METHODS ARTICLE

Comparison group selection in the presence of rolling entry for health services research: Rolling entry matching

Allison Witman PhD¹ | Christopher Beadles MD, PhD² | Yiyan Liu PhD³ |
Ann Larsen MS³ | Nilay Kafali PhD⁴ | Sabina Gandhi PhD⁵ | Peter Amico PhD⁶ |
Thomas Hoerger PhD²

¹Cameron School of Business, University of North Carolina Wilmington, Wilmington, North Carolina

²RTI International, Research Triangle Park, North Carolina

³RTI International, Waltham, Massachusetts

⁴Department of Economics, Boston University, Boston, Massachusetts

⁵RTI International, Washington, District of Columbia

⁶Amico Consulting, Orlando, Florida

Correspondence

Allison Witman, PhD, Cameron School of Business, University of North Carolina Wilmington, Wilmington, NC 28403-5945. Email: witmana@uncw.edu

Funding information

Centers for Medicare and Medicaid Services, Grant/Award Number: HHSM-500-2010-00021I; RTI International

Objective: To demonstrate rolling entry matching (REM), a new statistical method, for comparison group selection in the context of staggered nonuniform participant entry in nonrandomized interventions.

Study Setting: Four Health Care Innovation Award (HCIA) interventions between 2012 and 2016.

Study Design: Center for Medicare and Medicaid Innovation HCIA participants entering these interventions over time were matched with nonparticipants who exhibited a similar pattern of health care use and expenditures during each participant's baseline period.

Data Extraction Methods: Medicare fee-for-service claims data were used to identify nonparticipating, fee-for-service beneficiaries as a potential comparison group and conduct REM.

Principal Findings: Rolling entry matching achieved conventionally-accepted levels of balance on observed characteristics between participants and nonparticipants. The method overcame difficulties associated with a small number of intervention entrants.

Conclusions: In nonrandomized interventions, valid inference regarding intervention effects relies on the suitability of the comparison group to act as the counterfactual case for the intervention group. When participants enter over time, comparison group selection is complicated. Rolling entry matching is a possible solution for comparison group selection in rolling entry interventions that is particularly useful with small sample sizes and merits further investigation in a variety of contexts.

KEYWORDS

causal inference, comparison groups, matching, methods, propensity scores

1 | INTRODUCTION

Although the gold standard of experimental research is the randomized control trial, many health care interventions are implemented without a randomized control group for practical or ethical reasons. For interventions without a control group, a rigorous evaluation of

intervention effects often requires a carefully selected comparison group to serve as the counterfactual case for outcomes in the absence of the intervention. Without a randomized control group, intervention effects are often estimated through difference-in-difference or regression adjustment. In either case, comparison individuals are typically selected based on their similarity to intervention

participants during a baseline period prior to the intervention; however, comparison group selection becomes complicated when there is rolling entry of participants into the intervention over multiple time periods.

The primary challenge of comparison group selection for rolling entry interventions is defining a baseline period for potential comparison group individuals who do not enter the intervention. Instead of one baseline period for all participants, rolling entry generates a baseline period for each participant entry into the intervention. For example, if quarters are the unit of time for the analysis and participants enter over eight calendar quarters, then there are eight different baseline periods corresponding to each entry quarter.

In some cases, the entry criteria for the intervention can be applied to the potential comparisons to delineate the baseline and intervention period, such as when a hospitalization or diagnosis triggers entry. For example, in an intervention to test a new hip implant, the comparison group would comprise individuals undergoing hip implant surgery but not receiving the new implant. For comparisons, the baseline period would be the time before the surgery and the intervention period would start at the surgery.

Unfortunately, entry criteria are not always clear or available for comparison group selection. Entry criteria that are difficult to replicate for comparison group selection include community outreach (e.g., neighborhood associations, community-based organizations, social services agencies), advertising (e.g., brochures, radio programs), community events (e.g., health fairs, festivals), and referrals (e.g., friends and family, social worker, health care provider). When entry criteria for an intervention cannot be replicated for the comparison group, an "entry" date must be determined for nonparticipants who do not enter the intervention.

In this paper, we present a new methodology for selecting comparison groups for rolling entry interventions: rolling entry matching (REM). Rolling entry matching selects individuals for the comparison group who are similar to intervention participants with respect to both static, unchanging characteristics (e.g., race, date of birth) and dynamic characteristics that change over time (e.g., health conditions, health care use). Accounting for observable dynamic variables is critical to comparison group selection in health care interventions because a participant's health and utilization often predicts entry into the intervention. For each participant, REM identifies potential comparison individuals with similar patterns of static and dynamic variables during the intervention participant's unique baseline period. Rolling entry matching then utilizes a matching algorithm to consider potential intervention-comparison matches for every entry period, selecting comparison individuals to act as a comparison in the entry period for which they are the best match.

We demonstrate REM using a case study of four Health Care Innovation Award (HCIA) grantees whose unique challenges for impact evaluation inspired the method. Health Care Innovation Award grantees were funded by the Center for Medicare and Medicaid Innovation (CMMI) to implement interventions aimed at improving health, delivering better care, and reducing costs among Medicare and Medicaid beneficiaries.¹ Most HCIA grantees enrolled participants on a rolling basis and some grantees had a relatively small number of participants from an evaluation perspective. Relative to other methods of comparison group selection for rolling entry, REM is best suited to interventions with small numbers of participants entering during each time period and intervention entry criteria that are not easily replicated for comparison group selection. Figure 1 provides a conceptual example of how REM selects a comparison group in

FIGURE 1 Rolling entry with health care expenditure spikes at entry

Notes: Top panel presents a hypothetical spending trend for two intervention entrants. The bottom panel presents hypothetical spending trends for comparisons. Dashed lines indicate the best-matched comparisons.

the absence of a replicable entry criteria. The top panel illustrates expenditure trends for two hypothetical intervention participants that enter at different times. Each participant has a different level of baseline health care expenditures and each participant experiences a spike in expenditures immediately preceding entry. As a simplification for illustrative purposes, assume that health care expenditure is the only dimension on which we are interested in matching participants and comparison nonparticipants. The second panel of Figure 2 illustrates expenditures trends for five potential comparison nonparticipants. The two best comparison nonparticipants are indicated by the dashed lines. Control 3 is a good match for entrant 1 because control 3's level of expenditure and expenditure spike is similar to entrant 1's during participant 1's baseline period before entry. For the same reasons, control 4 is the best match for participant 2. From the set of nonparticipants, REM selects the comparison(s) that are most similar to intervention participants prior to entry and assigns comparisons the entry date of their best-matched participant. Thus, REM solves the problem of how to determine an "entry" date for nonparticipants.

Rolling entry matching is an addition to a handful of statistical methodologies designed for selecting a comparison group for rolling entry interventions including the stepwise matching framework with outlier first matching,² balanced risk set matching,³ and sequential cohort.⁴⁻⁶ Each methodology takes a different approach.

In a graduate dissertation, Sun used generalized estimating equations to construct time-variant propensity scores for matching purposes, then demonstrates a matching process to select the comparison group. Balanced risk set matching uses not-yet or never-treated participants as the comparison group for participants already receiving treatment, estimating the effect of treating now vs later because untreated units generally become treated units. As in REM, outlier first matching and balanced risk set matching estimate

a single model to calculate distance for all participants who enter the intervention over time. Outlier first matching selects matches by searching for units with the smallest maximum distance while balanced risk set matching selects treatment-control pairs such that the total distance between pairs is minimized. In contrast, REM's matching algorithm is focused on finding at least one comparison match for all treated units in order to preserve sample size.

Sequential cohort matching matches blocks of participants who enter the treatment group around the same time to nonparticipants from a pool of currently available nonparticipants, repeating a separate matching process for each cohort. Rolling entry matching differs from sequential cohort matching in two important ways. First, REM estimates a single propensity score model for all cohorts. This is useful for studies with a small number of entrants in a block for which a propensity score model would begin to suffer from small-sample bias or perfect classification. Second, REM matches nonparticipants based on the entry period for which their baseline period is most similar to their matched participant. In sequential cohort matching, the order in which the cohorts are matched affects which entry period nonparticipants are selected as comparisons. For example, a nonparticipant selected as a comparison for cohort 1 might be a better match to a participant who enters in cohort 4, but the participant is no longer available to be matched in cohort 4 because they were selected in cohort 1. Rolling entry matching considers all entry periods when matching nonparticipants to participants. Sequential cohort matching is widely used and conceptually similar to REM; therefore, we implement our case study using cohort matching and compare its performance to REM.

The next section provides a detailed overview of implementing REM. Section 3 presents a case study of using REM and sequential cohort matching to select a comparison group for 4 HCIA grantees. Section 5 discusses limitations and Section 6 concludes.

ID	TIME	TREAT	SPENDING T-1	SPENDING T-2
1	3	1	4	2
2	8	1	2	1
3	1	0	2	2
3	2	0	2	2
3	3	0	4	2
3	4	0	1	4
3	5	0	1	1
3	6	0	1	1
3	7	0	1	1
3	8	0	1	1

FIGURE 2 Example rolling entry matching (REM) matching dataset [Color figure can be viewed at wileyonlinelibrary.com] Note. Figure shows example of a REM matching dataset for two participants and one nonparticipant.

2 | ROLLING ENTRY MATCHING

Rolling entry matching is a three step process. First, the matching dataset is constructed. Second, a measure of distance between participants and nonparticipants is generated (e.g., propensity score, Euclidean distance, or Mahalanobis distance). Third, the REM matching algorithm matches participants to nonparticipants using an algorithm that assesses every potential entry period for a nonparticipant and determines a single counterfactual "entry" period according to the best match (on static and dynamic characteristics) with a participant.

2.1 | REM matching dataset

Rolling entry matching requires a quasi-panel matching dataset containing one observation per participant and multiple observations (one for each potential entry period) per nonparticipant. Participants appear once in the data at their entry date, when the baseline variables to be matched on (e.g., demographics, spending,

utilization) are observed. Nonparticipants do not enter the intervention; therefore, a nonparticipant observation is created for each potential entry period. If participants enter the intervention in one of eight calendar quarters, eight observations of each nonparticipant are created with each observation corresponding to one of the entry quarters.

Consider a hypothetical intervention in which participants enter over eight quarters. Participants and nonparticipants will be matched using health care expenditure (in thousands) during the two quarters prior to entry. Figure 2 illustrates an example of a REM matching dataset for this hypothetical intervention for two participants and one nonparticipant. Participant 1 enters the intervention in quarter three and participant 2 enters the intervention in quarter eight, as noted in the TIME variable. Participants are denoted by their treatment status ($TREAT = 1$) and lagged expenditures in the two periods prior to the participant's entry quarter are included ($SPENDING_{T-1}$, $SPENDING_{T-2}$). The nonparticipant (ID = 3) appears in the data eight times, once for each potential entry quarter. Since spending is a dynamic variable, it changes for nonparticipant 3 depending on the potential entry quarter.

Rolling entry matching allows for estimation of a single propensity score model for participants who enter an intervention at different times by duplicating nonparticipants in the data. Participants and nonparticipants are then able to be matched in a single step rather than sequentially for each entry period or cohort. Although nonparticipants do not enroll in the intervention, a version of the nonparticipant exists for each potential entry period. As described in detail in Section 2.3, the matching algorithm will select the version of the nonparticipant from the quarter that he is best suited to act as a comparison.

2.2 | Distance calculation

The second step in REM is to generate a measure of distance between intervention participants and nonparticipants, usually the difference in propensity scores. There are two noteworthy

features of propensity scores in the context of REM. First, propensity scores will be smaller when estimated using the REM matching dataset. Smaller magnitude propensity scores are generated because the comparison group size is multiplied by the number of entry periods, making the probability of participation mechanically lower in the data.

The second noteworthy feature of REM is that intervention participants who enroll over several entry periods are pooled into a single propensity score model. This allows the researcher to overcome difficulties associated with small sample size in logit and probit models. Propensity score models may fail to converge in small samples due to perfect classification. Additionally, inclusion of too many independent variables can cause bias in the propensity score when the outcome has few events relative to the number of predictors; therefore, the number of independent matching variables often limits models to a fraction of the number of events in the data.⁷⁻⁹ By pooling participants from all entry periods into a single model and duplicating nonparticipants, the sample size and is larger and the number of matching variables can be increased.

2.3 | REM matching algorithm

After a measure of distance between participants and nonparticipants is generated, the REM matching algorithm selects nonparticipant comparison(s) for each intervention participant. Matching is conducted with the dual goals of: (1) maximizing the number of participants who receive a matched comparison and (2) selecting nonparticipants to act as comparisons during the entry period that they are most similar to their matched participants.

Figure 3 illustrates the idea behind the matching algorithm. Participant 1 enters during period three and can potentially be matched with nonparticipant 3 in period three. Likewise, participant 2 enters during period eight and can be matched with nonparticipant 3 in period eight. The difference in propensity scores between participant 1 and comparison 3 is 0.03. Likewise, the difference in propensity scores between participant 2 and comparison 3 is 0.05. Because the

ID	TIME	TREAT	SPENDING T-1	SPENDING T-2	PROPENSITY SCORE
1	3	1	4	2	0.13
2	8	1	2	1	0.15

ID	TIME	TREAT	SPENDING T-1	SPENDING T-2	PROPENSITY SCORE
3	1	0	2	2	0.14
3	2	0	2	2	0.13
3	3	0	4	2	0.10
3	4	0	1	4	0.11
3	5	0	1	1	0.17
3	6	0	1	1	0.18
3	7	0	1	1	0.14
3	8	0	1	1	0.10

FIGURE 3 Example rolling entry matching (REM) matching [Color figure can be viewed at wileyonlinelibrary.com]

Notes: Spending is in thousands. Arrows indicate which version of the comparison group each treatment participant could be matched with. A solid arrow indicates a match. A dashed arrow indicates a nonmatch. Comparison individuals are allowed to act as a match in 1 time period and are selected as a control in the period they are best matched. In this example, the comparison person 3 acts as a control for treatment person 1 because the propensity scores are the same.

TABLE 1 Grantee descriptions

Grantee	Grantee 1	Grantee 2	Grantee 3	Grantee 4
Target population	Patients diagnosed with a chronic disease, people at risk of developing diabetes, vulnerable seniors, homebound individuals, young children, and hard-to-reach county residents in a rural county. Additionally, patients over 65 with asthma and newly eligible adult Medicaid patients with diabetes, hypertension, or asthma	Adults, children, and pregnant women with chronic diseases, including hypertension, diabetes, and asthma and living in a high-risk zip code who do not have a primary care provider (e.g., frequent users of emergency departments)	Patients with chronic diseases or complex patients of a particular community health center	High-risk patients living in a rural county
Intervention description	CHWs provide health education and preventive care services, including immunization campaigns and participation in community events. They also offer intensive case management to support effective chronic disease management, including home visits	CHWs and PSRs are part of a 5-member blended clinical/nonclinical care team that provides home- and clinic-based primary care services and self-management education to frequent ED users. They work out of existing clinical sites and newly established micro-clinics in the targeted high-risk community	CHWs help identify patients at a high risk for frequent ED use, deliver individualized health coaching, establish linkages with support services, and facilitate health planning. The innovation involves both care coordination changes within the clinic and increased linkages with community resources	HNs offer health education, transportation, care coordination, benefit management, nutrition support, and active lifestyle training to high risk patients. HN's receive training through a newly-established CHW degree program
Entry criteria/ Recruitment	Patients recruited for home visits are identified in several ways: door-to-door recruitment, community events, referrals from community agencies, and disease registry reports	Recruitment and enrollment is the responsibility of the CHWs who are assigned to specific public housing units within the targeted zip code area, where they focus their enrollment efforts through a collaborative relationship with the housing managers. The CHWs implement targeted community outreach (e.g., neighborhood associations, community-based organizations, social services agencies), use various communication channels (i.e., tailored brochures, radio programs), and participate in community events (e.g., health fairs, festivals) to increase awareness about Innovations Health and identify high-need patients. Often, residents learned about the program from enrolled patients who share information with their neighbors and family members. Finally, other community organizations make referrals	A comprehensive health assessment is conducted at the time of an initial appointment to assign a patient to a risk category. There are no strict eligibility requirements to receive health coaching services, but patients with chronic diseases or high needs are prioritized. The provider identifies patients either by using a risk stratification scale that quantifies a patient's health risk by several domains and classifies patients by risk level ("low-complexity risk," "moderate risk," "high risk," and "super high risk") and refers the high- and super-high-risk patients to receive health coaching, or by judging that the patient could benefit from health coaching services	Patients who have listed grantee as their primary care provider, who do not have a primary care provider, or who need substance abuse or mental health services
Number of entrants	180	76	53	106
Entry period	2012Q4-2014Q3	2012Q4-2014Q2	2013Q1-2014Q3	2012Q3-2015Q2

difference in propensity scores is smallest, nonparticipant 3 would be selected as a comparison for participant 1 and not participant 2.

RTI International has developed an R package named *rollmatch* that estimates a propensity score model and implements the REM matching algorithm for users. R is an open source, statistical analysis software and the *rollmatch* package, documentation, and help files are downloadable for free.⁷ The *rollmatch* matching algorithm proceeds in three steps.

First, the pool of potential comparisons for each participant is limited to observations of nonparticipants with the same entry period. Thus, participants who enter in the first intervention period are only able to be matched with nonparticipant observations from the same period. The same holds for all entry periods. Next, a list of the best nonparticipants for each participant is created. Comparison(s) are selected for each participant to minimize the absolute value of the distance between the participant and the pool of eligible comparisons with the same entry period.

Matching for all entry periods is done at the same time to reduce overall imbalance between participants and nonparticipants. The *rollmatch* package allows one-to-one, one-to-many, or one-to-variable matching with or without a caliper. Weights can also be calculated to account for 1 to variable matching and matching with replacement. Matching with replacement allows a nonparticipant to act as a comparison multiple times in one entry period only. In this case, the nonparticipant acts as a comparison for multiple participants entering in the period, for which the average distance between the nonparticipant and matched participants is smallest.

Applying a caliper may result in some participants having no or few comparisons in their list of matches. Often, participants with no comparisons are dropped from the analysis sample. Because REM was designed with small interventions in mind, the matching algorithm contains a step to preserve participants who have one

comparison within the caliper. Matches for these participants are given priority. Participants with one comparison inside the caliper are assigned the comparison regardless of whether that comparison is a closer match to a participant from another entry period.

3 | CASE STUDY

Rolling entry matching was developed to evaluate the impact of CMMI-funded HCIA interventions aimed at improving health, delivering better care, and reducing costs among Medicare and Medicaid beneficiaries. We demonstrate REM using the four grantees with the smallest number of Medicare fee-for-service enrollees in the RTI International evaluation portfolio of HCIA grantees (Table 1). All four grantees targeted participants that were high-risk or had chronic conditions. Interventions included a combination of home visits, patient education, case management, care coordination, referrals to appropriate health care and other community resources, and transportation. Participants were recruited door-to-door, at community events, through provider and peer referral, at medical appointments, or from established patient lists. The number of Medicare fee-for-service entrants ranged from 53 to 180 over 7 or 8 entry quarters.

Figure 4 illustrates entry of fee-for-service Medicare participants into the HCIA grantees over time. The mean number of participants across interventions was 104 and entry occurred over 7.5 quarters on average, resulting in an average of 14 entrants per quarter. As shown in Figure 4, enrollment started low, ramped up, and then steadily declined for the four grantees. These HCIA grantees posed several challenges for comparison group selection: (1) participants entered over multiple entry periods, (2) a small number of participants entered in each entry period, and (3) the entry criteria were not replicable for selecting a comparison nonparticipant's

FIGURE 4 Participant entry into Health Care Innovation Award (HCIA) grantees by quarter

Note: Figure presents the number of Medicare fee-for-service entrants in each quarter between quarter 3 of 2012 and quarter 4 of 2015. Each series represents a different HCIA grantee.

“entry” period. Rolling entry matching was developed to solve these challenges.

In the following sections, we demonstrate the implementation of REM and compare its performance to an alternative comparison group selection methodology, sequential cohort matching. This example is illustrative for demonstrating the REM methodology and is not intended to be an example of program evaluation.

3.1 | Rolling entry matching

We implemented REM using Chronic Conditions Data Warehouse (CCW) fee-for-service Medicare Parts A and B claims data. Using information provided by grantees, participants were matched to the CCW data by social security number, health insurance claim number, or a combination of name, date of birth, gender, and geographic identifiers. Participants who were enrolled in fee-for-service Medicare during the quarter of entry and at least one quarter after entry were included in the sample.

For each grantee, the potential comparison group was drawn from nonparticipants in the geographic area of the intervention. Participating and nonparticipating individuals were matched using a logit model predicting the likelihood that a beneficiary was enrolled in the innovation. The propensity score model varied across grantees because different variables were relevant for predicting entry into different interventions. Furthermore, grantees with fewer participants were matched with a smaller number of variables. Matching variables included traditionally fixed variables (gender, race) and dynamic variables that have the potential to change over time (age, disability, end-stage renal disease status, number of dual-eligible months, number of chronic conditions, hypertension diagnosis, diabetes diagnosis, asthma diagnosis, number of ED visits and inpatient stays in the calendar quarter and year prior to entry, and total Medicare payments in the calendar quarter and calendar year prior to entry). Participants in grantee 1’s intervention were matched using all the aforementioned variables and participants in the other three interventions were matched using a subset. We suggest that researchers choose matching variables that are most relevant and appropriate for the application they are considering. Specific guidance on selecting matching variables is provided in Stuart.¹⁰

We used one-to-variable matching with replacement, matching each innovation participant with up to three nonparticipants with the closest propensity score. A caliper of 20 percent of the standard deviation of the logit of the propensity scores was applied and participants with no comparisons inside the caliper were dropped from the analytic sample. Cochran and Rubin¹¹ estimate that a caliper of 20 percent of the standard deviation will remove 98 percent of bias in a normally distributed covariate, while Rosenbaum and Rubin¹² suggest that a caliper of 25 percent is acceptable. We implemented the REM matching algorithm to match comparison individuals to participants. First, participants with one comparison individual inside the caliper were assigned that comparison person. We then considered participants with multiple comparisons. Among the set of first-choice comparisons,

we assigned nonparticipants to act as comparisons in the quarter with the smallest average difference between the comparison and his matched participants. Once all participants were assigned their first comparison, we repeated the process to select up to three comparisons within the caliper for each participant. Nonparticipants could act as a comparison for multiple participants who enter during the same quarter, but not for participants entering in different quarters.

3.2 | Sequential cohort matching

Sequential cohort matching was implemented as a comparison to REM. We divided each grantee’s participants into two entry cohorts constructed such that cohort size was as close to equalized as possible. For each grantee, we estimated a propensity score model using participants in the first cohort and the full set of nonparticipants. The variables in the propensity score model were the same variables used in the REM propensity score model. Participants were matched with up to three nonparticipants using one-to-variable matching with replacement. As in REM, we applied a caliper of 20 percent of the standard deviation of the logit of the propensity score, eliminating matches outside of the caliper distance. After completing matching for the first cohort, nonparticipants that were selected to act as a comparison were removed from the pool of potential comparisons. The propensity score model and matching process were then repeated for the second cohort.

4 | RESULTS

The goal of matching is to increase similarity between the intervention and comparison group, thereby reducing bias in the estimated intervention outcomes. In the case of the HCIA grantees, the small number of entrants into each intervention means that an analysis of outcomes will be underpowered to detect significant treatment-comparison differences unless they are large. For this reason, we report the absolute standardized difference after matching for variables in the propensity score model as a measure of balance between the intervention and comparison group. Balance between the intervention and matched comparison group imitates a randomized experiment, reduces bias in the estimated treatment effect, and improves confidence in a causal interpretation of outcome effects. Standardized differences close to zero indicate better matches and a threshold of 0.20 for most variables is considered adequate balance.^{13,14}

Figure 5 summarizes the results of implementing REM and cohort matching for the four grantees. For each variable in the propensity score model, the absolute standardized difference between the participants and nonparticipants after REM and cohort matching is shown. Additionally, the average standardized difference across all variables in the model is presented at the top of each subfigure. Standardized differences for REM are represented by circles and by crosses for cohort matching.

REM achieved better balance between the intervention and comparison group for three out of four grantees. For grantee 1, sequential cohort matching outperformed REM as measured by the average

FIGURE 5 Absolute standardized difference after matching: rolling entry matching and sequential cohort matching

Notes: Graph shows absolute standardized differences matching for four Health Care Innovation Award grantees. Each grantee is represented in a separate panel with the variables used in the propensity score model listed on the vertical axis. Hollow circles represent rolling entry matching and plus signs represent cohort matching.

standardized difference across variables in the propensity score model. For grantees 2, 3, and 4, the average standardized difference between the intervention and comparison group is smaller when using REM than sequential cohort matching. Across all variables for the four grantees, matching with REM resulted in one variable with a postmatching standardized difference >0.2 and sequential cohort matching resulted in three variables with a standardized difference >0.2 .

In this example, REM performs well relative to sequential cohort matching when there are a small number of entrants into an intervention. Sequential cohort matching outperformed REM for grantee 1, which was the largest intervention with 180 participants. Grantees 2, 3, and 4 had 76, 53, and 106 entrants, respectively and were better-matched using REM.

5 | LIMITATIONS

There are drawbacks to REM that researchers should consider. First, REM's method of running a single propensity score model for all entry periods is potentially less flexible at the propensity score

model stage than sequential cohort matching's method of running a separate propensity score model for each cohort. Sequential cohort matching allows for the possibility that characteristics predicting entry will change over time through a separate propensity score model for each cohort. Rolling entry matching, on the other hand, applies the same model to all cohorts. Although REM could theoretically achieve the same flexibility by interacting all matching variables with period of entry, this would increase the number of variables in the propensity score model considerably. The need for a different model across cohorts can be empirically assessed by estimating propensity score models across cohorts and testing the equality of coefficients across cohorts. In cases where equality cannot be rejected, implementing REM using a single propensity score model is an option for the researcher. Given the previously discussed challenges with model convergence and sample size, REM is unlikely to be able to flexibly capture changing characteristics predicting entry over time. By running a single model for all cohorts, REM overcomes small sample size issues at the expense of having a unique model for each cohort.

Second, REM did not universally outperform sequential cohort matching. In this case study, REM outperformed sequential cohort matching for three out of four grantees. Rolling entry matching performed better for the three grantees with the smallest enrollment. For this reason, REM may be best for use in interventions with relatively small enrollment. Future work should further evaluate the performance of REM relative to other matching methods in a variety of study settings.

Third, the REM matching algorithm does not minimize the global distance between treatment and comparison units. The rollmatch R package minimizes the distance for a participant within each time period, but not globally across all participants and time periods. Minimizing the global distance is an area for future development of the rollmatch R package.

6 | CONCLUSION

For interventions lacking random assignment, rolling entry complicates comparison group selection because a counterfactual entry date for comparison individuals is not observed but must be assigned in order to make pre/post intervention comparisons between the intervention and comparison group. Rolling entry matching is a new methodology for comparison group selection in the presence of rolling entry. Rolling entry matching selects the “entry” period of nonparticipants by assigning the entry period of the participant to the matched comparison. For each participant, comparison(s) are selected such that they have a similar pattern of static and dynamic variables during the participant’s baseline period. Rolling entry matching accommodates the dynamic nature of health care utilization by including data immediately preceding entry in the matching process, regardless of when the participant enrolled.

Rolling entry matching is best suited to interventions with certain characteristics. First, participants enter the intervention over time. Second, intervention entry criteria are not replicable in the comparison group, meaning an “entry” date must be selected for comparison individuals. Third, when entry depends on criteria that change over time such as health care utilization or conditions, REM is useful for selecting a comparison individual that has similar values for dynamic variables *at the same time* as the intervention participant. Lastly, REM is particularly useful for interventions in which a small number of participants enroll per period because all participants are pooled into a single propensity score model, circumventing small sample size issues with logit and probit models. Rolling entry matching’s matching algorithm is designed to maintain sample size by prioritizing comparison selection for participants with one comparison match. Users can implement REM using the rollmatch package for R.

In the changing health care landscape, interventions are being implemented in a flexible manner for practical reasons, to maximize enrollment, or even incorporate lessons learned into the intervention on an ongoing basis. Impact and evaluation methodology

must also advance to provide impact estimates. Reliable estimates hinge on valid comparison groups acting as the counterfactual case for the treatment group in the absence of the intervention. In this paper, we demonstrated REM as a new methodology for comparison group selection in the presence of rolling entry into an intervention. Ongoing research should test REM in a variety of intervention settings and compare it to existing methodologies used for rolling entry interventions.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: We gratefully acknowledge funding from the Center for Medicare and Medicaid Innovation (HHSM-500-2010-000211) and b International that inspired and directly supported this research. Kumar Jeyaraman, Amy Huebeler, Douglas Kendrick, Mahin Manley, and Robert Chew provided substantial input into the development of rolling entry matching.

CONFLICT OF INTEREST

None.

ENDNOTE

* <https://github.com/RTIInternational/rollmatch>

ORCID

Allison Witman <https://orcid.org/0000-0001-7467-2369>

REFERENCES

1. Rojas Smith L, Amico P, Hoerger T, Jacobs S, Payne J, Renaud J. Evaluation of the Health Care Innovation Awards: Community Resource Planning, Prevention, and Monitoring: Third Annual Report. Prepared for the Center for Medicare and Medicaid Innovation, Baltimore, MD; 2017.
2. Sun Y. New matching algorithm – outlier first matching and its performance on propensity score analysis under new stepwise matching framework. 2014. Doctoral dissertation. PQDT Open, <http://pqdtopen.proquest.com/pubnum/3633233.html>
3. Li YP, Propert KJ, Rosenbaum PR. Balanced risk set matching. *J Am Stat Assoc*. 2001;96(455):870-882.
4. Seeger JD, Williams PL, Walker AM. An application of propensity score matching using claims data. *Pharmacoepidemiol Drug Saf*. 2005;14:465-476.
5. Gagne JJ, Glynn RJ, Rassen JA, et al. Active safety monitoring of newly marketed medications in a distributed data network: application of a semi-automated monitoring system. *Clin Pharmacol Ther*. 2012;92:80-86.
6. Schneeweiss S, Gagne JJ, Glynn RJ, Ruhl M, Rassen JA. Assessing the comparative effectiveness of newly marketed medication: methodological challenges and implications for drug development. *Clin Pharmacol Ther*. 2011;90(6):777-790.
7. Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards

- analysis. I. Background, goals, and general strategy. *J Clin Epidemiol*. 1995;48:1495-1501.
8. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. accuracy and precision of regression estimates. *J Clin Epidemiol*. 1995;48:1503-1510.
 9. Peduzzi P, Concato J, Kemper E, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49:1373-1379.
 10. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1-21.
 11. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. *Sankhya Ser A*. 1973;35:417-446.
 12. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*. 1985;39:33-38.
 13. 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.
 14. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399-424.

SUPPORTING INFORMATION

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

How to cite this article: Witman A, Beadles C, Liu Y, et al. Comparison group selection in the presence of rolling entry for health services research: Rolling entry matching. *Health Serv Res*. 2019;54:492-501. <https://doi.org/10.1111/1475-6773.13086>