Paper AA02-2014

Comparing regression, propensity matching and coarsened exact matching in healthcare observational studies using SAS®: An example from the Medical Expenditure Panel Survey (MEPS)

Doug Thompson, Blue Cross Blue Shield of IL, MT, NM, OK & TX, Chicago, IL

ABSTRACT

Healthcare expenditures in the United States have increased at an unsustainable rate, reaching 18% of GDP in 2011. There is a need to find ways to reduce healthcare spend while maintaining healthcare quality. Numerous interventions to reduce healthcare spend have been proposed and implemented. It is crucial to accurately evaluate which interventions are most effective. The gold standard method for evaluation is the randomized controlled trial (RCT), but this is sometimes unfeasible. Observational studies are an alternative when RCTs are not feasible. One aspect of observational studies is to balance the groups being compared (e.g., an intervention group and a comparison group) with respect to measured baseline characteristics, using methods such as regression or matching. After balancing on baseline characteristics, post-intervention outcome measures are compared between the groups. This presentation compares three methods to balance groups on baseline characteristics (regression, propensity matching, and coarsened exact matching). The methods are illustrated in the context of a study comparing healthcare costs in private “gatekeeper” insurance plans (which typically involve care coordination through a primary care physician) with other private insurance plans, using data from the Medical Expenditure Panel Survey (MEPs). Differences between the methods are summarized and SAS® code to implement the methods is described.

INTRODUCTION

Healthcare expenditures in the United States have increased at an alarming rate, rising from 13% of GDP in 2000 to 18% in 2011 (data from World Bank). This rate of increase is unsustainable. There is a need to find ways to reduce healthcare spend while maintaining healthcare quality.

Numerous interventions to reduce healthcare spend have been proposed and implemented. These include disease management programs aimed at closing gaps in care; interventions to prevent avoidable utilization (e.g., avoidable hospital readmissions, unnecessary tests and scans, and high cost elective surgeries); and value based care models (including patient centered medical homes and accountable care organizations) aimed at incenting providers to administer healthcare efficiently while maintaining high quality and patient satisfaction.

It is important to accurately evaluate which interventions are most effective. This may be relatively straightforward in randomized, controlled tests (RCTs). However, barriers to randomizing patients to interventions (ethical, legal, and other) may prevent RCTs.

When an RCT cannot be conducted, other methods are needed to evaluate intervention outcomes. A primary challenge to evaluating outcomes of non-randomized interventions is self-selection bias (for discussions of self-selection specific to healthcare, see Hromadkova, 2009 and Reeve et al, 2008). Individuals who choose to participate in an intervention may differ from individuals who choose not to participate. To evaluate intervention outcomes, it may be necessary to compare a group of individuals choosing to participate with a group choosing not to participate (depending on the data available). Some differences between these groups may be measureable while other differences are not. Randomization tends to equalize differences between groups with respect to both measured and unmeasured characteristics. Without randomization, it is impossible to guarantee that groups are equivalent prior to the intervention (at “baseline”) with respect to unmeasured characteristics. If groups are not equivalent at baseline, group differences in outcomes (e.g., healthcare spend) may be due to pre-existing group differences rather than to the intervention itself.

Observational studies are an alternative when RCTs are not feasible. Observational studies typically capitalize on existing data (e.g., claims data, surveys, disease registries) to examine how variation in a characteristic is associated with variation in an outcome (Stroup et al, 2000). For example, the characteristic might be participation vs. non-participation in a disease management program and the outcome might be healthcare spend.

Observational studies attempt to approximate the design of RCTs as much as possible (Rubin, 2008; Rosenthal, 2009). One aspect of this is to equalize groups with respect to measured baseline characteristics, while acknowledging that the groups may still differ on unmeasured characteristics. To balance groups on baseline characteristics, observational studies use methods including: 1) regression adjusting for baseline characteristics, and 2) match intervention participants with one or more non-participants who have similar baseline characteristics.

There are advantages and disadvantages of every method for creating equivalence on measured baseline characteristics. Regression makes conceptual sense, it is familiar to researchers and it is relatively easy to implement. However, some studies have pointed out shortcomings of regression adjustment relative to matching techniques for creating baseline equivalence between groups (e.g., Dehejia & Wahba, 1999; Reeve et al, 2008; Hrmadkova, 2009). These include the finding that propensity matching produces estimates that more closely approximate results from a randomized design than regression adjustment alone (Dehejia & Wahba, 1999). Also, propensity matching can effectively adjust for more covariates than regression adjustment, using a 2-stage approach (Reeve et al, 2008). Standard errors for intervention effect estimates are smaller in propensity matched-designs than with regression adjustment (Reeve et al). AHRQ has recommended matched control designs as the next best alternative to randomized controlled tests for evaluating healthcare interventions such as patient centered medical homes (Meyers et al, 2011).

On the other hand, regression adjustment has some advantages over matching. In regression, all observations can be used, while matching may require discarding observations for which a reasonably close match cannot be found. In addition, matching may involve more choices (e.g., width of calipers, matching techniques such as greedy vs. optimal, number of matches to use such as 1:1 vs. 1:many) which could lead to subjectivity and manipulation of results.

Matching has several variants. The most common matching approach is to match on a propensity score (Austin et al, 2010; Wells et al, 2013). However, some researchers have more recently advocated coarsened exact matching (CEM; Iacus et al, 2011; Wells et al, 2013). Advantages of CEM relative to propensity matching are described by Iacus et al (2011). Briefly, advantages include the fact that increasing balance on one variable cannot increase imbalance on another in CEM (this can happen in propensity matching); ease of implementation; less sensitivity to measurement error; and greater computational efficiency. The intuitive appeal of exact matching is another advantage of CEM.

The goal of the present study is to illustrate the implementation of several observational study approaches using SAS, and to discuss practical tradeoffs of different approaches. An example is presented, showing how an observational study design might be used to evaluate the association of enrolling in a private “gatekeeper” health insurance plan with healthcare spend. “Gatekeeper” plans attempt to manage care through a single source (e.g., a primary care physician), in contrast to other “non-gatekeeper” plans which pay a set amount for specific services rendered without requiring a single source to manage the services. Gatekeeper plans sometimes incentivize healthcare efficiency by paying providers a fixed, pre-set amount per patient regardless of the services that the patient actually receives (capitation). Providers can maximize their financial returns by maintaining a high level of healthcare quality while minimizing the cost of services administered. Thus, relative to private non-gatekeeper plans, enrollment in a private gatekeeper plan is expected to result in lower spend due to providers’ financial incentive to manage care efficiently. More recent “value based care models” such as patient centered medical homes and accountable care organizations have sometimes been viewed as similar to gatekeeper arrangements, because providers are incentivized to efficiently manage a patient’s care.

The example uses data from the Medical Expenditure Panel Survey (MEPS). Several past studies used MEPS data to examine the association of type of private insurance with outcome measures (Shi et al, 2000; Escare et al, 2001; Newacheck et al, 2001; Hromadkova, 2009; Stanton et al, 2002). However, to our knowledge, Hromadkova (2009) was the only study to specifically compare "gatekeeper" private insurance plans (HMO or other private gatekeeper plans) with non-gatekeeper plans using MEPS data. The present study is similar to Hromadkova (2009), in that both used MEPS data and both looked at outcomes associated with enrollment in gatekeeper vs. non-gatekeeper private insurance plans. The studies differ in several respects: 1) the present study used more recent MEPS data; 2) the studies looked at different outcome variables (spend in the present study vs. initial provider choice and utilization measures in Hromadkova, 2009); 3) the present study looked at different design and matching variants not used by Hromadkova, 2009; and 4) the present study provides a detailed walk-through of the SAS® code required to execute the methods.

METHOD

DATA

Data were from the Medical Expenditure Panel Survey (MEPS), panel 14. This was a two year longitudinal survey of individuals who participated in interviews in both 2009 and 2010.

In MEPS, each respondent in a panel participates in 5 "rounds" of data collection across two consecutive years. For panel 14, these years were 2009 and 2010; rounds 1 (R1) and 2 (R2) were in 2009 and rounds 4 (R4) and 5 (R5) were in 2010. Round 3 (R3) was in 2009 for some respondents and 2010 for others. For purposes of this study, R2 measures and time-invariant measures (e.g., gender, race/ethnicity) were used as baseline measures. The intervention was defined as type of insurance at the end of 2009. Healthcare spend was measured across all of 2009 (baseline) and across all of 2010 (the study outcome).

Healthcare spend was the total paid by all sources, including patient out-of-pocket expenses, taking discounts into account; therefore, it similar to what payors would refer to as allowed amount. Payment information was collected from individual respondents as well as all hospitals (and associated physicians), emergency rooms, outpatient departments home health agencies and pharmacies, and a sample of office-based physicians. Information collected from providers was used if complete, otherwise information collected from individual respondents was used.

The definition of gatekeeper (GK) vs. non-gatekeeper (NGK) plans in MEPS data follows that of Hrmadkova (2009): GK plans were defined as either private HMO or other private GK plans, while NGK plans were defined as any other private insurance.

STUDY DESIGN

We followed guidance from Rubin (2008) and Rosenthal (2009) to design an observational study to approximate (to the extent possible) the design of an ideal RCT of the intervention. An ideal RCT might take a large sample of individuals who are eligible for private insurance and randomize them to either private GK or private NGK plans. At baseline, the individuals will represent a mix including ones who already have private GK or NGK plans, ones who have public insurance, and ones who have no health insurance. Due to random assignment, some individuals assigned to a given private insurance plan type will already have that same plan type at baseline, while others will not. However, random assignment will balance baseline plan type across the groups, therefore post-randomization group differences in outcomes will not be due to baseline group differences in plan type.

Although this appears to be a reasonable design for estimating outcomes associated with being enrolled in private GK vs. NGK plans, it does not necessarily reflect what one would expect to happen if individuals enrolled in a private NGK plan switched to a private GK plan (or vice versa). To study the effects of this, a different randomized test would be ideal – namely, take individuals enrolled in private NGK plans at baseline, and randomize them to either stay in the NGK plan or switch to a GK plan (or vice versa). Because switching plans per se might impact outcomes, rather than switching to a specific plan type, it is important to study plan switches in both directions.

Timing of measurement is critical in both RCTs and observational studies. In an RCT, an intervention is implemented at a specific time point and outcomes are measured at some time after the intervention; randomization typically ensures that groups are equivalent right before the intervention started (at baseline). In an observational study, groups are identified at a specific time point and outcomes are measured at some time point after group identification; methods (e.g., matching, regression) are used to create groups that have a similar distribution of measured baseline characteristics.

Three observational studies corresponding to the ideal RCT designs were conducted:

Study 1: *Outcomes associated with participating in private GK vs. NGK plans*. Identify individuals in private GK plans and private NGK plans at the end of 2009. Use methods to ensure that the two groups are equivalent on baseline medical spend and other characteristics measured earlier in 2009. Examine group differences in medical spend across 2010.

Study 2: *Outcomes associated with switching to a private GK plan*. Identify individuals in private NGK insurance plans earlier in 2009 (rounds 1 and 2). Identify ones who switched to private GK plans and ones who remained in private NGK plans at the end of 2009. Use methods to ensure that the two groups are equivalent on baseline medical spend and other characteristics measured earlier in 2009. Examine group differences in medical spend across 2010.

Study 3: *Outcomes associated with switching to a private NGK plan*. This is like Study 2, except the focus is on individuals who had private GK insurance plans at baseline, and look at those who switched to private NGK plans vs. remained in private GK plans.

CREATING BALANCE ON BASELINE VARIABLES

In each study, healthcare spend across 2010 was compared between groups. Separate analyses were conducted using three methods to create group equivalence on observed baseline characteristics: regression without matching, propensity matching, and coarsened exact matching (CEM).

Propensity matching is the most common matching method used in observational studies. The propensity score represents an individual's probability of being assigned to a specific group, conditional on the observed baseline measures. Logistic regression was used to create the propensity scoring algorithm; the dependent variable was group (e.g., GK vs. NGK) and the independent variables were measured baseline characteristics.

Instead of matching on a score that summarizes the baseline characteristics, coarsened exact matching (CEM) requires an exact match on all baseline characteristics. This method involves several steps. First, strata are formed by taking all possible levels of the measured baseline characteristics. Continuous variables are "coarsened" by categorizing prior to creating the strata. Then individuals in each group (e.g., GK="treatment" vs. NGK="control") are placed into the appropriate stratum. Strata including at least one individual in each group are retained in the analysis, while all other strata (and the individuals in them) are excluded. A weight is created for each unit in the retained strata. The weight is defined as follows. Let m_T be the count of all treatment units in all strata retained in the analysis (i.e., strata with at least one treatment and one control individual) and m_C the count of all control units across all strata. For stratum *s*, $m_{T}s$ is the count of treatment units in s and $m_{C}s$ the count of controls in *s*. Then for each unit i in *s*, the weight (w_i) is 1 if i is an treatment unit and $(m_C/m_T)*(m_{T}s/m_{C}s)$ if i is a control unit. Subsequent analyses use the matched units along with weights (see Iacus et al, 2011, for a more detailed description).

With matching methods, it is necessary to choose which baseline covariates to match on. While in theory there is no limit on the number of baseline covariates that can be used, there are practical limitations. This is most obvious for CEM – if the strata are too complex, many individuals can be excluded because exact matches cannot be found.

In this study, the full baseline covariate set consisted of household size, household income, employment status, geographic region, residence in MSA, race/ethnicity, age, gender, education, self-rated health, self-rated mental health, born in US, baseline total healthcare spend, has usual healthcare provider, usual healthcare provider is PCP, and end date of the R3 interview. In addition, Study 1 included baseline insurance type (GK or NGK).

In preliminary analyses, nearly all individuals were excluded in CEM using the full set of baseline covariates (6,695 of 6,765, or 99%, were excluded, leaving only 70 observations), even with very simple median splits to coarsen continuous variables. This occurred because the strata were very complex, thus it was rare to find a matching NGK member within any stratum containing a GK member. Individuals in a stratum without at least one comparison group member in the same stratum are excluded in CEM. Due to near total exclusion of observations using the full baseline covariate set, it was necessary to use a reduced set of baseline covariates for CEM. The reduced set included household income, race/ethnicity, age, gender, self-rated health, self-rated mental health, baseline total healthcare spend, has usual healthcare provider, usual healthcare provider is PCP, and end date of the R3 interview, as well as baseline insurance type (GK or NGK) in Study 1.

Using the full baseline covariate set created no apparent problems for propensity matching, therefore two separate propensity matching analyses were conducted: 1) using the full covariate set, and 2) using the reduced covariate set. CEM using the full covariate set was unfeasible as noted above, but CEM using the reduced covariate set was feasible and resulted in a substantial number of individuals being retained in the analyses.

Goodness of matching by each method was evaluated using three summary metrics. First, for each baseline variable, the standardized difference ("d") before vs. after matching was computed as recommended by Austin et al (2010). This measure was then averaged across all baseline variables to compute "average d." Although there are no hard-and-fast rules for determining what level of standardized difference is "large," 0.10 has sometimes been proposed as a cutoff for a noteworthy standardized difference between variables before vs. after matching. Therefore, a second measure of goodness of matching was percent of baseline variables where the standardized difference before vs. after matching was greater than 0.10 ("% d>0.10"). Both of these metrics ("average d" and "% d>0.10") give equal weight to each baseline variable in computing overall goodness of matching. However, one might want to give more weight to variables that have a stronger independent association with group membership, as measured (for example) by magnitude of the corresponding coefficient in multivariate logistic regression. Therefore, a third metric was computed by taking the average standardized difference ("d"), weighted by the magnitude of the standardized coefficient from a logistic regression model where group membership (GK vs. NGK) was the dependent variable and the independent variables were the baseline covariates. This third metric is denoted "weighted average d."

ANALYSIS

The natural logarithm of total healthcare spend across 2010 was compared between groups (GK vs NGK in Study 1, switch to GK vs remain NGK in Study 2, and switch to NGK vs remain GK in Study 3). Spend estimates are reported on the natural logarithm scale. Another option would be to transform the estimates back to the original scale and apply a smearing adjustment to account for the back-transformation bias, but this often results in unsatisfactory estimates. Other common analysis options include 2-part models and gamma regression (Diehr et al, 1999).

Matching analyses also adjusted for residual differences in baseline characteristics using regression, which is a common and recommended approach (Austin et al, 2010).

SAS CODE

The SAS code for propensity matching was based on Austin et al 2010 (code available from SAS website, http://support.sas.com/publishing/bbu/zip/61876.zip, downloaded in March, 2014). Propensity matching also involved the %gmatch macro developed by Erik Bergstralh and Jon Kosanke of the Mayo Clinic (macro available from Mayo Clinic website, http://www.mayo.edu/research/departments-divisions/department-health-sciences-research/division-biomedical-statistics-informatics/software/locally-written-sas-macros). Matches were 1:1 with a caliper width of 0.2 standard deviations of the logit of the propensity score.

A new macro, %cem_weights, was created to calculate the weights needed for coarsened exact matching (CEM). Prior to running the macro, a stratum variable was created by concatenating the variables to be used in matching (e.g., stratum=var_1||var_2||var_3||…||var_n). For each continuous variable, a binary indicator (coded 1=above median, 0=otherwise) was created as an input to defining the stratum variable. The %cem_weights macro takes as inputs the stratum variable (stratum=); an input data set (inset=); the group variable, assumed to be binary and taking values of 1 or 0 (group=), for example GK=1 or NGK=0; and a name to be assigned to the CEM weight variable that is created by the macro (weightname=). As a first step, the macro counts the number of observations within each stratum, separately for each group. Next, for each stratum, the macro flags whether the stratum has at least one observation within each group; if so, the stratum is retained for further analysis (SAS data set stratum_with_match) and if not, the stratum is discarded. In the next step, the macro counts the number of observations in each group retained in the analysis (after discarding observations in strata without at least one observation within each group) and places the resulting counts in the variables sum_freq0 and sum_freq1 for the groups denoted 0 and 1, respectively. The resulting counts are recorded in SAS data set sumcase. After this, the macro computes the CEM weights using the following steps:

```
data _null_;
set sumcase;
mc_mt=sum_freq0/sum_freq1;
call symput('mc_mt',mc_mt);
run;

data stratum_with_match_wt;
set stratum_with_match;
control_weight=&mc_mt*(freq1/freq0);
case_weight=1;
keep &stratum case_weight control_weight;
run;
```

These weights are then merged onto the original data set and each observation gets either its appropriate weight, or a missing weight (for observations excluded from the analysis due to being in a stratum without at least one observation per group), executed as follows:

```
data &inset;
merge &inset stratum_with_match_wt(in=a);
by &stratum;
if a then do;
   if &groupvar=1 then &weightname=case_weight;
   if &groupvar=0 then &weightname=control_weight;
end;
drop case_weight control_weight;
run;
```

After creating CEM weights using the %cem_weights macro, CEM analyses can proceed simply by using the CEM weight in any desired SAS procedure that has a weight statement, for example:

```
proc reg data=inscope_with_strata;
weight cem_weight;
model ln_totexpy2 =
prv_gatekeeper_y1
totexpy1
...
have_usc;
run;
quit;
```

To facilitate comparison of different methods with respect to goodness of matching on baseline variables, another macro, %compare_dist_by_grp, was created. This macro executes methods described by Austin et al (2010) for estimating standardized differences between groups on each baseline variable. The macro will not be described in detail here, but its key step is the following:

```
data newdata;
  length _NAME_ $ 50;
  merge g10 g11;

format mean_0 10.4;
format mean_1 10.4;

%if &curr_binary = 0 %then %do;
  d = (mean_1 - mean_0)/ sqrt((s_1*s_1 + s_0*s_0)/2);
  d = round(abs(d),0.001);
%end;
%else %if &curr_binary = 1 %then %do;
  d = (mean_1 - mean_0)/ sqrt((mean_1*(1-mean_1) + mean_0*(1-mean_0))/2);
  d = round(abs(d),0.001);
%end;
  _NAME_="&curr_var";
  _NAME_=upcase(_NAME_);

  keep d mean_1 mean_0 _NAME_ label;
run;
```

SAS data sets g10 and g11 contain means and standard deviations for the study groups denoted 0 and 1, respectively, for a specific baseline variable (the macro cycles through this process for each baseline variable, &curr_var, separately). A different calculation is used depending on whether the baseline variable is binary or continuous (the only options available in this macro).

RESULTS

Study 1: *Outcomes associated with participating in private GK vs. NGK plans*. This analysis was limited to individuals younger than age 65 who had either a private GK or NGK plan at the end of 2009. Prior to matching, the inscope sample consisted of 3,240 private gatekeeper (GK) insured individuals and 3,525 insured by private non-gatekeeper plans (NGK). Table 1 shows the distribution of baseline characteristics for the GK and NGK groups, in the original sample and after matching. Prior to matching, there were substantial differences on baseline characteristics between the GK and NGK groups (mean d=0.31, weighted mean d=0.26, % d >0.10=28%). These differences tended to be reduced by matching. The best match on baseline characteristics was achieved from propensity matching on the more extensive set of characteristics (mean d=0.03, weighted mean d=0.03, % d >0.10=4%). Propensity matching on the more limited set of characteristics resulted in the next best match (mean d=0.07, weighted mean d=0.06, % d >0.10=24%) and CEM the worst (mean d=0.11, weighted mean d=0.08, % d >0.10=44%).

Table 1. Mean or percent of baseline covariates and standardized difference (“Diff”) by group and matching method (Study 1).

	No matching			Propensity matched 1			Propensity matched 2			Coarsened exact		
Variable	**Non-GK**	**GK**	**Diff.**	**Non-GK**	**GK**	**Diff.**	**Non-GK**	**GK**	**Diff.**	**Non-GK**	**GK**	**Diff.**
Family size	3.4	3.5	0.06	3.3	3.5	0.11	3.3	3.4	0.06	3.1	3.6	0.28
Northeast	14%	16%	0.06	17%	13%	0.11	16%	15%	0.02	20%	15%	0.14
Midwest	25%	21%	0.09	18%	20%	0.03	20%	20%	0.01	21%	21%	0.00
South	40%	27%	0.27	40%	32%	0.16	39%	34%	0.11	33%	29%	0.08
In MSA	83%	93%	0.32	81%	91%	0.28	88%	89%	0.04	83%	92%	0.28
Age	33.2	33.6	0.02	31.7	31.3	0.02	31.4	31.8	0.02	32.8	32.6	0.02
Female	51%	52%	0.02	48%	49%	0.02	49%	48%	0.01	54%	54%	0.00
Asian	8%	12%	0.13	11%	11%	0.02	11%	12%	0.02	15%	10%	0.16
Black	16%	17%	0.02	16%	14%	0.04	17%	15%	0.06	7%	13%	0.22
White	75%	72%	0.08	74%	75%	0.03	73%	74%	0.02	79%	79%	0.00
Hispanic	12%	23%	0.29	12%	21%	0.23	13%	16%	0.09	13%	21%	0.21
Married	47%	44%	0.06	44%	40%	0.09	41%	41%	0.00	52%	44%	0.15
Educ. HS or less	43%	44%	0.02	44%	42%	0.04	43%	41%	0.06	27%	41%	0.31
Educ. College	26%	24%	0.05	26%	24%	0.03	25%	26%	0.02	43%	26%	0.36
Health	2.0	2.0	0.04	2.1	2.0	0.08	2.1	2.0	0.04	1.9	1.8	0.06
Mental health	1.8	1.8	0.01	1.9	1.8	0.08	1.9	1.8	0.05	1.7	1.6	0.11
Born in US	89%	81%	0.22	84%	79%	0.12	84%	81%	0.06	82%	85%	0.06
Have usual care provider	81%	84%	0.07	74%	75%	0.04	75%	76%	0.02	85%	85%	0.00
Usu. care prov. is PCP	38%	36%	0.04	35%	38%	0.06	36%	38%	0.04	33%	33%	0.00
Employed	60%	61%	0.01	61%	60%	0.00	60%	61%	0.02	57%	60%	0.05
Household income	$80,280	$75,701	0.09	$70,430	$69,348	0.02	$72,322	$71,549	0.02	$86,579	$82,911	0.07
Healthcare spend	$3,281	$3,192	0.01	$2,779	$2,971	0.02	$2,880	$3,109	0.03	$2,700	$3,176	0.06
Round 3 end date	18337	18337	0.01	18338	18337	0.01	18338	18337	0.02	18340	18334	0.15
Had GK R1-2	4%	85%	2.86	21%	21%	0.00	23%	23%	0.00	74%	74%	0.00
No GK R1-2	87%	5%	2.90	25%	25%	0.00	27%	27%	0.00	11%	11%	0.00

Note: “Propensity matched 1” involved matching on the reduced set of baseline characteristics (the same ones used in CEM) while “Propensity matched 2” involved matching on the full set of baseline characteristics.

For propensity matching, good matches were bought at the expense of excluding more respondents from the analysis, whereas CEM resulted in a poorer match but retained more respondents. The matching algorithms resulted in markedly different mixes of respondents – for propensity matching, the resulting groups included about a quarter with each type of private insurance (GK and NGK) at baseline and about half with neither. In contrast, the CEM group consisted mostly of individuals with NGK insurance at baseline (74%). Only 366 individuals were retained by both CEM and the full propensity match (394 for the reduced propensity match). For propensity matching (full covariate set), there was little overlap in distribution of the propensity score for the GK group (25th pctl=0.93, median=0.95, 75th pctl=0.97) and the NGK group (25th pctl=0.03, median=0.05, 75th pctl=0.07). Although CEM retained more individuals in the analysis, CEM and propensity matching both excluded the majority of the sample. This is referred to as a situation of minimal "common support," meaning that few individuals with similar baseline characteristics were available in the sample. While discarding observations outside the region of common support restores internal validity, it does so at the potential expense of generalizeability. Although this may be an uncomfortable tradeoff, there is nothing wrong with focusing on observations in the region of common support and in fact this is often what happens in RCTs where only potential participants meeting a specific set of inclusion criteria are included in the RCT (Iacus et al, 2011).

As shown in Table 2, being in a GK plan was only significantly associated with spend for the basic regression. The parameter estimates were negative but non-significant for all other methods. The negative coefficients indicate that enrollment in GK plans is associated with reduced spend in 2010 compared with NGK plans, which is consistent with the hypotheses that GK plans tend to reduce spend.

Table 2. Parameter estimate for the estimated effect of being in a GK plan by method.

Method	Sample size used (intervention group)	Beta (log dollars scale)	SE	p-value
Basic regression[1]	3,240 (3,525 control)	-\$0.38	0.13	<0.01
Regression full[2]	3,240 (3,525 control)	-\$0.19	0.12	0.13
Propensity reduced[3]	612 (612 control)	-\$0.16	0.16	0.34
Propensity full[4]	569 (569 control)	-\$0.19	0.16	0.23
Coarsened exact[5]	1,319 (1,523 control)	-\$0.04	0.10	0.69

Note: Estimates are reported on the scale of natural logarithm of individuals' annual healthcare spend.

[1] This used OLS regression without matching. Being in a GK plan (coded 1=GK plan vs. 0=NGK plan) is the estimate of primary interest ("Beta" in the table). Regression adjusted for baseline insurance type, baseline healthcare spend and end date of the R3 interview.

[2] This used OLS regression without matching. Being in a GK plan (coded 1=GK plan vs. 0=NGK plan) is the estimate of primary interest ("Beta" in the table). Regression adjusted for household size, household income, employment status, geographic region, residence in MSA, race/ethnicity, age, gender, education, self-rated health, self-rated mental health, born in US, baseline total healthcare spend, has usual healthcare provider, usual healthcare provider is PCP, end date of the R3 interview, and baseline insurance type.

[3] This used the propensity-matched sample based on the reduced set of baseline covariates, i.e., household income, race/ethnicity, age, gender, self-rated health, self-rated mental health, baseline total healthcare spend, has usual healthcare provider, usual healthcare provider is PCP, end date of the R3 interview, and baseline insurance type. After matching, OLS regression adjusting for the same set of baseline covariates was conducted; "Beta" in the table is the estimated effect of being in a GK plan (coded 1=GK plan vs. 0=NGK plan), adjusting for the other covariates.

[4] This used the propensity-matched sample based on the full set of baseline covariates, i.e., the same ones listed for "Regression full". After matching, OLS regression adjusting for the same set of baseline covariates was conducted; "Beta" in the table is the estimated effect of being in a GK plan (coded 1=GK plan vs. 0=NGK plan), adjusting for the other covariates.

[5] This used the coarsened exact matching sample based on the reduced set of baseline covariates, i.e., the same ones listed for "Propensity reduced". After matching, OLS regression adjusting for the same set of baseline covariates was conducted; "Beta" in the table is the estimated effect of being in a GK plan (coded 1=GK plan vs. 0=NGK plan), adjusting for the other covariates.

Study 2: *Outcomes associated with switching to a private GK plan*. Prior to matching, the inscope sample consisted of 143 individuals switching to private GK plans and 2,958 remaining in private NGK plans. Table 3 shows the distribution of baseline characteristics for the group switching to GK and the group remaining in NGK plans, in the original sample as well as after matching. There were notable group differences in the sample prior to matching (mean d=0.10, weighted mean d=0.10, % d >0.10=43%). Propensity matching on the reduced set of characteristics resulted in no improvement in balance on baseline characteristics (mean d=0.12, weighted mean d=0.11, % d >0.10=43%). The best match was achieved via coarsened exact matching (mean d=0.06, weighted mean d=0.06, % d >0.10=26%) and propensity matching on the more extensive set of characteristics achieved a match that was nearly as good (mean d=0.07, weighted mean d=0.07, % d >0.10=26%).

Table 3. Mean or percent of baseline covariates and standardized difference ("Diff") by group and matching method (Study 2).

	No matching			Propensity matched 1			Propensity matched 2			Coarsened exact		
Variable	**Non-GK**	**GK**	**Diff.**	**Non-GK**	**GK**	**Diff.**	**Non-GK**	**GK**	**Diff.**	**Non-GK**	**GK**	**Diff.**
Family size	3.4	3.6	0.15	3.7	3.6	0.04	3.6	3.6	0.02	3.5	3.6	0.07
Northeast	14%	14%	0.00	9%	14%	0.15	11%	14%	0.09	17%	15%	0.07
Midwest	26%	24%	0.05	27%	24%	0.06	23%	24%	0.02	25%	24%	0.04
South	39%	38%	0.03	42%	38%	0.09	35%	38%	0.06	38%	38%	0.01
In MSA	83%	90%	0.21	78%	90%	0.33	94%	90%	0.13	86%	90%	0.11
Age	33.8	30.4	0.19	28.8	30.4	0.09	30.9	30.5	0.02	31.6	30.8	0.05
Female	52%	48%	0.08	44%	48%	0.07	38%	48%	0.20	49%	49%	0.00
Asian	7%	11%	0.14	5%	11%	0.23	11%	11%	0.00	9%	12%	0.10
Black	16%	9%	0.20	15%	9%	0.19	12%	9%	0.09	13%	10%	0.12
White	77%	77%	0.01	80%	77%	0.07	73%	77%	0.11	78%	78%	0.00
Hispanic	12%	13%	0.05	9%	13%	0.13	14%	13%	0.02	15%	14%	0.04
Married	49%	39%	0.20	40%	39%	0.01	41%	39%	0.03	46%	40%	0.14
Educ. HS or less	42%	43%	0.01	33%	43%	0.20	44%	43%	0.01	44%	43%	0.03
Educ. College	27%	25%	0.03	24%	25%	0.02	29%	25%	0.08	22%	25%	0.08
Health	2.0	1.9	0.05	2.0	1.9	0.06	2.0	1.9	0.08	1.9	1.9	0.01
Mental health	1.8	1.8	0.01	1.7	1.8	0.02	1.7	1.8	0.10	1.7	1.7	0.07
Born in US	90%	83%	0.19	94%	83%	0.36	77%	83%	0.16	87%	82%	0.13
Have usual care provider	82%	81%	0.03	80%	81%	0.04	81%	81%	0.00	82%	82%	0.00
Usu. care prov. is PCP	39%	53%	0.29	50%	53%	0.07	49%	53%	0.09	53%	53%	0.00
Employed	61%	60%	0.01	47%	60%	0.27	60%	61%	0.01	60%	60%	0.00
Household income	$82,659	$75,741	0.15	$69,799	$75,741	0.15	$84,172	$75,838	0.17	$83,801	$76,317	0.16
Healthcare spend	$3,403	$2,985	0.05	$2,993	$2,985	0.00	$2,889	$3,005	0.02	$2,021	$3,100	0.18
Round 3 end date	18337	18346	0.22	18341	18346	0.12	18351	18346	0.11	18345	18345	0.01

Note: "Propensity matched 1" involved matching on the reduced set of baseline characteristics (the same ones used in CEM) while "Propensity matched 2" involved matching on the full set of baseline characteristics.

As shown in Table 4, switching to a GK plan was associated with significant reductions in spend using regression (both types) as well as propensity matching on a reduced set of characteristics. Switching to a GK plan was not significantly associated with spend by the other two methods, although the parameter estimates were negative.

Table 4. Parameter estimate for the estimated effect of switching to a GK plan by method.

Method	Sample size used (intervention group)	Beta (log dollars scale)	SE	p-value
Basic regression[1]	143	-$0.69	0.23	<0.01
Regression full[2]	143	-$0.47	0.22	0.03
Propensity reduced[3]	143	-$0.87	0.34	0.01
Propensity full[4]	142	-$0.47	0.33	0.16
Coarsened exact[5]	136	-$0.36	0.24	0.13

Note: Estimates are reported on the scale of natural logarithm of individuals' annual healthcare spend.

[1] This used OLS regression without matching. Switching to a GK plan (coded 1=switched to GK plan vs. 0=stayed in NGK plan) is the estimate of primary interest ("Beta" in the table). Regression adjusted for baseline healthcare spend and end date of the R3 interview.

[2] This used OLS regression without matching. Switching to a GK plan (coded 1=switched to GK plan vs. 0=stayed in NGK plan) is the estimate of primary interest ("Beta" in the table). Regression adjusted for household size, household income, employment status, geographic region, residence in MSA, race/ethnicity, age, gender, education, self-rated health, self-rated mental health, born in US, baseline total healthcare spend, has usual healthcare provider, usual healthcare provider is PCP, and end date of the R3 interview.

[3] This used the propensity-matched sample based on the reduced set of baseline covariates, i.e., household income, race/ethnicity, age, gender, self-rated health, self-rated mental health, baseline total healthcare spend, has usual healthcare provider, usual healthcare provider is PCP, and end date of the R3 interview. After matching, OLS regression adjusting for the same set of baseline covariates was conducted; "Beta" in the table is the estimated effect of switching to a GK plan (coded 1=switched to GK plan vs. 0=stayed in NGK plan), adjusting for the other covariates.

[4] This used the propensity-matched sample based on the full set of baseline covariates, i.e., the same ones listed for "Regression full". After matching, OLS regression adjusting for the same set of baseline covariates was conducted; "Beta" in the table is the estimated effect of switching to a GK plan (coded 1=switched to GK plan vs. 0=stayed in NGK plan), adjusting for the other covariates.

[5] This used the coarsened exact matching sample based on the reduced set of baseline covariates, i.e., the same ones listed for "Propensity reduced". After matching, OLS regression adjusting for the same set of baseline covariates was conducted; "Beta" in the table is the estimated effect of switching to a GK plan (coded 1=switched to GK plan vs. 0=stayed in NGK plan), adjusting for the other covariates.

Study 3: *Outcomes associated with switching to a private NGK plan*. Prior to matching, the inscope sample consisted of 113 individuals switching to private NGK plans and 2,661 remaining in private GK plans. There were substantial group differences prior to matching (mean d=0.13, weighted mean d=0.09, % d >0.10=43%). Like Study 2, propensity matching on the reduced set of characteristics resulted in no improvement in balance on baseline characteristics (mean d=0.13, weighted mean d=0.12, % d >0.10=52%). The best match was achieved via propensity matching on the more extensive set of characteristics (mean d=0.07, weighted mean d=0.07, % d >0.10=30%) followed by coarsened exact matching (mean d=0.11, weighted mean d=0.08, % d >0.10=43%).

As shown in Table 5, switching to NGK plans was not significantly associated with spend for any of the methods. Parameter estimates tended to be positive with the exception of full propensity matching.

Table 5. Parameter estimate for the estimated effect of switching to an NGK plan by method.

Method	Sample size used (intervention group)	Beta (log dollars scale)	SE	p-value
Basic regression[1]	113	$0.42	0.27	0.12
Regression full[2]	113	$0.11	0.25	0.66
Propensity reduced[3]	113	$0.60	0.39	0.12
Propensity full[4]	113	-$0.03	0.32	0.92
Coarsened exact[5]	108	$0.15	0.26	0.55

Note: Estimates are reported on the scale of natural logarithm of individuals' annual healthcare spend.

[1] This used OLS regression without matching. Switching to a NGK plan (coded 1=switched to NGK plan vs. 0=stayed in GK plan) is the estimate of primary interest ("Beta" in the table). Regression adjusted for baseline healthcare spend and end date of the R3 interview.

[2] This used OLS regression without matching. Switching to a NGK plan (coded 1=switched to NGK plan vs. 0=stayed in GK plan) is the estimate of primary interest ("Beta" in the table). Regression adjusted for household size, household income, employment status, geographic region, residence in MSA, race/ethnicity, age, gender, education, self-rated health, self-rated mental health, born in US, baseline total healthcare spend, has usual healthcare provider, usual healthcare provider is PCP, and end date of the R3 interview.

[3] This used the propensity-matched sample based on the reduced set of baseline covariates, i.e., household income, race/ethnicity, age, gender, self-rated health, self-rated mental health, baseline total healthcare spend, has usual healthcare provider, usual healthcare provider is PCP, and end date of the R3 interview. After matching, OLS regression adjusting for the same set of baseline covariates was conducted; "Beta" in the table is the estimated effect of switching to a NGK plan (coded 1=switched to NGK plan vs. 0=stayed in GK plan), adjusting for the other covariates.

[4] This used the propensity-matched sample based on the full set of baseline covariates, i.e., the same ones listed for "Regression full". After matching, OLS regression adjusting for the same set of baseline covariates was conducted; "Beta" in the table is the estimated effect of switching to a NGK plan (coded 1=switched to NGK plan vs. 0=stayed in GK plan), adjusting for the other covariates.

[5] This used the coarsened exact matching sample based on the reduced set of baseline covariates, i.e., the same ones listed for "Propensity reduced". After matching, OLS regression adjusting for the same set of baseline covariates was conducted; "Beta" in the table is the estimated effect of switching to a NGK plan (coded 1=switched to NGK plan vs. 0=stayed in GK plan), adjusting for the other covariates.

DISCUSSION

This was a case study illustrating results obtained from using different methods to balance baseline covariates between groups in an observational study. All of these methods are relatively easy to implement in SAS.

In this study, the best match tended to result from propensity matching on the full set of baseline covariates (Studies 1 and 3). However, CEM resulted in a slightly better match than full propensity matching in Study 2. There are many different ways to measure goodness of matching and results might differ if other measures were used. Also, it should be noted that goodness of matching was measured across the full set of baseline covariates, which may give an advantage to methods that attempt to match on all variables within the full baseline set. It was unfeasible to conduct CEM using the full set of baseline covariates -- when CEM was done using the full set, 99% of the observations were excluded due to inability to find matches within strata. When CEM was compared with propensity matching using the same set of baseline covariates (i.e., the reduced propensity match), results were mixed: CEM resulted in closer matches in Studies 2 and 3 while propensity matching resulted in a closer match in Study 1.

Overall, these results do not suggest that CEM consistently yields the best matches and it could be argued that propensity matching tended to yield closer matches (although not always). However, these results are on a single dataset and there is no reason to believe that the results will generalize to other datasets.

Directionally, the model coefficients tended to support the hypothesis that being in a private GK plan is associated with lower healthcare spend, compared with being in a private NGK plan, although in most cases the estimates were not significantly different from 0 at the alpha=0.05 level. However, lack of statistical significance may have been due in part to small sample sizes.

All methods (propensity matching, CEM and regression without matching but adjusting for the full baseline covariate set) tended to result in coefficients closer to 0 compared with regression adjusting only for baseline healthcare spend and end date of the R3 interview (and baseline insurance type in Study 1). This points to the importance of adjusting for a wide range of variables in studies of healthcare spend. CEM tended to result in estimates closer to 0 than propensity matching and regression adjusting for the full baseline covariate set (Studies 1 and 2, but not 3). The true effects of insurance plan type are unknown so it cannot be said that any of these estimates were more accurate than the others. At this point, these must be viewed as interesting but not necessarily general trends, to be explored in future research.

PAST STUDIES USING MEPS DATA TO ESTIMATE ASSOCIATION OF TYPE OF PRIVATE INSURANCE WITH OUTCOMES

This study fits within a larger body of work using MEPS data to examine the association of type of private insurance with outcomes such as patient satisfaction, utilization, healthcare spend, and quality measures (Shi et al, 2000; Escare et al, 2001; Newacheck et al, 2001; Hmadkova, 2009; Stanton et al, 2002). To our knowledge, all past studies on this topic looked at HMO vs. non-HMO plans, although non-HMO was characterized in different ways, perhaps reflecting the diversity of non-HMO plans. To our knowledge, Hrmadkova (2009) was the only study to specifically attempt to compare “gatekeeper” private insurance plans (HMO or other private gatekeeper plans) with non-gatekeeper plans using MEPS data.

Typically these studies found small, inconsistent or non-existent associations of type of private insurance plan (HMO vs. other) with outcome variables, which is also true of the present study. Escarce et al (2001) found no significant associations of plan type with healthcare spend after statistically adjusting for covariates. Newacheck et al (2001) found no significant associations of plan type with access, utilization, satisfaction or quality measures, after adjusting for covariates. However, Stanton et al (2002) found that HMOs reduce healthcare spend, but only in competitive markets (more HMOs in the market) – in competitive markets, HMOs reduce cost by substituting ambulatory with hospital care, in order to reduce costs to offer more attractive premiums. Hrmadkova (2009) focused on the impact of plan type on choice of type of first provider seen (PCP vs. specialist) and found that individuals in GK plans were more likely to see a PCP as the first provider type; however, after initial provider type was taken into account, GK vs. non-GK plans were not significantly associated with differences in utilization.

Most studies using MEPS data to look at the impact of private insurance type adjusted for other covariates (e.g., demographics, health) using regression alone. Escarce et al (2001) used 2-part models while Newacheck et al (2001) and Hrmadkova (2009) used single (one-part) regression models. Hrmadkova (2009) used a propensity-matched comparison group in addition to applying regression adjustment alone, similar to the present study.

LIMITATIONS

Individuals with missing data on the analytic variables were excluded from the analyses. Few individuals were excluded due to missing data. However, more appropriate approaches to handle missing data would be multiple imputation (Allison, 2002) or direct likelihood imputation (Allison, 2012).

Using a single panel of MEPS resulted in limited sample sizes, especially for analyses of individuals switching from private non-gatekeeper to private gatekeeper plans (and vice versa), limiting the statistical power of these analyses. Sample size could be increased by combining panels of MEPS, although this would have the downside of combining less with more recent data.

MEPS survey weights and survey design variables were intentionally not used. There were several reasons for this. First, coarsened exact matching assigns weights to observations and it is unclear how these should be combined with survey weights; perhaps simple multiplication of the weights would be appropriate but the implications of this are unknown. Second, matching techniques, by selecting certain observations, compromise the integrity of the complex survey design of MEPS in ways that would need to be explored in detail (beyond the scope of this paper). Third, this is primarily a methodology study rather than an attempt to use MEPS data for accurate estimation of intervention effects. In reality, these techniques would typically be applied to claims data as opposed to survey data, so adjustment for complex survey design is not relevant to the most likely applications of the methods illustrated.

No attempt was made to adjust for non-independence of observations within matched pairs. Techniques such as mixed models could be used to do this.

REFERENCES

Allison PD. (2002). Missing data. Thousand Oaks, CA: Sage.

Allison PD. (2012). Handling missing data by maximum likelihood. Proceedings of SAS Global Forum.

Austin PC, Chiu M, Ko DT, Goeree R, Tu JV. (2010). Propensity score matching for estimating treatment effects. In Faries DE, Leon AC, Haro JM, Obenchain RL (Eds.), Analysis of observational healthcare data using SAS. Cary NC: SAS Institute.

Dehejia RH, Wahba W. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. J Am Stat Assoc 94, 1053-1062.

Diehr P, Yanez D, Ash A, Hornbrook M, Lin DY. (1999). Methods for analyzing health care utilization and costs. Annu Rev Public Health. 1999;20:125-44.

Escarce JJ, Kapur K, Joyce GF, Van Vorst KA. (2001). Medical care expenditures under gatekeeper and point-of-service arrangements. Health Serv Res. 36(6 Pt 1):1037-57.

Hromadkova E. (2009). Gatekeeping – open door to effective medical care utilization? Working paper series (ISSN 1211-3298), CERGE-EI, Prague, Czech Republic.

Iacus SM, King G, Porro G. (2011). Causal inference without balance checking: Coarsened exact matching. Political Analysis.

Meyers D, Peikes D, Dale S, Lundquist E, Genevro J. (2011). Improving Evaluations of the Medical Home. AHRQ Publication No. 11-0091. Rockville, MD: Agency for Healthcare Research and Quality.

Newacheck PW, Hung YY, Marchi KS, Hughes DC, Pitter C, Stoddard JJ. (2001). The impact of managed care on children's access, satisfaction, use, and quality of care. Health Serv Res. 36(2):315-34.

Reeve BB, Smith AW, Arora NK, Hays RD. (2008). Reducing bias in cancer research: application of propensity score matching. Health Care Financ Rev. 29(4):69-80.

Rosenthal PR. (2009). Design of observational studies. New York: Springer.

Rubin DB. (2008). For objective causal inference, design trumps analysis. Ann Appl Stat. 2(3), 808-840.

Shi L. (2000). Type of health insurance and the quality of primary care experience. Am J Public Health. 90(12):1848-55.

Stanton MW, Rutherford MK. (2002). Reducing costs in the health care system: learning from what has been done. Rockville (MD): Agency for Healthcare Research and Quality, Research in Action Issue 9. AHRQ Pub. No. 02-004.

Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB. (2000). Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA. 283(15):2008-12.

Wells AR, Hamar B, Bradley C, Gandy WM, Harrison PL, Sidney JA, Coberley CR, Rula EY, Pope JE. (2013). Exploring robust methods for evaluating treatment and comparison groups in chronic care management programs. Popul Health Manag. 16(1):35-45.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Doug Thompson
Blue Cross Blue Shield of IL, MT, NM, OK & TX
300 E Randolph
Chicago, IL 60601
doug_thompson@bcbsil.com