CrossMark

ORIGINAL RESEARCH ARTICLE

Incorporating Linked Healthcare Claims to Improve Confounding Control in a Study of In-Hospital Medication Use

Jessica M. Franklin[1] · Wesley Eddings[1] · Sebastian Schneeweiss[1] · Jeremy A. Rassen[2]

Abstract

Introduction The Premier Perspective hospital billing database provides a promising data source for studies of inpatient medication use. However, in-hospital recording of confounders is limited, and incorporating linked healthcare claims data available for a subset of the cohort may improve confounding control. We investigated methods capable of adjusting for confounders measured in a subset, including complete case analysis, multiple imputation of missing data, and propensity score (PS) calibration.

Methods Methods were implemented in an example study of adults in Premier undergoing percutaneous coronary intervention (PCI) in 2004–2008 and exposed to either bivalirudin or heparin. In a subset of patients enrolled in UnitedHealth for at least 90 days before hospitalization, additional confounders were assessed from healthcare claims, including comorbidities, prior medication use, and service use intensity. Diagnostics for each method were evaluated, and methods were compared with respect to the estimates and confidence intervals of treatment effects on repeat PCI, bleeding, and in-hospital death.

Results Of 210,268 patients in the hospital-based cohort, 3240 (1.5 %) had linked healthcare claims. This subset was younger and healthier than the overall study population. The linked subset was too small for complete case evaluation of two of the three outcomes of interest. Multiple imputation and PS calibration did not meaningfully impact treatment effect estimates and associated confidence intervals.

Conclusions Despite more than 98 % missingness on 24 variables, PS calibration and multiple imputation incorporated confounders from healthcare claims without major increases in estimate uncertainty. Additional research is needed to determine the relative bias of these methods.

✉ Jessica M. Franklin
JMFranklin@partners.org

[1] Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont St., Suite 3030, Boston, MA 02120, USA

[2] Aetion, Inc., New York, NY, USA

Key Points

Even with two large, nationally representative databases, the proportion of patients from the inpatient cohort with linked healthcare claims was small and not representative of the full cohort. Complete case analysis therefore led to highly imprecise and non-generalizable estimates of treatment effect.

Propensity score calibration and multiple imputation did not greatly impact treatment effect estimates, but these methods also did not lead to inflated estimator variance, despite more than 98 % missingness on 24 variables.

Performance of these methods was similar regardless of whether the linked subset was a representative subsample of the population.

1 Introduction

The Premier Perspective database is a promising data source for studies of inpatient medication use and health outcomes. These data include information on diagnoses, procedures, and medications from approximately 15 % of US hospitalizations [1], allowing for the evaluation of routine care inpatient medication use that is not available in standard healthcare claims databases. However, recording of patient characteristics that confound the association between medication use and outcomes may be incomplete. For example, intensity of prior healthcare services utilization, considered to be one of the most important confounders measured in administrative claims data [2], cannot be assessed from inpatient data alone. Incorporating additional data on study subjects from healthcare claims prior to hospitalization may improve confounding control, but linked claims data are often accessible only for a small subset of the main study population.

Several methods are available for incorporating additional confounder data available in a subset of the main study population, also known as a 'validation subset'. The complete case method restricts analyses to patients in the validation subset that have all confounders measured. While this approach simplifies analyses, it often results in a large reduction in study size and may bias treatment effect estimates [3–5]. Propensity score (PS) calibration is an alternative approach that was developed for use with healthcare claims data; it incorporates confounders measured in the validation subset by 'calibrating' the PS values in the main study population, thereby preserving study size [6, 7].

The problem of integrating external data available on a subset may also be viewed as a missing data problem; the additional confounders measured in the validation subset are missing for all patients who do not have linked external data [8]. Through that lens, all of the missing data methods become available, including multiple imputation of external confounders [9]. This approach has not often been used in the context of incorporating linked confounding information in administrative healthcare data [10, 11], likely due to the perceived weakness of this method when there is a large proportion of patients with missing data on many variables. However, a large proportion of missing data on some confounders is likely to become increasingly common as comparative-effectiveness and drug safety studies seek to use data from multiple sources in order to effectively control confounding. In addition, there is no theoretical upper bound to the amount of missing data that can be imputed as long as the number of imputations used is sufficient to achieve estimator efficiency [12]. Thus, with careful exploration, the tools of multiple imputation may be applicable to this scenario and could provide better confounding control than PS calibration.

In this study, we focused on the example of the comparative safety of two anticoagulants in the routine care of patients hospitalized with acute coronary syndrome and undergoing percutaneous coronary intervention (PCI). We used this example to compare the advantages and disadvantages of approaches for adjusting for confounders from healthcare claims data that are measured for a subset of the primary inpatient population. Randomized trials indicated that bivalirudin provides protection from thrombotic events that is similar to that from heparin, but a significantly lower risk of major bleeds and potentially death [13–18]. In routine care, many patients receiving these drugs may be at considerably higher risk of adverse outcomes than patients included in randomized trials [19, 20], and bivalirudin is preferred for these high-risk patients [21].

2 Methods

2.1 Data Sources

The primary inpatient cohort was drawn from the Premier Perspective Comparative Database, a repository of hospital administrative data that includes approximately one-sixth of all hospitalizations in the USA. Premier provides data services to hospitals, including tabulation and benchmarking against the performance of other institutions. Service-level data that are recorded include charges for medications, procedures, and laboratory tests, allowing for assessment of in-hospital medication use that is generally not available in healthcare claims data. Other data including hospital characteristics, patient demographic characteristics, discharge diagnoses, and discharge status (including death, but not its cause) are also available [22, 23].

Pre-hospitalization healthcare claims were derived from the UnitedHealth Research Database, a large, nationally representative database. Cross-sectionally, more than 15 million patients are enrolled in commercial health plans through UnitedHealth and are accruing claims in the UnitedHealth database. For covered patients, it contains a longitudinal record of all claims for physician visits, hospitalizations, nursing home stays, and outpatient prescription medication dispensings. Claims information includes inpatient and outpatient diagnoses and procedures, eligibility, and date of death. However, there are no data available from UnitedHealth on inpatient dispensings of medications.

The institutional review board of Brigham and Women's Hospital approved this study.

2.2 Primary Inpatient Cohort

We identified patients aged 18 years or older in Premier who were admitted to hospital and underwent PCI between 1 January 2004 and 31 December 2008 [20]. We excluded patients whose index PCI was at a rural hospital or at any hospital with an average PCI volume of less than one PCI per day in the quarter. In the remaining patients, we evaluated inpatient medication charges on the day of the index PCI. Patients receiving bivalirudin with or without glycoprotein IIb/IIIa inhibitor (GPI) on the day of PCI and no exposure to heparin that day were considered exposed to bivalirudin. The comparison group comprised patients receiving at least 1000 units of heparin plus GPI on the day of PCI, but no exposure to bivalirudin on that day. Patients with any other exposure pattern on the day of PCI, such as exposure to both bivalirudin and heparin, were excluded.

Outcomes included a repeat PCI procedure, blood transfusion, or in-hospital death. Blood transfusion was defined as a charge for any blood product and was intended to proxy for major bleed. Follow-up for repeat PCI began the day after PCI and continued until hospital discharge or 30 days post-PCI. Follow-up for transfusion and death was similar, beginning on the day of PCI.

We extracted potential confounders from Premier inpatient data, including demographics, admission characteristics, comorbid diagnoses, and hospital characteristics. Demographic information included age, race, low-income status, and marital status. Admission characteristics included year of admission, whether the admission was urgent with a primary cardiovascular diagnosis (*International Classification of Diseases, ninth edition* [ICD-9] code: 410, 411, 414.01), whether the PCI occurred within 1 day of admission, and the number of stents used in PCI. Comorbid diagnoses were assessed based on up to 100 discharge codes for the index admission.

2.3 Subset with Linked Healthcare Claims

Within the primary inpatient cohort, we identified patients who were continuously enrolled in a health insurance plan from UnitedHealth for at least 90 days prior to their index hospitalization. For these patients, we assessed 24 additional confounders (Table 1) from UnitedHealth claims during the 90 days prior to hospitalization for PCI. Specifically, we reassessed the presence of several comorbidities that may be recorded more completely in healthcare claims, including diabetes mellitus, hypertension, and prior PCI. In addition, we assessed the use of cardiovascular medications and intensity of health services utilization during the 90 days before hospitalization for PCI.

2.4 Cohort Matched to Linked Patients

Because the subset of patients with linked healthcare claims differed from the primary inpatient cohort on many important patient characteristics, we identified an additional cohort within the primary inpatient cohort that mimicked the characteristics of the linked subset. The matched cohort was created so that the linked patients formed a representative subset of the matched cohort, and it allowed us to compare the performance of methods when applied to a cohort that contains a representative linked subset versus a non-representative linked subset as in the primary inpatient cohort. In order to identify the matched population, we developed a model for the propensity to be selected into the linked subset using logistic regression on the inpatient confounders in the primary cohort. We then performed 10:1 fixed ratio matching on the predicted probability of selection within each exposure group, which matched ten patients who did not have linked healthcare claims to each patient who did. This process creates balance on inpatient characteristics between patients in the linked subset and their corresponding matches, together forming the matched subset. A large, fixed matching ratio (10:1) was possible because of the large pool of potential matches in the primary inpatient cohort.

2.5 Statistical Analyses

As detailed below, we estimated crude and PS-adjusted risk ratios (RRs) using the inpatient confounders only in the full cohort, the matched cohort, and the linked subset. Adjusting for both inpatient and healthcare claims confounders, we performed a complete case analysis in the linked subset and performed PS calibration and multiple imputation to estimate fully-adjusted RRs in the primary inpatient and matched cohorts.

2.5.1 Ordinary Propensity Score (PS) Adjustment with Inpatient Confounders Only

In the primary inpatient cohort, we used logistic regression to estimate the propensity to receive bivalirudin rather than heparin, based on the inpatient confounders (PS_{in}). We then performed marginal mean weighting through stratification to estimate the PS-adjusted RR of each outcome in bivalirudin versus heparin patients [24]. This method combines the benefits of stratification and weighting and is a good general approach for creating PS-adjusted estimates [25]. In our example study, we created 25 PS strata and calculated average treatment effect among the treated weights based on the number of patients in each exposure group in the stratum [26]. These weights were used to calculate the weighted mean or weighted proportion of

Table 1 Patient characteristics from in-patient data and pre-admission claims in the full cohort and linked subpopulation

Covariates	Cohort					
	Primary inpatient		Matched to linked subset		Linked subset	
In-patient variables	Bivalirudin $N = 78{,}918$	Heparin $N = 131{,}350$	Bivalirudin $N = 9207$	Heparin $N = 26{,}433$	Bivalirudin $N = 837$	Heparin $N = 2403$
Age (mean)	66.13	62.49	59.19	56.43	58.47	56.46
Male	64.73	68.53	74.88	77.73	73.00	77.57
White	73.80	74.23	74.25	77.23	73.36	76.90
Low income	4.63	5.49	0.85	0.48	1.31	0.33
Married	58.38	57.54	70.46	67.38	70.37	67.71
Smoking	14.63	13.29	14.43	11.73	15.41	12.32
Number of stents received						
1–2	87.60	85.70	90.49	89.19	93.67	88.06
≥3	1.83	2.45	1.85	2.13	1.08	2.37
PCI within 1 day	64.29	66.19	67.80	71.18	64.16	71.79
Urgent CV admission	65.85	77.80	60.03	79.15	60.93	78.28
Comorbidities						
Prior MI	24.64	58.66	23.55	60.89	22.46	60.13
Ischemic heart disease	34.65	25.26	32.74	23.58	32.50	24.34
Hypertension	63.97	59.34	61.02	56.32	61.17	56.47
Prior PCI	24.05	15.54	26.31	15.25	25.21	16.40
Diabetes mellitus	28.13	24.15	23.93	19.85	25.33	20.56
Liver disease	0.46	0.41	0.67	0.44	0.96	0.37
Prior stroke	3.67	3.02	1.86	1.13	2.39	1.12
Peripheral artery disease	8.10	6.07	5.26	3.75	5.26	3.91
COPD/asthma	2.37	2.49	1.34	1.39	1.43	1.41
Cancer	1.56	1.53	0.84	0.89	1.08	0.79
Prior VTE	1.48	1.26	1.14	0.65	1.19	0.62
Chronic kidney disease	4.68	3.06	2.62	1.60	2.87	1.62
Hospital characteristics						
Number of beds						
<399	26.94	33.35	25.67	33.62	19.47	36.00
400–649	43.93	41.59	36.02	40.78	31.42	42.61
≥650	29.13	25.05	38.32	25.60	49.10	21.39
Teaching hospital	70.09	52.75	62.67	38.47	68.46	36.12
High-volume	26.14	10.26	24.95	7.24	32.38	4.33
Northeast	27.86	12.91	15.52	5.03	14.46	5.62
Midwest	7.30	26.44	16.48	40.70	17.68	40.74
South	48.41	48.84	54.84	47.45	54.96	46.36
West	16.43	11.81	13.16	6.82	12.90	7.28
Healthcare claims variables						
Diagnoses and procedures						
Charlson score (mean)					1.19	0.75
Diabetes mellitus					30.70	19.14
Hypertension					56.51	39.12
Cardiovascular disease					69.89	41.78
Hyperlipidemia					55.91	38.29
Prior CABG					2.27	0.96
Prior PCI					9.32	5.16

Table 1 continued

Covariates	Cohort					
	Primary inpatient		Matched to linked subset		Linked subset	
In-patient variables	Bivalirudin $N = 78,918$	Heparin $N = 131,350$	Bivalirudin $N = 9207$	Heparin $N = 26,433$	Bivalirudin $N = 837$	Heparin $N = 2403$
Medications						
Beta blocker					42.29	28.30
Calcium channel blocker					13.98	10.11
Statin					42.17	30.13
Clopidogrel					34.05	15.77
Fibrate					7.29	5.49
Digoxin					3.46	1.08
Insulin					6.81	3.58
Other antidiabetic medication					18.52	13.90
Thiazide diuretic					6.69	5.45
Loop diuretic					10.27	4.33
ACE inhibitor					28.08	18.23
ARB					11.71	6.87
Aldosterone antagonist					0.60	0.17
Health services intensity (mean)						
Unique generics					5.67	4.06
Physician visits					2.89	1.94
Hospitalizations					0.29	0.15
Total hospital LOS					1.90	0.88

All values are % unless otherwise indicated

ACE angiotensin-converting enzyme, *ARB* angiotensin receptor blocker, *CABG* coronary artery bypass graft, *COPD* chronic obstructive pulmonary disease, *CV* cardiovascular, *LOS* length of stay, *MI* myocardial infarction, *PCI* percutaneous coronary intervention, *VTE* venous thromboembolism

each variable within exposure groups in order to compare PS-adjusted covariate balance between exposure groups. These weights were also used to estimate the adjusted RR. This analysis was repeated in the subset of patients with linked healthcare claims data and in the matched subset.

2.5.2 Complete Case Analysis

The complete case analysis was restricted to patients in the linked subset who had inpatient confounders available as well as confounders measured in healthcare claims. In this analysis, the propensity score was developed based on both sets of confounders (PS_{in+HC}). This propensity score was then used to estimate the adjusted RR in the linked subset, as described above.

In order to estimate a fully-adjusted RR in the primary cohort and matched subset, where healthcare claims were available for only a small portion of the sample, we employed PS calibration and multiple imputation.

2.5.3 PS Calibration

PS calibration was performed in two ways: the single imputation approach and the adjustment approach [7, 27]. In each approach, we estimated the linear measurement error model in the linked subset by regressing the gold standard PS (PS_{in+HC}) on the indicator for exposure and the error prone PS (PS_{in}). In the single imputation PS calibration approach, we used this model to impute a single value of the gold standard PS for each patient in the primary inpatient cohort. We used stratification on this calibrated PS to estimate the adjusted RR in the primary inpatient cohort, as described above. In the adjustment approach, we estimated a logistic regression model for outcome that included exposure and the error-prone PS as dependent variables. We extracted the coefficient on exposure as the error-prone log odds ratio (OR) estimate of treatment effect and then used the coefficients from the linear measurement error model to adjust the estimate and associated confidence interval, as implemented in the SAS (Cary, NC, USA) macro '%blinplus' [27, 28].

In order to assess the validity of the PS calibration approach in this example, we replicated the tests of the PS calibration surrogacy assumption described in Stürmer et al. [7]. Surrogacy requires that the error-prone PS does not independently predict outcome, given the gold standard PS and exposure. Surrogacy was assessed through a likelihood ratio test for an independent effect of the error-prone PS (PS_{in}) on the outcome, after adjusting for the gold standard PS and exposure and through the percentage of the variation in the outcome explained by the two PSs that is due to gold standard PS.

Two-stage calibration, an alternative calibration approach that does not require assumptions about the measurement error model, was also considered [29]. However, because this method involves modeling the outcome in both the primary cohort and the linked subset, it was not feasible in our data, where some outcomes were rarely observed in the linked subset.

2.5.4 Multiple Imputation

Multiple imputation was implemented using chained equations for monotone missing data [30] to impute 200 copies of each of the 24 variables derived from healthcare claims data (200 complete datasets). We used logistic regression to impute binary variables. Because the continuous variables in our study were right skewed and therefore not Normally distributed, we used predictive mean matching to generate imputed values. In this method, the linear imputation model is used to generate a predicted value for each patient. Each patient with missing data is matched to the three patients with observed data with the nearest predicted values, and an imputed value is randomly selected from those patients' observed values [31]. Imputation equations were based on inpatient confounders, exposure, and all outcomes [32].

After variables were imputed for all patients, we estimated the exposure PS using inpatient and healthcare claims confounders, separately within each imputation dataset. This process produced 200 PS values for each patient. In the across multiple imputation approach [33], we averaged the 200 PS values and proceeded with the ordinary PS stratified analysis, as described previously. In the within multiple imputation approach, we estimated the PS-adjusted RR separately in each imputation, yielding 200 treatment effect estimates, which we combined using the ordinary combination rules [31]. All analyses were completed in Stata, Version 13 (College Station, TX, USA), except for the '%blinplus' PS calibration macro for SAS and matching, which used the MatchIt package in R, Version 3.0.2 (Vienna, Austria). Both the PS calibration and the multiple imputation analyses were repeated in patients matched to the linked subset.

To assess the validity of multiple imputation in these data, we evaluated the predictive accuracy of the imputation models for the 24 variables that were imputed as measured by model pseudo-R^2 [34]. We compared the imputed values with observed values in the primary inpatient cohort, where imputations were based on a non-representative linked subset, and in the matched subset, where the linked subset was representative. We also evaluated the predictive accuracy of the model for the propensity to be selected into the linked subset. Because membership in the linked subset in these data can be thought of as our missing data mechanism, this model determines to what degree the missing data mechanism depended on observed inpatient characteristics. Finally, we calculated the fraction of missing information, which can be thought of as a measure of the proportion of estimator variance from the imputation analysis that is due to the missing data [31].

3 Results

We identified 210,268 patients who met all inclusion criteria from 177 hospitals in the primary inpatient cohort extracted from Premier, including 78,918 exposed to bivalirudin and 131,350 exposed to heparin. Within this cohort, 3240 (1.5 %) had linked claims data from UnitedHealth and were included in the linked subset. The proportion with linked data was higher among heparin patients than among bivalirudin patients (1.8 vs. 1.1 %).

Table 1 shows measured patient characteristics for the primary inpatient cohort, the linked subset, and the patients matched to the linked subset. Patients in the linked subset were younger and had fewer comorbidities than patients in the primary cohort because patients in the linked subset were required to be enrolled in a commercial health plan and were more likely to be employed. In contrast, the matched subset closely mimicked the linked subset on characteristics measured in inpatient data. In all cohorts, bivalirudin patients were slightly older and had more comorbidities than patients receiving heparin. However, bivalirudin patients were much less likely to have had a prior myocardial infarction, were less likely to have an urgent cardiovascular admission, and received fewer stents during PCI, indicating that bivalirudin patients may have had less severe cardiovascular disease.

3.1 Ordinary PS Adjustment with Inpatient Confounders Only

Figure 1 shows the balance on healthcare claims confounders in the linked subset before and after PS-adjustment using inpatient variables alone (PS_{in}) or using both
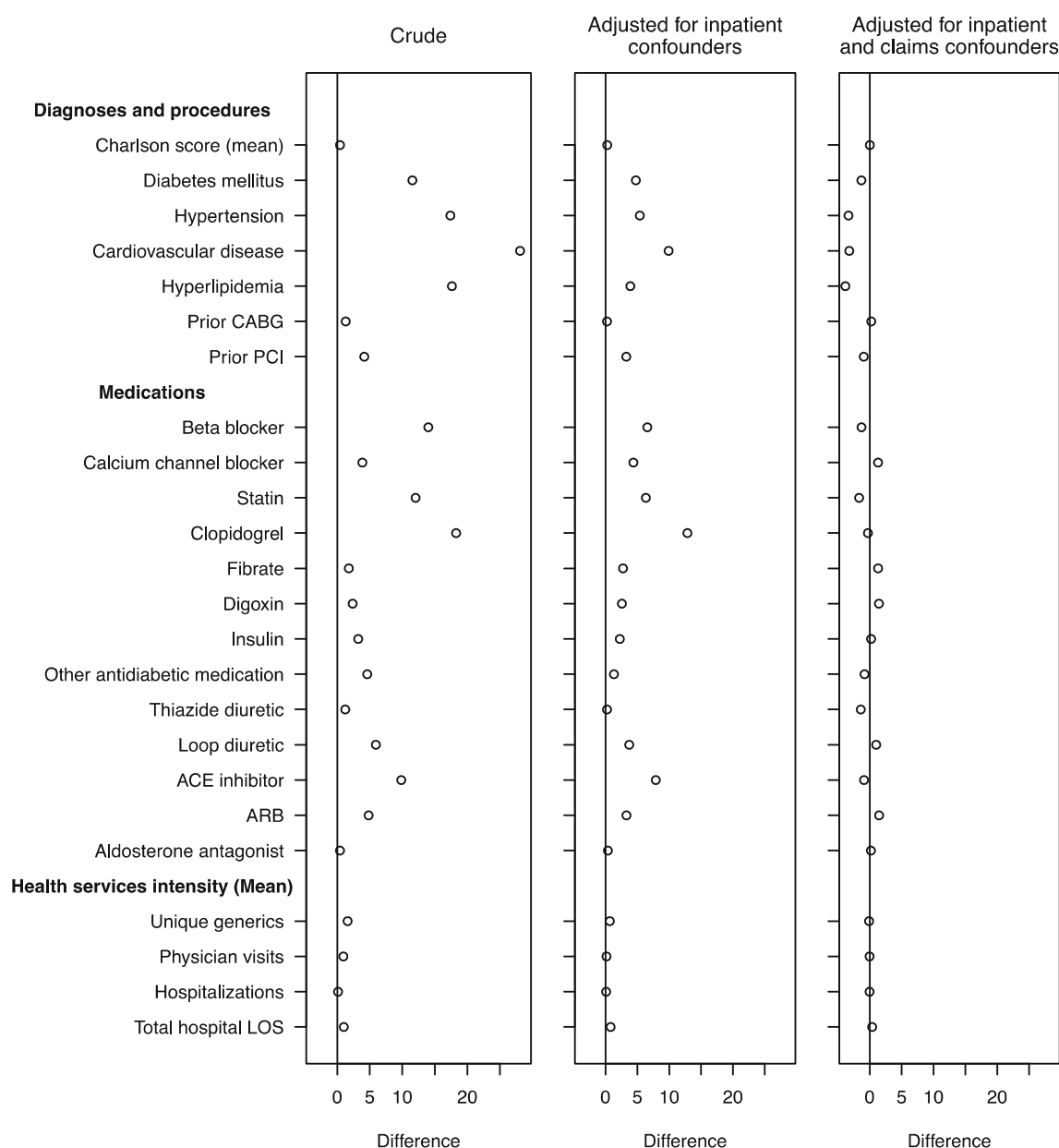
Fig. 1 Reduction in imbalance on healthcare claims covariates after propensity score adjustment in the linked subset. Imbalance is calculated as the difference in means and percentages between bivalirudin and heparin patients. *ACE* angiotensin-converting enzyme, *ARB* angiotensin receptor blocker, *CABG* coronary artery bypass graft, *LOS* length of stay, *PCI* percutaneous coronary intervention

inpatient and healthcare claims variables (PS$_{in+HC}$). Before adjustment, these variables were highly imbalanced and indicated greater comorbidity and medication use among bivalirudin patients. Adjustment for inpatient characteristics reduced imbalance on the measured outpatient variables, owing to correlations among inpatient and healthcare claims variables, but significant imbalances remained. Imbalances were largely removed after adjustment for all confounders.

Crude RR estimates indicated a strong protective effect of bivalirudin on all outcomes in the primary cohort and in each subset (Table 2). PS adjustment for confounders measured in inpatient data reduced estimated effects, but still indicated that bivalirudin was associated with a RR of 0.71 (95 % confidence interval [CI] 0.67–0.76) for repeat PCI procedures, 0.53 (95 % CI 0.49–0.57) for transfusion, and 0.40 (95 % CI 0.35–0.45) for in-hospital death in the full cohort. Results were similar in the matched subset. In

Table 2 Comparative effectiveness and safety of bivalirudin based on inpatient data only

	Outcome		
	Repeat PCI	Transfusion	Death
Full cohort (210,268 patients)			
Bivalirudin (78,918)	2068 (2.6)	1380 (1.7)	419 (0.5)
Heparin + GPI (131,350)	6836 (5.2)	4651 (3.5)	2536 (1.9)
Crude[a]	0.50 (0.48–0.53)	0.49 (0.47–0.52)	0.27 (0.25–0.30)
PS-adjusted[b]	0.71 (0.67–0.76)	0.53 (0.49–0.57)	0.40 (0.35–0.45)
Matched cohort (35,640 patients)			
Bivalirudin (9207)	238 (2.6)	67 (0.7)	25 (0.3)
Heparin + GPI (26,433)	1323 (5.0)	554 (2.1)	269 (1.0)
Crude	0.52 (0.45–0.59)	0.35 (0.27–0.47)	0.27 (0.18–0.40)
PS-adjusted	0.70 (0.58–0.83)	0.35 (0.26–0.47)	0.41 (0.25–0.68)
Linked subset (3240 patients)			
Bivalirudin (837)	21 (2.5)	<5	<5
Heparin + GPI (2403)	101 (4.2)	48 (2.0)	16 (0.7)
Crude	0.60 (0.38–0.95)	–	–
PS-adjusted	0.96 (0.45–2.03)	–	–

Cell sizes of <5 are suppressed. Data are presented as *n* (%) or RR (95 % CI)

CI confidence interval, *GPI* glycoprotein IIb/IIIa inhibitor, *PCI* percutaneous coronary intervention, *PS* propensity score, *RR* risk ratio

[a] Bivalirudin versus heparin

[b] The PS contains only inpatient characteristics listed in Table 1

Table 3 Comparative effectiveness and safety of bivalirudin after adjusting for covariates measured in both inpatient and healthcare claims data

	Repeat PCI	Transfusion	Death
Full cohort (210,268 patients)			
PSC: imputation	0.73 (0.67–0.80)	0.55 (0.49–0.61)	0.42 (0.36–0.51)
PSC: adjustment	0.76 (0.71–0.80)	0.51 (0.48–0.55)	0.42 (0.38–0.48)
MI: within	0.77 (0.67–0.88)	0.53 (0.44–0.64)	0.35 (0.25–0.48)
MI: across	0.82 (0.75–0.89)	0.55 (0.50–0.60)	0.38 (0.33–0.43)
Matched cohort (35,640 patients)			
PSC: imputation	0.74 (0.60–0.90)	0.35 (0.25–0.48)	0.42 (0.24–0.74)
PSC: adjustment	0.76 (0.64–0.90)	0.35 (0.26–0.48)	0.52 (0.32–0.84)
MI: within	0.75 (0.60–0.94)	0.35 (0.24–0.50)	0.37 (0.20–0.67)
MI: across	0.78 (0.64–0.96)	0.35 (0.25–0.48)	0.37 (0.22–0.63)
Linked subset (3240 patients)			
Complete case analysis[a]	1.17 (0.50–2.76)	–	–

MI multiple imputation, *PSC* propensity score calibration

[a] The complete case analysis adjusts for a propensity score that includes all inpatient and healthcare claims characteristics from Table 1

the linked subset, there were few observed events, and results varied.

3.2 Complete Case Analysis

The complete case analysis that adjusted for all measured variables from inpatient and healthcare claims data was restricted to the linked subset, where few events led to poor precision (Table 3).

3.3 PS Calibration

In the primary cohort and matched subset, adjusting for all confounders using PS calibration generally had little impact on estimated RRs or associated CIs for transfusion and in-hospital death. The estimated RR for repeat PCI was generally increased when using PS calibration; for example, in the adjustment approach, bivalirudin was associated with an estimated 24 % reduction in repeat PCI (RR 0.76

[95 % CI 0.71–0.80]). This estimated effect is closer to the results from previous non-randomized studies that controlled for unmeasured confounding with instrumental variable analysis [20]. Results in the primary cohort and matched subset were similar.

Within the two PS calibration approaches, the adjustment approach generally had a greater impact on estimated RRs than the imputation approach. Because a single PS imputation does not account for uncertainty in the calibration model while the adjustment approach does adjust standard errors to account for this uncertainty, we expected that the imputation approach would generally have narrower CIs. However, in these data, the imputation PS calibration approach consistently produced estimates with wider CIs than the adjustment approach.

When assessing surrogacy, we found that the error-prone PS was highly non-significant in models for all outcomes when adjusting for the gold standard PS and exposure. We also found that more than 98 % of the variance in each outcome that is explained by either the error-prone or gold standard PS is attributable to the gold standard PS, so there was no evidence against the surrogacy assumption in these data.

3.4 Multiple Imputation

Similar to PS calibration, both multiple imputation approaches had little impact on the estimates or CIs for transfusion or death. Multiple imputation had a slightly greater impact on the estimated RR for repeat PCI, particularly the across approach. The across approach also produced narrower CIs, which was expected since this approach does not account for the uncertainty associated with the imputation of missing data. Both approaches produced CIs that were comparable to the ordinary PS-adjusted approaches that did not attempt to use variables with missing data.

Figure 2 compares the mean or proportion of the observed versus imputed values for each variable in each of the 200 imputed datasets. This figure shows that, despite the fact that the patients with linked claims data constituted a non-representative subsample of the primary inpatient cohort, the imputed healthcare claims variables appropriately indicated increased comorbidity in the primary cohort. In contrast, the imputations in the matched subset were similar to the observed data from the linked subset (data not shown). The pseudo-R^2 values from the imputation models indicate a range of predictive ability across variables. Prediction accuracy was generally higher for the summary comorbidity and health services variables, such as Charlson score or number of unique generics ($R^2 = 0.52$ and 0.58, respectively). Predictions were less accurate for

use of specific medications, such as calcium channel blockers ($R^2 = 0.10$).

Membership in the linked subset was predicted well from inpatient confounders. The C-statistic that describes the ability of inpatient confounders to discriminate between patients in the linked subset and others was 0.803. Model coefficients, shown in the Electronic Supplementary Material, indicated that the strongest predictors of missing data (not having linked data) were administrative factors, including being Medicare eligible (age ≥65 years), hospitalization for PCI in 2004 (vs. 2005 or later), residence in northeastern or western USA, and low-income status. Despite the large proportion of study patients who were missing confounders from claims, the fraction of missing information was calculated to be 0.4 %, indicating that there should be only a small increase in estimator variance due to the presence of missing data.

4 Discussion

We explored the use of methods to integrate confounder data from linked healthcare claims in a study of inpatient medication use. We used the example of the comparative safety of bivalirudin during PCI, evaluated in the Premier Perspective database linked to UnitedHealth claims data. We found that, even when using two large, nationally representative databases for linkage, only a small proportion of patients could be linked, and these patients were systematically different from the full cohort of study patients. Thus, methods that can adjust for confounders measured in a subset must be capable of integrating data from small, non-representative linked subsets in order to be useful for supplementing confounding adjustment in studies of inpatient medication use. Complete case analysis is unlikely to be useful in such scenarios due to the very small study size and the fact that the complete cases do not generalize to the larger routine-care population of interest.

When exploring PS calibration and multiple imputation, we found that these methods did not meaningfully impact estimates of treatment effect as compared with estimates that used inpatient confounder data alone. However, calibration and imputation preserved study size and did not lead to issues of non-convergence or variance inflation. The differences among PS calibration and multiple imputation approaches were generally small, although some procedures more appropriately accounted for the uncertainty attributable to the missing data. Alternative methods exist for estimating the standard error for the imputation PS calibration approach and the across multiple imputation approach [7, 33]; however, those methods were outside the scope of this paper.
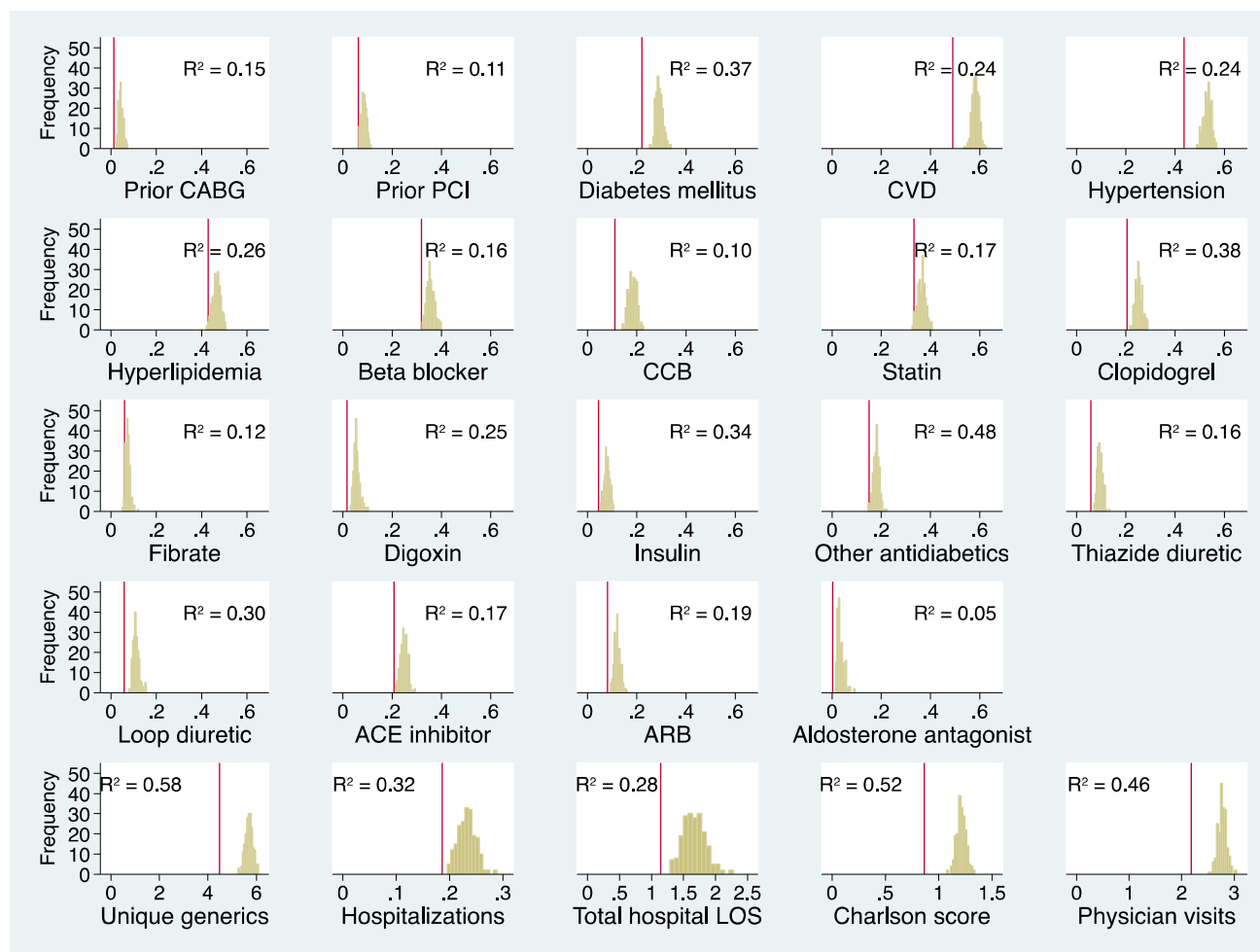
Fig. 2 Observed and imputed values for the 24 healthcare claims confounders. The mean or proportion for each variable in the observed data (linked subset) is shown with the *solid line*. Histograms display the means and proportions from each of the 200 imputations (full cohort). Pseudo-R^2 values for imputation of each variable are shown in the corresponding panels. Continuous variables are in the *bottom row*. *ACE* angiotensin-converting enzyme, *ARB* angiotensin receptor blocker, *CABG* coronary artery bypass graft, *CCB* calcium channel blockers, *CVD* cardiovascular disease, *LOS* length of stay, *PCI* percutaneous coronary intervention

Prior research has compared the imputation PS calibration approach with sample reweighting for incorporating confounders measured in a validation subset [8], but there has been limited work comparing PS calibration with multiple imputation [35]. In studies where the subset with linked data contains many additional measured confounders but a small proportion of study patients, investigators may assume that multiple imputation will fail. In our example, despite more than 98 % missing data on 24 variables, multiple imputation performed similarly to PS calibration, and both approaches increased estimator variance only slightly from the ordinary PS approach that did not attempt to incorporate external confounders. Therefore, additional research should evaluate the relative ability of these approaches to eliminate confounding from incompletely observed confounders across varying data-generating scenarios.

Our study evaluated methods in the context of a representative linked subset, when making inference in the matched subset, and a non-representative linked subset, when making inference in the full inpatient cohort. Results in both cases were similar across the methods under study. Although PS calibration does not require a representative validation subset in order to yield unbiased estimation of treatment effects, it does require that the measurement error model estimated in the validation subset accurately estimates the calibration factor needed in the full cohort [36]. In other words, the linear relation between the error prone and gold standard PSs should be constant in the full cohort and the linked subset. This assumption cannot be tested based on observed data, since the calibration factor cannot be measured in the full cohort, and in cases where these samples are very different, this assumption may be questionable.

The validity of multiple imputation also does not strictly require that the patients with complete data (the linked subset) are representative of the full cohort. Instead, the missing at random assumption necessary for unbiased inference from imputation requires that the likelihood of missing data depends only on variables that are observed for everyone; thus, it should not depend on the variables from outpatient claims or other variables not measured in either database. This assumption also may be questionable when the validation subset is very different from the full cohort on observed variables, particularly if investigators believe that differences in observed variables may be indicative of differences in other unmeasured variables.

In the context of the example presented here, we found that there were large differences between the linked subset and the full cohort. However, the fact that inclusion in the linked subset could be predicted well from completely observed variables and inclusion was predicted most strongly by administrative variables that are captured well in the inpatient data provides some confidence that the missing at random assumption may be appropriate in these data. Furthermore, diagnostic plots indicated that the multiple imputation procedure appropriately accounted for the fact that the full cohort was older and sicker than the linked subset.

Unbiasedness of all approaches also requires that the treatment effect is unconfounded after conditioning on both inpatient and claims covariates. In our study, this assumption was likely violated, as all treatment effect estimates appeared to be negatively biased compared with results from randomized trials. For example, in one meta-analysis [37], the estimated OR for major bleeds was 0.58 (95 % CI 0.49–0.69), whereas our estimates of the RR of transfusion (our proxy for major bleeds) ranged from 0.35 to 0.55. Similarly, from meta-analysis there appears to be no significant effect on death (OR 0.94 [95 % CI 0.78–1.14]), but our estimates for the RR of death ranged from 0.27 to 0.52. Although some differences are to be expected due to differing populations between a randomized trial and a routine care observational study, the magnitude of the difference in estimated effects on death indicate that patients receiving bivalirudin in our cohort were likely healthier than patients receiving heparin in ways that may not have been measured in either the inpatient or healthcare claims variables.

The ability of claims data to augment confounding information from inpatient databases will depend on the specific example, and in some cases, these data may not be sufficient to capture all relevant confounders. In that scenario, investigators may seek other data sources. In this example, important variables available in claims, for example, health services intensity variables, were predicted well by variables available in the inpatient data, indicating that the incorporation of claims variables provided little additional confounder control. In other words, the claims variables were no longer important sources of confounding after adjusting for the inpatient variables. In addition, we observed a very small fraction of missing information in the multiple imputation analysis, despite a large proportion of patients with missing data. This finding indicates that, under the assumptions of the multiple imputation, estimator variance would change little if the missing data had been observed, which suggests that the missing data are unlikely to add much information to the analysis.

5 Conclusions

Based on the results in this study, we conclude that PS calibration and multiple imputation may be useful for adjusting for confounders measured in healthcare claims databases when studying the comparative safety and effectiveness of inpatient medication use, but additional research is needed. Simulation studies that investigate the relative performance of these methods should be designed with careful attention to producing realistic confounding and missing data mechanisms. Additional research on method diagnostics is also needed in order to educate investigators on each method's strengths and weaknesses and to allow for assessment of method assumptions in a given study. Outside of this particular application, the increasing availability of linkage across data sources will increase the importance of methods such as these that can make use of all existing data for improved confounding adjustment in comparative-effectiveness studies.

References

1. Schneeweiss S, Seeger JD, Landon J, Walker AM. Aprotinin during coronary-artery bypass grafting and risk of death. N Engl J Med. 2008;358(8):771–83.

2. Schneeweiss S, Seeger JD, Maclure M, Wang PS, Avorn J, Glynn RJ. Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. Am J Epidemiol. 2001;154(9):854–64.

3. Knol MJ, Janssen KJ, Donders ART, Egberts AC, Heerdink ER, Grobbee DE, Moons KG, Geerlings MI. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. J Clin Epidemiol. 2010;63(7):728–36.

4. van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol. 2006;59(10):1102–9.

5. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med. 2010;29(28):2920–31.

6. Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. Am J Epidemiol. 2005;162(3):279–89.

7. Stürmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration—a simulation study. Am J Epidemiol. 2007;165(10):1110–8.

8. Nelson JC, Marsh T, Lumley T, Larson EB, Jackson LA, Jackson ML. Validation sampling can reduce bias in health care database studies: an illustration using influenza vaccination effectiveness. J Clin Epidemiol. 2013;66(8):S110–21.

9. Little R, Rubin D. Statistical analysis with missing data. 2002.

10. Schneeweiss S, Rassen JA, Glynn RJ, Myers J, Daniel GW, Singer J, Solomon DH, Kim S, Rothman KJ, Liu J. Supplementing claims data with outpatient laboratory test results to improve confounding adjustment in effectiveness studies of lipid-lowering treatments. BMC Med Res Methodol. 2012;12(1):180.

11. Toh S, García Rodríguez LA, Hernán MA. Analyzing partially missing confounder information in comparative effectiveness and safety research of therapeutics. Pharmacoepidemiol Drug Saf. 2012;21(S2):13–20.

12. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci. 2007;8(3):206–13.

13. Kastrati A, Neumann F-J, Schulz S, Massberg S, Byrne RA, Ferenc M, Laugwitz K-L, Pache J, Ott I, Hausleiter J. Abciximab and heparin versus bivalirudin for non-ST-elevation myocardial infarction. N Eng J Med. 2011;365(21):1980–9.

14. Lincoff AM, Bittl JA, Harrington RA, Feit F, Kleiman NS, Jackman JD, Sarembock IJ, Cohen DJ, Spriggs D, Ebrahimi R. Bivalirudin and provisional glycoprotein IIb/IIIa blockade compared with heparin and planned glycoprotein IIb/IIIa blockade during percutaneous coronary intervention. JAMA J Am Med Assoc. 2003;289(7):853–63.

15. Lincoff AM, Kleiman NS, Kereiakes DJ, Feit F, Bittl JA, Jackman JD, Sarembock IJ, Cohen DJ, Spriggs D, Ebrahimi R. Long-term efficacy of bivalirudin and provisional glycoprotein IIb/IIIa blockade vs heparin and planned glycoprotein IIb/IIIa blockade during percutaneous coronary revascularization. JAMA J Am Med Assoc. 2004;292(6):696–703.

16. Schulz S, Mehilli J, Ndrepepa G, Neumann F-J, Birkmeier KA, Kufner S, Richardt G, Berger PB, Schömig A, Kastrati A. Bivalirudin vs. unfractionated heparin during percutaneous coronary interventions in patients with stable and unstable angina pectoris: 1-year results of the ISAR-REACT 3 trial. Eur Heart J. 2010;31(5):582–7.

17. Stone GW, McLaurin BT, Cox DA, Bertrand ME, Lincoff AM, Moses JW, White HD, Pocock SJ, Ware JH, Feit F. Bivalirudin for patients with acute coronary syndromes. N Eng J Med. 2006;355(21):2203–16.

18. Stone GW, Witzenbichler B, Guagliumi G, Peruga JZ, Brodie BR, Dudek D, Kornowski R, Hartmann F, Gersh BJ, Pocock SJ.

19. Hibbert B, MacDougall A, Labinaz M, O-Brien ER, So DY, Dick A, Glover C, Froeschl M, Marquis J-F, Wells GA. Bivalirudin for primary percutaneous coronary interventions outcome assessment in the Ottawa STEMI Registry. Circ Cardiovasc Interv. 2012;5(6):805–12.

20. Rassen JA, Mittleman MA, Glynn RJ, Brookhart MA, Schneeweiss S. Safety and effectiveness of bivalirudin in routine care of patients undergoing percutaneous coronary intervention. Eur Heart J. 2010;31(5):561–72.

21. Popma JJ, Berger P, Ohman EM, Harrington RA, Grines C, Weitz JI. Antithrombotic therapy during percutaneous coronary intervention: the seventh ACCP conference on antithrombotic and thrombolytic therapy. Chest J. 2004;126(3 Suppl):576S–99S.

22. Lindenauer PK, Pekow P, Wang K, Mamidi DK, Gutierrez B, Benjamin EM. Perioperative beta-blocker therapy and mortality after major noncardiac surgery. N Eng J Med. 2005;353(4):349–61.

23. Lindenauer PK, Pekow P, Wang K, Gutierrez B, Benjamin EM. Lipid-lowering therapy and in-hospital mortality following major noncardiac surgery. JAMA. 2004;291(17):2092–9.

24. Hong G. Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. J Educ Behav Stat. 2010;35(5):499–531.

25. Linden A. Combining propensity score-based stratification and weighting to improve causal inference in the evaluation of health care interventions. J Eval Clin Pract. 2014;20(6):1065–71

26. Hansen BB. Full matching in an observational study of coaching for the SAT. J Am Stat Assoc. 2004;99(467):609–18.

27. Spiegelman D, Carroll RJ, Kipnis V. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. Stat Med. 2001;20(1):139–60.

28. Logan R, Spiegelman D. The SAS %BLINPLUS Macro. 2012.

29. Lin H-W, Chen Y-H. Adjustment for missing confounders in studies based on observational databases: 2-stage calibration combining propensity scores from primary and validation data. Am J Epidemiol. 2014;180(3):308–17.

30. Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv Methodol. 2001;27(1):85–96.

31. Rubin DB. Multiple imputation for nonresponse in surveys, vol. 307. USA: Wiley; 2009.

32. Moons KG, Donders RA, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59(10):1092–101.

33. Mitra R, Reiter JP. A comparison of two methods of estimating propensity scores after multiple imputation. Stat Methods Med Res. 2012;1–17. doi:10.1177/0962280212445945

34. Veall MR, Zimmermann KF. Pseudo-R2 measures for some common limited dependent variable models. J Econ Surv. 1996;10(3):241–59.

35. Stürmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Comparison of performance of propensity score calibration (PSC) and multiple imputation (MI) to control for unmeasured confounding using an internal validation study. Pharmacoepidemiol Drug Saf. Wiley 2006;S39–S40.

36. Rosner B, Spiegelman D, Willett W. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol. 1990;132(4):734–45.

37. Lipinski MJ, Lhermusier T, Escarcega RO, Baker NC, Magalhaes MA, Torguson R, Suddath WO, Satler LF, Pichard A, Waksman R. Bivalirudin versus heparin for percutaneous coronary intervention: an updated meta-analysis of randomized controlled trials. Cardiovasc Revascularization Med. 2014;15(6):315-22.