Sample Selection for Medicare Risk Adjustment Due to Systematically Missing Data

Savannah L. Bergquist , Thomas G. McGuire, Timothy J. Layton , and Sherri Rose

Objective. To assess the issue of nonrepresentative sampling in Medicare Advantage (MA) risk adjustment.

Data Sources. Medicare enrollment and claims data from 2008 to 2011.

Data Extraction. Risk adjustment predictor variables were created from 2008 to 2010 Part A and B claims and the Medicare Beneficiary Summary File. Spending is based on 2009–2011 Part A and B, Durable Medical Equipment, and Home Health Agency claims files.

Study Design. A propensity-score matched sample of Traditional Medicare (TM) beneficiaries who resembled MA enrollees was created. Risk adjustment formulas were estimated using multiple techniques, and performance was evaluated based on R^2 , predictive ratios, and formula coefficients in the matched sample and a random sample of TM beneficiaries.

Principal Findings. Matching improved balance on observables, but performance metrics were similar when comparing risk adjustment formula results fit on and evaluated in the matched sample versus fit on the random sample and evaluated in the matched sample.

Conclusions. Fitting MA risk adjustment formulas on a random sample versus a matched sample yields little difference in MA plan payments. This does not rule out potential improvements via the matching method should reliable MA encounter data and additional variables become available for risk adjustment.

Key Words. Risk adjustment, medicare, machine learning, regression

Missing data is a pervasive obstacle in health services research and may have particularly important implications for health plan payment policy. In some cases, variables may be systematically missing—unavailable for all subjects in a particular dataset, but available in another data source, possibly drawn from a different population. Two major payment systems in the United States,

Medicare Advantage (MA) and the Affordable Care Act Health Insurance Marketplace systems, incorporate estimates of health care spending parameters with systematically missing data. Health plans in these systems receive risk-adjusted payments, which involves predicting individual health spending, based on formulas estimated in data samples chosen in part because of their availability and that differ nontrivially from their respective target populations. This is not only a statistical problem; it is also a potential policy problem. Setting per person plan payments based on data not characterizing the target population runs the risk of incorrectly characterizing the expected costs of that target population and thereby providing health plans with unintended adverse incentives.

Much of the literature addressing systematically missing data has focused on methods for generalizing causal effect estimates from randomized trials across intervention sites (e.g., Rudolph and van der Laan 2016) or to more general populations (e.g., Stuart et al. 2011; Stuart, Bradshaw, and Leaf 2015). While there are many important causal inference questions to be asked regarding MA versus Traditional Medicare (TM)—such as examining the differential impact of MA enrollment on utilization or spending (Landon et al. 2012; Baicker and Robbins 2015; Curto et al. 2017)—in the context of MA plan payment, we seek to accurately predict what MA enrollees would have spent had they enrolled in TM. Accordingly, addressing systematically missing data in health plan payment requires techniques for prediction rather than effect estimation, and draws from a much smaller literature. Previously, Rose et al. (2015) combined propensity score matching and imputation across two nonoverlapping data sources in an effort to draw an improved sample for Marketplace plan payment. They found that propensity score matching used for sample selection produced a more balanced sample, but small gains in risk adjustment performance as measured by R^2 .

Medicare claims data are a rich source of information about Medicare enrollees, yet these data are absent for individuals who choose to enroll in MA, the private managed care portion of Medicare. Enrollees in TM have individual bills for each procedure or service (“fee-for-service”), and these claims are invaluable for studying patterns of care for TM enrollees. In contrast, MA enrollees have their health care paid for via a lump sum from the federal government to a private managed care plan on an annual basis. Thus,

Address correspondence to Savannah L. Bergquist, M.Sc., Health Policy Doctoral Program, Harvard University, 180A Longwood Ave., Boston, MA 02115; e-mail: bergquist@g.harvard.edu. Thomas G. McGuire, Ph.D., Timothy J. Layton, Ph.D., and Sherri Rose, Ph.D., are with the Department of Health Care Policy, Harvard Medical School, Boston, MA.

they do not have a comparable source of data available to researchers, although efforts to release limited versions of MA data are ongoing. Over 30 percent of Medicare enrollees participate in MA, and these individuals differ from TM enrollees in terms of health status, race/ethnicity, age, and income (Newhouse and McGuire 2014; Curto et al. 2017). The 2003 Medicare Modernization Act introduced payment increases to encourage plan participation and thereby increase beneficiary choice; MA enrollment has steadily increased over the past decade, even with the payment reductions in the 2010 Affordable Care Act (Jacobson et al. 2017). For many Medicare enrollees, MA is a less expensive option than TM, and the trade-off between a restricted network and lower out-of-pocket costs tends to attract individuals who are healthier, younger, lower income, and racial/ethnic minorities.

Both the viability and the efficiency of the MA program depend critically on the system used to pay MA plans for each Medicare beneficiary they enroll. In order to ensure budget neutrality of the MA program and to provide MA plans with incentives to engage in efficient competition rather than simply selecting the healthiest beneficiaries, the government attempts to set plan payments equal to the amount that the enrollee is expected to have spent in TM. To accomplish this goal, the government “risk-adjusts” payments to MA plans according to the chronic conditions of the enrollee, with additional payments for each condition equal to the estimated incremental cost of an average person with the condition. There is a substantial literature investigating risk selection and the MA program, examining issues such as coding differences between MA and TM (e.g., Frogner et al. 2011; Geruso and Layton 2015) and plan behavior to select healthier enrollees (e.g., McWilliams, Hsu, and Newhouse 2012; Brown et al. 2014; Newhouse et al. 2015). A related body of work focuses on assessing remedies for under- and overpayment by risk adjustment systems (e.g., Layton, McGuire, and van Kleef 2018; Shrestha et al. 2018); this paper seeks to add to that literature.

The MA risk adjustment formula coefficients are currently estimated using TM claims data. These coefficients are then used to calculate individual enrollee risk scores from beneficiary data submitted by MA plans, which are translated into payments. Therefore, MA payments are based on patterns of care generated by individuals who choose not to enroll in MA. The Centers for Medicare and Medicaid Services (CMS) is in the midst of a multiyear effort to utilize MA encounter data for risk adjustment purposes. In 2012, CMS began collecting MA enrollee encounter data through a new system, the Encounter Data System, which aims to capture a more comprehensive set of data points. Starting with 2015 payments, CMS began calculating a portion of

plans' risk scores based on encounter data, but there is no timeline to update or re-estimate the MA risk adjustment formula itself using the Encounter Data System, and the Government Accountability Office has recommended CMS continue validating these data before using them for payment purposes (United States Government Accountability Office 2017).

If the goal is to pay MA plans the amount MA enrollees would have cost had they enrolled in TM, then it is reasonable to use the TM data because it is free from indirect plan selection effects—such as offering or restricting certain services—and other factors that influence utilization. However, the current system for estimating risk adjustment formulas to pay MA plans does not address potential sources of bias from observable differences generated by endogenous consumer behavior-based selection that leads some individuals to enroll in MA and others to choose TM. This paper explores an alternative to standard practice: instead of estimating coefficients using the full population of TM enrollees, selecting a sample of TM enrollees who resemble MA enrollees in terms of observable demographic characteristics obtained from the same administrative data source used to estimate risk adjustment. Rather than modifying the risk adjustment formula with the aim of simply improving R^2 or predictive ratios (traditional goals), we view the data input into the risk adjustment system as a policy tool to be considered alongside the risk adjustment function.

Our strategy relies on flexible machine learning-based estimators to construct a matched sample of claims data and then estimate an MA risk adjustment formula using conventional and novel estimation methods. Because propensity score matching methods for prediction are an original contribution without established performance criteria, we compare multiple matching approaches and select among them based on prespecified metrics. Several risk adjustment algorithms are then estimated in the matched and random TM samples to explore the possibility of improved performance under less restrictive assumptions and more parsimonious formulas selected via data-adaptive methods. We compare overall predictive performance on metrics such as cross-validated R^2 , and we assess predictive ratios and formula coefficients to examine the impact of matching methods and algorithm choice on enrollee subgroups.

METHODS

Data

The Medicare Beneficiary Summary File contains basic demographic information on a 20 percent random sample of all Medicare beneficiaries,

including MA and TM enrollees. We required all subjects to have 12 months of continuous Part A and B coverage or 12 months of MA enrollment in the given calendar year, and we required TM subjects to have at least 1 month of Part A and B coverage in the prediction year. (This follows CMS risk adjustment practices regarding inclusion and exclusion criteria.) We excluded beneficiaries with ESRD and non-US residents, due to differences in plan characteristics and payment, and small sample sizes in these categories.

Because claims utilization data are not available for MA patients, we used basic demographic data from the Medicare Beneficiary Summary File to build the propensity score function needed to create a TM matched sample. To estimate the “propensity to be enrolled” in an MA plan, defined as belonging to a risk-based MA plan, we excluded variables likely to be affected by choice of insurance and selected those predictive of MA enrollment: sex, rural residence, region, race, and age. Nursing facility residence information is only available for TM enrollees in claims data (Yun et al. 2010); thus, we excluded institutionalized enrollees after conducting the match. We also drew a random sample of TM individuals from the same pool of subjects used to create the matched sample, equal in size to the matched sample, and enforced the same exclusion criteria. This random sample approximates standard CMS practice of using an unadjusted sample of TM beneficiaries to estimate risk adjustment formulas and assess their performance (Pope et al. 2011).

Medicare inpatient, outpatient, home health, and durable medical equipment claims files for both the matched and random TM samples were combined to estimate risk adjustment formulas. The outcome variable, spending in the following year, was constructed by summing Medicare payments (except those paid by beneficiaries, hospice payments, and indirect medical education payments). We summed payments for the months in the prediction year that beneficiaries met the sample inclusion criteria and then divided by the fraction of months each beneficiary was in our sample; all analyses were weighted by this fraction.

To construct the 87 Hierarchical Condition Category (HCC) indicator variables in the risk adjustment formulas, we used the CMS-HCC risk adjustment software (version 21), which maps ICD-9 codes to HCCs. The remaining risk adjustment variables were 24 age–sex cells and interaction terms indicating disability, Medicaid status, and reason for Medicare entitlement. We also mapped ICD-9 codes to Clinical Classification Software (CCS) groups, which capture a more comprehensive set of diagnoses and individuals than the HCCs (Montz et al. 2016; Shrestha et al. 2018). The CCS groups were used to evaluate risk adjustment performance.

Our primary period of analysis is the 2010–2011 calendar years, where 2010 contains the matching variables and risk adjusters used to estimate spending for 2011. For sensitivity analyses, we examine 2009–2010 and 2008–2009.

Propensity Score Estimation and Matching

We relied on basic demographic variables populated for both MA and TM subjects in the 20 percent Medicare Beneficiary Summary File to compare MA and TM enrollees and perform the matching. The goal of the matching exercise was to select TM subjects who resemble MA subjects. To accomplish this, we performed one-to-one matching comparing ten matching methods: exact matching and nine implementations of greedy within-caliper matching.

Our exact matching approach selected subjects based on four key variables: age (<65, 65–69, 70–74, 75–79, 80+), race (white, black, other), sex, and rural residence. The greedy within-caliper approaches selected the first match for each MA observation within a given caliper. We prespecified three sizes of caliper: $SD(\text{propensity score})/\epsilon$, with $\epsilon = \{1, 2, 4\}$ (Rosenbaum and Rubin 1985). Additionally, we estimated the propensity score for each caliper choice using three algorithms: main terms logistic regression, LASSO penalized regression, and random forest with a minimum node size of 250 and 500 trees. All algorithms were implemented using 10-fold cross-validation and considered the same set of demographic variables described earlier. We direct the interested reader to Rose (2016) for further description of machine learning methods in risk adjustment.

The ten matching techniques selected matches with replacement. Matching with replacement is often desirable due to decreased bias, but typically leads to samples where the same subject is matched multiple times (Stuart 2010). To maintain a large sample while balancing this issue of subject overlap, we chose to target a sample size of 1,000,000 enrollees, which had an expected yield of <10 percent overlap.

We defined the best performing matching approach as that which minimized the number of variables with absolute standardized difference in means under 10 percent (a commonly used threshold for sample balance (Stuart, Lee, and Leacy 2013; Austin and Mamdani 2006), while also factoring in the number of overlapping enrollees and simplicity of implementation (in cases of ties or near ties in performance). We evaluated the absolute standardized difference in means between the MA and TM groups for each variable, divided by the square root of the mean of the MA and TM variances (i.e., group standard

deviations) before matching: $ASDM = (\bar{X}_{MA} - \bar{X}_{TM})/\sigma_{MA, TM}$. For the sensitivity analyses in 2009 and 2008, we implemented the propensity score matching approach that performed best in the 2010 sample.

Risk Adjustment Estimation

To comprehensively compare the performance of risk adjustment in our matched sample versus the random sample, we fit several risk adjustment algorithms to both samples. Using an ensembling framework, we ran six implementations of different algorithms with cross-validation and then selected one implementation based on a prespecified metric. The individual algorithms were: ordinary least-squares (OLS) regression, LASSO, and random forest. Each of these three algorithms considered both the full set of 122 risk adjusters described earlier and a subset selected data-adaptively by another LASSO within the cross-validation folds (Rose, Bergquist, and Layton 2017). We limited this “variable screening” LASSO to select no more than 15 variables to investigate the role of parsimony in the risk adjustment formula.

We select the single best algorithm according to the cross-validated MSE: $CV\ MSE = \sum (Y_i - Z_i)^2 / N$, where Y_i is observed spending, Z_i is the vector of cross-validated predicted values from a given algorithm, and N is the number of observations in the sample. We also compare cross-validated R^2 : $CV\ R^2 = 1 - \sum (Y_i - Z_i)^2 / \sum (Y_i - \bar{Y})^2$, where $\bar{Y}$ is the mean observed spending. Lastly, we examine predictive ratios based on mental health diagnoses (as an example of a chronic illness; one that is often poorly captured by U.S. risk adjustment schemes (Shrestha et al. 2018; Montz et al. 2016; McGuire et al. 2014), number of HCCs, and observed expenditure deciles: $CV\ PR = (\sum (Z_{ig})/n_g) / (\sum (Y_{ig})/n_g)$, where g is the group of interest and n is the number of observations in g . We use observed expenditure because it allows fit assessment along the distribution of observed spending—for example, to examine how well the algorithm predicts spending for the highest decile of observed spending.

Lastly, we apply the OLS fit based on the random sample to the matched sample and obtain predicted values. This is useful to determine how well the conventional approach (using a random TM sample to fit the formula) performs on an MA-like population. Results from this component of our analysis are therefore meant to mimic current practice—that is, nonrepresentative data are used to fit the risk adjustment formula coefficients, but then MA data are run through the fixed formula to generate predicted values and payments.

RESULTS

After applying our exclusion criteria, we had a sample of 2,095,128 MA subjects and 5,567,744 TM subjects in 2010–2011. Table 1 contains summary statistics describing our samples in terms of the basic demographic features in common between MA and TM enrollees. Columns (1) and (2) describe the MA and TM enrollees in the sample prior to matching. Compared to TM, more MA enrollees live in an urban area, more live in the West, and fewer are white.

Propensity Score Estimation and Matching

Our final matched sample of TM enrollees contained 987,563 observations after excluding 12,437 institutionalized enrollees from the one-to-one propensity-score-based match on 1,000,000 MA observations, as outlined earlier. Column (3) in Table 1 describes the matched TM sample, and column (4) summarizes the randomly drawn TM sample. The matched sample means generally fall between the prematch MA and TM means, while the random sample means are similar to the larger, prematch TM sample. Our matched sample had 9.1 percent overlap.

The propensity-score-based matching methods performed similarly in terms of balance on key variables, while the exact matching methods balanced

Table 1: Medicare Variable Summaries by Sample, 2010

	<i>Prematch MA</i>	<i>Prematch TM</i>	<i>Matched TM</i>	<i>Random TM</i>
<i>N</i>	2,095,128	5,567,744	987,563	987,563
Age (years)	73	71	72	71
Female (%)	57	56	56	56
Rural (%)	12	24	17	24
Region (%)				
Northeast	21	18	19	18
West	27	17	21	17
South	34	40	39	40
Midwest	18	25	21	24
Race (%)				
White	81	85	85	85
Black	11	9	9	9
Other	8	6	6	6

Note: MA is Medicare Advantage, and TM is Traditional Medicare. Prematch refers to samples prior to performing propensity score matching or random selection; Matched refers to the sample post-propensity score matching; Random refers to the sample randomly selected.

fewer variables. As described in the Methods section, we selected the algorithm that performed best with respect to variable balance, number of overlapping enrollees, and simplicity. This was the logistic regression with a caliper of $SD(\text{propensity score})/4 = 0.02$. Figure 1 shows the absolute standardized difference in means for this method, demonstrating that the matched sample improved on or was the same as the unmatched sample for all variables. All variables except “rural” and “West” had differences reduced to <10 percent. See Web Figures S1 and S2 in online Appendix for variable imbalance among a selection of the alternative matching methods that were not chosen.

Risk Adjustment

R^2 results for the risk adjustment algorithms containing all 122 variables across the samples used for fit and prediction (i.e., matched or random sample) are compared in Table 2. The Match-Match and Random-Random columns follow the typical practice of fitting a risk adjustment algorithm on a sample and using the fitted algorithm to generate predicted values in the same sample. The Random-Match column contains results for algorithms fit on the random sample that were used to predict spending in the matched sample. Recall, this Random-Match comparator aims to approximate current practice. Within each algorithm, there is little difference in the R^2 generated by using the matched fit or random fit to predict for the matched sample. The random forest generates an improvement of approximately 7 percent in R^2 over the OLS in both settings, although all R^2 were low. The LASSO did not shrink any coefficients to zero and used all 122 variables, generating identical R^2 results to the OLS. When we compare the difference in predicted values for the

Figure 1: Variable Balance after Matching on Propensity Score

Note. Variables are sorted by magnitude of the absolute standardized difference in means.

Table 2: R^2 and Mental Health Predictive Ratios across Fit and Prediction Samples

<i>Algorithm</i>	<i>Match-Match</i>	<i>Random-Match</i>	<i>Random-Random</i>
R^2			
OLS	10.9	11.0	10.9
LASSO	10.9	11.0	10.9
Random forest	11.7	11.8	10.4
Predictive ratios			
OLS	0.96	0.95	0.96
LASSO	0.95	0.95	0.96
Random forest	0.94	0.92	0.94

Note: Match-Match refers to the combination of fitting an algorithm on the propensity-score matched sample and performing prediction using the same matched sample. Random-Match refers to the combination of fitting an algorithm on the randomly selected sample and using that fitted algorithm to perform prediction for the propensity-score matched sample. Random-Random refers to the combination of fitting an algorithm on the randomly selected sample and performing prediction using the same randomly selected sample.

Match-Match versus the Random-Match OLS, we see the Match-Match fit is systematically overpredicting compared to the Random-Match version (Web Figure S3).

Predictive ratios for individuals with a mental health disorder, as defined by the CCS categories, are presented in Table 2. Mental health conditions are systematically underpaid across all algorithms, with random forest slightly worse than OLS and LASSO, although all are above the typical 0.90 performance cutoff (Kautter et al. 2014). Individuals with mental health conditions make up 21 percent of our matched sample (22 percent in random sample) and average \$16,435 in observed spending (compared to \$8,356 for those without a mental health condition). The small underprediction error seen here is unlikely to indicate large selection incentives for plans to avoid individuals with a mental health condition. Web Tables S1–S4 in the online Appendix contain R^2 and predictive ratio results for the risk adjustment algorithms that considered a limited subset of variables chosen by the “variable screening” LASSO. These algorithms performed consistently worse across all benchmarks.

We also examined sample proportions and predictive ratios based on number of HCCs from the OLS with all risk adjustors, which is the standard practice approach for MA risk adjustment (Table 3). A greater proportion of individuals in the matched sample have no HCCs (39 percent) compared to in the random sample (9 percent). Approximately half (51 percent) of both the matched and random sample have 1–3 HCCs, while a greater proportion of

Table 3: Sample Proportions and Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for OLS with All Risk Adjustors

<i>Sample Proportions (%)</i>			<i>Predictive Ratios</i>		
	<i>Matched Sample</i>	<i>Random Sample</i>	<i>Match-Match</i>	<i>Random-Match</i>	<i>Random-Random</i>
Number of HCCs					
0	39	9	0.91	0.89	0.89
1–3	51	51	1.03	1.02	1.03
4–6	2	38	1.03	1.03	0.98
7–9	9	2	0.97	0.96	0.88
10+	<1	<1	0.88	0.88	1.48
Spending Decile					
1	10.0	10.3	583.54	569.59	549.58
2	10.0	10.4	22.80	22.45	22.21
3	10.0	10.0	10.28	10.16	10.15
4	10.0	10.0	6.61	6.54	6.55
5	10.0	9.9	4.62	4.59	4.60
6	10.0	9.8	3.38	3.36	3.35
7	10.0	9.9	2.40	2.39	2.38
8	10.0	9.9	1.64	1.63	1.63
9	10.0	10.0	0.91	0.90	0.90
10	10.0	9.9	0.31	0.31	0.31

Note: Matched refers to the sample post-propensity score matching; Random refers to the sample randomly selected. Match-Match refers to the combination of fitting an algorithm on the propensity-score matched sample and performing prediction using the same matched sample. Random-Match refers to the combination of fitting an algorithm on the randomly selected sample and using that fitted algorithm to perform prediction for the propensity-score matched sample. Random-Random refers to the combination of fitting an algorithm on the randomly selected sample and performing prediction using the same randomly selected sample.

the random sample (38 percent vs. 2 percent) have 4–6 HCCs. Despite these differences, OLS performs similarly across the prediction combinations for the 0, 1–3, and 4–6 HCC groups. The Match-Match and Random-Match combinations produce more stable predictive ratios in the 7–9 and 10+ HCC groups. In the 10+ HCC group, making up <1 percent of the data, the Random-Random sample predictive ratio is 1.48, compared to 0.88 in the Match-Match and Random-Match samples.

Table 3 additionally displays the matched sample observed spending deciles from the OLS with all risk adjustors. The random sample is relatively evenly distributed when ranked by the matched sample observed spending decile. Examining predictive ratios by observed spending decile shows that across all prediction combinations, the predictive performance is similar, with the best performance occurring around mean spending (\$10,039 in the

matched sample; \$9,881 in the random sample), which falls in the ninth decile. Spending deciles 1–8 all feature predictive ratios >1.0 , indicating that revenue for the majority of beneficiaries exceeds their expected TM costs, whereas only for deciles 9 and 10 do predictive ratios fall below 1.0. Because most of the spending is concentrated in these last two deciles, the predictive ratios “balance out” to a value of 1.0 for the population overall. Web Tables S3–S6 in online Appendix contain results from the LASSO and random forest risk adjustment algorithms.

Figure 2 compares the coefficients from the OLS with all risk adjustors fit on the matched sample versus fit on the random sample. The five risk adjustor coefficients that increase the most and the five that decrease the most between regression fits are among the rarest conditions, each occurring in <1 percent of either sample. The five most prevalent conditions are the same in both the matched and random samples: diabetes without complication, specified heart arrhythmias, COPD, vascular disease, and congestive heart failure. The coefficients for these conditions remain relatively constant between samples and across years (Web Figures S5–S7).

In our sensitivity analyses for 2008 and 2009, greedy within-caliper matching on a logistic regression-estimated propensity score similarly improved balance across observables (Web Figures S7 and S8). Risk adjustment performance was also similar to the primary year results (Web Tables S7–S19).

DISCUSSION

The observable differences in demographics and care utilization between the MA and TM populations have raised concerns regarding the estimation of MA plan payments based on an unadjusted TM sample. Specific concerns include creating unintended incentives for plans to select against groups generating lower revenues compared to groups overcompensated or accurately compensated by the risk adjustment formula. Standard practice is to fit an OLS on a random sample of TM enrollees and use the parameter estimates from that fit to predict spending among MA beneficiaries. We examined this issue of nonrepresentative sampling in MA risk adjustment by creating a propensity-score-based sample of TM beneficiaries who more closely resembled MA enrollees. We then compared performing both risk adjustment fit and prediction on our matched sample to (1) fit on a random sample and prediction using a matched sample, approximating current practice, and (2) fit and prediction on a random sample.

Figure 2: Change in OLS Coefficients between Random Sample Fit and Matched Sample Fit

Evaluation metrics when performing both fit and prediction on the matched sample (Match-Match) compared to fit on the random sample and prediction in the matched sample (Random-Match) were similar. Our results imply that little gain in overall performance can be expected by fitting the MA risk adjustment algorithm on a matched sample derived from basic administrative claims variables. Given the scope of data availability issues surrounding the MA program, as well as the importance of the risk adjustment system in shaping plan behavior and enrollee options, our results provide empirical evidence that a random sample for risk adjustment estimation does not appear to be a large contributor to problematic selection incentives, when compared

to a matched sample based on the limited demographic variables available for MA enrollees.

Because the existing statistics literature has focused primarily on systematically missing data methodology for effect estimation, we explored several matching approaches for this prediction problem and a composite performance criterion. While exact matching was demonstrably worse—due to constraints imposed by finding exact matches for the specified variables—the propensity-score-based methods performed similar to each other. We also compared OLS to two other algorithms for risk adjustment. We saw modest improvements in R^2 when the random forest algorithm was applied to the matched sample, with a 7 percent improvement over the OLS when fit on either the matched or random sample. R^2 results for the OLS and LASSO were similar (10.9 percent or 11.0 percent) across all three of the matched and random sample fit and prediction combinations. Our OLS results, using a simplified version of the CMS-HCC version 21 algorithm, were slightly lower than the full sample R^2 , 12.5 percent, reported in the official evaluation for the CMS-HCC version 21 formula for aged beneficiaries (the largest subgroup model) (Pope et al. 2011).

Collectively, the mental health predictive ratio results across the three fit and prediction sample combinations all yielded slight underprediction, likely driven by the difference in how well the HCC risk adjusters and CCS groups capture individuals with a mental health diagnosis. Focusing on OLS, the predictive ratios based on the number of HCCs were similar across the two matched sample combinations for the 0, 1–3, 4–6, and 7–9 HCC groups. In the 10+ HCC group for these samples, the predictive ratios indicated underprediction. The predictive ratios by observed spending decile were poor across all fit and prediction samples. The systematic undercompensation produced in both the matched and random samples suggests there is space to improve incentives for plans to serve vulnerable subgroups.

We find common conditions have similar coefficients in the OLS formula fit on the matched and random samples. However, rare conditions necessarily occur in a small number of enrollees and estimating coefficients for conditions based on a small number of subjects leads to changing coefficients across samples and from year to year. Balancing the desire to base payment on a comprehensive set of conditions with adequate sample size is a difficult task. Whether to include rare conditions in the risk adjustment formula is a choice the regulator must make, either based on a prespecified empirical objective function or policy considerations. If rare conditions are to be included, widely varying coefficients may be addressed by constraining the coefficients of less

populous conditions to be equal to more stable condition coefficients; this is the approach CMS uses. Another potential solution is to use nonparametric methods like random forests, which may be less susceptible to gaming by insurers. While prior studies suggest parsimonious algorithms might reduce the possibility of gaming and address the issue of unstable coefficients for small groups (Rose 2016), we find the algorithms estimated with fewer risk adjusters identified by our “variable screening” LASSO yielded consistently lower R^2 and predictive ratio results. The practical impact of varying coefficients for groups depends both on the magnitude of the prediction error and on the size of the group; an insurer may care more about a smaller error for a much larger group than a larger error for an extremely rare group.

There are limitations to our study methodology. Notably, matching can create or exacerbate imbalance on unmatched variables. Here, we were unable to observe and therefore match directly on health status or comorbidity indicators; our matched sample was based on only a limited set of demographic variables available for both MA and TM enrollees. Matching improved balance on all observed variables, but both “rural” and “West” retained >10 percent absolute standardized differences in means, which may have reduced the similarity of our matched TM sample to the unobserved MA data. Despite using all available variables, this set of variables is unlikely to explain much of the residual difference between MA and TM enrollees after conditioning on the variables included in the risk adjustment formula. Given a more detailed set of variables, the improvements possible via the matching methods we implement here may be more substantial and able to target improvements for diagnosis-based subgroups. Although we cannot firmly assert whether our matched TM sample is a suitable proxy for the MA population without being able to capture these characteristics, which may be influenced by plan strategies to favorably select healthier beneficiaries, we did see a larger proportion of individuals in the matched sample with fewer HCCs than the random sample, as expected. If this is reflective of MA being generally healthier than TM, it points toward exploring alternative methods to address systematic undercompensation.

Although there is currently no widely available source of data that contains diagnoses and procedures for all MA enrollees, there are survey data and commercial claims that could be used for matching or weighting exercises prior to risk adjustment estimation. The Medicare Current Beneficiary Survey contains a variety of survey and administrative variables for both MA and TM enrollees, but is available for only a small subset of the Medicare population. The Health Care Cost Institute has also recently expanded access to their

claims data for commercially insured individuals, including claims from three large insurers that cover approximately 40 percent of MA enrollees.

CMS collects diagnoses and procedures on MA enrollees through both the Risk Adjustment Processing System and the new Encounter Data System. These data are not commonly available to researchers—unlike the claims data—and although the Government Accountability Office has recommended CMS continue conducting quality assessment, CMS has announced plans to begin releasing encounter data in late Fall 2018, starting with data from 2015. Our study did not find that nonrepresentative data in MA risk adjustment greatly exacerbate inappropriate payments to MA plans compared to a matched sample. Importantly, if additional variables on MA and TM enrollee characteristics are available in the future, a matched sample incorporating those characteristics using the methods described in this paper may yield improvement in MA risk adjustment performance and enhance payment accuracy.

ACKNOWLEDGMENT

Joint Acknowledgment/Disclosure Statement: This research was supported by the National Institute of Aging (P01-AG032952; P30-AG024409), the National Institute of Mental Health (NIMH T32-MH019733), and AcademyHealth (New Investigator Small Grant Program).

Disclosures: None.

Disclaimer: None.

REFERENCES

- Austin, P. C., and M. M. Mamdani. 2006. “A Comparison of Propensity Score Methods: A Case-Study Estimating the Effectiveness of Post-AMI Statin Use.” *Statistics in Medicine* 25 (12): 2084–106.
- Baicker, K., and J. A. Robbins. 2015. “Medicare Payments and System-Level Health-Care Use: The Spillover Effects of Medicare Managed Care.” *American Journal of Health Economics* 1 (4): 339–431.
- Brown, J., M. Duggan, I. Kuziemko, and W. Woolston. 2014. “How Does Risk Selection Respond to Risk Adjustment? New Evidence from the Medicare Advantage Program.” *American Economic Review* 104 (10): 3335–64.

- Curto, V., L. Einav, A. Finkelstein, J. D. Levin, and J. Bhattacharya. 2017. "Healthcare Spending and Utilization in Public and Private Medicare." NBER Working Paper, No. 23090.
- Frogner, B. K., G. F. Anderson, R. A. Cohen, and C. Abrams. 2011. "Incorporating New Research into Medicare Risk Adjustment." *Medical Care* 49 (3): 295–300.
- Geruso, M., and T. Layton. 2015. Upcoding: Evidence from Medicare on Squishy Risk Adjustment. NBER Working Paper, No. 21222.
- Jacobson, G., A. Damico, T. Neuman, and M. Gold. 2017. "Medicare Advantage 2017 Spotlight: Enrollment Market Update. Kaiser Family Foundation." [accessed on December 14, 2017]. Available at <https://www.kff.org/medicare/issue-brief/medicare-advantage-2017-spotlight-enrollment-market-update/>
- Kautter, J., G. C. Pope, M. Ingber, S. Freeman, L. Patterson, M. Cohen, and P. Keenan. 2014. "The HHS-HCC Risk Adjustment Model for Individual and Small Group Markets under the Affordable Care Act." *Medicare & Medicaid Research Review* 4 (3): E1–46.
- Landon, B. E., A. M. Zaslavsky, G. C. Pawlson, J. P. Newhouse, and J. Z. Ayanian. 2012. "Analysis of Medicare Advantage HMOs Compared with Traditional Medicare Shows Lower Use of Many Services During 2003–2009." *Health Affairs* 31 (12): 2609–17.
- Layton, T. J., T. G. McGuire, and R. van Kleef. 2018. "Deriving Risk Adjustment Payment Weights to Maximize Efficiency of Health Insurance Markets." *Journal of Health Economics* 61: 93–110.
- McGuire, T. G., J. P. Newhouse, S.-L. Normand, and J. Z. Shi. 2014. "Assessing Incentives for Service-Level Selection in Private Health Insurance Exchanges." *Journal of Health Economics* 35 (1): 47–63.
- McWilliams, J. M., J. Hsu, and J. P. Newhouse. 2012. "New Risk-Adjustment System was Associated with Reduced Favorable Selection in Medicare Advantage." *Health Affairs* 31 (12): 2630–40.
- Montz, E., T. Layton, A. B. Busch, R. P. Ellis, S. Rose, and T. G. McGuire. 2016. "Risk-Adjustment Simulation: Plans May Have Incentives to Distort Mental Health and Substance Use Coverage." *Health Affairs* 35 (6): 1022–8.
- Newhouse, J. P., and T. G. McGuire. 2014. "How Successful is Medicare Advantage?" *Milbank Quarterly* 92 (2): 351–94.
- Newhouse, J. P., M. Price, J. M. McWilliams, J. Hsu, and T. G. McGuire. 2015. "How Much Favorable Selection is Left in Medicare Advantage?" *American Journal of Health Economics* 1 (1): 1–26.
- Pope, G. C., J. Kautter, M. J. Ingber, S. Freeman, R. Sekar, and C. Newhart. *Evaluation of the CMS-HCC Risk Adjustment Model*. Baltimore, MD: The Centers for Medicare & Medicaid Services Office of Research; 2011.
- Rose, S. 2016. "A Machine Learning Framework for Health Plan Payment Risk Adjustment." *Health Services Research* 51 (6): 2358–74.
- Rose, S., S. Bergquist, and T. Layton. 2017. "Computational Health Economics for Identification of Unprofitable Health Care Enrollees." *Biostatistics* 18 (4): 682–94.

- Rose, S., J. Shi, T. G. McGuire, and S. L. Normand. 2015. "Matching and Imputation Methods for Risk Adjustment in the Health Insurance Marketplaces." *Statistics in Biosciences* 9 (2): 525–42.
- Rosenbaum, P. R., and D. B. Rubin. 1985. "The Bias Due to Incomplete Matching." *Biometrics* 41 (1): 103–16.
- Rudolph, K., and M. van der Laan. 2016. "Robust Estimation of Encouragement-Design Interventions Effects Transported across Sites." *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 79 (5): 1509–25.
- Shrestha, A., S. Bergquist, E. Montz, and S. Rose. 2018. "Mental Health Risk Adjustment with Clinical Categories and Machine Learning." *Health Services Research* 53 (4): 3189–206.
- Stuart, E. A. 2010. "Matching Methods for Causal Inference: A Review and Look Forward." *Statistical Science* 25 (1): 1–21.
- Stuart, E., C. Bradshaw, and P. Leaf. 2015. "Assessing the Generalizability of Randomized Trial Results to Target Populations." *Prevention Science* 16 (3): 475–85.
- Stuart, E. A., B. K. Lee, and F. P. Leacy. 2013. "Prognostic Score-Based Balance Measures can be a Useful Diagnostic for Propensity Score Methods in Comparative Effectiveness Research." *Journal of Clinical Epidemiology* 66 (8 0): S84–S90.ei.
- Stuart, E. A., S. Cole, C. Bradshaw, and P. Leaf. 2011. "The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials." *Journal of the Royal Statistical Society. Series A, (Statistics in Society)* 174 (2): 369–86.
- United States Government Accountability Office. "Medicare Advantage: Limited Program Made to Validate Encounter Data Used to Ensure Proper Payments." Report to Congress; 2017.
- Yun, H., M. L. Kilgore, J. R. Curtis, E. Delzell, L. C. Gary, K. G. Saag, M. A. Morrissey, D. Becker, R. Matthews, W. Smith, J. L. Locher. 2010. "Identifying Types of Nursing Facility Stays Using Medicare Claims Data: An Algorithm and Validation." *Health Services and Outcomes Research Methodology* 10 (1–2): 100–10.

SUPPORTING INFORMATION

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

Appendix SA1: Author Matrix.

Appendix SA2: Sensitivity Analyses.

Figure S1: Variable Balance after Matching on Propensity Score Estimated by LASSO and Small Caliper.

Figure S2: Variable Balance after Exact Matching on Urban/Rural, Geographic Region, and Race.

Figure S3: Predicted Value Difference between Match-Match and Random-Match Samples.

Figure S4: Change in OLS Coefficients between Random Sample Fit and Matched Sample Fit for Five Most Common Conditions.

Figure S5: Change in OLS Coefficients across 2008, 2009, and 2010 Matched Samples for Five Most Common Conditions.

Figure S6: Change in OLS Coefficients across 2008, 2009, and 2010 Random Samples for Five Most Common Conditions.

Figure S7: Variable Balance after Matching on Propensity Score Estimated by Logistic Regression with Small Caliper, 2008.

Figure S8: Variable Balance after Matching on Propensity Score Estimated by Logistic Regression with Small Caliper, 2009.

Table S1: R^2 and Mental Health Predictive Ratios across Fit and Prediction Samples, with LASSO Variable Selection.

Table S2: Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for OLS with LASSO Selected Risk Adjustors.

Table S3: Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for LASSO with LASSO Selected Risk Adjustors.

Table S4: Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for Random Forest with LASSO Selected Risk Adjustors.

Table S5: Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for LASSO with All Risk Adjustors.

Table S6: Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for Random Forest with All Risk Adjustors.

Table S7: R^2 and Mental Health Predictive Ratios across Fit and Prediction Samples, 2008.

Table S8: R^2 and Mental Health Predictive Ratios across Fit and Prediction Samples, 2009.

Table S9: Sample Proportion and Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for OLS with All Risk Adjustors, 2008.

Table S10: Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for OLS with LASSO Selected Risk Adjustors, 2008.

Table S11: Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for LASSO with All Risk Adjustors, 2008.

Table S12: Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for LASSO with LASSO Selected Risk Adjustors, 2008.

Table S13: Predictive Ratios by Number of HCCs and Matched Sample

Observed Spending Decile for Random Forest with All Risk Adjustors, 2008.

Table S14: Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for Random Forest with LASSO Selected Risk Adjustors, 2008.

Table S15: Sample Proportion and Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for OLS with All Risk Adjustors, 2009.

Table S16: Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for OLS with LASSO Selected Risk Adjustors, 2009.

Table S17: Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for LASSO with All Risk Adjustors, 2009.

Table S18: Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for LASSO with LASSO Selected Risk Adjustors, 2009.

Table S19: Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for Random Forest with All Risk Adjustors, 2009.

Table S20: Predictive Ratios by Number of HCCs and Matched Sample Observed Spending Decile for Random Forest with LASSO Selected Risk Adjustors, 2009.