Title: Machine learning for improving high-dimensional proxy confounder adjustment in healthcare database studies: an overview of the current literature.

Short running title: High-dimensional proxy confounder adjustment

Authors: Richard Wyss[1]*, Chen Yanover[2]*, Tal El-Hay[3], Dimitri Bennett[4], Robert W. Platt[5], Andrew R. Zullo[6], Grammati Sari[7], Xuerong Wen[8], Yizhou Ye[9], Hongbo Yuan[10], Mugdha Gokhale[11], Elisabetta Patorno[1], Kueiyu Joshua Lin[1,12]

*co-first authors

Author Affiliations:
[1] Division of Pharacoepidemioogy and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.
[2] KI Research Institute, Kfar Malal, Israel
[3] IBM Research–Haifa Labs, Haifa, Israel
[4] Global Evidence and Outcomes, Takeda Pharmaceutical Company Ltd., Cambridge, MA, USA
[5] McGIl University, Montreal, Canada
[6] Department of Health Services, Policy, and Practice, Brown University School of Public Health and Center of Innovation in Long-Term Services and Supp orts, Providence Veterans Affairs Medical Center, Providence, RI, USA
[7] Real World Evidence Strategy Lead, Visible Analytics Ltd, Oxford, UK
[8] Health Outcomes, Pharmacy Practice, College of Pharmacy, University of Rhode Island, Kingston, RI, USA
[9] Global Epidemiology, AbbVie Inc. North Chicago, IL, USA
[10] Canadian Agency for Drugs and Technologies in Health, Ottawa, Canada
[11]Pharmacoepidemiology, Center for Observational and Real-world Evidence, Merck, PA, USA
[12] Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA.

Corresponding Author Contact Information:
Richard Wyss (rwyss@bwh.harvard.edu)

Target Journal: Pharmacoepidemiology and Drug Safety

Conflict of Interest statement:
RWP has consulted for Amgen, Biogen, Merck, Nant Pharma, and Pfizer. DB is an employee of Takeda. GS is employed by Visible Analytics Ltd. HY is an employee of CADTH. ARZ receives research grant funding from Sanofi Pasteur to support research on infections and vaccinations in nursing homes unrelated to this manuscript. MG is a full time employee of Merck and owns stocks in Merck. EP is is supported by a career development grant K08AG055670 from the National Institute on Aging. She is researcher of a researcher-initiated grant to the Brigham and Women’s Hospital from Boehringer Ingelheim, not directly related to the topic of the submitted work.

Word Count: 4,080 (4,000 limit)

Funding and Competing Interests: Funding to support this manuscript development was provided by the International Society for Pharmacoepidemiology (ISPE).

Prior Presentation
No prior submissions or presentations have been done on this paper.

Abstract

Controlling for large numbers of variables that collectively serve as 'proxies' for unmeasured factors can often improve confounding control in pharmacoepidemiologic studies utilizing administrative healthcare databases. There is a growing body of evidence showing that data-driven machine learning algorithms for high-dimensional proxy confounder adjustment can supplement investigator-specified variables to improve confounding control compared to adjustment based on investigator-specified variables alone. Consequently, there has been a recent focus on the development of data-driven methods for high-dimensional proxy confounder adjustment. In this paper, we discuss the considerations underpinning three areas for data-driven high-dimensional proxy confounder adjustment: 1) feature generation—transforming raw data into covariates (or features) to be used for proxy adjustment; 2) covariate prioritization, selection and adjustment; and 3) diagnostic assessment. We discuss current approaches and recent advancements within each area, including the most widely used approach to proxy confounder adjustment in healthcare database studies (the high-dimensional propensity score or hdPS). We also discuss limitations of the hdPS and outline recent advancements that incorporate the principles of proxy adjustment with machine learning extensions to improve performance. We further discuss challenges and areas of future development within each area.

Key Points

- To improve confounding control in healthcare database studies, data-driven algorithms can be used to leverage the large volume of information in healthcare databases to generate and identify features that indirectly capture information on unmeasured or unspecified confounding factors (proxy confounders).
- Three areas to consider for data-driven high-dimensional proxy confounder adjustment include: 1) feature generation—transforming raw data into covariates (or features) to be used for proxy adjustment; 2) covariate prioritization, selection and adjustment; and 3) diagnostic assessment.
- The areas of feature generation and diagnostic assessment have particular limitations and challenges when applying machine learning algorithms for high-dimensional proxy confounder adjustment.

1. Introduction

Routinely-collected healthcare data are increasingly being used to generate real-world evidence (RWE) to inform decision making in clinical practice, drug development, and health policy.[1] However, unmeasured confounding from non-randomized treatment allocation and poorly measured information on comorbidities, disease progression, and disease severity remains a fundamental obstacle to effectively utilizing these data sources for RWE generation.[2] Statistical methods should therefore be used to extract the maximum possible information on confounding from the data to minimize the effects of unmeasured confounding so that accurate comparative estimates of treatments' effectiveness and safety can be obtained. Approaches to mitigate confounding bias would ideally be based on causal diagrams and expert knowledge for confounder selection.[3] However, adjustment based on researcher-specified variables alone is not always adequate because some confounders are either unknown to researchers or not directly measured in these data sources.

To improve confounding control in healthcare database studies, data-driven algorithms can be used to leverage the large volume of information in these data sources to generate and identify features that indirectly capture information on unmeasured or unspecified confounding factors (proxy confounders).[4] Proxy confounder adjustment is based on the concept that unmeasured confounding can be mitigated by adjusting for large numbers of variables that collectively serve as proxies for unobserved factors.[5] For example, donepezil use (captured in any claims database) could be used as a proxy for cognitive impairment since cognitive impairment and early Alzheimer's disease and related disorders (ADRD) are often unmeasured in administrative data (Figure 1). For more on the concept of proxy confounder adjustment see VanderWeele[3] and Schneeweiss[4].

While researcher-specified confounders are identified using expert background knowledge, empirical or proxy confounders are identified using empirical associations and coding patterns observed in the data. There is a growing body of evidence showing that complementing

researcher-specified variables with empirically-identified proxy confounders improves confounding control compared to adjustment based on researcher-specified confounders alone.[4,6-9] Consequently, there has been a recent focus on the development of data-driven methods to empirically identify high-dimensional sets of proxy variables for adjustment in healthcare database studies.[6,10-18]

In this paper, we discuss the considerations underpinning three areas for data-driven high-dimensional proxy confounder adjustment: 1) feature generation—transforming raw data into covariates (or features) to be used for proxy adjustment; 2) covariate prioritization, selection and adjustment; and 3) diagnostic assessment (Figure 2). We discuss current approaches and recent advancements within each area, including the most widely used approach to proxy confounder adjustment in healthcare database studies (the high-dimensional propensity score or hdPS). We also discuss limitations of the hdPS and outline recent advancements that incorporate the principles of proxy adjustment with machine learning (ML) extensions to improve performance. We further discuss challenges and areas of future development within each area. We give particular focus to diagnostics for validity assessment as this has received the least attention when performing high-dimensional proxy confounder adjustment in the pharmacoepidemiology literature.

2. Generating features for proxy confounder adjustment

The first challenge for proxy confounder adjustment is determining how to best leverage the full information content in healthcare databases to generate features (or proxy variables) that best capture confounder information. Several approaches for feature generation of proxy confounders have been applied in the pharmacoepidemiologic literature. These have ranged from very simple approaches that generate binary indicators representing whether or not a given code occurs during a pre-defined exposure assessment period,[19] to more sophisticated approaches that transform information from healthcare databases into a common data model format with

common terminologies and coding schemes representing health concepts.[20-22] Examples include the Observational Medical Outcomes Partnership (OMOP) Common Data Model, maintained by the open-science Observational Health Data Sciences and Informatics (OHDSI) network and also used in the European Health Data and Evidence Network (EHDEN) project, and the National Patient-Centered Clinical Research Network (PCORnet).[23-25] Generating features consistent with a common data model format can be advantageous for capturing relevant health concepts, but these approaches require more data pre-processing to extract and transform the original codes into variables representing health concepts.

Instead of generating features based on health concepts, an alternative approach is to generate features based on empirical associations and longitudinal coding patterns observed in the data. Such approaches can be more flexible since they can be independent of the coding system and do not rely on a common data model.[6] The hdPS has become the most widely used tool to generate features based on observed coding patterns in healthcare claims databases.[6] The hdPS generates features by transforming raw medical codes into binary indicator variables based on the frequency of occurrence of each code during a defined pre-exposure period.

By taking into account the frequency of occurrence of various codes during the covariate assessment period, the hdPS tries to capture information on the intensity of the medical event or drug dispensing. In theory, algorithms could consider more complex longitudinal coding patterns to try and capture additional confounder information. For example, recent work has proposed using neural networks to model a patient's full course of care to consider temporal sequences of a specific course of treatment.[26] The use of neural networks for extracting confounder information by modeling complex coding patterns is promising but examples are limited.[27,28]

2.1 Challenges in generating features for proxy adjustment from unstructured free-text electronic health records:

An important limitation of current high-dimensional proxy confounder adjustment approaches is that they can only use structured electronic healthcare information. However, much of the essential confounder information, such as patient-reported symptoms, severity, stage and prognosis of the disease, and functional status, is frequently recorded in free-text notes or reports in electronic health records (EHRs) that are substantially underutilized for confounding adjustment. This limitation has left some key research questions unaddressable, at least in some populations, with routine-care databases due to unmeasured confounding.[29,30]

Less is currently known about the impact of incorporating these data for confounding adjustment since unstructured data are not readily analyzable. Natural language processing (NLP) is a subfield of machine learning that can be used to generate variables from unstructured free text.[31] NLP methods are increasingly used to identify health outcomes from EHRs, but the application of NLP algorithms for purposes of identifying high-dimensional sets of confounding factors is limited.[32] More research is needed on the use of NLP algorithms for generating high-dimensional sets of proxy confounders and the value of unstructured EHR data in proxy adjustment. Future research could evaluate the use of more sophisticated NLP methods to convert unstructured data into structured features. Scalable NLP methods that could potentially be used to convert unstructured information into structured data could include named entity recognition (clinical and contextual information extraction and encoding),[33-36] distributional semantics models,[37-39] and word embeddings.[40,41] A detailed discussion on NLP is provided elsewhere.[31]

3. Covariate Prioritization, Selection, and Adjustment

Once proxy variables have been generated through transformations of the raw data, some degree of dimension reduction is needed to prioritize and select variables for adjustment. The hdPS prioritizes or ranks each generated variable based on its potential for bias by assessing the variable's prevalence and univariate, or marginal, association with the treatment and outcome

using the Bross bias formula.[6,42,43] From this ordered list, researchers then specify the number of variables to include for adjustment along with pre-specified variables such as age and gender.

3.1. Limitations of the hdPS prioritization:

While the hdPS has been shown to often improve confounding control when used to complement investigator-specified confounders,[6,8,9,44,45] there are cases where adjustment for hdPS generated variables had no impact or even harmed the properties of effect estimates beyond adjustment for researcher-specified confounders alone.[46,47] Limitations of the hdPS prioritization include: 1) the method assesses a variable's potential confounding impact through marginal, or univariate, associations with both treatment and outcome (ideally one would want to consider conditional (or joint) associations among variables); 2) the method requires researchers to subjectively determine how many "proxy" variables to include for adjustment. These limitations can lead to "*over adjusting*" for variables that can harm the properties of effect estimates without reducing bias (Figure 3).[48-50]

The choice of the number of proxy variables to include in an hdPS model to adequately control for confounding without "*over adjusting*" varies according to the properties and structure of a given dataset and cannot be identified by only evaluating marginal associations between variables. Determining how many empirically identified "proxy" confounders to include for adjustment is particularly challenging in studies with rare events-settings relevant to RWE studies. In these settings, previous work has shown unstable effect estimates where results are highly dependent on the number of "proxy" confounders included for adjustment.[9,44]

3.2. Machine learning extensions for covariate prioritization and selection

To overcome the limitations outlined above, recent studies have developed extensions for proxy confounder adjustment that combine the principles of proxy confounder adjustment with ML tools for prediction modeling and variable selection. These tools have largely focused on incorporating principles for proxy confounder adjustment with regularized regression and

Targeted Learning tools, including Super Learning and Collaborative Targeted variable selection. While other machine learning tools for variable prioritization and selection are available (e.g., principal components, random forests, feature importance selection with neural networks), here we focus on targeted learning tools and regularized regression as these have been the most widely used approaches in the pharmacoepidemiology literature.

Regularized Regression for high-dimensional proxy confounder adjustment: Regularized regression models use penalized maximum likelihood estimation to shrink imprecise model coefficients toward zero. LASSO is the most commonly used regularized regression model for variable selection in high-dimensional covariate datasets.[45,51] Previous work[20-22,52] found that LASSO regression can be used to select a subset of generated "proxy" confounders to supplement researcher-specified confounders to form the adjustment set for confounding control. To improve the performance of regularized regression for high-dimensional confounder selection, several studies have developed variations of LASSO that consider covariate associations with both treatment and outcome when penalizing the likelihood function. These recent extensions include: 1) Outcome adaptive LASSO[17], 2) Group LASSO,[16] 3) Highly Adaptive LASSO,[53] 4) Highly Adaptive Outcome LASSO,[11] and 5) Collaborative Controlled LASSO.[54] Other versions of regularized regression, including ridge regression and elastic net, have also been shown to perform well for confounder selection and can be preferable to the LASSO penalization in certain settings.[52]

Combining the hdPS with Super Learning: Super Learning is an ensemble ML algorithm for prediction modeling that forms a set of predicted values based on the optimal weighted combination of a set of user-specified prediction models.[55,56] The flexibility of *super learning* can be utilized to identify a small number of optimally performing prediction algorithms that generally perform best for a given data structure. Previous work has combined Super Learning with proxy confounder adjustment in high-dimensional covariate spaces.[18] Super Learning can simplify

propensity score estimation in high dimensions and has been shown to perform well in a number of simulations.[13,18]

High-dimensional proxy adjustment with scalable versions of Collaborative Targeted Maximum Likelihood Estimation (CTMLE): CTMLE is an extension of the doubly robust targeted maximum likelihood estimation (TMLE) method.[57,58] TMLE consists of fitting an initial outcome model to predict the counterfactual outcomes for each individual, then using the estimated propensity score to fluctuate this initial estimate to optimize a bias/variance tradeoff for a specified causal parameter (i.e., the treatment effect). CTMLE extends TMLE by using an iterative forward selection process to construct a series of TMLE estimators, where each successive TMLE estimator controls for one additional variable to consider how a variable relates to both treatment and outcome after conditioning on a set of previously selected variables.[57,58] By taking into account a variable's conditional association with both treatment and outcome, CTMLE avoids "*over-adjustment*" to improve the properties of effect estimates by reducing the likelihood of controlling for variables that are conditionally independent of the outcome after adjusting for a set of previously identified confounders. Recent work has developed adaptations of CTMLE that are computationally scalable to large healthcare databases.[14] These adaptations modify the standard version of CTMLE by including a pre-ordering of variables to avoid the iterative process of searching through each variable in the selection procedure. Simulations indicate that computational gains are substantial and that combining scalable CTMLE with methods for proxy adjustment work well relative to the standard instantiations of CTMLE, hdPS, and TMLE.[14,18]

3.3. Adjustment for proxy confounders.

Once proxy confounders have been prioritized and selected, researchers must determine a method for adjustment and causal estimation. Propensity score methods (e.g., propensity score matching, inverse probability weighting) using logistic regression for estimation of the propensity score function have become the most common approach for adjustment of selected proxy

confounders in the medical literature.[59,60] Some evidence suggests that improvements can be gained in both predictive performance and bias reduction when using more flexible ML models for propensity score estimation.[28,61-63] Another avenue for improving estimations is to adapt ML algorithms to casual inference. Two important examples are the adaptation of random forest to *causal forest* and *X-learner*, a meta-algorithm that uses ML methods as an intermediate step in an efficient estimation algorithm.[64,65]

Machine learning with doubly robust estimation for improved adjustment. Widely used doubly robust methods include TMLE, augmented inverse probability weighting (AIPW), and double ML (e.g., R-learner [ref]).[58,66-68] These approaches use a model for both the outcome and the propensity score, requiring only one of the two to be correctly specified for consistent estimation of average treatment effects. Theory and simulations have shown that doubly robust approaches are asymptotically efficient and more robust than conventional singly robust methods like propensity score matching and inverse probability weighting.[69]

Recent work has further shown that the use of flexible nonparametric ML models for the estimation of nuisance functions (i.e., the propensity score or outcome model) comes at a cost of slow convergence rates. This slow convergence is particularly problematic within singly robust estimation methods and can yield effect estimates with poor statistical properties with performance deteriorating as the dimension of the data increases (the 'curse of dimensionality').[70] This work has further demonstrated that doubly robust methods allow for slower converging nuisance models and, therefore, can mitigate or even resolve such problems. Consequently, recent literature suggests that ML-based methods for estimation of nuisance functions should be applied within doubly robust frameworks rather than more commonly used singly robust methods. For more on machine learning in causal inference see Kennedy[70], Naimi et al.[71,72], and Zivich et al.[73]

4. Diagnostic Validity Assessment of Causal Estimations

Evaluating the validity of causal analyses for high-dimensional proxy adjustment remains challenging but is essential to improving robustness and validity of estimated effects.[74] While held-out sets and cross-validation allow a direct comparison of ML predictions to observed target variables, such a straightforward evaluation is infeasible in causal inference and the role of prediction diagnostics for purposes of causal inference is less clear.[48,75-77] Below, we survey a list of standard ML diagnostics for model prediction and diagnostic for causal inference with a focus on assessing the consistency and accuracy of models for high-dimensional proxy adjustment. We highlight their underlying assumptions and limitations.

4.1. Diagnostics for Model Prediction

Standard ML diagnostics, such as cross validation, are often recommended to assess model robustness and generalizability and to examine the characteristics of the inferred models to verify the importance of domain-relevant variables. Below we focus on additional measures with specific importance to causal model diagnostics.

4.1.1. Dichotomous and categorical models

Calibration plots depict the average predicted versus observed (empirical) probability of the studied event in subsets of entities (typically, deciles), to evaluate the accuracy of the predicted probabilities.[78,79] Probability estimation accuracy is essential for causal inference, more than it typically is for ML classification tasks, as downstream calculations, e.g., inverse probability weighting, may rely on these values as being "true" probabilities. Various metrics can be used to quantitatively measure calibration quality, e.g., Hosmer-Lemeshow goodness of fit test[80], but these have several drawbacks[81]; visual inspection of the calibration plots or characterization of its slope and intercept is thus recommended.

C-statistic (or area under the receiver operating characteristic, ROC, curve), a measure of classification accuracy, is commonly used in standard ML applications. For outcome models, it can be used to assess prediction accuracy over the observed treatment assignment (and

assuming, but not verifying, that the causal assumptions hold). For propensity models its utility is less straightforward: an extreme (close to 0 or 1) value, corresponding to a highly discriminative model, may indicate a potential violation of positivity; and, conversely, a value around 0.5, suggesting the model cannot discriminate between treatment groups, is not necessarily a sign for inaccurate model, but potentially good covariate overlap. As a result, some researchers recommended to avoid using C-statistic in propensity model diagnostics.[76] We note that post-matching C-statistic may be used to evaluate covariate balance; see below.

4.1.2. Continuous models

The performance of continuous outcome models can be assessed in each observed treatment group (and observed outcomes) and assuming causal assumptions are met, using standard measures such as the coefficient of determination (R^2) or mean squared error[79]. A poorly performing model for a specific treatment group, e.g., over or underestimating outcomes, may subsequently lead to biased effect estimation. As with binary outcome models, poor performance may suggest an inadequate prediction model and guide its improvement.

4.2. Diagnostics for Causal Inference

Previous work has shown that the use of prediction model diagnostics alone to guide model selection and validity assessment can lead to suboptimal performance for causal inference.[48-50,76,82] We next survey diagnostic methods to assess more directly assumptions and model validity for purposes of causal inference.

Positivity. An important usage for propensity models for high-dimensional proxy adjustment is to examine the positivity assumption. This assumption states that every individual has a non-zero probability to be assigned to any treatment. A comparison of propensity score distributions can help in identifying (and potentially excluding) sub-populations where violations or near violations of the positivity assumption occur.[83-85] While high-dimensional proxy adjustment assumes that unconfounded treatment effects are more plausible when controlling for large

numbers of variables, covariate overlap can be more difficult when adjusting for high-dimensional sets of variables.[86] Therefore, positivity should be tested at the initial stages of analyses for high-dimensional proxy adjustment.

Balancing. Propensity score modeling aims to facilitate matching, reweighting or stratification to emulate a random assignment of individuals to treatment groups. Therefore, several studies explored methods to directly evaluate balancing of covariates among these groups.[78,79,87] In a simulation study, Franklin et al[87] compared several metrics to assess covariate balance and observed that two had consistently strong associations with bias in estimated treatment effects. The first metric, post-matching C-statistic, re-trains a treatment model on the propensity score matched (similarly, stratified or weighted) sample and assesses its (preferably, lack of) ability to discriminate between patients in different treatment groups using C-statistic. The second recommended metric, general weighted difference, computes a weighted sum of absolute difference in all individual covariates, all covariate squares, and all pairwise interactions.

The application of balance diagnostics for high-dimensional propensity scores is more challenging as it is unclear on which set of variables balance should be assessed. A large literature has shown that balancing variables that are independent of the outcome except through treatment (instrumental variables) harms the properties of effect estimates.[48,50,82] In high-dimensional settings, however, identifying instrumental variables is difficult and previous work has argued that priority should be given to controlling for all confounders at the expense of balancing instruments.[20,21,49] This has led to some researchers assessing balance on all variables in the database when using propensity scores for high-dimensional proxy adjustment.[20-22] More research is needed on the best use of balance diagnostics for high-dimensional propensity score adjustment.

4.2.1. Estimand diagnostics (Simulation-based approaches and negative controls)

Recent studies have suggested methods to assess the overall accuracy of effect estimation using control and synthetic control studies.[20,21,88-90] These frameworks have largely

been based on the use of simulation methods to generate synthetic datasets under constraints where certain relations among variables are known (e.g., the simulated treatment effect) while maintaining much of the complexity and statistical properties of the observed data structure.

Parametric bootstrap ('Plasmode' simulation). Simulation frameworks for model validation in causal inference have largely been based on use of the parametric bootstrap. Such approaches bootstrap subjects from the observed data structure, then use modeled relationships from the original data to inject causal relations between a subset of variables while leaving all other associations among variables unchanged. With treatment-outcome associations known by design and patterns of confounding that mimic the observed data structure, synthetic datasets have become increasingly popular to provide a benchmark for comparing statistical methods for causal inference.

Franklin et al.[89] proposed using a parametric bootstrap approach to compare causal inference methods in settings specific to healthcare database studies and high-dimensional propensity scores. Franklin's approach, termed 'plasmode simulation', bootstraps subjects from the full study cohort, then uses a model for the outcome-generating distribution to simulate synthetic control outcomes while leaving baseline covariates and their joint association with treatment unchanged. Schuler et al.[90] and others[88,91,92] have proposed variations and extensions of plasmode simulation for model validation in healthcare database studies. Schuemie et al.[20,21] use a plasmode simulation-based approach for generating positive control outcomes to quantify bias due to measured confounders when calibrating effect estimates and confidence intervals. Peterson et al.[93] apply a similar parametric bootstrap method as a diagnostic to assess bias due to violations of positivity. Alaa and van der Schaar[88] developed a validation method that uses the parametric bootstrap and influence functions, which are a key technique in robust statistics.

Wasserstein Generative Adversarial Networks (WGANs) is an alternative approach to generating synthetic data for simulation-based model validation in causal inference.[94] GANs estimate the distribution of a particular dataset using a 'generator' and a 'discriminator'.[95] The

generator is a flexible neural network to create synthetic data while the discriminator is a competing neural network model that attempts to distinguish between the synthetic and real data. The process is repeated in an iterative fashion until the discriminator is no longer able to distinguish between the synthetic and real data. This technique has become very powerful for supervised and unsupervised ML.[95] WGANS have recently been shown to be useful for generating synthetic datasets that closely approximate the joint correlation structure of an actual dataset for purposes of model validation in causal inference.[94]

Negative controls: Another approach that has become increasingly popular for evaluating high-dimensional models for confounder adjustment is the use of real negative controls—exposure-outcome pairs that are not, as far as we know, causally related.[78,96] Such controls can be used to detect residual biases, e.g., confounding, in the estimation process. Replicating a known association through use of positive controls can also increase confidence in primary estimates' validity. However, some researchers have argued that identifying positive controls is difficult since the magnitude of known effects is rarely known.[20-22]

4.2.2. Sensitivity analyses

Quantitative bias analysis: Estimating an effect from observational data involves multiple, at times somewhat arbitrary, modeling decisions and assumptions, e.g., with respect to the definition of confounders, exposures, and outcomes or the statistical analysis.[97] Sensitivity analysis re-computes the estimated effect under various sets of such decisions[98] or using multiple data sources to verify its robustness.[84,99] Sensitivity analyses can also quantify the change that an unmeasured confounder would have on the studied estimand and thus assess its sensitivity to violations of the assumption of no unmeasured confounding.[98] This can be particularly useful when applying methods for high-dimensional proxy adjustment as researchers can never be certain how well a set of features captures information on unmeasured factors. Bias analysis helps quantify the impact of various systematic errors to estimated effects to increase confidence that

the estimated effects are robust to violations of various assumptions. Lash et al. provide a detailed discussion on methods for quantitative bias analysis.[100]

5. Discussion

In this paper, we have provided an overview of high-dimensional proxy confounder adjustment in studies utilizing electronic healthcare databases. We have focused on three areas for proxy adjustment: 1) feature generation, 2) covariate prioritization, selection, and adjustment, and 3) validity assessment. We have discussed recent ML extensions and areas for future research within each area. Much attention has been given to the development of ML tools for confounder selection and adjustment for high-dimensional proxy adjustment. These tools have great potential to improve confounding control in healthcare database studies. However, less attention has been given to advancing methods for feature generation and validity assessment for proxy confounder adjustment. Future research is warranted to investigate the optimal methods that extract the relevant confounding information to generate features for proxy adjustment while preserving scalability and data-adaptability to large healthcare databases. Future research is also needed in the development of diagnostic methods to evaluate and compare the validity of alternative approaches to high-dimensional proxy adjustment in healthcare database studies.

Author contributions: RW, CY, and TEH drafted the manuscript. All authors revised the manuscript for important intellectual content and approved the final manuscript to be submitted for publication

Ethics statement: The authors state that no ethical approval was needed

Patient consent statement: Not applicable.

Acknowledgements:
This manuscript is endorsed by the International Society for Pharmacoepidemiology (ISPE). Also, we would like to thank Ehud Karavani and Itay Manes, IBM Research – Haifa, for insightful discussions and comments. In addition, we also thank the ISPE members for reviewing and providing comments on our manuscript.

REFERENCES

1. Corrigan-Curay, J., Sacks, L. & Woodcock, J. Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. *Jama* **320**, 867-868 (2018).
2. Streeter, A.J.*, et al.* Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. *Journal of clinical epidemiology* **87**, 23-34 (2017).
3. VanderWeele, T.J. Principles of confounder selection. *Eur J Epidemiol* **34**, 211-219 (2019).
4. Schneeweiss, S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clinical epidemiology* **10**, 771-788 (2018).
5. Schneeweiss, S. & Avorn, J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of clinical epidemiology* **58**, 323-337 (2005).
6. Schneeweiss, S.*, et al.* High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20**, 512-522 (2009).
7. Guertin, J.R., Rahme, E., Dormuth, C.R. & LeLorier, J. Head to head comparison of the propensity score and the high-dimensional propensity score matching methods. *BMC medical research methodology* **16**, 22 (2016).
8. Guertin, J.R., Rahme, E. & LeLorier, J. Performance of the high-dimensional propensity score in adjusting for unmeasured confounders. *European journal of clinical pharmacology* **72**, 1497-1505 (2016).
9. Patorno, E., Glynn, R.J., Hernandez-Diaz, S., Liu, J. & Schneeweiss, S. Studies with many covariates and few outcomes: selecting covariates and implementing propensity-score-based confounding adjustments. *Epidemiology* **25**, 268-278 (2014).
10. Ertefaie, A., Asgharian, M. & Stephens, D.A. Variable selection in causal inference using a simultaneous penalization method. *Journal of causal inference* **6**(2018).
11. Ju, C., Benkeser, D. & van der Laan, M. Flexible collaborative estimation of the average causal effect of a treatment using the outcome-highly-adaptive Lasso. *arXiv:1806.06784 [stat.ME]* (2018).
12. Ju, C., Benkeser, D. & van der Laan, M.J. Robust inference on the average treatment effect using the outcome highly adaptive lasso. *arXiv preprint* (2019).
13. Ju, C.*, et al.* Propensity score prediction for electronic healthcare databases using Super Learner and High-Dimensional Propensity Score methods. *arXiv preprint* (2017).
14. Ju, C.*, et al.* Scalable collaborative targeted learning for high-dimensional data. *Statistical methods in medical research*, 962280217729845 (2017).
15. Ju, C.*, et al.* Collaborative-controlled LASSO for constructing propensity score-based estimators in high-dimensional data. *Statistical methods in medical research* (2017).
16. Koch, B., Vock, D.M. & Wolfson, J. Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics* **74**, 8-17 (2018).
17. Shortreed, S.M. & Ertefaie, A. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics* **73**, 1111-1122 (2017).
18. Wyss, R.*, et al.* Using Super Learner Prediction Modeling to Improve High-dimensional Propensity Score Estimation. *Epidemiology* **29**, 96-106 (2018).
19. Wyss, R.*, et al.* Matching on the disease risk score in comparative effectiveness research of new treatments. *Pharmacoepidemiology and drug safety* (2015).
20. Schuemie, M.J., Hripcsak, G., Ryan, P.B., Madigan, D. & Suchard, M.A. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc Natl Acad Sci U S A* **115**, 2571-2577 (2018).

21. Schuemie, M.J., Ryan, P.B., Hripcsak, G., Madigan, D. & Suchard, M.A. Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philos Trans A Math Phys Eng Sci* **376**(2018).
22. Tian, Y., Schuemie, M.J. & Suchard, M.A. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *Int J Epidemiol* **47**, 2005-2014 (2018).
23. Overhage, J.M., Ryan, P.B., Reich, C.G., Hartzema, A.G. & Stang, P.E. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* **19**, 54-60 (2012).
24. Kent, S.*, et al.* Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment. *Pharmacoeconomics* **39**, 275-285 (2021).
25. Fleurence, R.L.*, et al.* Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc* **21**, 578-582 (2014).
26. Rassen, J. A neural-network driven longitudinal propensity score for evaluation of drug treatment effects. in *International Conference on Pharmacoepidemiology & Therapeutic Risk Management* (Berlin, Germany, 2020).
27. Weberpals, J.*, et al.* Deep Learning-based Propensity Scores for Confounding Control in Comparative Effectiveness Research: A Large-scale, Real-world Data Study. *Epidemiology* **32**, 378-388 (2021).
28. Louizos, C.*, et al.* Causal effect inference with deep latent-variable models. *In Proceedings of teh 31st International Conference on Neural Information Processing Systems* **of NeurIPS'17**(2017).
29. Dong, Y.H.*, et al.* Evidence of potential bias in a comparison of beta blockers and calcium channel blockers in patients with chronic obstructive pulmonary disease and acute coronary syndrome: results of a multinational study. *BMJ Open* **7**, e012997 (2017).
30. Setoguchi, S.*, et al.* Influence of healthy candidate bias in assessing clinical effectiveness for implantable cardioverter-defibrillators: cohort study of older patients with heart failure. *BMJ* **348**, g2866 (2014).
31. Nadkarni, P.M., Ohno-Machado, L. & Chapman, W.W. Natural language processing: an introduction. *J Am Med Inform Assoc* **18**, 544-551 (2011).
32. Afzal, Z., Masclee, G.M.C., Sturkenboom, M., Kors, J.A. & Schuemie, M.J. Generating and evaluating a propensity model using textual features from electronic medical records. *PloS one* **14**, e0212999 (2019).
33. Zhou, L.*, et al.* Drug allergies documented in electronic health records of a large healthcare system. *Allergy* **71**, 1305-1313 (2016).
34. Lai, K.H., Topaz, M., Goss, F.R. & Zhou, L. Automated misspelling detection and correction in clinical free-text records. *J Biomed Inform* **55**, 188-195 (2015).
35. Zhou, L.*, et al.* Representation of information about family relatives as structured data in electronic health records. *Appl Clin Inform* **5**, 349-367 (2014).
36. Zhou, L.*, et al.* Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes. *AMIA Annu Symp Proc* **2011**, 1639-1648 (2011).
37. Tang, C., Zhou, L., Plasek, J., Rozenblum, R. & Bates, D. Comment Topic Evolution on a Cancer Institution's Facebook Page. *Appl Clin Inform* **8**, 854-865 (2017).
38. AK., M. MALLET: A Machine Learning for Language Toolkit 2002 [cited 2018 Feburary 5]. Available from: http://mallet.cs.umass.edu/.
39. Shao, Y.*, et al.* Identification and Use of Frailty Indicators from Text to Examine Associations with Clinical Outcomes Among Patients with Heart Failure. *AMIA Annu Symp Proc* **2016**, 1110-1118 (2016).

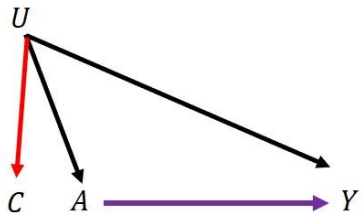
40. Kumamaru, H., Gagne, J.J., Glynn, R.J., Setoguchi, S. & Schneeweiss, S. Comparison of high-dimensional confounder summary scores in comparative studies of newly marketed medications. *Journal of clinical epidemiology* (2016).
41. Mikolov T, C.K., Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. *ArXiv13013781 Cs [Internet] 2013; Available from: http://arxiv.org/abs/1301.3781* (2013).
42. Bross, I.D. Spurious effects from an extraneous variable. *Journal of chronic diseases* **19**, 637-647 (1966).
43. Wyss, R., Fireman, B., Rassen, J.A. & Schneeweiss, S. Erratum: High-dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data. *Epidemiology* **29**, e63-e64 (2018).
44. Rassen, J.A., Glynn, R.J., Brookhart, M.A. & Schneeweiss, S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *American journal of epidemiology* **173**, 1404-1413 (2011).
45. Franklin, J.M., Eddings, W., Glynn, R.J. & Schneeweiss, S. Regularized Regression Versus the High-Dimensional Propensity Score for Confounding Adjustment in Secondary Database Analyses. *American journal of epidemiology* (2015).
46. Toh, S., Garcia Rodriguez, L.A. & Hernan, M.A. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. *Pharmacoepidemiology and drug safety* **20**, 849-857 (2011).
47. Austin, P.C., Wu, C.F., Lee, D.S. & Tu, J.V. Comparing the high-dimensional propensity score for use with administrative data with propensity scores derived from high-quality clinical data. *Statistical methods in medical research*, 962280219842362 (2019).
48. Brookhart, M.A.*, et al.* Variable selection for propensity score models. *American journal of epidemiology* **163**, 1149-1156 (2006).
49. Myers, J.A.*, et al.* Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology* **174**, 1213-1222 (2011).
50. Bhattacharya, J. & Vogt, W.B. Do Instrumental Variables Belong in Propensity Scores? *(NBER Technical Working Paper no. 343)* **Cambridge, MA: National Bureau of Economic Research.**(2007).
51. Low, Y.S., Gallego, B. & Shah, N.H. Comparing high-dimensional confounder control methods for rapid cohort studies from electronic health records. *Journal of comparative effectiveness research* **5**, 179-192 (2016).
52. Karim, M.E., Pang, M. & Platt, R.W. Can We Train Machine Learning Methods to Outperform the High-dimensional Propensity Score Algorithm? *Epidemiology* **29**, 191-198 (2018).
53. Benkeser, D. & van der Laan, M. The Highly Adaptive Lasso Estimator. *Proceedings of the ... International Conference on Data Science and Advanced Analytics. IEEE International Conference on Data Science and Advanced Analytics* **2016**, 689-696 (2016).
54. Ju, C.*, et al.* Collaborative-controlled LASSO for constructing proopensity score-based estimators in high-dimensional data. *Statistical methods in medical research*.
55. van der Laan, M.J., Polley, E.C. & Hubbard, A.E. Super Learner. *Statistical applications in genetics and molecular biology* **6**(2007).
56. Polley, E.C., Rose, S. & van der Laan, M.J. Super Learning. in *Targeted Learning* 43-66 (Springer, 2011).
57. van der Laan, M.J. & Gruber, S. Collaborative double robust targeted maximum likelihood estimation. *The international journal of biostatistics* **6**, Article 17 (2010).
58. Van der Laan, M.J. & Rose, S. *Targeted learning: Causal inference for observational and experimental data*, (Springer, Berlin, Heidelberg, New York, 2011).

59. Brookhart, M.A., Wyss, R., Layton, J.B. & Sturmer, T. Propensity score methods for confounding control in nonexperimental research. *Circulation. Cardiovascular quality and outcomes* **6**, 604-611 (2013).
60. Sturmer, T., Wyss, R., Glynn, R.J. & Brookhart, M.A. Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. *Journal of internal medicine* **275**, 570-580 (2014).
61. Lee, B.K., Lessler, J. & Stuart, E.A. Improving propensity score weighting using machine learning. *Statistics in medicine* **29**, 337-346 (2010).
62. Pirracchio, R., Petersen, M.L. & van der Laan, M. Improving propensity score estimators' robustness to model misspecification using super learner. *American journal of epidemiology* **181**, 108-119 (2015).
63. Westreich, D., Lessler, J. & Funk, M.J. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology* **63**, 826-833 (2010).
64. Kunzel, S.R., Sekhon, J.S., Bickel, P.J. & Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* **116**, 4156-4165 (2019).
65. Wager, S. & Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical ASsociation* **113**, 1228-1242 (2018).
66. Chernozhukov, V.*, et al.* Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21**, C1-C68 (2018).
67. van der Laan, M.J. & Gruber, S. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The international journal of biostatistics* **8**(2012).
68. Nie, X. & Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108**, 299-319 (2021).
69. Funk, M.J.*, et al.* Doubly robust estimation of causal effects. *American journal of epidemiology* **173**, 761-767 (2011).
70. Kennedy, E.H. Semiparametric theory and empirical processes in causal inference. in *Statistical causal inferences and their applications in public health research* 141-167 (Springer, 2016).
71. Naimi, A.I. & Kennedy, E.H. Nonparametric double robustness. *arXiv preprint arXiv:1711.7137* (2017).
72. Naimi, A.I., Mishler, A.E. & Kennedy, E.H. Challenges in obtaining valid causal effect estimates with machine learning algorithms. *arXiv preprint arXiv:1711.07137* (2017).
73. Zivich, P.N. & Breskin, A. Machine learning for causal inference: on the use of cross-fit estimators. *Epidemiology* (2021).
74. Hernan, M.A., Hsu, J. & Healy, B.A. A second chance to get causal inference right: A classification of data science tasks. *CHANCE* **32**, 42-49 (2019).
75. Holland, P.W. Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945-960 (1986).
76. Westreich, D., Cole, S.R., Funk, M.J., Brookhart, M.A. & Sturmer, T. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and drug safety* **20**, 317-320 (2011).
77. Wyss, R.*, et al.* The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *American journal of epidemiology* **180**, 645-655 (2014).
78. OHDSI. *The book of OHDSI: Observational health data sciences and informatics*, (2019).
79. Shimoni, Y., Karavani, E., Ravid, S. & al., e. An evaluation toolkit to guide model selection and cohort definition in causal inference. *arXiv preprint* (2019).

80. Hosmer, D.W., Hosmer, T., Le Cessie, S. & Lemeshow, S. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine* **16**, 965-980 (1997).
81. Steyerberg, E.W.*, et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128-138 (2010).
82. Wooldridge, J. Should Instrumental Variables Be Used As Matching Variables? *East Lansing, MI: Michigan State University* **(http://www.msu.edu/~ec/faculty/wooldridge/current%20research/treat1r6.pdf)**(2009).
83. Beaudoin, F.L.*, et al.* Persistent pain after motor vehicle collision: comparative effectiveness of opioids vs nonsteroidal antiinflammatory drugs prescribed from the emergency department-a propensity matched analysis. *Pain* **158**, 289-295 (2017).
84. Ozery-Flato, M., Goldschmidt, Y., Shaham, O. & al., e. Framework for identifying drug repurposing candidates from observational healthcare data. *medRxiv.20018366* (2020).
85. Ozery-Flato, M., Goldschmidt, Y., Shaham, O., Ravid, S. & Yanover, C. Framework for identifying drug repurposing candidates from observational healthcare data. *JAMIA Open* **3**, 536-544 (2020).
86. D'Amour, A., Ding, P., Feller, A., Lei, L. & Sekhon, J. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics* **221**, 644-654 (2021).
87. Franklin, J.M., Rassen, J.A., Ackermann, D., Bartels, D.B. & Schneeweiss, S. Metrics for covariate balance in cohort studies of causal effects. *Statistics in medicine* **33**, 1685-1699 (2014).
88. Alaa, A. & Van Der Schaar, M. Validating causal inference models via influence functions. *PMLR* **97**, 191-201 (2019).
89. Franklin, J.M., Schneeweiss, S., Polinski, J.M. & Rassen, J.A. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational Statistics & Data Analysis* **72**, 219-226 (2014).
90. Schuler, A., Jung, K., Tibshirani, R., Hastie, T. & Shah, N. Synth-Validation: Selecting the best causal inference method for a given dataset. *arXiv:1711.00083* (2017).
91. Bahamyirou, A., Blais, L., Forget, A. & Schnitzer, M.E. Understanding and diagnosing the potential for bias when using machine learning methods with doubly robust causal estimators. *Statistical methods in medical research*, 962280218772065 (2018).
92. Lenis, D., Ackerman, B. & Stuart, E.A. Measuring Model Misspecification: Application to Propensity Score Methods with Complex Survey Data. *Comput Stat Data Anal* **128**, 48-57 (2018).
93. Petersen, M.L., Porter, K.E., Gruber, S., Wang, Y. & van der Laan, M.J. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research* **21**, 31-54 (2012).
94. Athey, S., Imbens, G.W., Metzger, J. & Munro, E.M. Using Wasserstein generative adversarial networks for the design of Monte-Carlo simulations. *NBER Working Paper No. 26566* (2019).
95. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M. & al., e. Generative adversarial networks. *arXiv preprint arXiv:1406.2661* (2014).
96. Lipsitch, M., Tchetgen Tchetgen, E. & Cohen, T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* **21**, 383-388 (2010).
97. Sarri, G.*, et al.* Framework for the synthesis of non-randomised studies and randomised controlled trials: a guidance on conducting a systematic review and meta-analysis for healthcare decision making. *BMJ Evid Based Med* (2020).
98. Delaney, J.A.C. & Seeger, J.D. Sensitivity Analysis. in *In: Developing a protocol for observational comparative effectiveness research: A user's guide* 145-159 (Agency for Healthcare Research and Quality (US). Rockville (MD), 2013).

99. Suchard, M.A., *et al.* Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *Lancet* **394**, 1816-1826 (2019).
100. Lash, T.L., *et al.* Good practices for quantitative bias analysis. *Int J Epidemiol* **43**, 1969-1985 (2014).

Figures



U = unobserved confounder
C = observed variable highly correlated with U (serves as proxy)
A = exposure
Y = outcome

Unobserved Confounder	Observed Proxy Measurement
Cognitive Impairment	donepezil use (captured in any claims database)
Healthy seeking behavior	Frequency of ICD-9 and CPT-4 codes for regular check up and screening visits and #PCP visits

Figure 1. Illustration and examples for 'proxy confounder' adjustment.

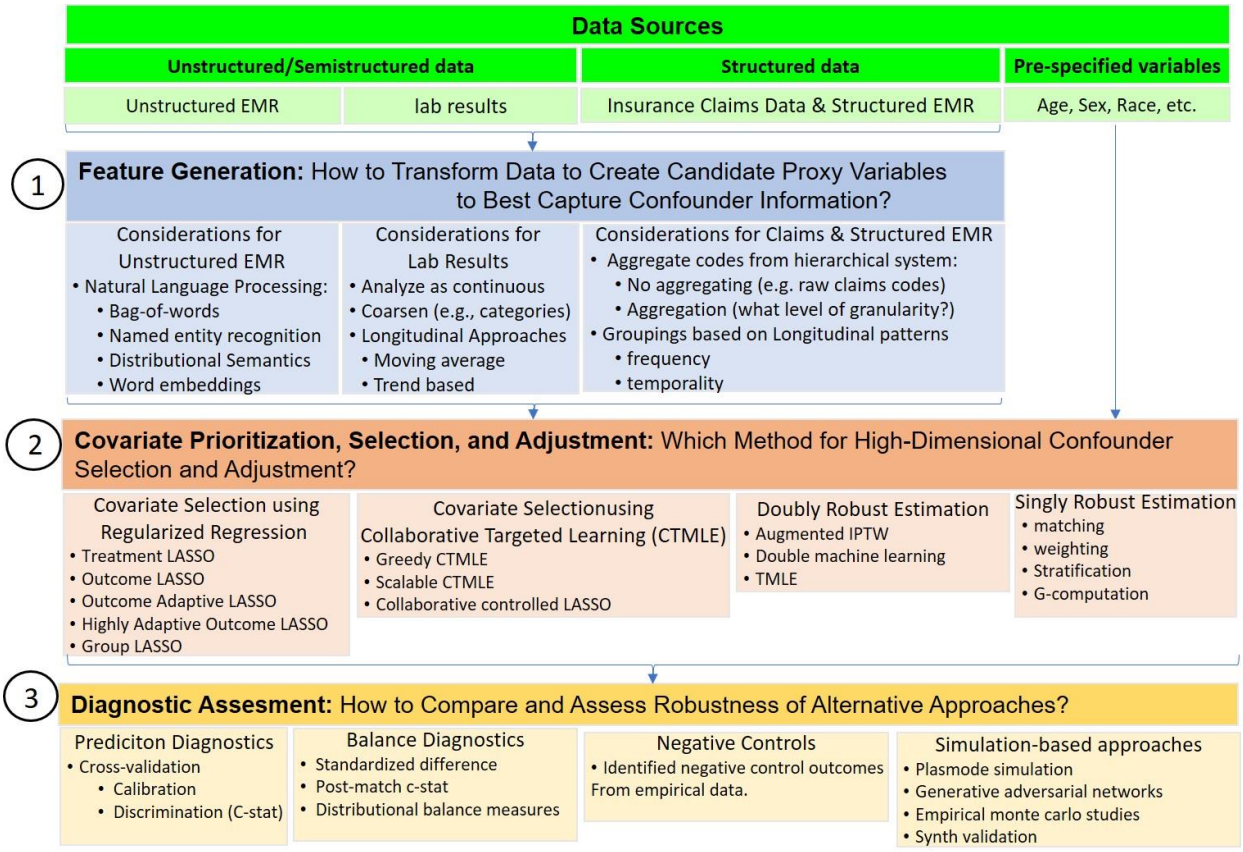


Figure 2. Considerations for different phases of high-dimensional proxy confounder adjustment.

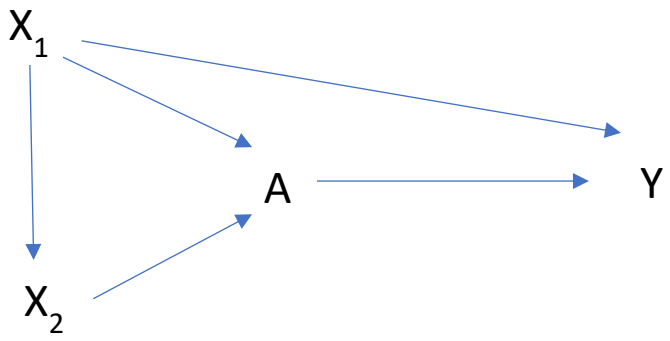


Figure 3. Causal diagram illustrating one scenario where the use of marginal empirical associations for confounder selection can result in over-adjusting for instrumental variables. In this causal structure, X_2 is marginally associated with both treatment and outcome, but is independent of the outcome after conditioning on X_1