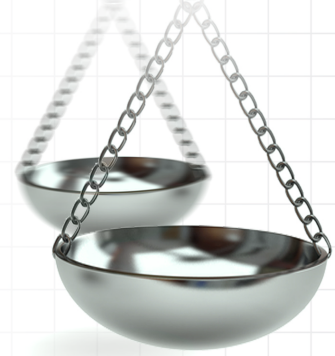
Research Article



Journal of **Comparative Effectiveness Research**

Comparing high-dimensional confounder control methods for rapid cohort studies from electronic health records

Yen Sia Low*[,1], Blanca Gallego[2] & Nigam Haresh Shah[1]

[1]Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA 94305, USA

[2]Center for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia

*Author for correspondence: Tel.: +1 919 428 7413; yenlow@gmail.com

Aims: Electronic health records (EHR), containing rich clinical histories of large patient populations, can provide evidence for clinical decisions when evidence from trials and literature is absent. To enable such observational studies from EHR in real time, particularly in emergencies, rapid confounder control methods that can handle numerous variables and adjust for biases are imperative. This study compares the performance of 18 automatic confounder control methods. **Methods:** Methods include propensity scores, direct adjustment by machine learning, similarity matching and resampling in two simulated and one real-world EHR datasets. **Results & conclusion:** Direct adjustment by lasso regression and ensemble models involving multiple resamples have performance comparable to expert-based propensity scores and thus, may help provide real-time EHR-based evidence for timely clinical decisions.

First draft submitted: 9 September 2015; Accepted for publication: 8 October 2015; Published online: 4 December 2015

Keywords: bias • clinical decision support • cohort studies • confounding • electronic health records • machine learning • propensity scores

Background

To realize the promise of electronic health records (EHR) for enabling learning health systems, there is increasing demand for evidence generated from EHR in real-time for clinical decision support at the point of care [1–4]. The scale and depth of EHR data, representing large patient populations and their rich clinical histories, offer new opportunities to learn practice patterns from real-world patients, signaling trends which may be impossible to detect in clinical trials, or providing guidance when randomization is not possible.

Indeed, there have been several precedents for clinicians using the EHR to learn from past patient records. Feinstein *et al.* created an electronic 'library of clinical experience' consisting of 678 highly similar lung cancer patients from whom to obtain personalized prognosis [5]. Frankovich *et al.* [6], in the absence of evidence from existing literature, analyzed the EHR of patients similar to their pediatric lupus patient. Within a few hours, they estimated the increased risk of thrombosis and promptly decided on prophylactic anticoagulant therapy. However, extracting reliable and valid evidence from EHR through observational studies remains a specialized endeavor [4], requiring careful design and controls. To enable such use of EHR data especially for clinically urgent decisions at the bedside, we have proposed a 'green button' [3] (analogous to 'info buttons' [7]) that will generate practice-based evidence from EHR in real-time by automatically selecting relevant patients, aggregating their characteristics and evaluating their outcomes [3,8,9].

There have been encouraging efforts enabling such a button. The first step is cohort selection, which can be performed by electronic phenotyping approaches [10–13]. These methods can automatically select patients with desired characteristics with high accuracy, obviating the need for laborious manual

Future Medicine part of fsg

chart review, the current gold standard for identifying patients. However, the next step, confounder adjustment – required for validity of observational studies in which biases cannot be randomized away [14] – is difficult to automate because it relies on expert knowledge to select appropriate confounder variables. Typical approaches such as direct adjustment (DA) by multivariate regression following Cepada's rule of 8 [15] (i.e., at least eight events per variable in model) or propensity scores (PS) [16] both require extensive expert consultation.

Several heuristics have been developed for the automatic handling of confounders [17–23]. To date, automated methods include: first, filters to select confounders by predefined criteria (e.g., prevalence) for modeling PS [17]; second, machine-learning (ML) [19–21,23,24] algorithms with automatic variable selection for modeling PS and third, matching by nearest neighbors based on multivariate similarity [18]. These methods, while comparable [19–21,25,26] or better [17,20,22,24,25,27] than expert-based PS, have mostly been confined to the PS framework. However, PS whether generated by experts or automated means are not foolproof. PS has been shown to approximate random matching at times and its utility as the default mechanism for confounder control has been questioned [28]. Hence, there is a need to explore methods other than PS such as DA by machine learning or repeated matching resembling ensemble models.

To this end, using three datasets, this study will compare first, PS-based approaches; second, DA by ML; third, matching by patient similarity and fourth, matching by multiple resamples akin to ensemble modeling (Table 1). This comprehensive comparison includes high-dimensional non-PS methods such as DA by machine learning for the purpose of assessing various automatic confounder control methods for quick and accurate cohort studies to facilitate timely clinical decisions.

Methods

Datasets

To ensure comparability with other benchmarking studies [20,21,29], we included a widely used simulated dataset [19], which mimics the presence of various types of confounding and exposure-outcome associations. However, since it contains only ten variables, it does not test the strengths of the high-dimensional methods. Therefore, we included another simulated dataset with 100 variables. A real-world clinical dataset, representative of the increased dimensionality and complexity of EHR data including clinical text, was also included.

Our synthetic datasets contain a combination of pre-exposure variables (expected to contribute to the estimation of a propensity score) together with instrumental variables (IV) and colliders. An IV is a pre-exposure variable affecting only exposure but not independently affecting the outcome. However, an IV could become related to the outcome via unobserved or residual confounding and thereby introduce bias. A collider is a variable affected by two independent variables such that the pathways originating from the two independent variables collide at the collider variable. Controlling for a collider could open a path between the exposure and outcome and create a spurious association [30]. Hence, both IV and colliders should be excluded from PS formulation [29–33].

Simulated dataset 1: 2000 patients × 10 variables

A small dataset (2000 patients by 10 variables x_1–x_{10},) was simulated as described in Setoguchi *et al.* (Scenario E) [19]. Supplementary Data A describes the variables in this dataset: binary variables (x_1, x_3, x_5, x_6, x_8 and x_9) and continuous variables (x_2, x_4, x_7 and x_{10}) of which x_7 was an IV. Weak (r = 0.2) and strong (r = 0.9) correlations were introduced among some variables (Supplementary Data A Figure 1). From the x_i variables, we generated corresponding exposure (E = 0 or 1) and outcome (Y = 0 or 1) using logistic models (Supplementary Data A Equations 2 & 3, respectively) such that the beta coefficient of exposure β was set at -0.4 (i.e., odds ratio [OR] = 0.67) as in Setoguchi *et al.* (Scenario E) [19]. This was repeated until 1000 datasets were generated. To avoid situations in which there would be no cases after matching, we ensured that there were at least 40 positive outcome events as in Setoguchi *et al.* [19].

Simulated dataset 2: 2000 patients × 100 variables

An extended dataset with additional variables, including colliders and noise was simulated. Supplementary Data A Figure 2 describes how the variables in this dataset are related and generated. Among the additional variables, binary variables were x_{11}, x_{13}, x_{16}, x_{17}, x_{51}–x_{100}, while continuous variables were x_{12}, x_{14}, x_{15}, x_{18}–x_{50}. Colliders were x_{19}–x_{22}. Continuous noise variables x_{23}–x_{50} were randomly generated from a uniform distribution between 0 and 1, while binary noise variables were randomly generated from a Bernoulli distribution (i.e., equivalent to a coin toss). Then, exposure and outcome statuses were generated by using logistic regression models (Supplementary Data A Equations 4 & 5, respectively). Beta coefficient of exposure β was set to +0.4 (i.e., OR = 1.5). As before, we generated 1000 such datasets, each with at least 80 positive outcome events.

Dataset 3: 5757 patients × 447 variables

This dataset consists of a real-world cohort of patients with peripheral arterial disease inclusive of 232 cilostazol-users and 5525 nonusers. Their 20 cardiovascular out-

Table 1. Description of the 18 confounder control methods being compared.

Confounder control method	Variable selection method	Matching method	Evaluation method
Baseline			
Random matching	–	Random	Logistic regression
ExpertPS_match	Expert selection	By closest PS	Conditional logistic regression
ExpertPS_adjust	Expert selection	–	Logistic regression
PS methods			
hdPS_match	Above minimum prevalence and ranked by univariate association with outcome	By closest PS	Conditional logistic regression
hdPS_adjust	As for hdPS	–	Logistic regression
lassoPS_match	Lasso regularization	By closest PS	Conditional logistic regression
lassoPS_adjust	Lasso regularization	–	Logistic regression
rfPS_match	Multiple random subspaces	By closest PS	Conditional logistic regression
rfPS_adjust	Multiple random subspaces	–	Logistic regression
lassoMV	Lasso regularization	–	Lasso logistic regression
Euclidean	–	By closest distance	Conditional logistic regression
Jaccard	–	By closest distance	Conditional logistic regression
Dice	–	By closest distance	Conditional logistic regression
Cosine	–	By closest distance	Conditional logistic regression
Pearson	–	By closest distance	Conditional logistic regression
Spearman	–	By closest distance	Conditional logistic regression
Ensemble			
Bootstrap	–	Random	Logistic regression
Jackknife	–	Random	Logistic regression

comes after adjustment using expert selection (expertPS) matching have been published [34]. We reproduced the results from expertPS and compared them against other confounder control methods. Variables included demographic factors and concepts identified from clinical text using a validated text-processing pipeline [35,36] based on the Unified Medical Language System [37] biomedical controlled vocabularies. We distinguished concepts before and after index time (i.e., first mention of cilostazol). Rare pre-exposure concepts present in less than 10% of the cohort were eliminated, resulting in a total of 447 baseline variables. These concepts were numerically represented as binary variables where presence of the concept is denoted as 1 and 0 otherwise. Because age and gender were known confounders, they were forced into the PS or outcomes models where possible (i.e., hdPS, lassoPS, lassoMV, see the 'Algorithms' section).

Algorithms

We compared a total of 18 algorithms which can be grouped into: first, PS methods; second, DA method; third, similarity matching methods and fourth, ensemble resampling methods (Table 1). PS methods require that PS be computed prior to fitting an outcome model. In contrast, DA methods fit an outcome model taking into account the large number of variables using high-dimensional ML approaches. Similarity matching methods require that interpatient similarities be computed for selecting matched controls prior to fitting the outcome model. Lastly, controls could also be matched randomly multiple times by ensemble resampling approaches.

PS methods

The four PS methods varied in the way confounder variables were selected: expertPS, filtering by some criteria (high-dimensional PS, hdPS [17]) and automatic variable selection built into the ML algorithm (e.g., lasso [38] regression, random forest [RF] [39]). Baseline variables for the calculation of the PS should be selected depending on their position in the causal pathway, which is rarely fully known in practice. Therefore, expert knowledge for formulating a causal model such that IV and colliders are appropriately identified and excluded can be critical.

In expertPS [16], variables were selected based on prior knowledge. In the simulated datasets, the PS

were generated from known relationships, in other words, known confounders, colliders and IV (see Supplementary Data A Equation 2 for dataset 1 and Supplementary Data A Equation 4 for dataset 2). However, in line with best practices for formulating expertPS, IV and colliders, where known, were excluded [30,32]. In the real-world dataset, expert-based PS came from a previous analysis [34]. PS was generated from a logistic regression model of the selected variables. Note that expertPS, being a reference for comparison, is grouped under the baseline reference group.

In hdPS (Pharmacoepidemiology Toolbox version 2.15), baseline variables above a minimum prevalence (5% patients) and ranked by univariate association with the outcome were selected and then fitted to a logistic model for PS estimation [17]. Because hdPS was designed for categorical variables, only categorical variables were considered for automatic selection. To handle the continuous variables, we adopted an inclusive approach to include them in the PS model as long as they correlated with the outcome even if marginally ($|r| > 0.05$) [33]. This helped to exclude possible IV that were correlated with only exposure. Automatically selected variables and continuous variables selected by the above correlation filter were then entered into a logistic regression model calculating PS.

In addition, we computed PS using ML algorithms with automatic variable selection, namely lasso logistic regression [38] and RF [39] (R packages glmnet [40] and randomForest [41], respectively). Lasso logistic regression is a penalized form of logistic regression where a penalty factor shrinks low-weight variables toward zero such that these variables are essentially eliminated from the model, providing built-in variable selection [38]. The penalty factor was tuned by fivefold cross-validation using the 'one-standard-error' rule [40].

Here, RF was an ensemble model that aggregated predictions from 100 decision trees, each of which predicted a PS [20,42] from a bootstrap sample [39] of randomly selected variables [43]. Variables highly contributory to the RF model have highly positive importance scores [39]. We used the default parameters in the randomForest package [41] except that the number of trees was set to 100 and the minimum node size was set to 5% of the sample size [44].

After estimating the PS, the second stage of the analysis used the PS either by covariate adjustment or matching. In covariate adjustment, the PS was included as a covariate in the final logistic regression relating exposure to outcome. In matching, each subject in the minor class (i.e., the smaller of the treated or untreated classes) was matched to one in the major class (i.e., the larger of the two treatment classes) with the most similar PS value within a caliper threshold. Unmatched subjects whose nearest neighbor exceeded the caliper were discarded. We used 1:1 greedy matching without replacement, transformed PS to its logit form and set the caliper to 0.2 standard deviation of logit(PS) as recommended [45,46]. Finally, the matched samples, no longer independent observations after matching, were analyzed by conditional logistic regression.

DA method

Instead of calculating a PS for confounder adjustment, (e.g., lassoPS), lasso multivariate logistic regression [38] (lassoMV) captures the relationship between outcome and exposure while directly adjusting for many variables automatically adjusted by shrinkage, by passing the need to calculate a PS. Its 95% CI is generated by bootstrapping [47] 100 times and taking the 2.5th and 97.5th percentiles as CI limits and the mean as the estimate value.

Similarity matching methods

Similarity methods match patients by closeness as defined by a distance function. In this study, patient similarity was determined by six widely used distance or similarity indices: Euclidean distance, Jaccard, Dice, cosine similarities, Pearson and Spearman rank correlations. To avoid distant matches, we set the caliper for a minimum of 0.1 similarity (or mean distance + 3 standard deviations if Euclidean distance). The matched pairs, no longer independent observations after matching, were then analyzed by conditional logistic regression.

Ensemble resampling methods

Instead of a single model, multiple logistic models from multiple resamples were pooled such that the effect size is estimated from the average of the multiple beta coefficients. We performed multiple 1:1 resampling with replacement (bootstrap) or without replacement (jackknife) for 100 times. Each sample was analyzed by a logistic regression model creating an ensemble of 100 models.

Baseline references

Additionally, we provided several baseline references for comparison: first, expertPS (see 'PS methods' section, above) and second, random matching (without replacement) where each subject in the minor class was randomly matched 1:1 to a subject in the major class and then analyzed by logistic regression.

Assessment metrics

All methods were assessed on: difference in baseline variables before and after matching (standardized mean difference (SMD) [48] and p-value), bias

(i.e., estimated OR – true OR), standard error (SE) of estimated effect β and computing time (Table 1). Additionally, PS methods were qualitatively assessed by the overlap of their PS distributions before and after matching; methods with automatic variable selection (hdPS, lassoPS, rfPS and lassoMV) were assessed for correct selection of baseline variables.

Results

We present results on the efficacy of 18 methods of automated confounder control including: first, propensity score methods; second, direct adjustment; third, patient similarity methods and fourth, ensemble resampling on two simulated and one real-world EHR dataset.

Simulated datasets 1 & 2

Performance summary

Tables 2 & 3 show the performance of the 18 methods on the two simulated datasets. The means and standard deviations (in parentheses) from 1000 simulations are reported. Random matching produced a small bias (0.09–0.12) compared to the crude bias of 0.41 to 0.44 without any adjustment. ExpertPS in which PS was used as a covariate for adjustment had the least bias (0.07), while expertPS used for matching instead of covariate adjustment had a relatively large bias (0.25–0.26). ExpertPS's relatively poor performance was due to a poor linear fit in modeling PS when the generative model contained quadratic terms and two-way interactions (Supplementary Data A Equations 2 & 4). The reported times for expertPS refer to the computing times of the logistic models and did not include time for expert consultation, which was not necessary as the causal structure was known for the simulated datasets.

All PS methods had similar performance, although hdPS stood for its low bias (0.09–0.13), low SE (0.20–0.32), large sample size (1626–1748) and reasonably fast computing time (1.1–5.1 s). When PS was used for covariate adjustment instead of matching, all the PS methods (expertPS, hdPS, lassoPS, rfPS) consistently resulted in smaller biases.

The performance of similarity methods did not vary much (with biases between 0.11 and 0.21), although Jaccard, Dice and cosine similarities were slow (8.4–12.4 s) as they were implemented off-the-shelf and had not been optimized for speed.

Ensemble resampling methods had small biases (0.09–0.11) but were the slowest due to resampling 100 times. Bootstrapping suffered from a larger loss of subjects (1553–1643), which is a known drawback of sampling with replacement [49].

Overall, top performers were hdPS, lassoMV DA and ensemble resampling which were consistently closest to the true OR (bias = 0.09–0.12) with the narrowest CI (SE = 0.20–0.34) and the least loss of subjects (n = 1553–1998). All methods generated 95% CI that always contained the true OR.

Balance of baseline variables

Baseline variables were considered balanced between the exposed and matched controls groups, if they had high p-values (i.e., low $-\log_{10}$ p-values) and low absolute SMD (Figure 1A, B, D & E). As expected, PS methods balanced the baseline variables well unlike the other methods. An exception was hdPS whose p-values and SMD indicated that the baseline variables were almost similar before and after matching. Note that IV x_7 was left out by expertPS and hdPS by design; hence, x_7 was not balanced after matching.

Balance of baseline variables between exposed and matched controls in simulated datasets 1 and 2 for dataset 2.

Variables selection by automated methods

Recall that in the theoretical models for dataset 1, variables x_1–x_7 (Supplementary Data A Equation 2), and in dataset 2, variables x_1–x_7, x_{11}, x_{12}, x_{13} and x_{14} (Supplementary Data A Equation 4) were used to calculate the PS. Of the PS methods with automatic variable selection (i.e., hdPS, lassoPS and rfPS), lassoPS and rfPS selected the correct variables most of the times (Figure 1C & F). Although lassoPS correctly left out the binary noise variables x_{51}–x_{100}, it occasionally picked up continuous noise variables x_{23}–x_{50}. rfPS often selected additional variables (e.g., x_9, x_{19}–x_{22}), especially those strongly correlated with important variables – a known artifact of RF [50].

hdPS was the least selective, frequently selecting noise variables x_{51}–x_{100} (Figure 1F). Because hdPS could automatically select only categorical variables, continuous variables were separately handled by a correlation filter that we introduced ($|r| > 0.05$, see 'Methods' section). Our correlation filter had correctly excluded the continuous noise variables x_{23}–x_{50}, IV x_7 and colliders x_{19}–x_{22} (Figure 1C & F).

lassoMV was the most selective, selecting fewer variables than expected but often correctly excluding noise variables x_{23}–x_{100} (Figure 1C & F).

Comparison of PS distributions

After matching, the PS distributions of controls compared well with those of the exposed group (Supplementary Data B).

Dataset 3: cilostazol users versus nonusers

We applied all 19 methods to a real-world dataset where cilostazol users were followed for increased odds of developing 20 major adverse cardiovascular (MACE)

and major adverse limb events (MALE) compared with nonusers [34]. The PS methods except hdPS balanced the baseline variables relatively well (Figure 2A & B). All PS methods produced comparable PS distributions after matching (Figure 2C).

Balance of baseline variables between exposed and matched controls in dataset 3.

For the 20 outcomes followed, the OR and their 95% CI (Supplementary Data D) are available as individual forest plots (Supplementary Data C) as well as summarized into a bubble plot (Figure 3), where each bubble is colored by their estimated effect size (i.e., $\beta = \ln[OR]$) and sized by their CI. Pronounced outcomes with large effect sizes and narrow CI (e.g., MALE and revascularization, shown as intensely colored small bubbles), were less affected by the choice of method, retaining the same color and size. In contrast, ambiguous outcomes with small effect sizes and wide CI (e.g., sudden cardiac death, ventricular fibrillation, shown as faintly colored large bubbles), had different effects (i.e., colors) depending on the method used. Compared with the results obtained by conventional expertPS used in the previous study [34] (first row) or unadjusted OR (second row), hdPS, Pearson, lassoMV and ensemble sampling (bootstrap and jackknife) produced similar results (i.e., bubbles were colored and sized similarly). lassoPS and rfPS had the worst performance with highly negative beta coefficients when positive values were expected (Figure 3; Supplementary Data C & D). One possible explanation may be the misestimated PS used for covariate adjustment or matching.

Discussion

In this study, we demonstrated that several high-dimensional methods provide comparable alternatives to the current standard, expertPS, for confounder adjust-

Table 2. Performance in means (standard deviations) of the 18 confounder control methods in simulated dataset 1.

Confounder control method	Estimated OR	Standard error	Bias	Sample size used	Computing time
Baseline	True OR ≈ 0.70				
Random matching	0.74 (0.23)	0.31 (0.02)	0.12 (0.09)	1863 (44)	0.2 (0.05)
ExpertPS_match	0.65 (0.30)	0.45 (0.10)	0.26 (0.23)	1333 (48)	0.9 (0.2)†
ExpertPS_adjust	0.67 (0.22)	0.33 (0.02)	0.07 (0.05)	2000 (0)	0.9 (0.2)†
PS methods					
hdPS_match	0.70 (0.24)	0.32 (0.03)	0.13 (0.10)	1748 (104)	1.1 (0.4)
hdPS_adjust	0.69 (0.21)	0.30 (0.02)	0.09 (0.07)	2000 (0)	1.1 (0.4)
lassoPS_match	0.67 (0.26)	0.39 (0.04)	0.20 (0.16)	1199 (47)	1.0 (0.3)
lassoPS_adjust	0.66 (0.22)	0.34 (0.02)	0.12 (0.09)	2000 (0)	1.0 (0.3)
rfPS_match	0.68 (0.26)	0.39 (0.04)	0.19 (0.15)	1237 (44)	0.7 (0.3)
rfPS_adjust	0.66 (0.22)	0.34 (0.02)	0.11 (0.08)	2000 (0)	0.7 (0.3)
Direct adjustment					
lassoMV	0.72 (0.23)	0.30 (0.04)	0.10 (0.08)	2000 (0)	1.1 (0.3)
Similarity methods					
Euclidean	0.70 (0.26)	0.37 (0.04)	0.17 (0.14)	1336 (31)	0.2 (0.07)
Jaccard	0.66 (0.22)	0.34 (0.02)	0.11 (0.08)	1361 (32)	10.4 (1.3)
Dice	0.75 (0.29)	0.37 (0.04)	0.20 (0.15)	1361 (32)	8.4 (0.9)
Cosine	0.74 (0.28)	0.37 (0.04)	0.19 (0.15)	1359 (32)	8.5 (0.8)
Pearson	0.75 (0.28)	0.37 (0.03)	0.20 (0.15)	1340 (40)	1.2 (0.2)
Spearman	0.76 (0.30)	0.37 (0.04)	0.21 (0.17)	1316 (40)	1.3 (0.2)
Ensemble					
Bootstrap	0.73 (0.23)	0.31 (0.02)	0.11 (0.08)	1863 (44)	18.2 (1.37)
Jackknife	0.75 (0.23)	0.34 (0.02)	0.11 (0.08)	1553 (26)	16.1 (1.6)

†The reported times for expertPS refer to the computing times of the logistic models and did not include time for expert consultation, which was not necessary as the causal structure was known for the simulated datasets.
OR: Odds ratio.

Table 3. Performance in means (standard deviations) of the 18 confounder control methods in simulated dataset 2.

Confounder control method	Estimated OR	Standard error	Bias	Sample size used	Computing time
Baseline	True OR ≈ 1.63				
Random matching	1.55 (0.33)	0.20 (0.01)	0.09 (0.07)	1998 (7)	0.2 (0.07)
ExpertPS_match	1.76 (2.01)	0.38 (0.11)	0.25 (0.25)	1231 (47)	1.7 (0.4)†
ExpertPS_adjust	1.48 (0.36)	0.23 (0.01)	0.07 (0.05)	2000 (0)	1.7 (0.4)†
PS methods					
hdPS_match	1.39 (0.32)	0.22 (0.01)	0.13 (0.10)	1626 (108)	5.1 (1.0)
hdPS_adjust	1.38 (0.28)	0.20 (0.01)	0.11 (0.08)	2000 (0)	5.1 (1.0)
lassoPS_match	1.41 (0.40)	0.27 (0.02)	0.16 (0.12)	1117 (46)	2.2 (0.5)
lassoPS_adjust	1.39 (0.33)	0.23 (0.01)	0.12 (0.09)	2000 (0)	2.2 (0.5)
rfPS_match	1.39 (0.36)	0.25 (0.02)	0.15 (0.12)	1256 (47	2.4 (0.5)
rfPS_adjust	1.38 (0.31)	0.22 (0.01)	0.12 (0.08)	2000 (0)	2.4 (0.5)
Direct adjustment					
lassoMV	1.54 (0.33)	0.19 (0.02)	0.08 (0.06)	2000 (0)	5.6 (0.9)
Similarity methods					
Euclidean	1.47 (0.39)	0.25 (0.02)	0.15 (0.11)	1328 (22)	0.4 (0.09)
Jaccard	1.56 (0.36)	0.22 (0.01)	0.12 (0.09)	1604 (18)	12.4 (2.2)
Dice	1.56 (0.37)	0.22 (0.01)	0.12 (0.09)	1604 (17)	10.2 (1.7)
Cosine	1.55 (0.39)	0.22 (0.01)	0.12 (0.09)	1606 (18)	10.3 (1.8)
Pearson	1.56 (0.38)	0.23 (0.01)	0.12 (0.09)	1582 (20)	1.6 (0.3)
Spearman	1.56 (0.39)	0.23 (0.01)	0.12 (0.09)	1555 (21)	1.9 (0.3)
Ensemble					
Bootstrap	1.57 (0.34)	0.23 (0.01)	0.09 (0.07)	1643 (10)	19.9 (2.6)
Jackknife	1.55 (0.33)	0.20 (0.01)	0.09 (0.07)	1998 (7)	22.9 (2.8)

†The reported times for expertPS refer to the computing times of the logistic models and did not include time for expert consultation, which was not necessary as the causal structure was known for the simulated datasets.
OR: Odds ratio.

ment. In particular, lassoMV DA and ensemble models based on multiple random resampling can adjust for large number of confounders without the use of PS and in some instances, even outperform expertPS, which sometimes assumes an overly simplistic linear model.

lassoPS estimated highly negative effect sizes, underscoring a known PS limitation that a misspecified PS may introduce bias particularly when IV and colliders are present [21,32,51,52]. Moreover, a misspecified PS model can also reduce sample size due to poor matching and further reduce efficiency [53]. Standard errors from PS matching also tended to be larger than those from PS adjustment, possibly due to the reduced sample size. If use of PS-based methods is desired, then fully automating PS calculation without expert involvement remains a challenge. There have been recent developments that generate covariate-balancing PS [21,54] automatically. Another solution may be an interactive interface that allows expert assessment and input of baseline variables for real-time sensitivity analysis and iterative corrections.

Similarity methods provided middling performance. One key drawback of using similarity is its poor performance in high-dimensional space. As the number of dimensions increases, subjects become increasingly equidistant and thus, similarity methods become increasingly indiscriminant at selecting closest neighbors [55]. Variable selection and caliper tuning may optimize the performance of similarity methods. In using unweighted similarity metrics, a downside is each variable having equal contribution to similarity instead of having more important variables weigh more in favor of less important ones. More sophisticated distance metrics including weighted

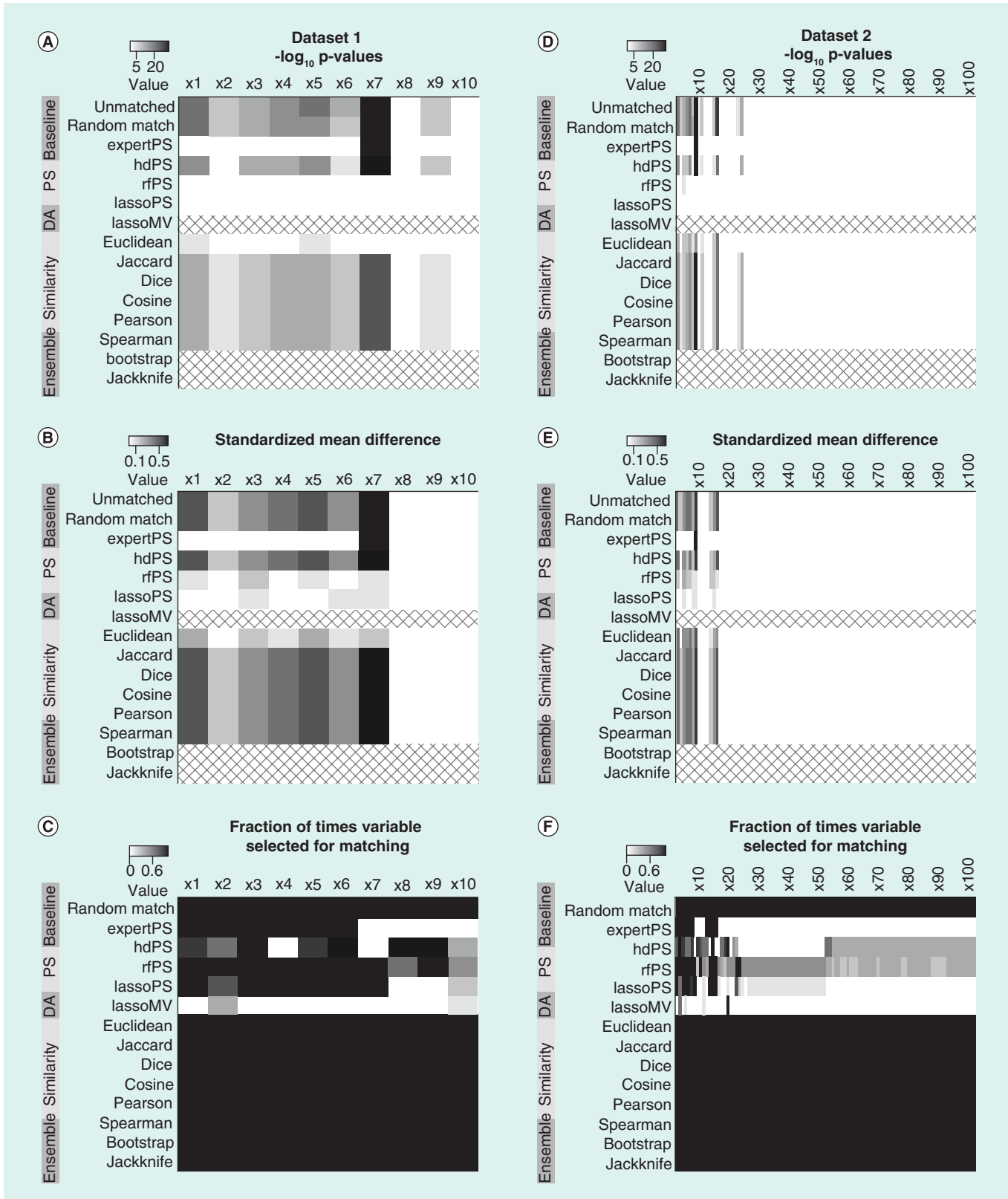


Figure 1. Balance of baseline variables between exposed and matched controls in simulated datasets 1 and 2. Heatmaps showing the balance of baseline variables between exposed and matched controls groups in terms of **(A)** -log $_{10}$ p-values and **(B)** standardized mean difference in dataset 1. Lighter cells indicate smaller values while darker cells indicate bigger values. **(C)** Heatmap showing fraction of times the variables were considered for confounder control method in dataset 1. Darker cells show variables that were selected more frequently. Heatmaps **(D–F)** show the respective equivalent of heatmaps **(A–C)**

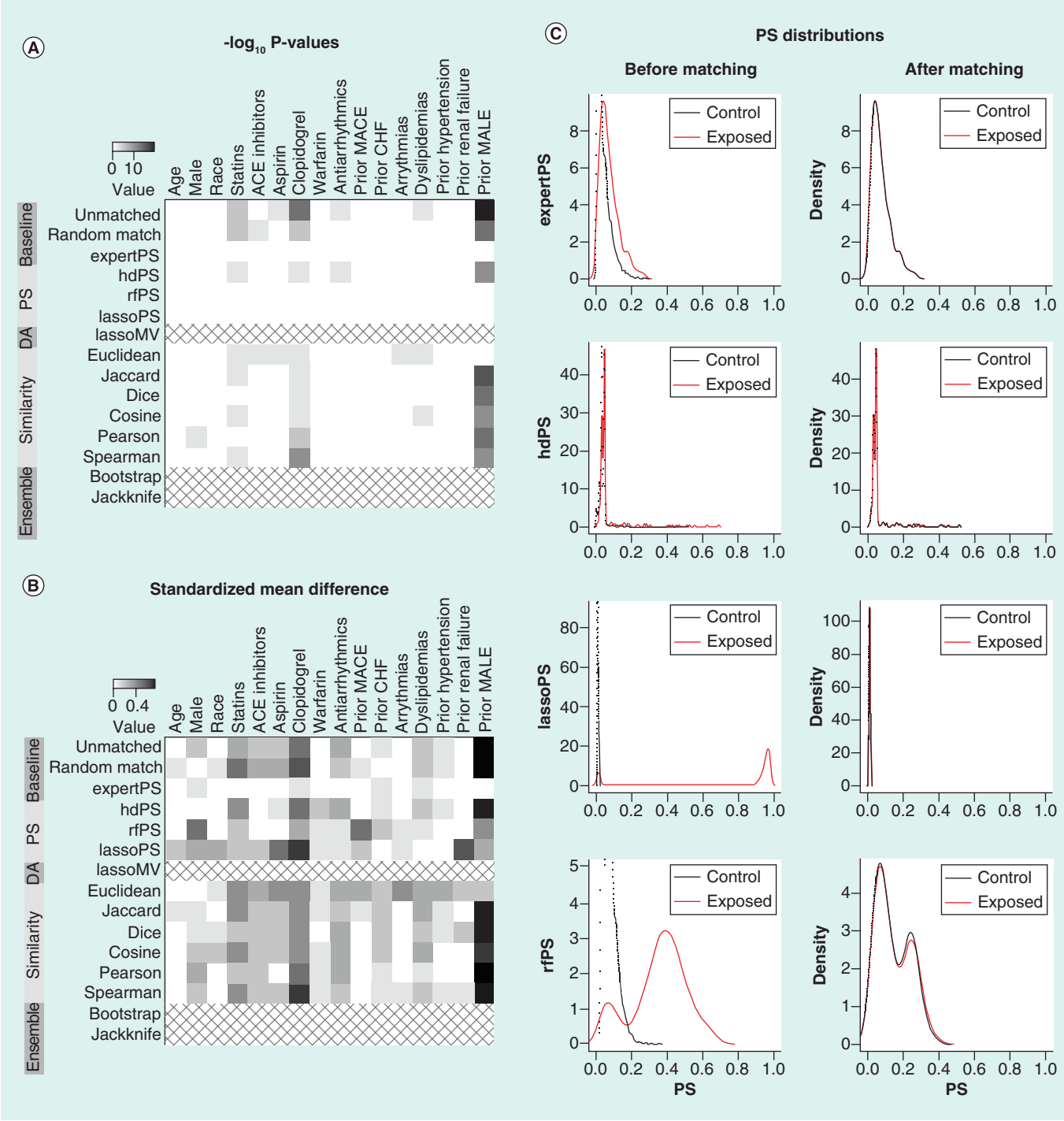


Figure 2. Balance of baseline variables between exposed and matched controls in dataset 3. Heatmaps showing the balance of baseline variables between exposed and matched controls groups in terms of **(A)** -log $_{10}$ p-values and **(B)** standardized mean difference in dataset 3. Lighter cells indicate smaller values while darker cells indicate bigger values. **(C)** PS distributions of exposed (red) and control (black) groups before and after matching by PS methods in dataset 3.

distances learned from the data may perform better, especially when the number of variables is large [18]. Thus, lassoMV may be an overall good alternative. By adopting a linear model framework and allowing for DA, lassoMV is highly amenable to interpretation and is already widely used in genetic epidemiology for handling millions of genetic variables [56,57].

Ensemble methods may be less interpretable due to bundling of multiple models and are computationally intensive (less so with parallel computing). Variables

are 'weighted' depending on how frequently they are selected in each model making up the ensemble. Random resampling procedures fared well despite not balancing the confounders. This may occur because multiple random sampling mimics a Monte Carlo process in which multiple estimates from multiple random samples are aggregated such that the aggregated value approximates the true parameter of interest [58]. An alternate explanation is that the aggregate of multiple models may also be viewed as an ensemble model in which errors from multiple constituent models are eventually averaged out [59–61] such that the overall variance of the ensemble model decreases as the number of constituent models increases [60]. In other words, the Monte Carlo process or ensemble model may be viewed as a meta-analysis of multiple studies where one may arrive at an accurate estimate of effect size given a large number of studies. We name such approaches as aggregate of random matched samples (ARMS). While ARMS is a promising choice in this study, additional research (e.g., parameter tuning) to optimize performance and increase our understanding of its strengths and weaknesses is necessary.

Among the base cases, random matching performed as well as expertPS (covariate adjustment). While this result may appear surprising, we note that expertPS has been shown to approximate random matching [28]. Although our expertPS model could have been improved by accommodating the nonlinear variables given the known nonlinear causal structure, we wanted to emulate the common practice of assuming a linear PS model when the causal structure is unknown, similar to the previous comparison study [19]. Consequently, the relatively large bias associated with expertPS in both simulated datasets demonstrates the limitations

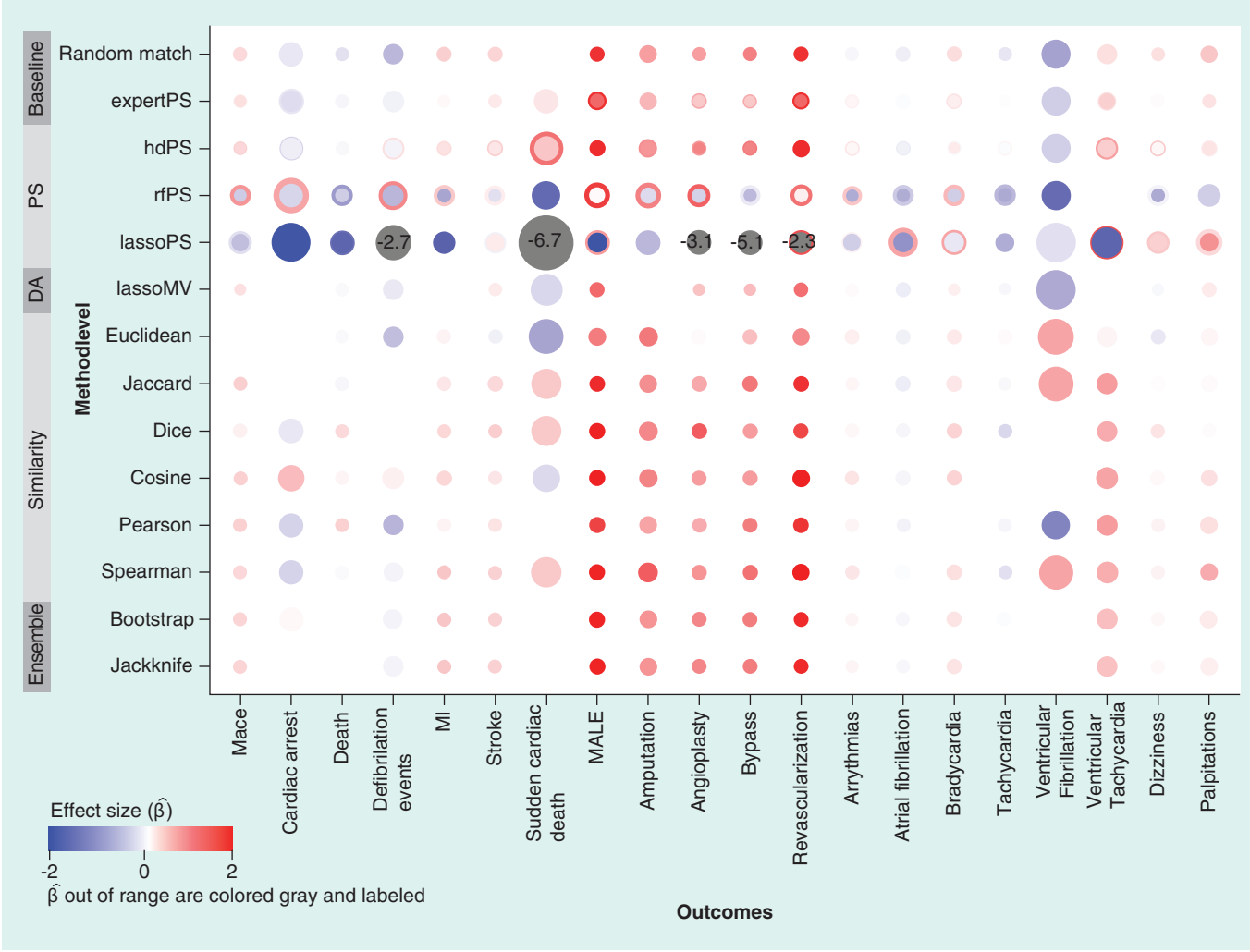


Figure 2. Bubble plot of estimated effect size β (bubble color) and their 95% CI (bubble size) across 14 methods (rows) and 20 outcomes (columns) in dataset 3. Small intensely colored bubbles indicate significant effects with narrow 95% CI. Because PS methods used both covariate adjustment and matching, results from covariate adjustment are overlaid on the results from matching (See Supplementary Data C for related forest plots and Supplementary Data D for numerical values).

of expertPS even when the causal structure and true effect sizes are known. In light of calls to reconsider and objectively assess PS [28], we also explore methods alternate to PS in this study.

There are several study limitations that warrant future research. First, this study's scope is limited to cohort studies because case–control studies will require different handling [62]. Second, we assumed strong ignorability, a condition for PS to have unbiased estimates whereby the expected error due to unmeasured confounding is zero [63]. However, given the presence of unmeasured confounders especially in real-world datasets, such an assumption may not always be valid. One solution may be to detect and assess unmeasured confounding using maximal ancestral graphs and sensitivity analysis [31,64]. Third, we re-used a previously simulated dataset [19–21], which had only ten variables. However, we included a second simulated dataset with 100 variables to assess performance of methods in the presence of a large number of variables. Fourth, both simulated datasets used nonlinear and nonadditive generative models. Additional settings with other data-generating models will need to be investigated. Fifth, we only studied two ways of using PS (matching and covariate adjustment) and did not investigate other approaches such as Inverse Probability of Treatment Weight estimators [29,65], which would be a fruitful area of further study.

Conclusion & future perspective

To leverage the scale and richness of EHR for clinical decisions, particularly in emergencies and in the absence of evidence from randomized control trials, timely and accurate synthesis of evidence through automated methods is necessary. Toward that vision, advances have been made with automatic cohort selection via electronic phenotyping [13], natural language processing [66,67], patient similarity [8] and automatic confounder control by PS methods [17,19–22,24–27]. This study supplements automation efforts by demonstrating that there exist automated alternatives to expert-based PS, including non-PS-based methods, for confounder control. Actual choice may depend on user preference for interpretable linear models (e.g., lassoMV DA) or 'meta-analysis' of multiple models of matched samples used in ensemble resampling.

We emphasize that our end goal is not to automate clinical decisions but to facilitate the process of extracting personalized evidence-based decisions from locally relevant and readily available EHR. a valuable alternative when evidence is lacking or inaccessible from established sources. The success of real-time EHR-based evidence for clinical decision support will depend on additional factors such as usability, transparency, interpretability and interactivity. Thus, we envision a transparent and interactive system that will allow real-time sensitivity analysis and iterative corrections for the user to assess the quality and generalizability of the evidence. The need for an expert review, even if subjective at times, cannot be overstated and it is possible that in the future, we may have 'epidemiology consultations' analogous to specialty consultations today.

By demonstrating, there exist automated confounder control methods for binary and continuous variables, our study is generalizable to many datasets including structured codes and unstructured text in EHR and other health databases. Extraction of features from clinical text – though challenging, and not widespread yet – is likely to unlock massive amounts of clinical information in the near future. As we expect more data to become available especially with advances in information retrieval, automated confounder control methods will be critical to faster and smarter clinical decision support.

Supplementary data

To view the supplementary data that accompany this paper, please visit the journal website at: www.futuremedicine.com/doi/full/10.2217/cer.15.53

Acknowledgements

The authors thank Jeremy Rassen for support with hdPS, Anna Bauer-Mehran for the cilostazol dataset and Alan M Brookhart for advice.

Financial & competing interests disclosure

This work was supported in part by grants R01 LM011369, U54 HG004028 and R01 GM101430. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Executive summary

- Clinical urgency requires quick and accurate evidence for clinical decision support at the point of care. Real-time observational studies from electronic health records may be enabled if automated methods exist for cohort selection, confounder control and statistical analysis.
- We compare 18 automated confounder control methods including non-propensitiy score (PS) methods for handling large number of variables.
- Automated methods accommodating numerous variables such as high-dimensional PS, direct adjustment lasso logistic regression and ensemble models yielded comparable or better performance than expert PS, potentially enabling real-time cohort studies for timely decisions.

References

Papers of special note have been highlighted as:
• of interest; •• of considerable interest

1 Schneeweiss S. Learning from big health care data. *N. Engl. J. Med.* 370(23), 2161–2163 (2014).

2 Dahabreh IJ, Kent DM. Can the learning health care system be educated with observational data? *JAMA* 312(2), 129–130 (2014).

3 Longhurst CA, Harrington RA, Shah NH. A "green button" for using aggregate patient data at the point of care. *Health Aff.* 33(7), 1229–1235 (2014).

• **Provides a vision of leveraging patient data at the point for care for clinical decisions.**

4 Frakt AB. An observational study goes where randomized clinical trials have not. *JAMA* 313(11), 1091–1092 (2015).

5 Feinstein AR, Rubinstein JF, Ramshaw WA. Estimating prognosis with the aid of a conversational-mode computer program. *Ann. Intern. Med.* 76(6), 911–921 (1972).

6 Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. *N. Engl. J. Med.* 365(19), 1758–1759 (2011).

7 Cimino JJ, Li J. Sharing infobuttons to resolve clinicians' information needs. *AMIA Annu. Symp. Proc.* 815 (2003).

8 Gallego B, Walter SR, Day RO *et al.* Bringing cohort studies to the bedside: framework for a "green button" to support clinical decision-making. *J. Comp. Eff. Res.* 11, 1–7 (2015).

•• **Illustrates how cohort studies can be performed at the bedside.**

9 Gallego B, Dunn AG, Coiera E. Role of electronic health records in comparative effectiveness research. *J. Comp. Eff. Res.* 2(6), 529–532 (2013).

10 McDavid A, Crane PK, Newton KM *et al.* Enhancing the Power of Genetic Association Studies through the use of silver standard cases derived from electronic medical records. *PLoS ONE* 8(6), e63481 (2013).

11 Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J. Am. Med. Informatics Assoc.* 20(e2), e206–e211 (2013).

12 Shivade C, Raghavan P, Fosler-Lussier E *et al.* A review of approaches to identifying patient phenotype cohorts using electronic health records. *J. Am. Med. Inform. Assoc.* 21(2), 221–230 (2014).

13 Agarwal V, Lependu P, Podchiyska T *et al.* Using narratives as a source to automatically learn phenotype models. Presented at: *Workshop on Data Mining for Medical Informatics, AMIA Annual Symposium Proceedings.* The American Medical Informatics Association (AMIA), DC, USA, 1–6 (2014).

14 Heinze G, Jüni P. An overview of the objectives of and the approaches to propensity score analyses. *Eur. Heart J.* 32(14), 1704–1708 (2011).

15 Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am. J. Epidemiol.* 158, 280–287 (2003).

16 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55 (1983).

17 Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 20(4), 512–522 (2009).

•• **Is one of the first automated high-dimensional propensity scores (PS) methods made available.**

18 Sekhon J. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *J. Stat. Softw.* 42(7) 1–52 (2011).

19 Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol. Drug Saf.* 17(6), 546–555 (2008).

•• **Provides the simulated data set widely used to benchmark various PS methods including the high-dimensional propensity score method (see Reference 17).**

20 Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat. Med.* 29(3), 337–346 (2010).

21 Wyss R, Ellis AR, Brookhart MA *et al.* The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *Am. J. Epidemiol.* 180(6), 645–655 (2014).

22 McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* 9, 403–425 (2004).

23 McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat. Med.* 32(19), 3388–3414 (2013).

24 Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J. Clin. Epidemiol.* 63(8), 826–833 (2010).

25 Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *Am. J. Epidemiol.* 173(12), 1404–1413 (2011).

26 Toh S, García Rodríguez LA, Hernán MA. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. *Pharmacoepidemiol. Drug Saf.* 20(8), 849–857 (2011).

27 Garbe E, Kloss S, Suling M, Pigeot I, Schneeweiss S. High-dimensional versus conventional propensity scores in a comparative effectiveness study of coxibs and reduced upper gastrointestinal complications. *Eur. J. Clin. Pharmacol.* 69(3), 549–557 (2013).

28 King G, Nielsen R, Coberley C, Pope JE. Comparative effectiveness of matching methods for causal inference. Harvard (2011). http://gking.harvard.edu/files/psparadox.pdf

- **Illustrates the limitations of propensity scores and offers suggestions for improved methods.**

29 Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat. Med.* 26(4), 734–753 (2007).

30 Brookhart MA, Stürmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Med. Care* 48(Suppl. 6), S114–S120 (2010).

- **Provides an overview of various confounder control approaches and discusses their pitfalls and potential solutions to overcome them.**

31 Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat. Sci.* 25(1), 1–21 (2010).

32 Sauer BC, Brookhart MA, Roy J, VanderWeele T. A review of covariate selection for non-experimental comparative effectiveness research. *Pharmacoepidemiol. Drug Saf.* 22(11), 1139–1145 (2013).

- **Describes the different types of variables in a causal diagram and reviews the various ways of handling them for observational studies.**

33 Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am. J. Epidemiol.* 163(12), 1149–1156 (2006).

34 Leeper NJ, Bauer-Mehren A, Iyer SV, LePendu P, Olson C, Shah NH. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. *PLoS ONE* 8(5), e63499 (2013).

35 Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen MA. Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics* 10(Suppl. 9), S14 (2009).

36 Lependu P, Iyer SV, Fairon C, Shah NH. Annotation analysis for testing drug safety signals using unstructured clinical notes. *J. Biomed. Semantics* 3(Suppl. 1), S5 (2012).

37 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32, D267–D270 (2004).

38 Tibshirani R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58(1), 267–288 (1996).

39 Breiman L. Random forests. *Mach. Learn.* 45, 5–32 (2001).

40 Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33(1), 1–22 (2010).

41 Liaw A, Wiener M. Classification and regression by randomForest. *R News.* 2, 18–22 (2002).

42 Austin PC. Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based G-computation. *Multivariate Behav. Res.* 47(1), 115–135 (2012).

43 Ho TK. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 20(8), 832–844 (1998).

44 Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A. Probability machines: consistent probability estimation using nonparametric learning machines. *Methods Inf. Med.* 51(1), 74–81 (2012).

45 Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm. Stat.* 10(2), 150–161 (2011).

46 Lunt M. Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. *Am. J. Epidemiol.* 179(2), 226–235 (2014).

47 Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* 1(1), 54–75 (1986).

48 Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat. Med.* 33(6), 1057–1069 (2014).

49 Kleiner A, Talwalkar A, Sarkar P, Jordan MI. A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B.* 76(4), 795–816 (2014).

50 Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. No Title. *BMC Bioinformatics* 8(1), 25 (2007).

51 Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. *J. Econ. Surv.* 22(1), 31–72 (2008).

52 Shadish WR, Steiner PM. A primer on propensity score analysis. *Newborn Infant Nurs. Rev.* 10(1), 19–26 (2010).

53 Kupper LL, Karon JM, Kleinbaum DG, Morgenstern H, Lewis DK. Matching in epidemiologic studies: validity and efficiency considerations. *Biometrics* 37(2), 271–291 (1981).

54 Imai K, Ratkovic M. Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B* 76(1), 243–263 (2014).

55 Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is "nearest neighbor" meaningful? In: *Database Theory.* Springer, Berlin, Heidelberg, Germany 217–235 (1999).

56 Dasgupta A, Sun YV, König IR, Bailey-Wilson JE, Malley JD. Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. *Genet. Epidemiol.* 35(Suppl. 1), S5–S11 (2011).

57 Heidema AG, Boer JMA, Nagelkerke N, Mariman ECM, van der ADL, Feskens EJM. The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. *BMC Genet.* 7, 23 (2006).

58 Metropolis N, Ulam S. The Monte Carlo method. *J. Am. Stat. Assoc.* 44(247), 335–341 (1949).

59 Dietterich TG. Ensemble methods in machine learning. In: *Multiple Classifier Systems.* Springer, Berlin, Heidelberg, Germany 1–15 (2000).

60 Biau G. Analysis of a random forests model. *J. Mach. Learn. Res.* 13, 1–31 (2010).

61 Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Mach. Learn.* 36(1), 105–139 (1999).

62 Rose S, van der Laan MJ. Why match? Investigating matched case-control study designs with causal effect estimation. *Int. J. Biostat.* 5(1), Article 1 (2009).

63 Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *J. R. Stat. Soc. Ser. A.* 171(2), 481–502 (2008).

64 Kalisch M, Bühlmann P. Causal structure learning and inference: a selective review. *Qual. Technol. Quant. Manag.* 11(1), 3–21 (2014).

65 Hade EM, Lu B. Bias associated with using the estimated propensity score as a regression covariate. *Stat. Med.* 33(1), 74–87 (2014).

66 Tamang S, Patel MI, Blayney DW *et al.* Detecting unplanned care from clinician notes in electronic health records. *J. Oncol. Pract.* 11(3), e313–e319 (2015).

67 Shah NH. Mining the ultimate phenome repository. *Nat. Biotechnol.* 31(12), 1095–1097 (2013).