D. Wennberg, and A. Devries. 2004. "Estimating the Return on Investment in Disease Management Programs Using a Pre-post Analysis." *Disease Management: DM7* (1): 5–23.
- Grossmeier, J., E. L. Seaverson, D. J. Mangen, S. Wright, K. Dalal, C. Phalen, and D. B. Gold. 2013. "Impact of a Comprehensive Population Health Management

- Program on Health Care Costs." *Journal of Occupational and Environmental Medicine* 55 (6): 634–43.
- Handley, M. A., D. Schillinger, and S. Shiboski. 2011. "Quasi-experimental Designs in Practice-based Research Settings: Design and Implementation Considerations." *Journal of the American Board of Family Medicine: JABFM* 24 (5): 589–96.
- Hawkins, K., P. M. Parker, C. E. Hommer, G. R. Bhattarai, J. Huang, T. S. Wells, and C. S. Yeh. 2015. "Evaluation of a High-Risk Case Management Pilot Program for Medicare Beneficiaries with Medigap Coverage." *Population Health Management* 18 (2): 93–103.
- Heckman, J. J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153–61.
- . 1992. "Randomization and Social Policy Evaluation." In *Evaluating Welfare and Training Programs*, edited by C. Manski, and I. Garfinkel, pp. 201–30. Cambridge, MA: Harvard University Press.
- Heckman, J., J. Smith, and C. Taber. 1998. "Accounting for Dropouts in Evaluations of Social Programs." *Review of Economics and Statistics* 80 (1): 1–14.
- Heckman, J., H. Ichimura, J. Smith, and P. Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66 (5): 1017–98.
- Hsiao, Y. L., E. B. Bass, A. W. Wu, M. B. Richardson, A. Deutschendorf, D. J. Brotman, M. Bellatoni, E. E. Howell, A. Everett, D. Hickman, L. Purnell, R. Zollinger, C. Sylvester, K. Lyketsos, L. Dunbar, and S. A. Berkowitz. 2017. "Implementation of a comprehensive program to improve coordination of care in an urban academic health care system." Unpublished manuscript.
- Jiang, H., and X. H. Zhou. 2004. "Bootstrap Confidence Intervals for Medical Costs With Censored Observations." *Statistics in Medicine* 23 (21): 3365–76.
- Key Features of the Affordable Care Act. 2014. [accessed on December 16, 2016]. Available at <https://www.hhs.gov/healthcare/facts-and-features/key-features-of-aca/index.html>
- Liang, K., and S. L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73 (1): 13–22.
- Lin, W. C., H. L. Chien, G. Willis, E. O'Connell, K. S. Rennie, H. M. Bottella, and T. G. Ferris. 2012. "The Effect of a Telephone-Based Health Coaching Disease Management Program on Medicaid Members with Chronic Conditions." *Medical Care* 50 (1): 91–8.
- McGlynn, E. A., and M. McClellan. 2017. "Strategies for Assessing Delivery System Innovations." *Health Affairs (Project Hope)* 36 (3): 408–16.
- Ryan, A. M., J. F. Burgess Jr, and J. B. Dimick. 2015. "Why We Should Not Be Indifferent to Specification Choices for Difference-in-Differences." *Health Services Research* 50 (4): 1211–35.
- Shadish, W. R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-experimental Design for Generalized Causal Inference*. Boston, MA: Houghton Mifflin Company.

- Stuart, E. A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 25 (1): 1–21.
- Stuart, E. A., and N. S. Ialongo. 2010. "Matching Methods for Selection of Subjects for Follow-up." *Multivariate Behavioral Research* 45 (4): 746–65.
- Weiner, J. P., B. H. Starfield, D. M. Steinwachs, and L. M. Mumford. 1991. "Development and Application of a Population-Oriented Measure of Ambulatory Care Case-Mix." *Medical Care* 29 (5): 452–72.
- Weir, S., G. Aweh, and R. E. Clark. 2008. "Case Selection for a Medicaid Chronic Care Management Program." *Health Care Financing Review* 30 (1): 61–74.
- Wilson, T. W. 2003. "Evaluating ROI in State Disease Management Programs." *State Coverage Initiatives Issue Brief: A National Initiative of the Robert Wood Johnson Foundation* 4 (5): 1–6.

SUPPORTING INFORMATION

Additional supporting information may be found online in the supporting information tab for this article:

Appendix SA1: Author Matrix.

Appendix SA2:

Exhibit S1. Study Participant Consort Diagram.

Exhibit S2. J-CHiP Participant Pre-enrollment Characteristics.