American Journal of Epidemiology
© The Author 2017. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Vol. 185, No. 6
DOI: 10.1093/aje/kww155
Advance Access publication:
February 17, 2017

Practice of Epidemiology

Conducting Privacy-Preserving Multivariable Propensity Score Analysis When Patient Covariate Information Is Stored in Separate Locations

Justin Bohn, Wesley Eddings, and Sebastian Schneeweiss*

* Correspondence to Dr. Sebastian Schneeweiss, Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02120 (e-mail: sschneeweiss@partners.org).

Initially submitted February 27, 2015; accepted for publication March 24, 2016.

Distributed networks of health-care data sources are increasingly being utilized to conduct pharmacoepidemiologic database studies. Such networks may contain data that are not physically pooled but instead are distributed horizontally (separate patients within each data source) or vertically (separate measures within each data source) in order to preserve patient privacy. While multivariable methods for the analysis of horizontally distributed data are frequently employed, few practical approaches have been put forth to deal with vertically distributed health-care databases. In this paper, we propose 2 propensity score–based approaches to vertically distributed data analysis and test their performance using 5 example studies. We found that these approaches produced point estimates close to what could be achieved without partitioning. We further found a performance benefit (i.e., lower mean squared error) for sequentially passing a propensity score through each data domain (called the "sequential approach") as compared with fitting separate domain-specific propensity scores (called the "parallel approach"). These results were validated in a small simulation study. This proof-of-concept study suggests a new multivariable analysis approach to vertically distributed health-care databases that is practical, preserves patient privacy, and warrants further investigation for use in clinical research applications that rely on health-care databases.

database linkage; databases; database studies; epidemiologic methods; pharmacoepidemiology; propensity scores

Abbreviations: ANOVA, analysis of variance; MSE, mean squared error; PS, propensity score.

Distributed systems for the analysis of electronic health-care data are increasingly being used to support large-scale analyses of medical interventions and products (1–3). Such systems may be horizontally distributed—that is, data on separate patients are combined across distinct databases to increase study size and population heterogeneity. An example is the Food and Drug Administration's Sentinel Initiative, which pools analyses across multiple data partners, with the data residing locally (1). Increasingly researchers deal with vertically distributed databases, in which different covariates on the same set of patients reside in multiple databases that are physically separated to reduce the risk of identifying patients. For example, information on important medication-use variables may be available in insurance claims data, while detailed medical test results might be available only in electronic health records. Analysis of vertically distributed databases attempts to make use of the increased depth of clinical information available on patients from multiple sources, and thus improve confounding adjustment in multivariable analyses. An example of this scenario is the linkage of administrative claims to genomic data (Figure 1). A defining feature of many distributed networks is that all data contributors control their own data, storing it behind their own firewall and performing as much analysis as possible on their own hardware before submitting summary results to a coordinating center (1, 4–6). This arrangement is intended to protect both the privacy of patients (by avoiding transfer of potentially identifying data) and the proprietary data related to the business practices of data owners.

Several strategies have been proposed for performing multivariable analysis of horizontally distributed data, including meta-analytical methods to pool site-specific estimates (7, 8),

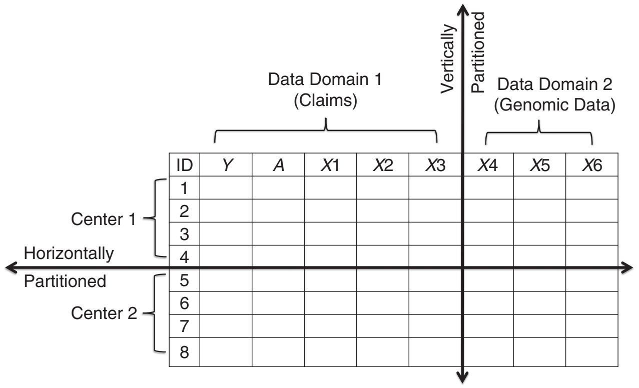


Figure 1. Structure of vertically and horizontally partitioned health-care databases. In this example, the analysis of interest concerns the effect of an exposure *A* on an outcome *Y*, wherein adjustment is needed for confounders *X*1–*X*6. In a horizontally partitioned system, different patient subsets are contributed by different sources (here, centers 1 and 2), while in a vertically partitioned system different patient covariates are contributed by different sources (here, medical insurance claims and a genomic database). ID, identification.

propensity score (PS)–based methods to reduce covariate-sharing (9, 10), and methods that share only limited stratified tabular data (7, 11), as well as strategies that allow for traditional computation with no physical data-sharing at all, such as distributed regression (12–14). However, few methods have been proposed that allow investigators to conduct multivariable adjusted analyses when covariate data on the same patients are in physically separate locations (13). Distributed regression routines may ultimately enable traditional analysis across distributed networks, but there are practical barriers to this approach. For instance, distributed regression approaches need to transmit data back and forth between sources at each iteration of the fitting routine, requiring repeated access to each data source, which data contributors may find objectionable (15, 16).

Methods for such vertically distributed multivariable analyses that are statistically valid, privacy-preserving, and operationally practical need to be further developed as more patient-level electronic data sources become available for distributed analyses, as privacy protection limits the physical pooling of all covariates (17). We developed a suite of PS-based approaches for multivariable analysis of vertically distributed data and demonstrated their performance using data from 5 previously published cohort studies on medication-outcome associations. We then assessed variations on how these approaches were applied.

METHODS

Building vertically distributed data systems for 5 example studies

We used data from 5 previously published cohort studies (18–21) of drug exposures to illustrate vertically distributed analyses in a range of settings. Table 1 summarizes the key characteristics of each example study. In each study cohort, we used the high-dimensional PS algorithm's covariate selection procedure (18) to empirically identify up to 4,800 covariates for consideration as confounders. Using raw claims data, this procedure identifies potentially confounding characteristics (defined by the presence or absence of diagnosis, procedure, or drug codes) according to their associations with the outcome of interest and their prevalence in the exposed and unexposed populations. The high-dimensional PS procedure selects the top 1,200 covariates from each of 4 domains: in-hospital diagnoses and procedures, outpatient diagnoses and procedures, outpatient pharmacy prescription drug-filling, and demographic factors. The demographic information included age at cohort entry, sex, race, and calendar year of cohort entry. Although data from these 4 domains were all available in a single claims database, we mimicked a vertically distributed data system by treating the 4 domains as distinct data sources between which only select patient-level information, including patient identification number, exposure indicator, outcome indicator, index date, and estimated PSs, could be shared in order to preserve patient privacy (9, 10). We further assumed that age and sex information was available in each data domain, a realistic assumption for most data sources in health care.

PSs for multivariable adjustment in distributed systems concealing patient characteristics

A PS is a subject's estimated probability of receiving the exposure of interest, conditional on the measured covariates, and is generally estimated via logistic regression. It has been recognized in the horizontally distributed data setting that the dimension-reducing property of PSs can be utilized for privacy-preserving multivariable distributed analyses (9, 10). The key idea is that in each horizontally distributed study center, a PS is estimated from a logistic regression model that includes the full covariate vector. Each center then shares a nonidentifiable individual-level file containing, at a minimum, 3 variables: exposure status, outcome status, and the estimated PS—information that makes it impossible to identify a patient. Time-to-event variables and variables that identify broad subgroups may also be included without revealing patient identity (22). These individual-level data can then be pooled and analyzed centrally, stratifying by study center.

This approach can be expanded to accommodate vertically distributed data by separately estimating PSs within distinct data domains (e.g., claims, genetic data) and then combining these PSs into a single value. Such an approach requires a unique identifier in each data domain to allow linkage (Figure 1). This joined identifier could be deterministic (e.g., an insurance identification number) or probabilistic (e.g., defined through patterns of health-care utilization or other measures), but it should not contain any personal health information. Since PSs are estimated by modeling exposure status, which is usually available in only a single data domain (e.g., medication use in the pharmacy prescription-filling file), this approach additionally assumes that exposure information can be shared with each of the data domains.

The first step of the PS approach is therefore to sort each data domain by a joint patient identification number and

Table 1. Overview of 5 Example Studies Used in Empirical Assessment of Parallel and Sequential Propensity Score Approaches to Analysis of Vertically Distributed Data, 2009–2014

First Author, Year (Reference No.)	Data Source	Exposure	Comparator	Outcome	No. of Events	No. of Persons	No. of Covariates			Outcome Model	OR/HR[a]		Follow-up Model[b]
							Inpatient	Outpatient	Drugs		Crude	Adjusted	
Schneeweiss, 2009 (18)	PACE/ Medicare	COX-2 inhibitors	Nonselective NSAIDs	Gastrointestinal bleeding	552	49,653	455	936	496	Logistic	1.09	0.86	Fixed
Schneeweiss, 2009 (18)	PACE/ Medicare	Statins	Glaucoma drugs	Mortality	1,739	36,122	696	1,185	486	Logistic	0.56	0.88	Fixed
Schneeweiss, 2010 (19)	British Columbia PharmaNet	Tricyclics	SSRIs	Suicide or attempted suicide	166	13,942	10	461	85	Cox	0.59	0.78	As-treated
Patorno, 2010 (20)	HealthCore	Gabapentin	Topiramate	Suicide or attempted suicide	346	200,718	370	574	415	Cox	0.96	1.56	As-treated
Patorno, 2014 (21)	HealthCore	CYP450-inducing anticonvulsants	Other anticonvulsants	Ischemic coronary or cerebrovascular events	564	166,031	118	623	461	Cox	1.72	1.38	As-treated

Abbreviations: COX-2, cyclooxygenase 2; CYP450, cytochrome P-450; HR, hazard ratio; NSAIDs, nonsteroidal antiinflammatory drugs; OR, odds ratio; PACE, Pharmaceutical Assistance Contract for the Elderly; SSRIs, selective serotonin reuptake inhibitors.

[a] Reference treatment effects were estimated using unpartitioned data sets with high-dimensional propensity score–adjusted logistic regression models (example studies 1 and 2) or Cox models (example studies 3–5).

[b] Fixed follow-up refers to analyzing patients with respect to the exposure they *initiated* (as in an "intention-to-treat" analysis), whereas as-treated follow-up involves censoring patients when they stop their initial exposure.

share each patient's exposure information across all data domains (Figure 2). Sharing the exposure status alone without any additional patient data will not make patients identifiable and should thus be acceptable to all data contributors (though, should a contributor decline, their specific data domain could not be included in analysis). In our example studies, we further assumed that it would also be possible to share age and sex information between data domains, which seems reasonable given that data contributors are unlikely to consider this proprietary information. Once this data structure is in place, one can estimate the PS in each data domain separately (the parallel approach) or estimate the PS in one domain first and then pass that PS on to the next data domain for inclusion in a second PS model, iteratively working through all available domains (the sequential approach).

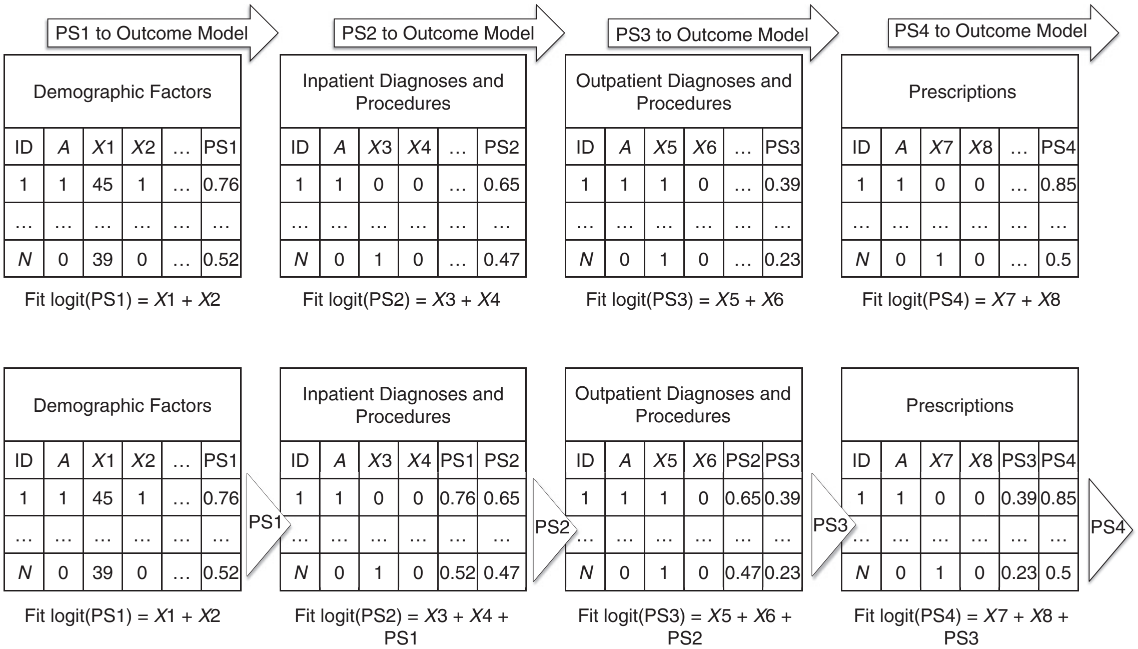
Parallel and sequential PS approaches for vertically partitioned data

In the parallel approach, a separate PS model was fitted within each data domain. Each patient in each example study had a demographic PS, an in-hospital PS, an outpatient PS, and a prescription drug PS (Figure 2, top row). To estimate the treatment effect, an outcome model using logistic or Cox regression (depending on example study; Table 1) was fitted including terms for exposure and some function of the 4 domain-specific PSs.

In the sequential approach, an initial PS model was fitted within the first domain (e.g., prescription drugs). The estimated PS was then passed to the second domain (e.g., inpatient diagnoses and procedures). In the second domain, a PS model was then fitted including terms for all domain-specific covariates and the estimated PS from the first domain. This PS was in turn passed to the third domain (Figure 2, bottom row). In the third domain, a PS model was fitted including terms for all domain-specific covariates and the estimated PS from the second domain. This process was repeated through all 4 domains until a final PS was produced. The treatment effect was estimated fitting a logistic or Cox regression, depending on example study (Table 1), including terms for exposure and the PS from the last domain.

Variations of the parallel and sequential approaches

We explored several variations of the parallel and sequential approaches. For the parallel approach, we considered 1) including age and sex in the calculation of all domain-specific PSs, 2) fitting a "last-stage" PS model (a logistic model regressing treatment on the domain-specific PSs, their squared terms, and pairwise interactions) to generate a single PS used in the outcome model, and 3) for variants with a last-stage model, including the final PS as a continuous term versus PS-decile stratification (i.e., using indicators in the



Demographic Factors

ID	*A*	*X*1	*X*2	…	PS1
1	1	45	1	…	0.76
…	…	…	…	…	…
N	0	39	0	…	0.52

Inpatient Diagnoses and Procedures

ID	*A*	*X*3	*X*4	…	PS2
1	1	0	0	…	0.65
…	…	…	…	…	…
N	0	1	0	…	0.47

Outpatient Diagnoses and Procedures

ID	*A*	*X*5	*X*6	…	PS3
1	1	1	0	…	0.39
…	…	…	…	…	…
N	0	1	0	…	0.23

Prescriptions

ID	*A*	*X*7	*X*8	…	PS4
1	1	0	0	…	0.85
…	…	…	…	…	…
N	0	1	0	…	0.5

Demographic Factors

ID	*A*	*X*1	*X*2	…	PS1
1	1	45	1	…	0.76
…	…	…	…	…	…
N	0	39	0	…	0.52

Inpatient Diagnoses and Procedures

ID	*A*	*X*3	*X*4	PS1	PS2
1	1	0	0	0.76	0.65
…	…	…	…	…	…
N	0	1	0	0.52	0.47

Outpatient Diagnoses and Procedures

ID	*A*	*X*5	*X*6	PS2	PS3
1	1	1	0	0.65	0.39
…	…	…	…	…	…
N	0	1	0	0.47	0.23

Prescriptions

ID	*A*	*X*7	*X*8	PS3	PS4
1	1	0	0	0.39	0.85
…	…	…	…	…	…
N	0	1	0	0.23	0.5

Figure 2. Schematic representation of the parallel and sequential approaches to analysis of vertically distributed data. The analytical goal is to estimate the effect of an exposure *A* on an outcome *Y*, wherein adjustment is needed for many covariates (*X*1–*X*8), on which data are available from 4 separate sources (domains) and cannot be pooled in a single analytical database. In the parallel approach (top row), separate propensity scores (PSs), PS1–PS4, are estimated within each domain, and the final analysis utilizes a function of the 4 domain-specific PSs—for example, in the model $Y = A + f\,(PS1 + PS2 + PS3 + PS4)$. In the sequential approach (bottom row), a PS (PS1) is estimated in the first domain and then passed to the second domain. In the second domain, a PS is estimated on the basis of covariates in that domain *and* the PS from the first domain (PS2). This process is repeated iteratively across all domains until a single final PS (PS4) is produced, which can be used in the final analysis—for example, in the model $Y = A + f\,(PS4)$. ID, identification.

outcome model), yielding 6 possible variations. For the sequential approach, we considered 1) varying the order of the domains in the process and 2) including the final PS as a continuous term versus PS-decile stratification (i.e., using indicators in the outcome model), producing 48 possible variants. In all analyses, PSs were used on the logit scale to allow for nonlinearity. Further mention of PSs should be taken to refer to the logit-scale PS. Additional information regarding these variations can be found in Web Table 1 and the Web Appendix (available at http://aje.oxfordjournals.org/).

In each study, the reference treatment effect estimate against which all method variations were compared was obtained from the unpartitioned analysis with a PS estimated using all variables in all domains. Logistic or Cox proportional hazards regression models were used, with the PS included as a linear term or in deciles, depending on the method variant being compared. Matching on the PS and inverse probability weighting were not considered here, as they could not be applied across some variations of the parallel approach. While adjustment for a PS as a linear term requires strong assumptions and is not often done in practice, it has been shown to produce confounding control comparable to that of other strategies, including adjustment for quintiles or spline functions of the PS, matching on the PS, and inverse probability weighting (23). The outcome models used in this study were the same as those used in the published reports of the example studies.

Assessing the relative performance of the parallel and sequential approaches

We defined the bias of a given variant of the parallel or sequential approach as the difference between the estimated treatment effect and the reference treatment effect on the log scale (i.e., $\beta_{\text{variant}} - \beta_{\text{reference}}$). We also examined bias on the absolute scale (i.e., $|\beta_{\text{variant}} - \beta_{\text{reference}}|$) and the mean squared error (MSE) (i.e., $E(\beta_{\text{variant}} - \beta_{\text{reference}})^2$), where relevant. We used analysis of variance (ANOVA) to determine the impact of each variation of the parallel and sequential approaches on bias, pooling results from the 5 example studies. ANOVA was conducted for the sequential approach, including a term for the 24 possible sequence permutations of the 4 domains, as well as a term for using the final PS as a continuous variable (vs. including deciles of the PS). ANOVA was conducted for the parallel approach, including a term for the inclusion of age and sex in the calculation of all domain-specific PSs and a term for fitting a last-stage PS model—that is, including a logistic model regressing treatment on the domain-specific PSs, their squared terms, and pairwise interactions. Among variants of the parallel approach using a last-stage PS model, a separate ANOVA was conducted including a term for using the last-stage PS as a continuous variable or in decile indicators. Each ANOVA was repeated once with bias as the outcome variable and once with the absolute value of the bias as the outcome variable. When conducting ANOVA, *P* values of 0.05 or smaller were considered to suggest statistical significance.

Plasmode simulation study

In order to validate our findings in a setting with a known treatment effect, we conducted a plasmode simulation study (24, 25) using the cohort from study 1 as the basis for simulation. To create each simulated cohort, we sampled 10,000 observations with replacement from the unpartitioned study 1 cohort, and their exposure and covariate data were retained. Outcomes were then simulated from a logistic model including the main effects of all of the covariates across all domains (whose coefficients were estimated from a logistic model in the full unpartitioned cohort) and a null term for the main effect of treatment. We simulated 2,000 such cohorts and performed all variations of the parallel and sequential approaches in each. Final treatment-effect estimates for each variation were averaged across the 2,000 simulations.

RESULTS

Overall, all tested variations produced point estimates with the same direction of effect as the reference estimate. Of the 270 variations of the parallel and sequential approaches considered, 203 (75%) produced effect estimates within 5% of the relevant reference estimate, and 246 (91%) of effect estimates were within 10% of the reference estimate. There was substantial heterogeneity in performance across data sets and approaches (Table 2). In 3 of 5 example studies, the MSE was lower among variants of the sequential approach than among variants of the parallel approach.

Figure 3 shows the performance of the sequential method. No substantial order effects were apparent. There was no evidence that average bias varied by domain sequence (ANOVA $F(23, 215) = 0.69$; 2-sided $P = 0.85$), nor was there evidence that variants using the final PS in deciles had a higher or lower average bias than variants using the final PS as a continuous variable (ANOVA $F(1, 215) = 1.91$; 2-sided $P = 0.17$). When absolute bias was examined, there was evidence that average absolute bias varied by domain sequence (ANOVA $F(23, 215) = 1.82$; 2-sided $P = 0.02$). There was also some evidence that variants using the final PS as decile indicators produced a lower average absolute bias than variants using the final PS as a continuous term (ANOVA $F(1, 215) = 3.12$; 2-sided $P = 0.08$) Table 3 shows a ranking of the sequences by average absolute bias. Sequences ending with the demographic domain tended to have the highest average absolute bias, followed by sequences ending with the drug domain, while sequences starting with either of these domains tended to have the lowest average absolute bias.

Among the 6 variations of the parallel approach (Web Figure 1), there was no evidence that variants including age and sex in every domain-specific PS model had a higher or lower average bias than those not doing so (ANOVA $F(1, 27) = 0.99$; 2-sided $P = 0.33$). There was no evidence that variants utilizing a last-stage PS model (i.e., a PS model including all domain-specific PSs, their squares, and all pairwise interactions between them) had a higher or lower average bias than those not doing so (ANOVA $F(1, 27) = 0.09$; 2-sided $P = 0.77$). Among those variants using a last-stage PS, there was no evidence that variants using the PS in deciles had a

Table 2. Estimated Treatment Effects in 5 Example Studies Using Variations of the Parallel and Sequential Approaches to Analysis of Vertically Distributed Data

Study and Method	No. of Variations	Variations Within 5% of Reference Estimate[a]		MSE[b]	Reference Estimates[c]		Estimated OR or HR[d]		
		No.	%		OR	HR	Minimum	Median	Maximum
Schneeweiss, 2009 (18)									
Parallel	6	6	100	0.0013	0.86/0.87		0.83	0.84	0.84
Sequential	48	48	100	0.0004	0.86/0.87		0.83	0.85	0.87
Schneeweiss, 2009 (18)									
Parallel	6	5	83	0.0011	0.88/0.89		0.87	0.90	0.94
Sequential	48	36	75	0.0012	0.88/0.89		0.86	0.89	0.96
Schneeweiss, 2010 (19)									
Parallel	6	5	83	0.0009		0.78/0.80	0.76	0.78	0.80
Sequential	48	35	73	0.0017		0.78/0.80	0.74	0.77	0.80
Patorno, 2010 (20)									
Parallel	6	1	17	0.0227		1.53/1.38	1.48	1.76	1.97
Sequential	48	17	35	0.0069		1.53/1.38	1.52	1.65	1.81
Patorno, 2014 (21)									
Parallel	6	2	23	0.0289		1.12/1.38	1.32	1.36	1.41
Sequential	48	48	100	0.0004		1.12/1.38	1.07	1.23	1.43

Abbreviations: HR, hazard ratio; MSE, mean squared error; OR, odds ratio; PS, propensity score.

[a] Percent change was calculated on the OR/HR scale—for example, as $\left|\frac{(\mathrm{OR}_{\mathrm{variant}} - \mathrm{OR}_{\mathrm{reference}})}{\mathrm{OR}_{\mathrm{reference}}}\right| \times 100\%$.

[b] The MSE was calculated on the log scale (i.e., $E\,[\beta_{\mathrm{variant}} - \beta_{\mathrm{reference}}]^2$).

[c] In each study, there were 2 reference estimates: 1) one generated by an outcome model including a continuous term for the full multivariable PS (estimated using all covariates across all domains) and 2) another generated by an outcome model using indicators for deciles of the full multivariable PS. These were applied in accordance with the treatment of the final PS(s) in each variant of the parallel or sequential approach (continuous or deciles). Depending on the study, the reference estimate was either an HR generated from a Cox proportional hazards model or an OR generated from a logistic regression model.

[d] For studies 1 and 2, these are ORs; for studies 3–5, these are HRs.

higher or lower average bias than those not doing so (ANOVA $F(1, 18) = 0.07$; 2-sided $P = 0.79$). When variants of the parallel approach were compared on the absolute bias scale, there was no evidence that those including age and sex in every domain-specific PS model had a higher or lower average absolute bias than those not doing so (ANOVA $F(1, 27) = 0.71$; 2-sided $P = 0.41$). There was no evidence that variants using a last-stage PS model had a higher or lower average absolute bias than those not doing so (ANOVA $F(1, 27) = 0.31$; 2-sided $P = 0.58$). Among variants using a last-stage PS, there was no evidence that variants using the PS in deciles had a higher or lower average absolute bias than those not doing so (ANOVA $F(1, 18) = 0.25$; 2-sided $P = 0.63$).

Figure 4 shows the results of the plasmode simulation study with treatment effect estimates averaged across 2,000 simulations and error bars displaying the 2.5th and 97.5th percentiles (empirical 95% confidence intervals) of the treatment effect estimate distributions. All 54 variants of the parallel and sequential methods produced treatment effect estimates within 10% of the true simulation parameter on the odds ratio scale. Among 48 variations of the sequential method, all treatment effect estimates were within 5% of the true simulation parameter on the odds ratio scale, as compared with one-sixth (16.7%) among variants of the parallel approach. Across the 2,000 simulated data sets, the MSE of variants of the sequential approach was 0.0109, as compared with 0.0134 among variants of the parallel approach (the MSE of the "ideal" approach utilizing all covariates across all domains in a single PS was 0.0105).

DISCUSSION

There is an imminent need in health-care database analytics for methods that can incorporate vertically distributed data into multivariable epidemiologic analyses; that need is likely to increase as more health-care database research networks are funded and become more elaborate. In this proof-of-concept study, we sought to determine whether existing PS methods can be used when distinct subsets (i.e., data domains) of covariates are available in separate locations but cannot be physically pooled into a single database.

We found that these PS-based approaches generally produce effect estimates that fall within a close margin of what can be achieved in a fully pooled database. This was validated in a plasmode simulation study, in which all variations of the parallel and sequential approach produced estimated

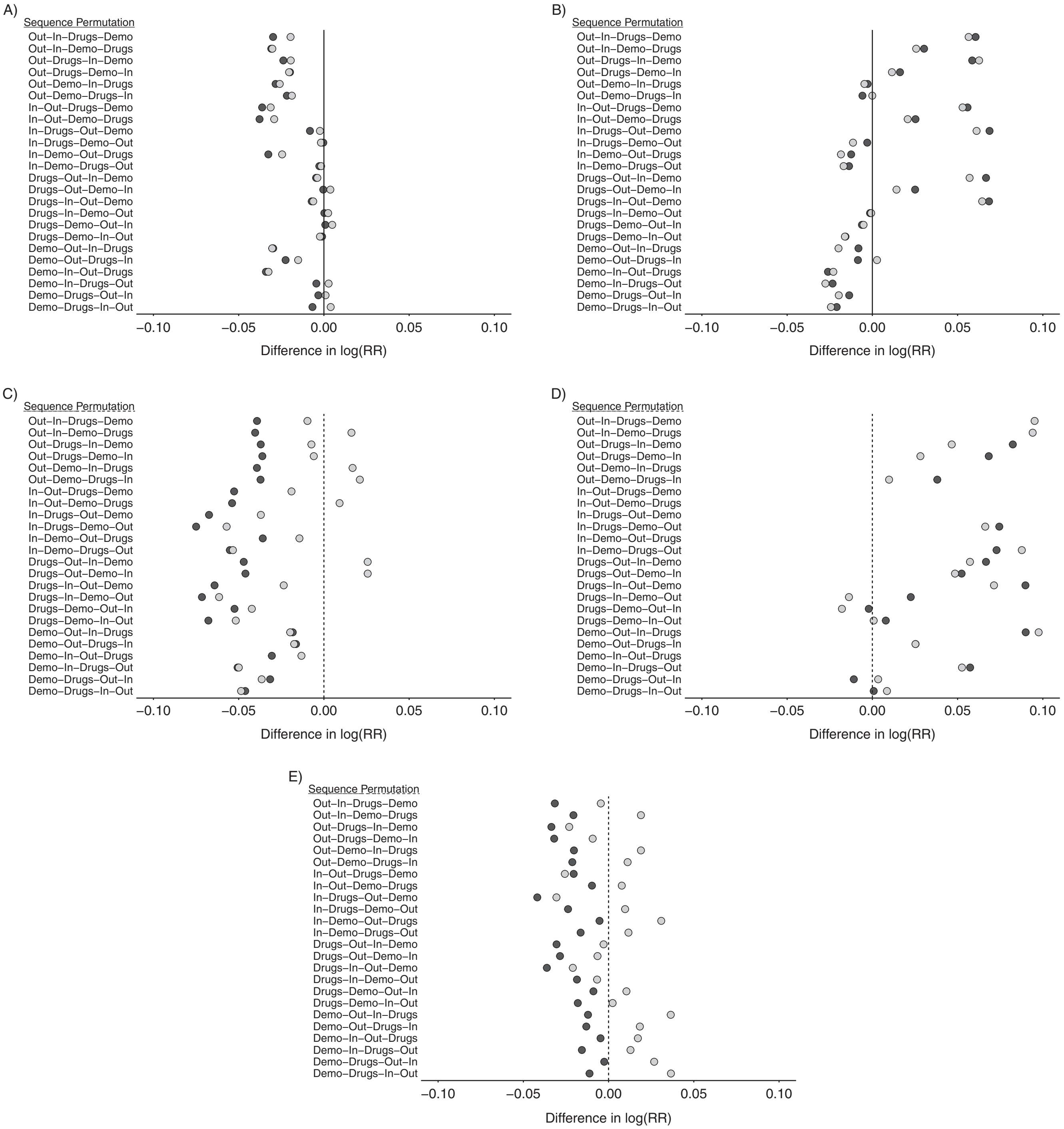


Figure 3. Variations of the sequential approach to analysis of vertically distributed data. The performance of the sequential approach to propensity score (PS) estimation when data are vertically distributed is demonstrated here, showing the influence of domain ordering (vertical axis) and inclusion of a single continuous term for the final PS in the outcome model (dark gray circles) versus decile-indicator (light gray circles) treatment of the final PS in the outcome model. The 4 domains are outpatient ("Out"), inpatient ("In"), demographic factors ("Demo"), and prescriptions ("Drugs"), giving 24 possible orderings. The horizontal axis shows the difference between the log hazard ratio or log odds ratio (both abbreviated as risk ratio (RR)) and its reference estimate. Results are given separately for each of the 5 example studies: Schneeweiss et al., 2009 (18) (analyses of cyclooxygenase 2 inhibitors (A) and statins (B)); Schneeweiss et al., 2010 (19) (C); Patorno et al., 2010 (20) (D); and Patorno et al., 2014 (21) (E).

Table 3. All Possible Orderings of 4 Domains in the Sequential Propensity Score Approach to Analysis of Vertically Distributed Data, Ranked According to Mean Absolute Bias

Rank	Sequence of Domains[a]	Mean Absolute Bias[b]	Mean Bias[b]
1	Demo-Drugs-Out-In	0.0149	−0.0087
2	Drugs-Demo-Out-In	0.0151	−0.0119
3	Demo-Out-Drugs-In	0.0165	−0.0021
4	Drugs-Demo-In-Out	0.0184	−0.0162
5	Out-Demo-Drugs-In	0.0185	−0.0025
6	Drugs-In-Demo-Out	0.0199	−0.0149
7	Demo-Drugs-In-Out	0.0208	−0.0108
8	Out-Drugs-Demo-In	0.0248	0.0001
9	Drugs-Out-Demo-In	0.0251	0.0088
10	Demo-In-Drugs-Out	0.0297	−0.0046
11	In-Drugs-Demo-Out	0.0323	−0.0022
12	In-Demo-Drugs-Out	0.0332	0.0012
13	Demo-Out-In-Drugs	0.0362	0.0086
14	Drugs-Out-In-Demo	0.0362	0.0184
15	Out-Demo-In-Drugs	0.0366	0.0123
16	Out-Drugs-In-Demo	0.0394	0.0106
17	Demo-In-Out-Drugs	0.0404	0.0076
18	In-Demo-Out-Drugs	0.0422	0.0135
19	Out-In-Demo-Drugs	0.0423	0.0179
20	In-Out-Demo-Drugs	0.0444	0.0183
21	Drugs-In-Out-Demo	0.0452	0.0136
22	Out-In-Drugs-Demo	0.0483	0.0215
23	In-Drugs-Out-Demo	0.0567	0.0192
24	In-Out-Drugs-Demo	0.0603	0.0232

[a] The 4 domains were outpatient (“Out”), inpatient (“In”), demographic factors (“Demo”), and prescriptions (“Drugs”), giving 24 possible orderings.

[b] For each domain, the mean absolute bias was calculated as $E(|\beta_{\mathrm{variant}} - \beta_{\mathrm{reference}}|)$, and the mean bias was calculated as $E(\beta_{\mathrm{variant}} - \beta_{\mathrm{reference}})$. Each mean was based on 10 observations: 2 per each of the 5 example studies (one with the outcome model including indicators for deciles of the final PS and one with the outcome model including a linear term for the logit PS). Because the direction of the bias differed between example studies, the absolute scale was used when ranking.

odds ratios within 10% of the true simulation parameter. Thus, it may be possible to perform internally valid epidemiologic investigations when privacy constraints prevent pooling. However, some performance heterogeneity between example studies suggests that these methods should undergo broader evaluation before being widely adopted.

Of the 2 main approaches tested, we observed a benefit (lower MSE in 3 out of 5 studies) for the sequential method, in which a PS is fitted in one domain and then iteratively passed through PS models fitted in the remaining domains, as compared with the parallel approach. Furthermore, in simulation, we observed that all variations of the sequential method outperformed all variations of the parallel method, having estimates closer to the true simulation parameter, and that variants of the sequential approach had a lower MSE than did variants of the parallel approach. The sequential method may better allow the final PS to reflect the joint effect of the covariates on treatment than the parallel method, which shares no information across domains. Furthermore, the sequential method has the advantage of producing a single final PS, which may be used by the analyst in the same ways as a PS estimated in the traditional manner, including matching, stratification, or inverse probability weighting, though at the expense of increased analytical complexity. When we examined bias on the absolute scale, there was some evidence that the domain sequence used in the sequential method did affect performance and that sequences ending with the demographic domain performed the worst; however, the difference in bias between orders was quite small. We also found evidence that, among sequential methods, those including indicators for deciles of the final PS in the outcome model performed better than those including a continuous term for the final (logit) PS alone. While this benefit of deciles was less apparent in the parallel approach, the strong assumptions inherent in adjusting for a continuous PS should caution readers against its use in practice.

While our results provide a proof-of-concept for the analysis of vertically distributed data, they do not account for several factors, including 1) the varying sample size, numbers of exposed patients and outcome events, and degree of confounding in each study; 2) the potential for imbalance in covariate informativeness across domains; 3) the potential for residual confounding within the reference estimates; 4) the uncertainty in the PS estimation procedure when estimating standard errors; and 5) additional ways to use the PS (e.g., matching, standardization, or inverse probability weighting).

Though it was not addressed in this study, missing data will be a critical issue in the application of the proposed approaches to real-world vertically distributed data, as many data sources (e.g., laboratory test-result databases) will not contain records for every patient in a cohort. The presented scenario is thus atypical in this regard, representing a serious limitation of the proposed approaches. Careful work is needed to determine whether it will be possible to overcome the missing-data problem, given data-partitioning constraints. In order to apply established imputation procedures in a vertically distributed setting, investigators would need to assume that data in a given domain are missing at random, conditional only on other covariates in that domain. If this assumption is not met, it is possible that limited, nonidentifying data could be shared across data sources to aid the imputation procedure, but this may be unreasonable to data owners. Extensive research is needed to determine the sensitivity of the proposed approaches to missing data.

Another critical issue in the implementation of the proposed approaches will be the specification of important effect modifiers. In this regard, privacy constraints should allow the sharing of information on a prespecified modifying variable, such that it can be identified from whichever domain it resides in and passed on to the final outcome model for inclusion as a product term or stratification factor. However, investigators must take care to ensure that modifying variables could not identify patients. This may be

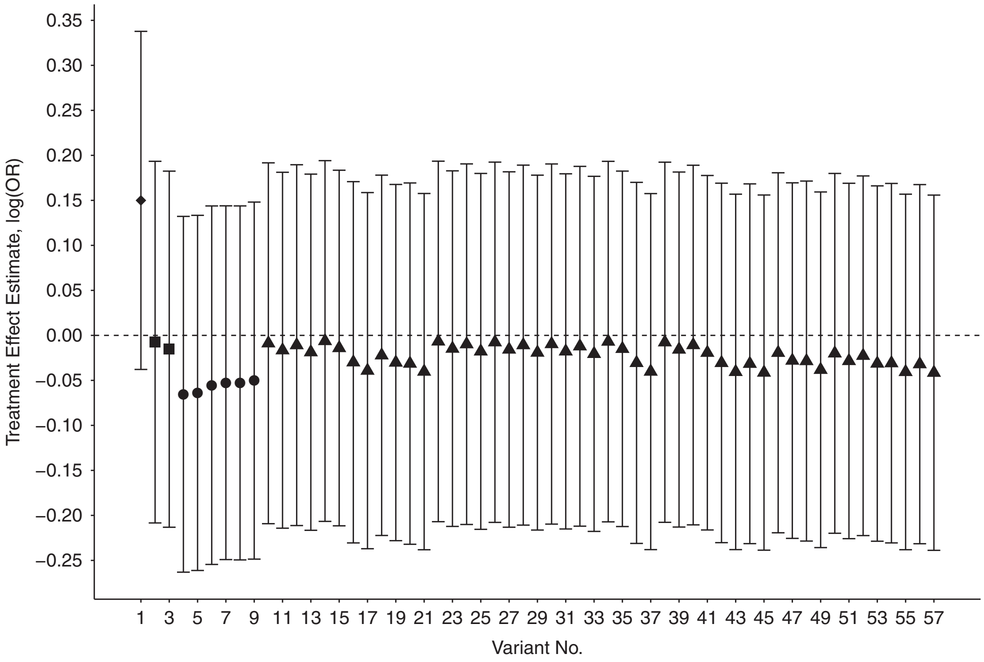


Figure 4. Performance of variations of the parallel and sequential propensity score approaches to analysis of vertically distributed data in simulation. The plotted treatment effect estimates are presented on the log odds ratio (OR) scale and have been averaged across the 2,000 simulations. All simulations were carried out under a true null treatment effect (log OR equal to 0). Error bars indicate the 2.5th and 97.5th percentiles (empirical 95% confidence intervals) of the treatment effect distributions. The horizontal axis shows an index of the variations of the parallel and sequential PS approaches. Symbol shape indicates the type of estimate: diamond, crude/unadjusted; squares, fully adjusted for all covariates across all domains; circles, parallel approach; triangles, sequential approach. Details on these variants can be found in Web Table 1.

particularly limiting for detailed clinical investigations in which researchers wish to examine results within care sites or providers.

Our results from 5 example studies suggest that it may be possible to perform multivariable confounding adjustment when patient covariates are distributed across separate, private domains. However, more extensive research is needed to determine the optimal method for such analysis, especially with regard to variance estimation, the handling of missing data, and applications to matching and weighting. In light of these results, investigators should be cautioned against employing any of the tested approaches in a vertically distributed data setting without consideration of several factors. A crucial consideration is the importance of the covariates contained outside the primary analytical database. The methods presented here are thus advisable when confounding control cannot be reasonably achieved without the inclusion of covariate information residing in a physically separated database. Investigators should also pay close attention to the modeling assumptions inherent in these approaches, since each requires multiple models, and some models depend on the correct specification of prior models. Finally, while the proposed approaches appear to perform adequately in some settings, their use for primary effect estimation at this point may be premature. Where possible, it may be advantageous to first perform analyses in the primary database only, the results of which can then be compared with the results of a parallel or sequential routine that includes additional data sources.

ACKNOWLEDGMENTS

Author affiliations: Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts (Justin Bohn, Wesley Eddings, Sebastian Schneeweiss).

All authors contributed equally to this work.

This work was funded by the Patient-Centered Outcomes Research Institute, with additional support provided by research grants from the National Library of Medicine (grant R01-LM010213) and the National Heart, Lung, and Blood Institute (grant RC4-HL102023). J.B. was partially supported by a student fellowship from the Spanish health insurance company ASISA.

We thank Drs. Kenneth Rothman, Robert Glynn, Joshua Gagne, and Jessica Franklin for their insightful comments.

S.S. is a consultant to World Health Information Science Consultants, LLC (Newton, Massachusetts) and to Aetion, Inc. (New York, New York), a software manufacturer in which he also owns shares. He is a Principal Investigator for investigator-initiated grants to the Brigham and Women's

Hospital from Novartis, Genentech, and Boehringer Ingelheim unrelated to the topic of this study.

REFERENCES

1. Platt R, Carnahan RM, Brown JS, et al. The US Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf*. 2012;21(suppl 1):1–8.
2. Califf RM. The Patient-Centered Outcomes Research Network: a national infrastructure for comparative effectiveness research. *NC Med J*. 2014;75(3):204–210.
3. Oliveira JL, Lopes P, Nunes T, et al. The EU-ADR Web Platform: delivering advanced pharmacovigilance tools. *Pharmacoepidemiol Drug Saf*. 2013;22(5):459–467.
4. Trifirò G, Coloma PM, Rijnbeek PR, et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? *J Intern Med*. 2014;275(6):551–561.
5. Curtis LH, Weiner MG, Boudreau DM, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf*. 2012; 21(suppl 1):23–31.
6. Selby JV, Krumholz HM, Kuntz RE, et al. Network news: powering clinical research. *Sci Transl Med*. 2013;5(182): 182fs13.
7. Toh S, Gagne JJ, Rassen JA, et al. Confounding adjustment in comparative effectiveness research conducted within distributed research networks. *Med Care*. 2013;51(8 suppl 3):S4–S10.
8. Toh S, Reichman ME, Houstoun M, et al. Comparative risk for angioedema associated with the use of drugs that target the renin-angiotensin-aldosterone system. *Arch Intern Med*. 2012; 172(20):1582–1589.
9. Rassen JA, Avorn J, Schneeweiss S. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. *Pharmacoepidemiol Drug Saf*. 2010;19(8):848–857.
10. Rassen JA, Solomon DH, Curtis JR, et al. Privacy-maintaining propensity score-based pooling of multiple databases applied to a study of biologics. *Med Care*. 2010;48(6 suppl):S83–S89.
11. Toh S, Reichman ME, Houstoun M, et al. Multivariable confounding adjustment in distributed data networks without sharing of patient-level data: confounding adjustment in distributed data networks. *Pharmacoepidemiol Drug Saf*. 2013;22(11):1171–1177.
12. Karr AF, Lin X, Sanil AP, et al. Secure regression on distributed databases. *J Comput Graph Stat*. 2005;14(2): 263–279.
13. Fienberg SE, Nardi Y, Slavković AB. Valid statistical analysis for logistic regression with multiple sources. In: Gal CS, Kantor PB, Lesk ME, eds. *Protecting Persons While Protecting the People*. Berlin, Germany: Springer-Verlag; 2009:82–94.
14. Lin X, Karr AF. Privacy-preserving maximum likelihood estimation for distributed data. *J Priv Confidentiality*. 2010; 1(2):213–222.
15. Karr AF, Fulp WJ, Vera F, et al. Secure, privacy-preserving analysis of distributed databases. *Technometrics*. 2007;49(3): 335–345.
16. Karr AF. Secure statistical analysis of distributed databases, emphasizing what we don't know. *J Priv Confidentiality*. 2010;1(2):197–211.
17. Schneeweiss S, Rassen JA, Glynn RJ, et al. Supplementing claims data with outpatient laboratory test results to improve confounding adjustment in effectiveness studies of lipid-lowering treatments. *BMC Med Res Methodol*. 2012; 12:180.
18. Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20(4): 512–522.
19. Schneeweiss S, Patrick AR, Solomon DH, et al. Comparative safety of antidepressant agents for children and adolescents regarding suicidal acts. *Pediatrics*. 2010;125(5):876–888.
20. Patorno E, Bohn RL, Wahl PM, et al. Anticonvulsant medications and the risk of suicide, attempted suicide, or violent death. *JAMA*. 2010;303(14):1401–1409.
21. Patorno E, Glynn RJ, Hernández-Díaz S, et al. Studies with many covariates and few outcomes: selecting covariates and implementing propensity-score-based confounding adjustments. *Epidemiology*. 2014;25(2):268–278.
22. Rassen JA, Glynn RJ, Rothman KJ, et al. Applying propensity scores estimated in a full cohort to adjust for confounding in subgroup analyses. *Pharmacoepidemiol Drug Saf*. 2012;21(7): 697–709.
23. Sturmer T. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *Am J Epidemiol*. 2005;161(9):891–898.
24. Franklin JM, Schneeweiss S, Polinski JM, et al. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*. 2014;72:219–226.
25. Franklin JM, Eddings W, Glynn RJ, et al. Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. *Am J Epidemiol*. 2015;182(7):651–659.