HSR Health Services Research

DOI: 10.1111/1475-6773.12848
METHODS ARTICLE

Robust Machine Learning Variable Importance Analyses of Medical Conditions for Health Care Spending

Sherri Rose

Objective. To propose nonparametric double robust machine learning in variable importance analyses of medical conditions for health spending.
Data Sources. 2011–2012 Truven MarketScan database.
Study Design. I evaluate how much more, on average, commercially insured enrollees with each of 26 of the most prevalent medical conditions cost per year after controlling for demographics and other medical conditions. This is accomplished within the nonparametric targeted learning framework, which incorporates ensemble machine learning. Previous literature studying the impact of medical conditions on health care spending has almost exclusively focused on parametric risk adjustment; thus, I compare my approach to parametric regression.
Principal Findings. My results demonstrate that multiple sclerosis, congestive heart failure, severe cancers, major depression and bipolar disorders, and chronic hepatitis are the most costly medical conditions on average per individual. These findings differed from those obtained using parametric regression.
Conclusions. The literature may be underestimating the spending contributions of several medical conditions, which is a potentially critical oversight. If current methods are not capturing the true incremental effect of medical conditions, undesirable incentives related to care may remain. Further work is needed to directly study these issues in the context of federal formulas.
Key Words. Risk adjustment, machine learning, regression

Health care spending has frequently been investigated in the context of longitudinal changes in total spending level or total spending growth (Chernew and Newhouse 2012). Simultaneously understanding the individual contributions of medical conditions to health care spending is typically examined in the context of "risk-adjusted" formulas for plan payment (Iezzoni 1997; Kautter et al. 2014). Risk adjustment in plan payment aims to redistribute funds according to the likely cost of enrollees, thereby encouraging health plan competition

based on quality and efficiency versus avoiding high-cost enrollees. Plans with costlier enrollees should receive larger payments than plans with less expensive enrollees (Iezzoni 1997). Properly attributing the relationship of specific medical conditions to total annual health care spending is critical to successful risk adjustment. Notably, however, the purpose is not to assess the impact of medical conditions on health care spending, but rather to use medical conditions to predict spending.

From a statistical standpoint, these risk-adjusted plan payment formulas make trade-offs for bias and variance for the conditional expectation of spending given adjustment variables in parametric regression models. The goal is to generate the best overall estimator for prediction; individual effects of medical condition categories on health care spending, as represented by coefficients in the parametric regression model, will only be unbiased if the parametric model is correctly specified, which is unlikely in practice. Because spending is a continuous outcome, this approach is also analogous to a parametric maximum-likelihood-based substitution estimator of the g-formula for cross-sectional data (Robins 1986) when no interactions are included.

Risk adjustment formulas for plan payment are common around the globe, and although the intricacies of their implementation differ, none feature machine learning-based methods. Other commonalities across the systems in Germany, the Netherlands, Belgium, and the United States, among others, include the use of provider-supplied diagnoses in parametric formulas. The choice of total annual spending versus payments and the services that comprise the spending outcome also vary, but typically the relationship between the possible spending outcomes is extremely highly correlated (Ellis, Martins, and Rose 2018). Some countries feature multiple risk adjustment formulas within the same health insurance program, such as the Netherlands, with four in their social health insurance system, including two focused on short-term and long-term mental health care spending (van Kleef et al. 2018).

As a specific example, the individual health insurance Marketplaces created after the establishment of the Patient Protection and Affordable Care Act (ACA) in the United States use ordinary least squares parametric regression to develop a risk adjustment formula. These parametric regressions include age cells and medical condition category variables (Kautter et al. 2014). Even if these health insurance Marketplaces change in form due to potential repeal of portions of the ACA, plan payment risk adjustment is needed in any type of

Address correspondence to Sherri Rose, Ph.D., Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115; e-mail: rose@hcp.med.harvard.edu.

regulated individual health insurance market. Medicare Advantage in the United States, a managed health care program for persons aged 65 and older and those with certain disabilities, also has a risk adjustment system that uses similar ordinary least squares parametric regression techniques for risk adjustment (Pope et al. 2011). Two studies in the United States have performed risk adjustment using decision trees in more narrowly focused spending prediction applications (Relles, Ridgeway, and Carter 2002; Drozd et al. 2006) with two other papers using ensemble learning to predict total health spending (Rose 2016) and mental health spending (Shrestha et al. 2018), although these methods have never been applied to plan payment in practice.

Other research in the United States has addressed related questions, such as estimating total spending or spending growth of medical conditions as survey means over time (Thorpe, Florence, and Joski 2004; Roehrig et al. 2009). In one study of average spending spanning 1996 to 2005, mental disorders and heart conditions were the most costly medical conditions (Roehrig et al. 2009). They found that mental disorders totaled $142.2 billion and heart conditions totaled $253.9 billion in 2005. An earlier analysis considered spending growth and found that the five most expensive conditions contributing to spending *growth* were heart disease, mental disorders, pulmonary disorders, cancer, and trauma (Thorpe, Florence, and Joski 2004). Both studies involved tabulated survey means and did not adjust for measured confounding. Importantly, these studies, as with the other studies, did not examine the average contribution of each medical condition to total health care spending for an individual.

Variable importance questions generally aim to assess the relationship of a set of variables on an outcome. With high-dimensional data, researchers are often interested in understanding the relative or absolute effect of the components in a list of covariates on an outcome. (I could also refer to these as measures of association.) The desired result is frequently an ordered list of effect estimates, ranked by effect size or p-value. These variable importance analyses have become commonplace in other applied literatures, such as in genetics and genomics, but not health economics. In quantitative trait loci mapping, for example, many techniques (e.g., univariate regression, interval mapping, composite interval mapping, multiple interval mapping, and Bayesian models) are parametric and also assume Gaussian errors, relying heavily on the correct specification of the functional form to have reliable performance (Lander and Botstein 1989; Haley and Knott 1992; Jansen 1993; Zeng 1994). Decision trees, in particular random forests, have been frequently used in varied genomic variable importance applications (Breiman et al. 1984;

Gromping 2012), although overfitting can be a concern, even with the use of cross-validation (Segal 2004; Strobl, Malley, and Tutz 2009).

Nonparametric double robust targeted learning methods for variable importance have been shown to outperform these other techniques. Much of this work, as with the other methods, has been in biomarker detection and genomics applications (Bembom et al. 2009; Wang, Rose, and van der Laan 2011; Chambaz, Neuvial, and van der Laan 2012; Wang et al. 2014). In this study, I propose implementing targeted learning methods to understand the impact of medical conditions on total annual health care spending, translating this scientific question as a variable importance problem for the first time. Targeted learning is an estimation framework featuring targeted maximum-likelihood estimators (TMLEs) that incorporate super learning, an ensembled machine learning technique, for effect estimation. I estimate the impact of medical condition categories based on their contributions to total health care spending, controlling for demographic information and other medical conditions. Thus, this study differs from previous work on medical condition spending as it (a) estimates how much more, on average, enrollees with each medical condition cost per year after controlling for measured confounders, (b) formulates the estimation problem as a variable importance analysis, and (c) provides double robust machine learning-based inference. Accurately assessing the contributions of medical conditions is critical. If current methods are not capturing the true incremental effect of medical conditions, undesirable incentives related to care may remain.

METHODS

I first translate the health care spending question as a variable importance problem. Thus, the statistical framework considered here examines an experiment where one randomly samples a unit from the population of interest and measure baseline features, a list of medical conditions, and total annual medical expenditures from the following year (the outcome variable). All medical conditions are binary flags, and I am interested in estimating the effect of each medical condition on the outcome, controlling for all other medical conditions and baseline features.

Data

The Truven MarketScan database, containing as many as 51 million enrollees annually, is the current source of subjects used in risk adjustment for

the state-level individual health insurance Marketplaces created by the ACA. This is despite the fact that MarketScan is an enrollment and claims database that includes records submitted by employers and private health plans, which means it contains enrollees ineligible for the Marketplaces (Adamson, Chang, and Hansen 2008; Kautter et al. 2014). Further discussion of the use of this database for the Marketplaces can be found elsewhere (Kautter et al. 2014; Layton, Ellis, and McGuire 2015). Here, I aim to make inferences about commercially insured enrollees only, although the generalizability of results for Marketplaces is an active area of continuing work. I also note that the current formulas for risk adjustment in the ACA are concurrent and use covariates and spending outcomes from the same year. I focus on so-called prospective risk adjustment where the spending variable is from the following year to ensure reasonable interpretation of the statistical target parameter discussed later in this section.

The study sample was constructed by drawing 1,000,000 subjects from the population of MarketScan enrollees with 2 years of sequential coverage in 2011–2012. Enrollees were eligible for insurance throughout this entire 24-month period, and thus, there is no drop-out due to death. This enrollment period requirement, which results in the feature that there is no drop-out due to death, is not a simplification of this real-world problem. In fact, it reflects current recommendations for performing risk adjustment when there is imperfect eligibility information available, which is why this requirement was enforced. Annualization of spending and then weighting by the portion of the year the enrollee is eligible is preferred when reliable eligibility information is obtainable (Ellis, Martins, and Rose 2018). Demographic baseline variables from 2011 included sex, age, metropolitan statistical area, region, and inpatient diagnoses. The 74 medical condition categories considered were constructed as Hierarchical Condition Category (HCC) variables using ICD-9 codes, which mimics the medical condition categories used in risk adjustment by the federal government (Kautter et al. 2014). Total annual medical spending from 2012 was the outcome variable.

Statistical Analysis Procedure

I formalize the observed data by defining $O = (W, M, Y)$ drawn from probability distribution P as the observational unit, which has $n = 1{,}000{,}000$ i.i.d. realizations in the study sample. Baseline variables discussed above are denoted by W, with M as the vector of medical condition categories, and Y as the outcome total annual medical expenditures. The additional notation A is used to

denote the current medical condition category under consideration, and M^- represents the vector M that excludes the variable A. I assume a nonparametric model and define the parameter of interest as: $\psi = E_{W,M^-}[E(Y|A = 1, W, M^-) - E(Y|A = 0, W, M^-)]$, which represents the effect of $A = 1$ versus $A = 0$ after adjusting for all other medical conditions M^- and baseline variables W. In this study, the parameter can be interpreted as the difference in total annual expenditures the following year when the enrollee has the medical condition under consideration (i.e., $A = 1$), adjusted for measured confounders. This parameter is equivalent to the coefficient in front of A in a parametric regression with a continuous outcome containing no interaction terms, except written nonparametrically.

While my nonparametric statistical model does not make restrictive assumptions about the functional form of the data-generating distributions, I do make two key statistical assumptions beyond i.i.d. mentioned above. These assumptions are important for both targeted learning and parametric regression. I make a time ordering assumption that requires that W and M occur before Y for proper interpretation of this parameter as an effect parameter, which is the case in this data. I also assume positivity as part of the statistical model; that variation in the medical condition under consideration occurs in the strata of W and M^-. I evaluated for and did not find positivity violations.

Of note, I do not augment my statistical model with additional causal assumptions (Pearl 2009). These assumptions are likely violated in this setting, particularly the randomization assumption (i.e., no unmeasured confounding). There are many variables that are not collected or available in the claims records used for risk adjustment, some with a plausibly confounding relationship between the medical conditions and health spending, such as socioeconomic status or education. However, I am specifically interested in informing *policy* and a *statistical* estimation question targeting the individual medical conditions typically used in risk adjustment formulas. Comorbid medical conditions are also considered contemporaneously as whether the enrollee had them at all in the base year. For example, to understand the impact of multiple sclerosis on next year's costs, I control for that individual's heart disease. Disentangling the causality of medical conditions within the base year (e.g., was the heart disease condition caused by multiple sclerosis) is not needed for this policy question and estimand. My approach therefore mimics the variables and considerations used in risk adjustment analyses, except with potentially improved estimation of the effects of each condition. The target parameter does not have a causal interpretation, nor do I seek to make causal assertions.

Targeted learning methods, as used in variable importance questions, estimate the effect of each variable from a list of variables on the outcome while adjusting for measured covariates. I focused on implementing TMLEs in this variable importance question for health spending due to its attractive statistical properties, particularly double robustness. With a consistent estimator of either the outcome regression *or* the probability of having the medical condition of interest given other covariates, the TMLEs for the target parameters will be consistent, and also efficient if both are consistently estimated. Other approaches commonly used for variable importance do not have this property. Machine learning methods, which can include decision trees and penalized regressions, are completely integrated into the estimation of the relevant components (e.g., outcome regression) of the TMLE algorithm.

The TMLE is a generalization of parametric maximum likelihood estimators for semiparametric and nonparametric models. With a target parameter for each medical condition category, the TMLE comprises two steps: an initial machine learning estimate of $E(Y|A, W, M^-)$ and a bias reduction step that uses information from a machine learning-based estimate of $P(A = 1|W, M^-)$. This bias reduction stage is necessary, as the initial estimator for $E(Y|A, W, M^-)$ is built based on the bias–variance trade-off for that conditional mean and not the effect parameter. Thus, the second stage focuses on a bias–variance trade-off for the parameter of interest instead, fluctuating the initial estimate of $E(Y|A, W, M^-)$ in a one-dimensional submodel described in the Supplementary Appendix.

I used the ensembling technique super learning to estimate both $E(Y|A, W, M^-)$ and $P(A = 1|W, M^-)$ (van der Laan, Polley, and Hubbard 2007). Super learning constructs a prediction function that is the optimal weighted average of all considered algorithms, based on an a priori selected loss function. As discussed above, the double robustness property of TMLE indicates that the estimator will be consistent if either component is consistently estimated, and efficient if both are consistently estimated. There are no performance guarantees in the setting that neither the outcome regression nor the probability of having the medical condition of interest given other covariates is estimated consistently. However, using ensemble machine learning to estimate these components, versus implementing a single a priori specified algorithm, one is more likely to obtain consistent estimates in practice. Additional technical details for TMLE, including a transformation for bounded continuous outcomes (Gruber and van der Laan 2010) and the influence curve definition used to calculate standard errors, are also available in the Supplementary Appendix.

Previous theoretical and empirical papers, including those with simulation studies, have demonstrated that TMLE will often outperform parametric regression when the regression is misspecified (e.g., omitted variable bias) and in high-dimensional settings with collinear variables (Bembom et al. 2009; van der Laan and Rose 2011; Wang, Rose, and van der Laan 2011; Wang et al. 2014). Parametric regression may perform better than TMLE in cases where $E(Y|A, W, M^{-})$ is well approximated by a parametric regression. In this case, TMLE will also consistently estimate the parameter of interest, but will have larger variance than the parametric regression. A straightforward tutorial on TMLEs for applied audiences featuring a continuous outcome, average treatment effect for a binary A, and the incorporation of machine learning has been published (Schuler and Rose 2017). The parameter estimated in that paper is therefore an analogous parameter to the one I study here, and it may be of interest to the reader, although I note that it is a single parameter and not a variable importance analysis. It also includes a reproducible simulation study with publicly available code to compare TMLE and parametric regression.

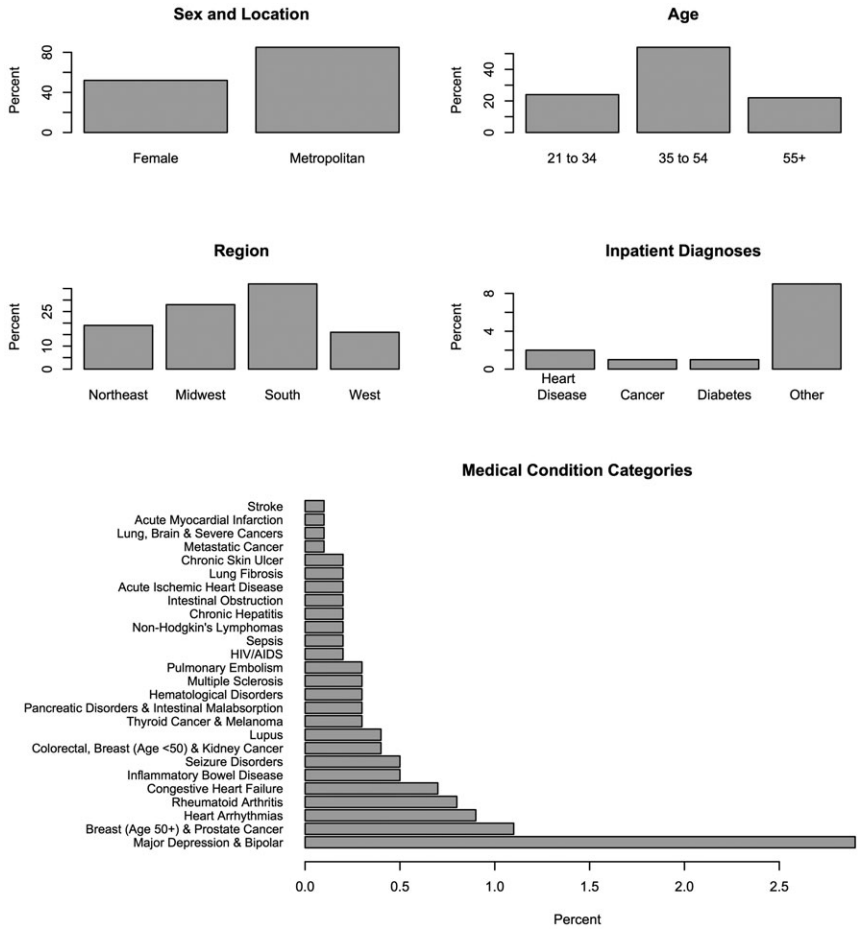
Given the rarity of the medical condition categories, I set a minimum threshold of 1,000 events in the sample to be considered as an A variable. Of the 74 medical condition categories, 26 met this threshold. While 48 medical condition categories were not rotated into consideration as A, they remained part of the vector M^{-} in all estimators to adjust for measured confounding along with the W vector. An initial estimate for the outcome regression $E(Y|A, W, M^{-})$ and the estimate for $P(A = 1|W, M^{-})$ were generated with super learning. For this study, the super learner contained three algorithms: main terms linear regression, lasso penalized regression, and a neural network with 2 units in the hidden layer.

The TMLE for each A was calculated separately, and influence curve-based estimates were used to calculate standard errors and corresponding p-values, adjusting for multiple testing using the false discovery rate at an 0.05 level (Benjamini and Hochberg 1995) as in previous work (Wang et al. 2014). The estimators were implemented in the *R* programming language, relying on the "tmle" (Gruber and van der Laan 2012) and "SuperLearner" (Polley and van der Laan 2013) packages. Scripts were parallelized such that each of the 26 effect estimators ran, on average, in under 5 hours.

RESULTS

The study variables from the MarketScan sample (n = 1,000,000) are summarized in Figure 1. The majority of enrollees were female, age 35–54, and 85

Figure 1: Demographic and Health Summary Information for MarketScan Sample



percent of enrollees were from a metropolitan statistical area. The most populous region was the South, with 37 percent of enrollees. Inpatient diagnoses for heart disease (2 percent), cancer (1 percent), and diabetes (1 percent) were rare, although 9 percent of enrollees had at least one other inpatient diagnosis. Because they represent only inpatient codes, they do not reflect, nor would one expect them to, overall prevalences of these conditions, which are higher. The medical condition categories, even among those with at least 1,000 events included in Figure 1, were also rare. The medical condition category

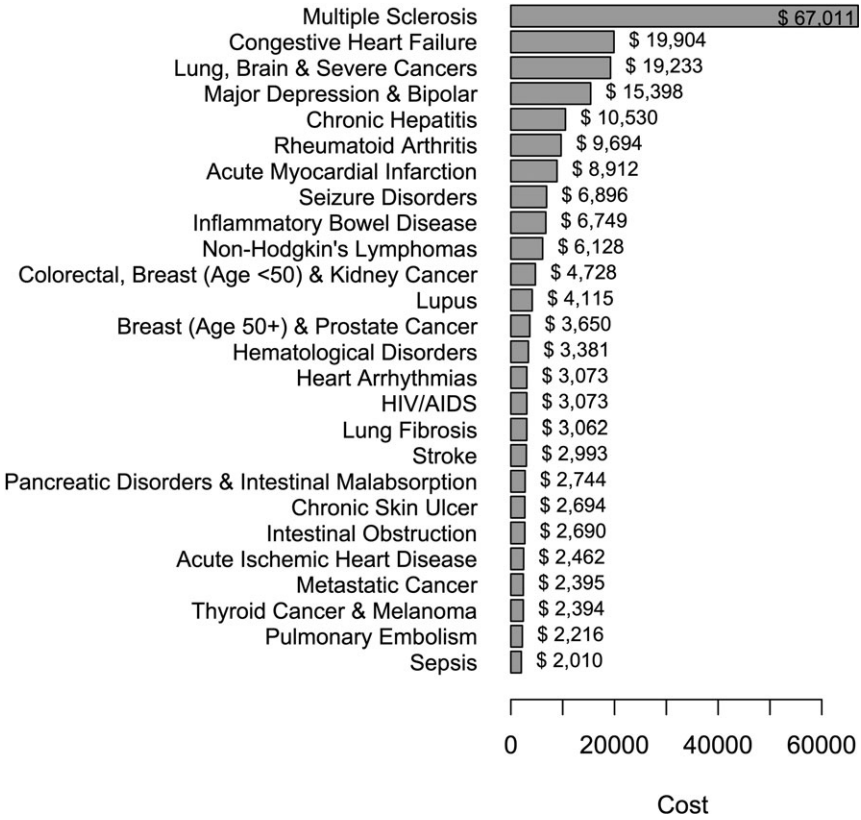
impacting the largest percentage of the sample was major depression and bipolar (2.9 percent) followed by the category that included breast cancer among women age 50 and older, prostate cancer, and benign or uncertain brain tumors (1.1 percent). All other medical condition categories had a percentage under 1.0 percent.

My evaluation focused on producing an ordered list of medical conditions based on how much more an enrollee with each medical condition would cost per year after controlling for other medical conditions and baseline variables. I found that the top five most expensive medical conditions were multiple sclerosis; congestive heart failure; lung, brain, and other severe cancers; major depression and bipolar disorders; and chronic hepatitis (Figure 2). All five conditions exceeded $10,000, with multiple sclerosis topping out at $67,011. Effect estimates presented in Figure 2 are all statistically significant at $p < .001$ after multiple testing correction with the false discovery rate, and I focus on effect size given the size of the data. (Confidence intervals are not displayed given the narrow bands obtained.) I note that these cost estimates do not necessarily reflect only the cost to treat that specific medical condition, but represent the increased overall costs of treating all conditions for individuals with the condition of interest.

In this work, I also include the results of main terms parametric regression, which is similar to current practice in plan payment risk adjustment and would be of interest to policy makers. The targeted learning results differed nontrivially from what would have been estimated in a standard risk adjustment formula with the same variables. Not only do the effect estimates differ, but the relative ranks as well (Figure 3). Parametric regression ranks metastatic cancer; multiple sclerosis; lung, brain, and other severe cancers; HIV/AIDS; and non-Hodgkin's lymphomas as the top five most expensive medical conditions. Only two of these conditions appeared in the top five ranking obtained using targeted learning. Although multiple sclerosis and severe cancers appear on both top five lists, the cost of multiple sclerosis estimated using targeted learning methods is twice that of parametric regression. Conversely, the effect estimate for metastatic cancer in parametric regression is more than 15-fold that estimated with targeted learning. When examining the top 10 largest effect estimates based on targeted learning, six were larger than the corresponding parametric regression estimates (Figure 4). Overall, 17 of the 26 effect estimates obtained using targeted learning were larger than those obtained with parametric regression (Figure 3).

Effect estimates in parametric regression were statistically significant at $p < .05$, except for sepsis. Given the concern related to parametric regression

Figure 2: Targeted Learning Effect Estimates in MarketScan Sample

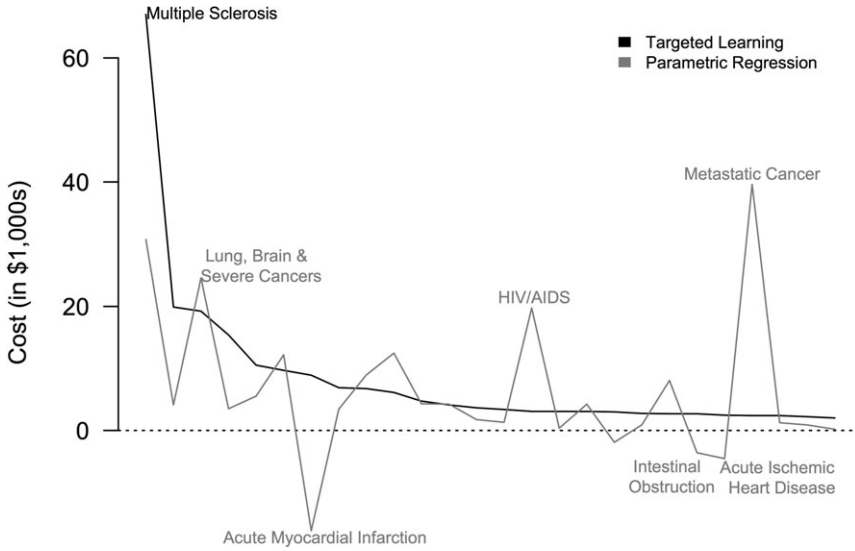


Note. Medical conditions and baseline variables were from 2011, and the outcome total health care spending was from 2012. Confidence intervals were suppressed from the plot for readability given the narrow bands obtained. All effects were statistically significant at $p < .001$ after multiple testing correction with the false discovery rate.

misspecification, I performed multiple fit diagnostics. The adjusted R^2 was 26.5 percent, which is low, although in line with other Marketplace estimates (Kautter et al. 2014; Rose 2016). I unsurprisingly found evidence of nonnormality, outliers, and heteroskedastic errors. (See Supplementary Appendix for detailed discussion of the parametric regression fit diagnostics.)

Parametric regression estimates for four medical condition categories were negative (Figure 3). This paradoxically indicates that the incremental effect of the medical condition is a cost savings (acute myocardial infarction, −$16,171; unstable angina and other acute ischemic heart

Figure 3: Comparison of Effect Estimates for Targeted Learning and Parametric Regression

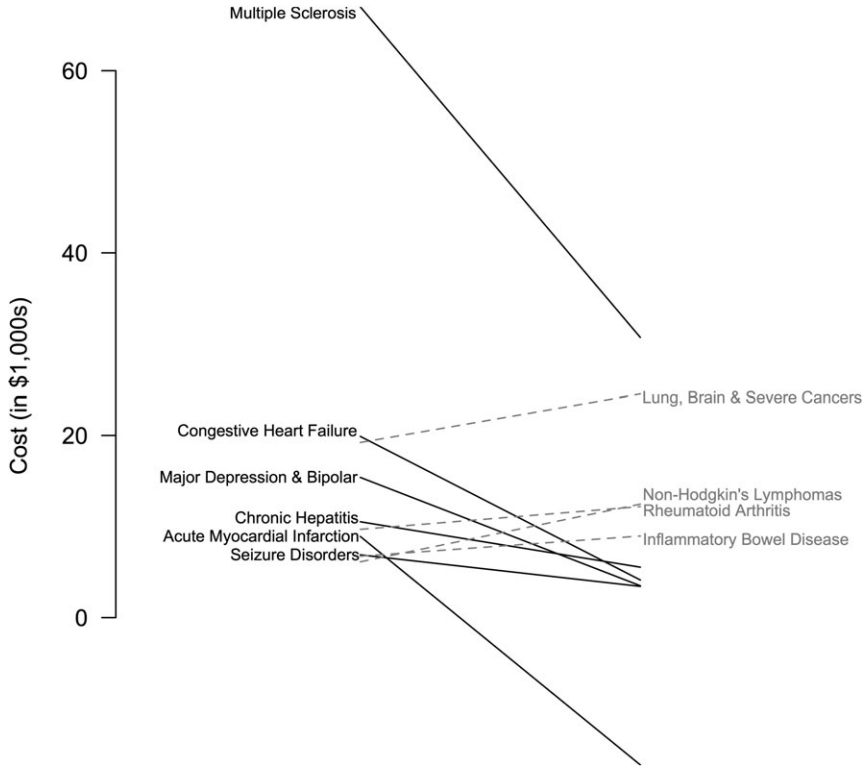


disease, −$4,539; intestinal obstruction, −$3,596; and ischemic or unspecified stroke, −$1,912). As this cohort of enrollees had no drop-out due to death and was continuous enrolled, negative estimates of these values are especially unexpected. All estimates obtained using targeted learning were positive. Employing a constrained parametric regression that did not allow for negative effect estimates merely shrank these four variables to zero while not substantially changing other effect estimates (results not shown). To even more closely emulate the federal formula, another sensitivity analysis for parametric regression added interaction terms between HCCs with at least 1,000 events and a "severe illness indicator" defined similarly to Kautter et al. (2014). Effect estimates were comparable to the main terms regression for most conditions.

DISCUSSION

Adjusting for health conditions is ubiquitous in health care. The federal government, as well as health plans and provider organizations, routinely relies on risk adjustment to estimate the contribution of medical conditions to

Figure 4: Top 10 Largest Targeted Learning Effect Estimates



Note. Left end of each line segment is targeted learning value; right end of each line segment is parametric regression value. Medical conditions with larger targeted learning effect estimates relative to parametric regression appear with solid black lines; those with smaller effect estimates appear with dashed gray lines.

overall health care spending. Additionally, risk adjustment also forms the cornerstone of efforts to improve quality in pay-for-performance and public reporting systems. Accurately assessing the impact of medical conditions is crucial, an endeavor that may be substantially improved by incorporating flexible statistical methods in nonparametric models. The degree to which these techniques more accurately estimate the impact of medical conditions on health care spending in comparison with standard risk adjustment was previously unexplored prior to this research and has important implications. This study evaluated how much more enrollees with each of the 26 most prevalent medical conditions cost per year after controlling for measured confounders,

including demographic information and other medical conditions, and yielded an ordered list of conditions.

This research reflects the first investigation of the impact of medical condition categories on total annual health care expenditures as a variable importance question using double robust machine learning-based estimators. I found that the expected incremental costs of specific medical condition categories deviated considerably from estimates in a standard parametric risk adjustment framework. These differences are likely due to the bias associated with misspecified parametric models, as described in other theoretical work as well as simulations (e.g., van der Laan and Rose 2011; Chambaz, Neuvial, and van der Laan 2012). Previous variable importance analyses have also demonstrated large differences between regression estimators and targeted learning performance in applications and simulation studies where targeted learning had lower bias (e.g., Bembom et al. 2009; van der Laan and Rose 2011; Wang, Rose, and van der Laan 2011; Wang et al. 2014). This can also be seen for the average treatment effect parameter described in the Schuler and Rose (2017) tutorial, as well as other papers.

My work has some limitations. The set of adjustment variables was not identical to those used by the United States Department of Health and Human Services (HHS). I note that the full HHS risk adjustment formula includes an additional 17 medical condition categories, a slightly different set of interaction variables compared to my sensitivity analysis, 18 demographic group variables defined differently than the baseline demographic information I include here, and excludes sex. The HHS estimates also reflect that each tier adjusts the total annual spending outcome variable so that it only includes plan-paid spending. As discussed earlier, Marketplace risk adjustment is performed concurrently, using a spending outcome from the same year as the covariate data. (Prospective risk adjustment, as considered in my study, is still highly relevant as it is used in the Medicare Advantage program and state Medicaid managed care programs.) Given the differences in variable definitions and variables included, I cannot make strong claims that my results differ from those estimates currently used by HHS, only the parametric regressions implemented in this study. However, I present select values in Table 1 for approximate comparisons as a point of discussion. The targeted learning estimates for the top five medical conditions mostly differed from the normalized coefficient values used by HHS for each plan tier (i.e., platinum, gold, silver, bronze, and catastrophic) (Kautter et al. 2014). The greatest differences are seen between the estimates for multiple sclerosis, severe cancers, and major depression and bipolar.

Table 1: Top Five Targeted Learning Effect Estimates with Alternative Values from HHS Platinum Plans, HHS Catastrophic Plans, and Parametric Regression

Condition	*Targeted Learning*	*HHS Platinum*	*HHS Catastrophic*	*Parametric Regression*
Multiple sclerosis	$67,011	$40,665	$37,435	$30,715
Congestive heart failure	19,904	20,712	19,641	4,131
Lung, brain, and severe cancers	19,233	64,438	61,399	24,528
Major depression and bipolar	15,398	10,220	7,848	3,498
Chronic hepatitis	10,530	7,498	5,716	5,539

Note. HHS estimates were multiplied by mean spending from the source population ($5,465) to approximately correct for the HHS normalization standard to divide by average cost.

Thus, an immediate area of possible future work would be to obtain the relevant additional data and directly replicate the exact data structure and regression formula used by HHS. It would then be possible to compare the HHS effect estimates formally to targeted learning estimators for this variable importance question. It may also be interesting to examine medical conditions based on other classification systems that have more clinically interpretable categories than those used by HHS, for example, implementing variable importance for diagnosis-related groups.

Notably, I identified major depressive and bipolar disorders and chronic hepatitis as two of the most costly conditions, whereas a standard risk adjustment approach did not. This is striking given the rising incidence of these conditions nationally in the United States and that these estimates represent the additional overall costs of treating all conditions for individuals with these conditions, not necessarily only the cost to treat those specific conditions. It is important to consider the national trends in incidence as well as improvements in mortality and other health outcomes associated with the medical conditions in the highest spending categories to inform a national picture of cost (Thorpe, Florence, and Joski 2004). If current risk adjustment methods are not capturing the true incremental effect of medical conditions, undesirable incentives related to prevention of disease and care may remain. For example, if major depression and bipolar have a marginal incremental cost of $15,398, but insurers are paying only $3,498, insurers are not incentivized to invest in preventing mental health disorders in their plan offerings.

Previous literature has also demonstrated that mental health and substance use disorders are substantially undercompensated in the current Marketplace risk adjustment system (McGuire 2016; Montz et al. 2016). Montz et al. 2016 found that 80 percent of Marketplace enrollees with mental health and substance use disorders have higher than average spending not compensated for by other comorbidities, but their mental health ICD-9 codes do not map to an HCC included in the risk adjustment formula. This provides incentives to plans to avoid enrollees with mental health and substance use disorders because they are unprofitable, possibly by designing their drug formulary or provider network to be undesirable to those with these conditions (Layton, Ellis, and McGuire 2015; Montz et al. 2016). This study provides additional evidence that the current risk adjustment system may underestimate the incremental costs of certain health conditions, including some mental health conditions. While further work is needed, I expand the evidence base that policy changes to risk adjustment formulas may be warranted. This could include altering risk adjustment methodology, such as adding HCCs or prescription drug flags to existing formulas (McGuire 2016; Montz et al. 2016). Of note, improvements in access to care and the health needs of individuals with mental health and substance use disorders over the last 50 years have been achieved partly through changes in financing rather than new treatments (McGuire 2016).

This study is one of very few studies in health services research to incorporate double robust machine learning-based estimation. This is important given that I found disparate results between targeted learning methods and standard parametric regression, as has been shown in other fields. The risk adjustment literature is potentially appreciably underestimating the spending contributions of numerous medical conditions, including multiple sclerosis, congestive heart failure, major depressive and bipolar disorders, and chronic hepatitis. These results provide early empirical evidence that these techniques may provide improved inferences for important health services research questions, and this work therefore has broader implications beyond studying the impact of medical conditions on health spending.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This work was supported by NIMH grant number R01-MH094290 and the Laura and John Arnold Foundation.

The author thanks Thomas McGuire, Bapu Jena, and the Health Policy Data Science Lab for helpful comments.

Disclosures: None.

Disclaimers: None.

REFERENCES

Adamson, D., S. Chang, and L. Hansen. 2008. *Health Research Data for the Real World: The MarketScan Databases*. New York: Thompson Healthcare.

Bembom, O., M. Petersen, S.-Y. Rhee, W. Fessel, S. Sinisi, R. Shafer, and M. van der Laan. 2009. "Biomarker Discovery Using Targeted Maximum Likelihood Estimation: Application to the Treatment of Antiretroviral Resistant HIV Infection." *Statistics in Medicine* 25 (1): 152–72.

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall.

Chambaz, A., P. Neuvial, and M. van der Laan. 2012. "Estimation of a Non-parametric Variable Importance Measure of a Continuous Exposure." *Electronic Journal of Statistics* 6: 1059–99.

Chernew, M., and J. Newhouse. 2012. "Health Care Spending Growth." In *Handbook of Health Economics*, edited by M. Pauly, T. McGuire, and P. Barros, pp. 1–43. Maryland Heights, MO: Elsevier.

Drozd, E., J. Cromwell, B. Gage, J. Maier, L. Greenwald, and H. Goldman. 2006. "Patient Casemix Classification for Medicare Psychiatric Prospective Payment." *American Journal of Psychiatry* 163 (4): 724–32.

Ellis, R., B. Martins, and S. Rose. 2018. Risk Adjustment for Health Plan Payment. In *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets: Theory and Practice*, edited by T. McGuire, and R. van Kleef. New York: Elsevier.

Gromping, U. 2012. "Variable Importance Assessment in Regression: Linear Regression Versus Random Forest." *The American Statistician* 63: 308–19.

Gruber, S., and M. van der Laan. 2010. "A Targeted Maximum Likelihood Estimator of a Causal Effect on a Bounded Continuous Outcome." *International Journal of Biostatistics* 6: Article 26.

———. 2012. "tmle: An R Package for Targeted Maximum Likelihood Estimation." *Journal of Statistical Software* 51: 1–35.

Haley, C., and S. Knott. 1992. "A Simple Regression Method for Mapping Quantitative Trait Loci in Line Crosses Using Flanking Markers." *Heredity* 69: 315–24.

Iezzoni, L. 1997. *Risk Adjustment for Measuring Healthcare Outcomes*. Chicago, IL: Health Administration Press.

Jansen, R. 1993. "Interval Mapping of Multiple Quantitative Trait Loci." *Genetics* 135: 205–11.

Kautter, J., G. 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).

van Kleef, R. C., F. Eijkenaar, R. van Vliet, and W. P. van de Ven. 2018. "Health Plan Payment in the Netherlands." In *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets: Theory and Practice*, edited by T. McGuire, and R. van Kleef. New York: Elsevier.

van der Laan, M., E. Polley, and A. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6: Article 25.

van der Laan, M., and S. Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data.* New York: Springer.

Lander, E., and D. Botstein. 1989. "Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps." *Genetics* 121: 185–99.

Layton, T., R. Ellis, and T. McGuire. 2015. "Assessing Incentives for Adverse Selection in Health Plan Payment Systems." Tech. rep., National Bureau of Economic Research.

McGuire, T. 2016. "Achieving Mental Health Care Parity Might Require Changes in Payments and Competition." *Health Affairs* 35 (6): 1029–35.

Montz, E. T., A. Layton, R. Busch, S. Rose Ellis, and T. McGuire. 2016. "Risk Adjustment Simulation: Plans May Have Incentives to Distort Mental Health and Substance Use Coverage." *Health Affairs* 35 (6): 1022–8.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*, 2d Edition. New York: Cambridge University Press.

Polley, E., and M. van der Laan. 2013. *SuperLearner: Super Learner for Prediction. R Package Version 2.0-21.*

Pope, G., J. Kautter, M. Ingber, S. Freeman, R. Sekar, and C. Newhart. 2011. "Evaluation of the CMS-HCC risk adjustment model." Tech. rep., RTI International and the Centers for Medicare and Medicaid Services.

Relles, D., G. Ridgeway, and G. Carter. 2002. "Data Mining and the Implementation of a Prospective Payment System for Inpatient Rehabilitation." *Health Services and Outcomes Research Methodology* 3: 247–66.

Robins, J. 1986. "A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods: Application to Control of the Healthy Worker Survivor Effect." *Mathematical Modelling* 7: 1393–512.

Roehrig, C., G. Miller, C. Lake, and J. Bryant. 2009. "National Health Spending by Medical Condition, 1996–2005." *Health Affairs* 28: w358–67.

Rose, S. 2016. "A Machine Learning Framework for Plan Payment Risk Adjustment." *Health Services Research* 51 (6): 2358–74.

Schuler, M., and S. Rose. 2017. "Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies." *American Journal of Epidemiology* 185 (1): 65–73.

Segal, M. 2004. Machine Learning Benchmarks and Random Forest Regression. Tech. rep.: Center for Bioinformatics and Molecular Biostatistics, University of California, San Francisco.

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 Pt 2): 3189–206. https://doi.org/10.1111/1475-6773.12818

Strobl, C., J. Malley, and G. Tutz. 2009. "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests." *Psychological Methods* 14 (4): 323–48.

Thorpe, K., C. Florence, and P. Joski. 2004. "Which Medical Conditions Account for the Rise in Health Care Spending?" *Health Affairs* 23: 437–45.

Wang, H., S. Rose, and M. van der Laan. 2011. "Finding Quantitative Trait Loci Genes with Collaborative Targeted Maximum Likelihood Learning." *Statistics and Probability Letters* 81: 792–6.

Wang, H., Z. Zhang, S. Rose, and M. van der Laan. 2014. "A Novel Targeted Learning Methods for Quantitative Trait Loci Mapping." *Genetics* 198: 1369–76.

Zeng, Z. 1994. "Precision Mapping of Quantitative Trait Loci." *Genetics* 136: 1457–68.

SUPPORTING INFORMATION

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

Appendix SA1: Targeted Learning.
Appendix SA2: Parametric Regression Fit Diagnostics.