Published in final edited form as:

Biometrics. 2019 March ; 75(1): 289–296. doi:10.1111/biom.12927.

Double Robust Estimation for Multiple Unordered Treatments and Clustered Observations: Evaluating Drug-Eluting Coronary Artery Stents

Sherri Rose^{1,*} and Sharon-Lise Normand^{1,2}

¹Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts, U.S.A

²Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, U.S.A

Summary

Post market comparative effectiveness and safety analyses of therapeutic treatments typically involve large observational cohorts. We propose double robust machine learning estimation techniques for implantable medical device evaluations where there are more than two unordered treatments and patients are clustered in hospitals. This flexible approach also accommodates a large number of covariates drawn from clinical databases. The Massachusetts Data Analysis Center percutaneous coronary intervention cohort is used to assess the composite outcome of 10 drug-eluting stents among adults implanted with at least one drug-eluting stent in Massachusetts. We find remarkable discrimination between stents. A simulation study designed to mimic this coronary intervention cohort is also presented and produced similar results.

Keywords

Clustered data; Comparative effectiveness research; Machine learning; Multiple treatments; Nonparametric methods

1. Introduction

Assessing the comparative effectiveness or safety of therapeutic treatments once released into the marketplace often involves the analyses of large observational cohorts. Data that inform such analyses are increasingly available due to the growing integration of health care delivery systems, dissemination of electronic medical records, and development of clinical registries. Such data present extraordinary opportunities for research aimed at improving quality and value in health care by providing information on multiple competing technologies, most of which have never been comparatively assessed in the randomized or observational setting. While a national medical device evaluation system has been proposed in the United States (Krucoff et al., 2015), information to uniquely distinguish devices is not

* rose@hcp.med.harvard.edu.

Supplementary Materials

The Web Appendices and Web Table 1 referenced in Sections 2, 3, and 5 are available with this paper at the Biometrics website on Wiley Online Library. An R code demonstration of the TMLE described in this paper, using simulated data, is also available there.

currently routinely collected, nor is it available in medical claims as it is for prescription drugs. Implantable medical devices represent high-risk treatments often evaluated in the premarket setting on the basis of smaller trials, are likely to change quickly over time, and have led to serious side effects, including deaths (Resnic and Normand, 2012). Additional postmarket evaluation tools are necessary for estimating true benefits and risks amidst the large number of reports of various health and safety outcomes. A recent report attributed approximately 3,000 deaths per year to medical devices, but this is likely an underestimate given incomplete tracking (Daniel et al., 2015). Additionally, the Department of Health and Human Services estimates that defective devices cost billions in taxpayer dollars (Frakt, 2016).

When appropriate data are available to evaluate medical devices from large databases, major statistical challenges emerge. First, dozens or hundreds of possibly confounding covariates are often available. Thus, this involves simultaneously identifying potential confounders amongst these covariates while also avoiding model misspecification. Despite the uncertainty involved in correctly specifying a parametric model that reflects the underlying unknown data distribution with such a large number of covariates, parametric estimation techniques remain the standard approach employed for comparative effectiveness research in device safety settings (Stettler et al., 2008; Sedrakyan et al., 2014).

In the statistical literature, parametric approaches for handling the large number of confounders have included Bayesian model averaging and high-dimensional propensity scores (Raftery et al., 1997; Schneeweiss et al., 2009). Given the regulatory and policy implications of postmarket device evaluations, concerns regarding model misspecification and its impact on inference are of increased practical importance. Additional statistical methods applicable to the comparative effectiveness of medical devices, especially from the area of causal inference, include a number of other choices, such as parametric and nonparametric implementations of propensity scores, inverse probability weighting, g-computation, and double robust estimators (Robins, 1986; Hernan et al., 2000; van der Laan and Robins, 2003; van der Laan and Rose, 2011).

The second major challenge is that much of this literature has focused heavily on binary treatments in parametric models. Postmarket device evaluation necessitates assessing the comparative effectiveness of multiple devices in larger semiparametric or nonparametric model spaces. Parametric approaches for multiple treatments include propensity scores (Imbens, 2000; Li et al., 2001; Lopez and Gutman, 2017) and discrete choice models using multinomial logit, nested logit, or multinomial probit models (Tchernis et al., 2005). Marginal structural models have been introduced as a flexible approach to define complicated parameters, including treatments with multiple levels (Robins, 1998; Hernan et al., 2000; Rosenblum and van der Laan, 2010).

A third barrier to robust analyses of large databases for device evaluations in health care lies in the inherent multilevel nature of the data (e.g., patients clustered in hospitals). There is an expansive literature on parametric approaches for clustered data (Laird and Ware, 1982; Raudenbush and Bryk, 2002; Gelman and Hill, 2007), with parametric mixed effects models commonly used and recommended for the public reporting of health outcomes (Normand et

al., 1997; Localio et al., 2001; Krumholz et al., 2006). Semiparametric methods largely focus on generalized estimating equations (Liang and Zeger, 1986; Diggle et al., 2002), including advances for observational data settings (Goetgeluk and Vansteelandt, 2008; Zetterqvist et al., 2016). Nonparametric approaches incorporating machine learning for clustered data have been underdeveloped to date with few papers in this area (Schnitzer et al., 2014, 2018; Balzer et al., 2017).

This paper is motivated by these intersecting challenges involved in the comparative effectiveness of drug-eluting coronary stents. The United States has the second-highest number of overall stent insertions per capita, behind only Germany (Cutler and Ly, 2011). Coronary artery stents, small metallic tubes used to enlarge coronary arteries that have narrowed due to plaque, are implantable devices that have substantially impacted the treatment of coronary heart disease. These devices have proliferated rapidly over time and have multiple competing versions supported by a few manufacturers.

Drug-eluting stents are coated with a time-released drug to prohibit cell proliferation in the stented artery. For most major health outcomes, the wide class of drug-eluting stents have been demonstrated to be superior to bare-metal stents (stents not coated with a drug) in reducing the risk of restenosis. However, a broad study of the comparative effectiveness of various models of drug-eluting stents, including different generations of the same device, has not been conducted. The exception includes network meta-analyses where multiple randomized trials were analyzed using parametric random effects models (Siontis et al., 2016). Differences among the stents include the polymer coating that holds and releases the drug (e.g., phosphorylcholine or polyethylene-co-vinyl acetate), the drug (i.e., sirolimus, paclitaxel, zotarolimus, everolimus), the stent platform (i.e., metallic, cobalt-chromium, cobalt alloy, platinum-chromium), and the stent delivery system.

We propose robust machine learning estimation techniques for device evaluations where there are more than two treatments that do not have a natural ordering and there is also clustering by hospital. We isolate the effects of individual drug-eluting stents for a composite outcome. This work represents the first presentation of double robust machine learning-based estimators that handle the combined problems of multiple unordered treatments, model misspecification with high dimensional data, and clustered observations. In Section 2 we describe the clinical registry data. Section 3 presents the statistical framework and methods for our analysis, with results reported in Section 4. A simulation study designed to reflect the data is described in Section 5. We end with a discussion in Section 6.

2. Coronary Stented Cohort

The Massachusetts Data Analysis Center (Mass-DAC) percutaneous coronary intervention (PCI) cohort contains detailed information on adult patients undergoing PCI from all 25 hospitals licensed to perform the procedure in Massachusetts non-federal hospitals. This data is prospectively collected using a common instrument, adjudicated by review panels annually, and includes detailed data entered by trained hospital personnel as well as patient-level information. Multiple stents can be implanted within a PCI procedure and multiple PCIs can be performed within a single admission. Mass-DAC data are also linked to medical

claims data from the Massachusetts Center for Health Information and Analysis, and mortality data from the Massachusetts Registry of Vital Records. Further details on Mass-DAC data sources are described elsewhere (Massachusetts Data Analysis Center, 2016).

Our study sample initially included all PCI procedures from January 1, 2008 to September 30, 2012 in the Mass-DAC all-year PCI data set. We removed: *a)* all subsequent PCIs in the same admission, *b)* PCIs involving bare metal stents or no stents, *c)* PCIs involving 2 or more different drug-eluting stents, and *d)* non-Massachusetts residents (for completeness of follow-up). This led to a cohort of unique patients with at least one drug-eluting stent ($n = 21,054$) clustered within 25 hospitals. The decision to limit our study to procedures performed in 2008 and forward was designed to reflect the recent trend that most procedures do not use the over the wire technique but rather the rapid exchange stent insertion technique. Additionally, this restriction removed older stents approved 2003–2005 from the cohort. These devices are no longer frequently used, thus improving the relevance of our analysis to current practice. Three manufacturers (denoted Manufacturers A, B, and C) associated with a total of $K = 10$ different stents were represented in our sample. Thus, each company is associated with multiple devices; for example, Manufacturer A contributes four stents in our study, some with differing anonymized FDA-approval years, although all of Manufacturer A's devices contain the anonymized Drug I. One stent, anonymized as D1, contained identical stents marketed separately by two of the three manufacturers. Table 1 provides details on manufacturers and stents.

We will explore the comparative effectiveness of these stents for a binary composite outcome, death or major adverse cardiac event (MACE), measured up to one year following the stent implant, including both in-hospital and preliminary 1 year outcomes (PCI, acute myocardial infarction, and coronary bypass grafting). Available covariates included age, sex, race, procedure date, compassionate use (an indicator of heightened risk), insurance type, comorbidities, medical history, operator (physician who implanted the device), and symptom indicators, among others. As in Schnitzer et al. (2014), we include hospital site in the set of adjustment covariates because patients in the same hospital may be similar and hospital features may be associated with stents and our composite outcome. In simulations, Schnitzer et al. (2014) found that treating hospital as a baseline covariate successfully removed the bias associated with unmeasured confounding at the hospital level. Additional details on the covariates in our data are provided in Web Table 1.

3. Statistical Framework

Our study setting has hospital-level “intervention” $\mathbf{T} = (T_1, T_2, \dots, T_K)$, with $K = 10$, where each T_k random variable is a binary indicator. We have $S = 25$ hospital sites, indexed with $s = \{1, \dots, S\}$. Patients within hospitals are indexed by $i = \{1, \dots, N_s\}$, where N_s is a site-specific sample size for each hospital, and n represents the total number of subjects pooled across hospitals. The observational unit is the hospital site, with observed data O_s drawn from a probability distribution $P_s \in \mathcal{M}_s$, where $\mathcal{M}_s$ is a nonparametric model. The marginal mixture probability distribution P for a randomly selected individual in a randomly selected site is a function of P_s : $P = R(P_s)$. P can be factorized according to the time-ordering of the data: $P(O) = P(Y | \mathbf{T}, \mathbf{W}) \times R(\mathbf{T} | \mathbf{W}) \times P(\mathbf{W})$, analogous to Schnitzer et al. (2018), where Y is

an individual-level binary outcome and $\mathbf{W}$ is a vector of individual-level 198 baseline covariates including hospital site. We assume a nonparametric model $\mathcal{M}$ for the observed distribution $O = (\mathbf{W}, \mathbf{T}, Y)$, with $P \in \mathcal{M}$, and constraints placed on $\mathcal{M}$ also restrict $\mathcal{M}_s$. For alternative formulations, including data structures where cluster-level covariates are measured, we refer to Balzer et al. (2017).

3.1 Considerations for the Parameter of Interest

We are interested in understanding the comparative effectiveness of 10 drug-eluting stents. Y_k^t is defined as the individual-level counterfactual outcome had the intervention assigned been $T_k = t$ with $t \in \{0, 1\}$. We use the convention that uppercase letters are random variables and lowercase letters represent particular values. Our parameter of interest is defined for each stent k as: $\Psi_k(P) = P(Y_k^1 = 1) = E_{\mathbf{W}}\{P(Y = 1 \mid T_k = 1, \mathbf{W})\}$, where we make a set of causal assumptions, such that we can express this parameter as $P(Y_k^1 = 1)$ (Petersen et al., 2007; Rosenblum and van der Laan, 2010; Schnitzer et al., 2018). These causal assumptions include no unmeasured confounding (randomization) and the stable unit treatment value assumption: no interference between subjects, except by operator, and consistency (Rosenblum and van der Laan, 2010; Balzer et al., 2017). We make the assumption of weak covariate interference to account for subjects who have the same operator, parallel to Balzer et al. (2017). $P(Y_k^1 = 1)$ is our causal parameter of interest as we seek to understand the expected outcome for the individual-level counterfactual outcomes if all hospital sites were assigned stent k . We also make the statistical positivity assumption $P(T = t \mid \mathbf{W} = \mathbf{w}) > 0$ (with the levels $\mathbf{w}$ defined in Web Table 1), which we can test with our observed data. We could have defined our parameter of interest in a saturated marginal structural model for each level in $\mathbf{T}$. The point estimates for that parameter would be the same, although additional computational efficiencies are possible with that implementation.

3.2 Targeted Learning

Targeted learning is a general framework that allows researchers to estimate causal or statistical quantities using machine learning while providing statistical inference with targeted maximum likelihood estimation (TMLE) (van der Laan and Rubin, 2006; van der Laan and Rose, 2011, 2018). The implementation of our TMLE involved the following steps:

Step 1. Estimating $P(Y = 1 \mid \mathbf{T}, \mathbf{W})$ using super learner (van der Laan et al., 2007), an ensemble machine-learning approach for prediction. We included three algorithms: main terms regression, ridge penalized regression, and classification trees.

Then, for each k :

Step 2. Obtaining a fit for $P(T_k = 1 \mid \mathbf{W})$ with super learner using the same algorithms as in Step 1. We bounded $P_{\hat{n}}(T_k = 1 \mid \mathbf{W})$ from below to 0.025, where we use the subscript n to denote estimates.

Step 3. Regressing Y on $X_n(T_k, \mathbf{W}) = I(T_k = 1)/P_n(T_k = 1 | \mathbf{W})$ using the initial fit $P_n^0(Y = 1 | T, \mathbf{W})$ as an offset (e.g., holding it fixed as an intercept). The coefficient in front of $X_n(T, \mathbf{W})$ is denoted ε .

Step 4. Updating the initial fit using the estimated coefficient ε_n : $\text{logit } P_n^*(Y = 1 | T, \mathbf{W}) = \text{logit } P_n^0(Y = 1 | T, \mathbf{W}) + \varepsilon_n X_n(T_k, \mathbf{W})$. This provides an estimate $P_n^*(Y = 1 | T_k = 1, \mathbf{W})$ for each subject.

Step 5. Estimating $\Psi_k(P)$ with $1/n \sum_{i=1}^n P_n^*(Y = 1 | T = 1, \mathbf{W})$.

A crucial component for describing our TMLE above was the construction of the efficient influence curve for our parameter $\Psi_k(P)$:

$$D_k(P, O) = I(T_k = 1) \left\{ \frac{Y - P(Y = 1 | T_k, \mathbf{W})}{P(T_k | \mathbf{W})} \right\} + P(Y = 1 | T_k = 1, \mathbf{W}) - \Psi_k(P). \quad (1)$$

We can then define the one-dimensional working submodel through our initial estimate: $\text{logit } P(\varepsilon)(Y = 1 | T, \mathbf{W}) = \text{logit } P_n^0(Y = 1 | T, \mathbf{W}) + \varepsilon X(T_k, \mathbf{W})$. This submodel appears in Step 3, and $X(T_k, \mathbf{W})$ is chosen to include the relevant component of the efficient influence curve such that the derivative condition defined in Web Appendix A is satisfied. The TMLE thereby solves its score equation (i.e., the efficient influence curve), and thus inherits several desirable statistical properties, including double robustness. This double robustness indicates that the targeting submodel step removes bias for the parameter if the initial estimator is not consistent. If the initial estimator from Step 1 is consistent, the targeting step will maintain this consistency. Additionally, if both components are estimated consistently the estimator will be asymptotically efficient. TMLEs are also loss-based substitution estimators. We note that $P_n(T_k = 1 | \mathbf{W})$ is estimated separately for each stent in this implementation, and thus treatment probabilities are not constrained to sum to one for a given individual.

As shown in Schnitzer et al. (2018), one can proceed with a TMLE defined at the individual level as above when using individual-level data for clustered observations, and the TMLE will maintain the derivative condition and solve the efficient influence curve equation. However, variance estimation must explicitly account for clustering. We pursued a similar approach here. Inference when components are estimated with machine learning can be obtained using influence curves or bootstrapping, and we used influence curves. Because our estimator is asymptotically linear, the variance of $\Psi_k(P)$ is well approximated by the sample variance of $D_k(P, O)$ given in Equation 1, divided by sample size. With an estimated value of $D_k(P, O)$ for each individual, which we write as $IC_{k,n}(O_i)$, we can reformulate at the hospital level to calculate $IC_{k,s}(O_s) = \sum_{i \in N_s} IC_{k,n}(O_i)(S/n)$. Thus, we use $\sqrt{\text{var} \{IC_{k,s}(O_s)\}/S}$ to estimate standard errors, although other approaches have been described (Schnitzer et al., 2014).

All estimators were constructed in the R programming language. We compare our TMLE, which relies on `SuperLearner` (Polley and van der Laan, 2013) and `tmle` (Gruber and van

der Laan, 2012) as well as additional custom coded cluster-based influence curves, to three g-computation-based substitution estimators. These comparator estimators first estimate $P(Y = 1 | T, W)$, as in Step 1 of the TMLE procedure, and then jump to Step 5, the “plug-in” estimator of our parameter $\Psi_k(P)$. Therefore, one key difference is that the comparison estimators do not make use of $P(T_k = 1 | W)$, and, unlike the TMLE, are not double robust, relying on consistent estimation of $P(Y = 1 | T, W)$ for consistent estimation of $\Psi_k(P)$.

The three comparison estimators vary in how they perform estimation of $P(Y = 1 | T, W)$. The first uses maximum likelihood estimation (MLE) with generalized linear mixed effects regression, which is arguably the most commonly used procedure for clustered data in the policy literature. This mixed effects regression does not use hospital site indicators as fixed effects in the estimator for $P(Y = 1 | T, W)$, but rather a vector of hospital sites as random effects using the `glmer()` function in `lme4` (Bates et al., 2017). The remaining two comparators use the same vector W as TMLE, and were ridge regression (`glmnet`, Friedman et al., 2010) and random forests (`randomForest`, Liaw and Wiener, 2002). They were selected based on their frequency of use in the statistical learning literature and their integration into a substitution estimator allows us to effectively target the same parameters while respecting the bounds of our statistical model. We implemented nonparametric bootstrapping for hierarchical data, resampling hospital sites with replacement, to calculate standard errors and 95% confidence intervals for these comparator estimators (Ren et al., 2010).

4. Mass-DAC PCI Cohort Results

The total number of drug-eluting stents implanted per individual ranged from one to nine, with 67% of subjects receiving one stent. Median age was 63 and 92% of the sample was white. Summary information for key variables is displayed in Figure 1 where we also plot the distribution of manufacturers by hospital. Two hospitals implanted stents from a single manufacturer and seven hospitals implanted stents from two different manufacturers. Our cohort had an overall event rate 15.8%. There were 147 operators where the median number of patients per operator was 132, and 122 operators (83%) treated at least 30 patients.

We found remarkable discrimination with our TMLE estimates for $\Psi_k(P)$, as shown in Figure 2. Stents C3 and A3 were the poorest performing among the 10 stents, with expected outcomes of 24% and 21%. Stents C1, A1, and A2 were the best performing, with expected outcomes of 6%, 8%, and 13%. Other stents were similar to the overall event rate. Recall that D1 contains identical stents simply marketed separately by two of the manufacturers. As a “falsification test” sensitivity analysis, we estimated the component stents within D1, and found no difference between them, as we would expect.

The results from the three comparison estimators differed widely from the TMLE results with respect to the point estimates, level of uncertainty, and the relative ranking of the stents. These comparator estimators generally did not distinguish estimates away from the overall event rate, but when they did, confidence intervals were large (Figure 2). The MLE found stent C3 worse than average with an expected outcome of 20% and ridge estimated that A1 was better than average (13%). In both these cases, the direction of the effect aligned with

the TMLE estimates. We note that our confidence intervals were not simultaneous or controlled for multiple testing, and these adjustments can be included. One might also wish to consider explicit comparisons between stents with respective confidence intervals as we did for the component stents within D1 (Rosenblum and van der Laan, 2010).

In exploring contributing factors for TMLE identifying expected outcome probabilities away from the overall rate in sensitivity analyses, we implemented a TMLE that used only main terms regression for both $P(Y = 1 \mid \mathbf{T}, \mathbf{W})$ and $P(T_k = 1 \mid \mathbf{W})$. This TMLE had similar performance (results not shown), except for stents A1 and C1 where estimates were 14% and 13%. The machine-learning-based TMLEs generally did not use substantial information from the main terms regression (i.e., small or zero weight in the super learner), and the ridge regression dominated in both $P(Y = 1 \mid \mathbf{T}, \mathbf{W})$ and $P(T_k = 1 \mid \mathbf{W})$. For stent A1, the ridge regression was 4.8 times more efficient compared to the main terms regression with respect to cross-validated risk for the estimation of $P(T_k = 1 \mid \mathbf{W})$. Thus, it is reasonable to surmise that it is the combination of targeting *and* machine learning (with the associated double robustness they yield) that led to our results, with much of it driven by the targeting step. These results would be in line with other work demonstrating that machine learning improved TMLE finite sample performance considerably compared to TMLEs with misspecified regressions (e.g., Porter et al., 2011).

We also assessed positivity violations given the structure of our data (clustering in hospitals where not all stents were available), multiple treatments, and the high-dimensional nature of our covariate vector $\mathbf{W}$ (Petersen et al., 2012). While we did see practical violations of the positivity assumption, with values near zero, we addressed this by including bounding from below at 0.025 for $P(T_k = 1 \mid \mathbf{W})$ in our TMLE. Both theoretical and practical considerations support bounding for consistent estimation of $\Psi_k(P)$, with a bounded TMLE particularly robust for practical positivity violations (van der Laan and Robins, 2003; Porter et al., 2011; van der Laan and Rose, 2011). Sensitivity analyses with the bound set at 0.01 and 0.05 yielded similar estimates (results not shown).

5. Simulation Study

We designed a simulation study inspired by the Mass-DAC PCI data. Variables were generated to partially mimic the distributions and relationships between variables in the Mass-DAC cohort. A population of 1,000 hospitals, each with 1,000 patients, was generated for a total of 1,000,000 subjects. From this population, 500 samples of 10,000 subjects were drawn by sampling 10 hospitals with replacement in each iteration. We limited the baseline vector to 10 variables $\mathbf{W}$ = (female, white, insurance type, age, height, weight, diabetes, hypertension, prior PCI, prior cerebrovascular disease). Female (W_1), white (W_2), and insurance (W_3) were drawn from Bernoulli distributions with probabilities 0.28, 0.92, and 0.64, respectively. Age (W_4), height (W_5), and weight (W_6) were generated from truncated Normal distributions as described in Web Appendix B. These distributions approximate those from our data analysis as parameter values were assigned based on summary measures in the Mass-DAC PCI cohort.

The four binary health variables diabetes (W_7), hypertension (W_8), prior PCI (W_9), and prior cerebrovascular disease (W_{10}) were generated dependent on the demographic baseline variables (except insurance status and procedure year) and are defined in Web Appendix B. The coefficients for the demographic variables were assigned by estimating a regression in the Mass-DAC PCI cohort, with modifications then made to those coefficients in the simulation as needed to obtain the approximate proportions found in the data. We then generated 10 separate stents for our simulation data. These stent variables T were generated similarly; all from Bernoulli distributions with probabilities b_T given in Web Appendix B. Here, coefficients were also adjusted to accommodate the mutually exclusive nature of stent assignment and to balance participants across stents to explore a variety of sample sizes, including those much smaller than in our PCI cohort. We eschew the complication of the dual manufacturers for the stent based on stent D1, and here refer to it as stent A4 belonging to Manufacturer A.

The outcome MACE after 1 year was $Y \sim \text{Bern}(b_Y)$, with b_Y given in Web Appendix B, where the coefficients were also based on a simple main terms regression of the outcome on covariates and treatment in the Mass-DAC PCI cohort data. The population proportion of MACE in the simulation was 18.8%. We analyzed this simulated data for our parameter $\Psi_k(P)$. The estimators followed the same implementation as that used in our analysis of the Mass-DAC PCI cohort data, and included g-computation-based substitution estimators with MLE, ridge regression, and random forests in comparison to the TMLE.

This simulation study found several results consistent with our data analysis, notably, that ridge and random forests largely did not estimate values away from the overall rate. (See Figure 3.) These two estimators had a majority of true values outside their respective confidence intervals (7 stents). The TMLE and MLE correctly estimated stents A1 and C1 as below the overall rate, and A2 and C3 as above the rate. All estimators correctly identified stent A1 as a poor performer compared to the overall rate, although the ridge and random forests estimators did so with nontrivial bias and their confidence intervals did not contain the truth. Stents B2, C4, and C2 all had average sample sizes under 100, and while the TMLE and MLE point estimates were close to the truth, the confidence intervals were wide.

As expected, the MLE performs better in our simulation setting than in our data analysis. This is due to the simplified data structure and the MLE's ability to capture the true relationship between the outcome and confounders because we estimated it with a correctly specified regression. The MLE was slightly inefficient compared to the TMLE for stents with small sample sizes. This appears to be due to approximation errors impacting efficiency, whereas the TMLE leverages its asymptotic linearity and double robustness to improve precision as $P(T_k = 1 | \mathbf{W})$ is reasonably well approximated. That said, these differences were trivial and we do not assert broad conclusions for this setting. We refer to additional literature for further simulation studies with TMLE estimator comparisons under a variety of settings (Porter et al., 2011; van der Laan and Rose, 2011, 2018; Schnitzer et al., 2014).

6. Discussion

In this paper, we discovered substantial differences for a 1-year MACE outcome in individual drug-eluting stents in a PCI cohort using TMLE. Other techniques did not capture these differences. The best and worse performing stents (C1 and C3) were from the same manufacturer and deliver the same drug. These variations are plausible as the better performing stent (C1) is a much newer device (approved 4 years after C3), and the devices have different polymers and delivery systems. There was a general trend in our analysis that newer devices performed better. While further work is needed to confirm our findings, these results could have important policy implications for patients, hospitals, device manufacturers, and regulators. Will patients wish to have a procedure performed at a hospital that only has devices from a poorer-performing manufacturer? Will hospitals reconsider their complex contracting with manufacturers to avoid poorer-performing devices? Should manufacturers consider pulling certain stents from the market? How should regulators respond to postmarket information that was not available at the time of device approval?

Assessing the performance of multiple devices has been a difficult challenge in practice, as randomized studies do not generally compare multiple devices, focus on efficacy, and are often under-powered for safety assessments. Thus, this paper adds to the literature on the performance of drug-eluting stents for a 1-year MACE outcome, and presents an application of double robust machine learning that can be computed in a computationally expedient procedure accounting for clustering. There are several limitations of our current study. We adjusted for 198 potential confounders related to the patient's health, procedure, and demographics, as well as operator and hospital. However, while our Mass-DAC PCI cohort has an incredibly rich collection of variables from clinical and claims data, we cannot say with certainty whether our untestable causal assumptions hold. Many critical *known* confounding variables, such as the acuity of the patient at the time of stent implant, are included. We did not have a comprehensive measure of operator experience, and opted to include the operator to capture this potential confounder and possible interference by operator. Previous work has demonstrated that this type of proxy is suitable for unmeasured confounding related to operator (Schnitzer et al., 2014) and weak interference (Balzer et al., 2017). We also do not generalize and study common support intervals for the setting of multiple treatments, which is an exciting area of future work. Even if our causal assumptions do not hold, the parameters we target represent interesting statistical quantities: the expected outcome under treatment adjusted for measured confounding.

There is currently an effort in the United States to routinely collect the information required to distinguish between different devices using unique identifiers, and advances to study effectiveness and safety of devices in population-wide cohorts using electronic information are anticipated. Other countries have such systems and have been faster to identify problems in devices that were eventually removed from the market in the United States (Frakt, 2016). There has been a call for statistical techniques to identify important causal relationships that can “increase confidence” in results from surveillance systems (Frakt, 2016). Thus, statistical learning methods, such as those applied here, that can handle both the challenge of many possible confounders *and* many treatments *and* clustered data in flexible nonparametric models will be increasingly relevant for postmarket safety analysis. In

particular, highlighting methods that can estimate effects of treatments that do not have a natural ordering will be distinctly important for assessing many types of medical devices.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are indebted to the Massachusetts Department of Public Health (Mass-DAC Database) and the Massachusetts Center for Health Information and Analysis (CHIA) (Case Mix Databases) for the use of their datasets. This work was supported by NIH grant number R01-GM111339 from the National Institute of General Medical Sciences in the United States.

References

- Balzer, LB; Zheng, W; van der Laan, MJ; Petersen, ML; , et al. [accessed September 1 2017] A New Approach to Hierarchical Data Analysis: Targeted Maximum Likelihood Estimation of Cluster-Based Effects Under Interference. 2017. <https://arxiv.org/abs/1706.02675>
- Bates D, Maechler M, Bolker B, Walker S, Christensen RHB, Singmann H, et al. 2017lme4: Linear Mixed-Effects Models.
- Cutler D, Ly D. 2011; The (paper) work of medicine: understanding international medical costs. The Journal of Economic Perspectives. 25:3. [PubMed: 21595323]
- Daniel, G; Colvin, H; Khaterzai, S; McClellan, M; Aurora, P. [accessed September 1, 2017] Strengthening Patient Care: Building an Effective National Medical Device Surveillance System. 2015. <http://bit.ly/2xEL2Wb>
- Diggle, P, Heagerty, P, Liang, K-Y, Zeger, S. Analysis of Longitudinal Data. 2002. Oxford Statistical Science Series
- Frakt, A. [accessed September 1, 2017] Why medical devices aren't safer. The New York Times. 2016. <http://www.nytimes.com/2016/04/19/upshot/why-medical-devices-arent-safer.html>
- Friedman, J, Hastie, T, Tibshirani, R. glmnet: lasso and elastic-net regularized generalized linear models. 2010. R package version 2.0-2
- Gelman, A, Hill, J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press; New York, NY, USA: 2007.
- Goetgeluk S, Vansteelandt S. 2008; Conditional generalized estimating equations for the analysis of clustered and longitudinal data. Biometrics. 64:772–780. [PubMed: 18047524]
- Gruber S, van der Laan M. 2012tmle: An r package for targeted maximum likelihood estimation. J Stat Softw. :51.
- Hernan M, Brumback B, Robins J. 2000; Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology. 11:561–570. [PubMed: 10955409]
- Imbens G. 2000; The role of the propensity score in estimating dose-response functions. Biometrika. 87:706–710.
- Krucoff M, Sedrakyan A, Normand SL. 2015; Bridging unmet medical device ecosystem needs with strategically coordinated registries networks. Journal of the American Medical Association. 314:1691–1692. [PubMed: 26302152]
- Krumholz HM, Brindis RG, Brush JE, Cohen DJ, Epstein AJ, Furie K, et al. 2006; Standards for statistical models used for public reporting of health outcomes. Circulation. 113:456–462. [PubMed: 16365198]
- Laird NM, Ware JH. 1982; Random-effects models for longitudinal data. Biometrics. 38:963–974. [PubMed: 7168798]
- Li Y, Propert K, Rosenbaum P. 2001; Balanced risk set matching. Journal of the American Statistical Association. 96:870–882.

- Liang KY, Zeger SL. 1986; Longitudinal data analysis using generalized linear models. *Biometrika*. 73:13–22.
- Liaw A, Wiener M. 2002; Classification and regression by randomforest. *R News*. 2:18–22.
- Localio AR, Berlin JA, Ten Have TR, Kimmel SE. 2001; Adjustments for center in multicenter studies: an overview. *Annals of Internal Medicine*. 135:112–123. [PubMed: 11453711]
- Lopez M, Gutman R. 2017; Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*. 32:432–454.
- Massachusetts Data Analysis Center. [Accessed September 1, 2017] Cardiac study annual reports. 2016. <http://www.massdac.org/index.php/reports/cardiac-study-annual>
- Normand SLT, Glickman ME, Gatsonis CA. 1997; Statistical methods for profiling providers of medical care: issues and applications. *Journal of the American Statistical Association*. 92:803–814.
- Petersen M, Deeks S, Martin J, van der Laan M. 2007; History-adjusted marginal structural models to estimate time-varying effect modification. *American Journal of Epidemiology*. 166:985–993. [PubMed: 17875580]
- Petersen M, Porter K, Gruber S, Wang Y, van der Laan M. 2012; Diagnosing and responding to violations in the positivity assumption. *Statistical Methods for Medical Research*. 21:31–54.
- Polley, E, van der Laan, M. SuperLearner. 2013. R package version 2.0-10
- Porter KE, Gruber S, van der Laan MJ, Sekhon JS. 2011; The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics*. 7:1.
- Raftery A, Madigan D, Hoeting J. 1997; Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*. 92:179–191.
- Raudenbush, SW, Bryk, AS. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage; 2002.
- Ren S, Lai H, Tong W, Aminzadeh M, Hou X, Lai S. 2010; Nonparametric bootstrapping for hierarchical data. *Journal of Applied Statistics*. 37:1487–1498.
- Resnic F, Normand S. 2012; Postmarketing surveillance of medical devices filling in the gaps. *New England Journal of Medicine*. 366:875–877. [PubMed: 22332950]
- 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*. 7:1393–1512.
- Robins J. 1998 Marginal structural models. 1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science. :1–10.
- Rosenblum M, van der Laan M. 2010; Targeted maximum likelihood estimation of the parameter of a marginal structural model. *International Journal of Biostatistics*. 6
- Schneeweiss S, Rassen J, Glynn R, Avorn J, Mogun H, Brookhart M. 2009; High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 20:512. [PubMed: 19487948]
- Schnitzer, M, van der Laan, M, Moodie, E, Platt, R. LTMLE with clustering. In: van der Laan, MJ, Rose, S, editors. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer; Berlin Heidelberg New York: 2018.
- Schnitzer ME, van der Laan MJ, Moodie EE, Platt RW. 2014; Effect of breastfeeding on gastrointestinal infection in infants: a targeted maximum likelihood approach for clustered longitudinal data. *The Annals of Applied Statistics*. 8:703. [PubMed: 25505499]
- Sedrakyan A, Graves S, Bordini B, Pons M, Havelin L, Mehle S, et al. 2014; Comparative effectiveness of ceramic-on-ceramic implants in stemmed hip replacement. *The Journal of Bone & Joint Surgery*. 96:34–41. [PubMed: 25520417]
- Siontis G, Piccolo R, Praz F, Valgimigli M, Räber L, Mavridis D, Jüni P, Windecker S. 2016; Percutaneous coronary interventions for the treatment of stenoses in small coronary arteries: a network meta-analysis. *Journal of the American College of Cardiology: Cardiovascular Interventions*. 9:1324–1334. [PubMed: 27318845]
- Stettler C, Allemann S, Wandel S, Kastrati A, Morice M, Schomig A, et al. 2008; Drug eluting and bare metal stents in people with and without diabetes: collaborative network meta-analysis. *The BMJ*. 337:a1331. [PubMed: 18757996]

- Tchernis R, Horvitz-Lennon M, Normand S. 2005; On the use of discrete choice models for causal inference. *Statistics in Medicine*. 24:2197–2212. [PubMed: 15887310]
- van der Laan M, Polley E, Hubbard A. 2007; Super learner. *Statistical Applications in Genetics and Molecular Biology*. 6
- van der Laan, M, Robins, J. *Unified Methods for Censored Longitudinal Data and Causality*. Springer; Berlin Heidelberg New York: 2003.
- van der Laan, M, Rose, S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer; Berlin Heidelberg New York: 2011.
- van der Laan, M, Rose, S. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer; Berlin Heidelberg New York: 2018.
- van der Laan M, Rubin DB. 2006; Targeted maximum likelihood learning. *International Journal of Biostatistics*. 2
- Zetterqvist J, Vansteelandt S, Pawitan Y, Sjölander A. 2016; Doubly robust methods for handling confounding by cluster. *Biostatistics*. 17:264–276. [PubMed: 26508769]

Figure 1.
Summary information for Mass-DAC PCI cohort

Figure 2.

Comparative effectiveness of drug-eluting stents in Mass-DAC PCI cohort. TMLE, targeted maximum likelihood estimation; MLE, maximum likelihood estimation with generalized linear mixed effects models; RF, random forests; Ridge, ridge regression. Confidence intervals are not simultaneous or corrected for multiple testing.

Figure 3.

Comparative effectiveness of drug-eluting stents in simulated data. TMLE, targeted maximum likelihood estimation; MLE, maximum likelihood estimation with generalized linear mixed effects models; RF, random forests; Ridge, ridge regression. Mean estimates over 500 simulations. Confidence intervals are not simultaneous or corrected for multiple testing.

Table 1**Drug-Eluting Stents by Manufacturer**

Manufacturer (N)	A (2,527)	B (458)	D (14,832)	C (3,237)
Approval Years	① ③	② ④	①	① ④
Drug	I	II	I	III
Stent (n)	A1 (891)	B1 (272)	D1 (14,832) *	C1 (793)
	A2 (1,316)	B2 (186)		C2 (1,273)
	A3 (320)			C3 (326)
				C4 (845)

* When D1 is separated into the manufacturers producing the stents, the total sample sizes for the other two manufacturers are 10,784 and 7,033, respectively.