Examination of the Treatment Selection Process in a Multicenter Observational Study

Kevin J. Anstrom, PhD; J. Matthew Brennan, MD, MPH; Eric L. Eisenstein, DBA; Jerome J. Federspiel, PhD; David Dai, PhD; Eric D. Peterson, MD, MPH; Pamela S. Douglas, MD

Background—Many multicenter clinical trials use permuted-block randomization to create balanced treatment allocations within clinical centers. Unlike randomized trials, observational studies do not control treatment allocation, and statistical models are used to adjust for measured confounders. For many observational data analyses, the variability in the treatment selection process within clinical centers is ignored. Furthermore, there is no consensus on the best approach for dealing with variability in the treatment selection process across clinical centers.

Methods and Results—Individuals aged ≥65 years receiving either drug-eluting stents or bare metal stents were included. A cohort of 262,700 patients from 650 CathPCI Registry sites was followed up for a median of 15 months. Propensity score models were estimated to describe the process used to select drug-eluting stents across the study population. Substantial variability in the use of drug-eluting stents at the clinical center level was observed—even after accounting for differences in patient and clinical center characteristics. By refitting and matching propensity scores within clinical centers, a balanced cohort on treatment allocation and prognostic factors was obtained. This approach generated an estimated hazard ratio that was qualitatively similar to standard regression models and other propensity score approaches.

Conclusions—Substantial variability in treatment selection existed between clinical centers. Matching recalibrated propensity scores within clinical centers has the potential to reduce a source of bias in multicenter observational studies. This methodology cannot eliminate all potential for biases; however, it removes the potential bias from site-level factors. (Circ Cardiovasc Qual Outcomes. 2014;7:00-00.)

Key Words: comparative effectiveness research • propensity score • selection for treatment

Comparative effectiveness research using existing large clinical databases has been touted as a promising tool for expanding the medical evidence base. One goal of comparative effectiveness research studies is to estimate treatment effects and to make comparisons among viable treatment alternatives for a particular population. Although studies based on observational data can be efficient and allow comparisons in broader patient populations and real-world settings, observational treatment comparisons may be subject to bias resulting from the lack of randomization (ie, confounding by indication). Because of this lack of randomization, correctly specified statistical models are required to estimate treatment effects from observational data accurately. These models attempt to adjust adequately for measured confounders. Propensity score methods are a commonly used approach involving construction of statistical models of treatment assignment. Although most propensity scores are designed to achieve balance in patient characteristics, rarely do propensity score models incorporate potential interfacility variation in how treatment decisions are made (ie, across facilities; providers may weigh clinical and demographic factors differently when determining treatment and may have different intrinsic propensity to choose particular treatments). As a result, such models can produce cohorts in which treated and control patients are not balanced with respect to treatment facility and in which there is no balance in potential confounding factors within individual facilities. In contrast, many multicenter clinical trials use permuted-block randomization to create balanced treatment allocations within clinical centers.

Our research team collaborated with the American College of Cardiology to use their National Cardiovascular Data Registry CathPCI Registry to estimate the comparative effectiveness of drug-eluting stents (DES) versus bare metal stents (BMS) in Medicare patients. Overall, this work concluded that DES was associated with a significant survival advantage, relative to BMS with an estimated hazard ratio (HR) of 0.75 (95% confidence interval, 0.72–0.79; P<0.0001). However, this publication did not include a formal sensitivity analysis.
Methods

Study Population

The CathPCI Registry is a component of the national database that collects information on patients undergoing diagnostic catheterization and percutaneous coronary intervention procedures—the study population is described in detail by Douglas et al. Within each selected site, we included all patients undergoing an intracoronary stent procedure who were aged ≥65 years on their procedure date and were admitted and discharged between January 1, 2004, and December 31, 2006. The Duke University Medical Center Institutional Review Board granted a waiver of informed consent and authorization for this study.

Follow-Up Information

Medicare’s 100% inpatient claims files were used to provide longitudinal patient follow-up. The following percutaneous coronary intervention codes were used to identify potential study patients: International Classification of Diseases, Ninth Revision, Clinical Modification procedure codes 06.66, 36.0x, 37.22, 37.23, and 88.5x, except 88.59. The linking process used to create the CathPCI Registry-Medicare database with ≤3 years of longitudinal follow-up has been described by Hammill et al.

Propensity Score Modeling and Propensity Score Matching

A common statistical tool used to adjust for treatment group differences within an observational study is the propensity score. Originally developed by Rosenbaum and Rubin, a propensity score is the probability of receiving a specific treatment conditional on a set of covariates. Several approaches for developing propensity score models have been suggested, and the optimal approach may depend on the research setting. For randomized clinical trials, the propensity score is usually known by design. For observational studies, the propensity score is unknown and must be estimated from available data.

We estimated 4 separate propensity score models with increasing complexity to account for differences within the clinical centers. The first propensity score (model 1) used the original model from Douglas et al that included 102 covariates. A second propensity score (model 2) was fitted using indicator variables for the 650 clinical centers and the logit of the original propensity score model. If all clinical centers used the same process for selecting stent type, we would expect that a high percentage of the 95% confidence intervals for the 650 site-specific log-odds parameter estimates would cover the null value of zero. A third propensity score (model 3) was fitted using site-specific terms and interactions between the site indicators and logit of the original propensity score. Under the hypothesis that all sites use the same selection process, the 650 site-specific intercepts and the 650 site-specific interaction terms would be estimated as zero.

A fourth propensity score (model 4) was constructed by refitting the logistic regression model predicting use of DES with the logit of the full propensity score model and the 102 individual variables from the original propensity score model. Within each clinical center, we fit a logistic regression model predicting use of DES with the logit of the full propensity score model and the 102 individual variables from the original propensity score model. Within each clinical center, the logistic regression model was refitted using a forward variable selection process (requiring 0.01 for entry) with only the logit of the original propensity score model forced into the model. Variable selection was used because many of the sites did not have large numbers of individuals receiving BMS (or DES).

For each of the 4 propensity scores, we constructed 2 types of propensity score matches—one matched DES and BMS patients across the entire cohort and the other matched DES and BMS patients within each of the 650 sites. We then applied a greedy-matching algorithm. As a result of this model building and matching, we created 8 propensity score matched cohorts.

Estimation of the Mortality HR

An objective of the analysis of Douglas et al was to estimate the HR for mortality comparing patients treated with BMS and DES. We compared those results with the HR for mortality based on the 8 propensity score matched cohorts. Estimates of the HR from the propensity score matched cohorts were obtained from a Cox model with an indicator variable for DES. All statistical analyses were conducted using SAS version 9.2 or higher (SAS Institute, Cary, NC).

Results

Population Characteristics

A total of 262,700 patients at 650 hospitals were included in the study population. The average follow-up time for this cohort was 15 months after percutaneous coronary intervention implantation. The median number of patients per site was 275.5, and the maximum number of patients at a site was 3924. Overall 82.9% of patients received DES and the median number of DES patients per site was 220. Among the 50 largest sites (those sites with >1152 patients), the proportion of patients receiving DES ranged from 38.1% to 96.9% (Figure 1A).

Propensity Score Models

On the basis of model 1, we found that the site-level mean estimated propensity score was relatively constant across the 50 largest sites (ranging from 78.5% to 88.5%). These results suggest substantial site-by-site variability in the use of DES, which was not accounted for using the original propensity score model from Douglas et al (Figure 1A). Model 2 addressed the lack of fit for the observed versus expected DES usage at the site level. As seen in Figure 1B, the proportion of patients receiving DES exactly matches the mean propensity score estimate at the site level. As expected (Figure 1C), there was a low correlation between the site-level mean propensity score estimates of model 1 and that of model 2, which includes additional terms for the site-specific intercepts (Pearson correlation, 0.24; P=0.10). The addition of the intercept terms to the propensity score adds flexibility, but the inclusion of the logit of the original propensity score model forces the within-site (rank) correlation between model 1 and model 2 to be 1.00 (Figure 2A).

Model 3 allows for site-specific intercept terms and interaction terms between the site-indicator variable and the logit of the original propensity score model. If every clinical site used the same decision rules for selection of stent type, we would expect the estimated interaction parameter to be close to the value 0. For this data set, we did not observe that result. Among the 50 largest clinical sites, only 9 of them had 95% confidence intervals for the interaction term that covered 0. This suggests that clinical centers place different weights on the observed covariates in the treatment selection process. Because model 3 does not include the individual covariates, we observed that the within-site (rank) correlation between model 1 and model 3 remains 1.00 (Figure 2B). For model...
Anstrom et al. Examination of Treatment Selection Process

4, we allowed for flexibility by fitting the logistic regression models separately for each of the clinical centers. All of the site-specific logistic regression models include intercept terms, the estimated slope from the logit of the original propensity score model, and other potential terms. In the model fitting process, there were 102 candidate variables for each of the site-specific models, and variables were included using the 0.01 level based on the forward selection process. Figure 3 shows the distribution of the number of added terms to the site-specific logistic regression models for the 50 largest sites. All 50 of the site-specific models added ≥1 additional covariate and the majority added between 2 and 6 additional covariates, suggesting that sites do have different approaches for selecting stent type. If every clinical site used the same treatment selection process, we would expect approximately one term to be added to each site-specific model.

Propensity Score Matching and Estimated Mortality HRs

The Table shows the results of the propensity score matching. Two types of matching were conducted—one forced
the matched pair to be from the same clinical site (restricted matching) and the other used unrestricted matching. When unrestricted matching was used for model 1, a total of 646 sites were included with a total of 89,694 patients. When matching was restricted to the same clinical center, the number of sites decreased to 625 and the number of patients decreased to 80,244. Similar patterns were observed for model 2 and model 3. Model #4 that included the site-specific models resulted in the smallest number of sites and patients for both the restricted and unrestricted cohorts. The results from the Cox regression models on all 8 matched cohorts were similar with HR estimates varying from 0.78 to 0.84 in every case favoring DES when compared with BMS. Notably, the 7 approaches that account for clinical center in some fashion (all but model 1 without restriction) have similar HR estimates varying from 0.78 to 0.80. Figure 4A to 4D illustrates the balance of treatment selection by site for models 1 and 4 with and without restriction. All models except 1 without restriction show excellent balance of treatment selection by site.

Discussion

Our analyses have yielded the following findings (1) there is considerable variability in the treatment selection process at the clinical center level and (2) application of different methods of building propensity scores resulted in similarly adjusted mortality HRs, despite only moderate correlation between propensity score estimates. Previous simulation results by Brookhart et al suggest that including terms in the propensity score model not related to outcome, but only related to treatment selection, can increase variability and bias. However, it seems difficult and perhaps impossible to know a priori whether an indicator variable for each clinical center should be included in the propensity score models, especially in observational studies with many sites. In addition, whether including the clinical centers in the propensity score model will reduce variability and bias may depend on the outcome variable. Our proposal to refit the propensity score model within clinical centers reduces the possible bias caused by variability in the treatment selection at the site level.

Previous studies have suggested that the clinical center is an important factor predicting outcomes for patients. There are several reasons that a clinical center could have an effect on the long-term outcome (independent of treatment choice), including hospital-based programs designed to promote adherence to evidence-based medicine and to reduce missed appointments. It is a well-known result that for a confounder to create bias on the estimated treatment effect, the factor must be differentially observed in the treatment groups and it

Table. Summary of PS Matched Cohorts and Estimated Mortality HRs

<table>
<thead>
<tr>
<th>Propensity Score Model</th>
<th>Unrestricted Matching</th>
<th>Restricted Matching*</th>
<th>Reduction in No. of Sites and Patients†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1: Original PS model</td>
<td>H=646</td>
<td>H=625</td>
<td>Sites=3.3%</td>
</tr>
<tr>
<td>n=89,694</td>
<td>n=80,244</td>
<td>Patients=10.5%</td>
<td></td>
</tr>
<tr>
<td>HR=0.84</td>
<td>HR=0.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% CI: 0.80–0.87</td>
<td>95% CI: 0.75–0.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 2: added site-specific intercepts to model 1</td>
<td>H=637</td>
<td>H=629</td>
<td>Sites=1.3%</td>
</tr>
<tr>
<td>n=81,966</td>
<td>n=79,676</td>
<td>Patients=2.8%</td>
<td></td>
</tr>
<tr>
<td>HR=0.80</td>
<td>HR=0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% CI: 0.76–0.83</td>
<td>95% CI: 0.76–0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 3: added site-specific intercepts and slopes to model 1</td>
<td>H=625</td>
<td>H=621</td>
<td>Sites=0.6%</td>
</tr>
<tr>
<td>n=80,978</td>
<td>n=79,396</td>
<td>Patients=2.8%</td>
<td></td>
</tr>
<tr>
<td>HR=0.78</td>
<td>HR=0.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% CI: 0.75–0.82</td>
<td>95% CI: 0.76–0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 4: refit of model 1 within clinical site</td>
<td>H=599</td>
<td>H=585</td>
<td>Sites=2.3%</td>
</tr>
<tr>
<td>n=73,740</td>
<td>n=72,024</td>
<td>Patients=2.3%</td>
<td></td>
</tr>
<tr>
<td>HR=0.79</td>
<td>HR=0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% CI: 0.75–0.83</td>
<td>95% CI: 0.76–0.84</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BMS indicates bare metal stents; CI, confidence interval for the hazard ratio comparing DES vs. BMS; DES, drug-eluting stents; H, number of clinical sites in the matched cohort; HR, estimated mortality hazard ratio comparing DES vs BMS; n=number of patients in the matched cohort; and PS, propensity score.

*Matched BMS and DES patients are from the same clinical site.

†Reduction in the number of clinical sites (patients) comparing the restricted matched cohort with the unrestricted matched cohort.

Figure 3. The number of variables added to the refitted propensity score (PS) models among the 50 largest sites.
must be associated with either increased or decreased risk of the outcome. By forcing balance of treatment groups within each clinical center, the possibility of a clinical center creating bias has been reduced. Arpino and Mealli have suggested that imposing matching within the same site (or cluster) automatically achieves balance for observed and unobserved site-level characteristics.

We think that there are several potential advantages of refitting propensity scores within clinical centers and then matching within site. First, this approach is guaranteed to produce balance of treatment within clinical centers. In randomized clinical trials, the potential bias resulting from treatment imbalance is a reason for using permuted-block designs. Second, in clinical trials, there is a requirement that investigators have equipoise to be willing to randomize a patient to one of several possible treatment strategies. If an investigator is unwilling to randomize the patient, then clearly, the patient is not included in the study population. Yet in observational studies, this protection does not exist. In general, it is worthwhile to consider how to include clinical centers that only treat patients with one treatment strategy. In such cases, it is not possible to estimate treatment effects within the clinical center. Our proposal to refit propensity scores within each center partially addresses this problem by automatically removing sites that do not have overlap in propensity scores across treatments. Although we have focused on site-level factors, similar approaches could be applied to any confounder.

**Limitations**

This study has several important limitations. First, extensive simulation studies would be required to understand the properties of the propensity score model building and matching strategies in multicenter and multilevel observational studies. Second, the data set for this application did not include information at the physician level. For some analyses, it may be more appropriate to balance patient characteristics and treatment allocations at the physician level, rather than the site level. Third, matching strategies other than the 1:1 greedy match may result in more efficient use of observational databases. For example, a 2 DES:1 BMS match might offer the similar bias reduction with more precise estimates of key parameters. Fourth, the proposed approach may reduce the cohort’s sample size relative to the usual propensity score matching approach. In the example presented when the matching was restricted, the number of clinical sites was reduced by ≈0.6% to 3.3% and the number of patients was reduced by 2.0% to 10.5%. Fifth, other possible unmeasured confounders may exist and our proposed approach does not correct for this concern.

**Conclusions**

The presence of omitted confounders can materially affect results; moreover, failing to account for factors, such as site-by-site variation and site-level unmeasured factors in the stent selection process, could lead to biased estimates of treatment effect. To eliminate this potential source of bias in large multicenter observational studies, a reasonable analytic strategy is to refit the propensity score model within the clinical center and then applying a propensity score match algorithm within the clinical center.

**Sources of Funding**

This project was sponsored by the Agency for Healthcare Research and Quality, United States Department of Health and Human Services, Rockville, MD, as part of the Cardiovascular Consortium and funded under project ID: 24-DKE-2 and Work Assignment Number:
Disclosures

Dr. Anstrom reports receiving funding for research grants from National Institutes of Health, Agency of Healthcare Research and Quality (AHRQ), AstraZeneca, Eli Lilly & Company, and Medtronic (all significant); and for consulting from Abbott Vascular (modest), AstraZeneca (modest), Bristol-Myers Squibb (modest), Gilead (modest), Pfizer (modest), GlaxoSmithKline (modest), Promedior (modest), and Ikaria (modest). Dr Anstrom has served on a Data Monitoring Board for NIH (modest), University of North Carolina (modest), University of Miami (modest), Forest (modest), GlaxoSmithKline (modest), Pfizer (modest), and Vertex (modest). Dr Anstrom reports an equity interest in Biscarodia. Dr Eisenstein reports receiving funding for research grants from Medtronic Vascular and Aptus Endosystems (both significant), Bristol-Myers Squibb, Eli Lilly & Company, Johnson & Johnson, Merck & Co., Sanofi-Aventis, American Heart Association, American College of Cardiology, and Society of Thoracic Surgeons (all significant); and consulting for Boehringer Ingelheim (modest). Dr Douglas reports receiving funding for research grants from Abiomed, Atritech, Edwards Lifesciences, Viacor, Inc., NIH and AHRQ (all significant); and consulting for Translation Research in Oncology and CardioDX (modest). The other authors report no conflicts.

References

Examination of the Treatment Selection Process in a Multicenter Observational Study
Kevin J. Anstrom, J. Matthew Brennan, Eric L. Eisenstein, Jerome J. Federspiel, David Dai,
Eric D. Peterson and Pamela S. Douglas

Circ Cardiovasc Qual Outcomes. published online August 12, 2014;
Circulation: Cardiovascular Quality and Outcomes is published by the American Heart Association, 7272
Greenville Avenue, Dallas, TX 75231
Copyright © 2014 American Heart Association, Inc. All rights reserved.
Print ISSN: 1941-7705. Online ISSN: 1941-7713

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://circoutcomes.ahajournals.org/content/early/2014/08/12/CIRCOUTCOMES.113.000482

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation: Cardiovascular Quality and Outcomes can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation: Cardiovascular Quality and Outcomes is online at:
http://circoutcomes.ahajournals.org//subscriptions/