Natural language processing for scalable feature engineering and ultra-high-dimensional confounding adjustment in healthcare database studies.

Richard Wyss[1], PhD, Jie Yang[1], PhD, Sebastian Schneeweiss[1], MD, PhD, Joseph M. Plasek[2], PhD, Li Zhou[2], PhD, Thomas Deramus[1], PhD, Janick G. Weberpals[1], PhD, Kerry Ngan[1], MS, Theodore N. Tsacogianis[1], MS, Kueiyu Joshua Lin[1,3], MD, PhD

Author Affiliations:
[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; [2]Division of General Internal Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; [3]Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.

Address for Correspondence:
Richard Wyss, PhD
Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School
1620 Tremont St. Suite 3030, Boston, MA 02120
Phone: (617) 278-0930; Fax: (617) 232-8602. Email: rwyss@bwh.harvard.edu

Source of Funding:
This project was funded by NIH RO1LM013204; additional funding was provided by PCORI ME-2022C1-25646.

Conflicts of Interest:
Dr. Schneeweiss is participating in investigator-initiated grants to the Brigham and Women's Hospital from Boehringer Ingelheim and UCB unrelated to the topic of this study. He is a consultant to Aetion Inc., a software manufacturer of which he owns equity. His interests were declared, reviewed, and approved by the Brigham and Women's Hospital in accordance with their institutional compliance policies. All other authors declare no competing interests for this work.

Running Head:
High-dimensional confounding control with NLP-generated features

Author Approval:
All authors of this paper have read and approved the final version submitted.

Data Availability:
Data used in the present study are not publicly available due to data use agreements.

ABSTRACT

Background: To improve confounding control in healthcare database studies, data-driven algorithms may empirically identify and adjust for large numbers of pre-exposure variables that indirectly capture information on unmeasured confounding factors ('proxy' confounders). Current approaches for high-dimensional proxy adjustment do not leverage free-text notes from EHRs. Unsupervised natural language processing (NLP) technology can scale to generate large numbers of structured features from unstructured notes.

Objective: To assess the impact of supplementing claims data analyses with large numbers of NLP generated features for high-dimensional proxy adjustment.

Methods: We linked Medicare claims with EHR data to generate three cohorts comparing different classes of medications on the 6-month risk of cardiovascular outcomes. We used various NLP methods to generate structured features from free-text EHR notes and used LASSO regression to fit several PS models that included different covariate sets as candidate predictors. Covariate sets included features generated from claims data only, and claims data plus NLP-generated EHR features.

Results: Including both claims codes and NLP-generated EHR features as candidate predictors improved overall covariate balance with standardized differences being <0.1 for all variables. While overall balance improved, the impact on estimated treatment effects was more nuanced with adjustment for NLP-generated features moving effect estimates further in the expected direction in two of the empirical studies but had no impact on the third study.

Conclusion: Supplementing administrative claims with large numbers of NLP-generated features for ultra-high-dimensional proxy confounder adjustment improved overall covariate balance and may provide a modest benefit in terms of capturing confounder information.

1. INTRODUCTION

Healthcare data generated from clinical practice, including electronic health records (EHR) and insurance claims, can complement randomized controlled trials to provide evidence on the effects of medical products to support clinical decisions.[1] However, estimating causal effects from these data sources, so called real-world evidence (RWE), can be challenging due to confounding caused by non-randomized treatment allocation and poorly measured information on comorbidities.[2,3] Approaches to mitigate confounding bias would ideally be based on causal diagrams and expert knowledge for variable selection.[4] Covariate adjustment based on expert knowledge alone, however, is not always adequate because some confounders may not be considered by researchers or not be directly measurable in such secondary healthcare data.

To improve confounding control in RWE studies, data-driven algorithms can be used to empirically identify and adjust for large numbers of pre-exposure variables that indirectly capture information on unmeasured or unspecified confounding factors ('proxy' confounders).[5-8] A growing literature has shown that supplementing investigator-specified variables with large numbers of empirically identified features can often improve confounding control compared to adjustment based on investigator-specified variables alone.[5-13] Current approaches for high-dimensional proxy adjustment, however, require data to be in a structured format (e.g., claims and structured EHR data), leaving unstructured EHR text information underutilized for confounding control. Leveraging this information can be challenging since patient-reported records are often recorded in free-text documents that are not readily analyzable at a large scale.

Recent work has demonstrated that unsupervised natural language processing (NLP) technology can scale to generate large numbers of structured features from unstructured clinical documents.[14,15] However, the added value of supplementing administrative claims data with large numbers of NLP-generated EHR features to improve high-dimensional proxy confounding

control in healthcare database studies remains unclear. Here, we use three empirical studies to investigate the added value of supplementing administrative claims with high-dimensional sets of NLP-generated features from time-indexed EHR notes for causal analyses. We consider several NLP methods for generating structured features from pre-exposure free-text clinical notes and observe how adjustment for different covariate sets impacts covariate balance and effect estimates after PS weighting. Our objective is to assess if unsupervised NLP tools can leverage information in EHR free-text notes to supplement claims data for improved high-dimensional proxy adjustment in healthcare database studies.

2. METHODS

Figure 1 summarizes the analytic pipeline we used for feature generation and high-dimensional proxy adjustment. Each step is described in detail below.

2.1. Data Source

We linked longitudinal claims data from the US Medicare system to the Research Patient Data Registry (RPDR) from 2007/01/01 to 2017/12/31. The RPDR data repository is based on all inpatient and outpatient activities of the Mass General Brigham (MGB), the largest healthcare delivery network in the greater Boston area including 2 academic medical centers, 14 community hospital and more than 100 satellite clinics. RPRD data records all medical records electronically, including diagnoses, procedures, test results (lab tests, imaging, biopsies, etc.), prescribing, and free text notes for all inpatient and outpatient services. Linking Medicare claims with the RPDR data repository helped reduce data leakage due to EHR-discontinuity (i.e., missing information from medical encounters provided outside the MGB network).[16]

2.2. Study population

Based on Medicare fee-for-service beneficiaries aged 65 years or older, we generated 3

cohorts:

1. *HTN Cohort*: A cohort of patients with hypertension comparing angiotensin-converting enzyme inhibitors (ACEi) vs beta blockers on the risk of hyperkalemia.
2. *Analgesics Cohort*: A cohort of patients using analgesics comparing NSAIDs vs Opioids in patients with a history of osteoarthritis (OA) in terms of risk of acute kidney injury (AKI).
3. *PPI Cohort*: A cohort of patients using proton pump inhibitors (PPIs) comparing high- vs low-dose PPI use in patients with a history of peptic ulcer in terms of risk of gastrointestinal (GI) bleeding.

For each empirical cohort, we identified individuals who initiated the treatment (or comparator) after no use of either the treatment or comparator medication in the previous year (new-user design).[17,18] The cohort entry date is the first record date of the medication use. To ensure that the study population had adequate information recorded in our data source, we required the study population to have at least 364 days of Medicare continuous enrolment in parts A (inpatient coverage), B (outpatient coverage), and D (prescription coverage). Characteristics for each study cohort are shown in Table 1. Further details of each generated cohort and study design are provided in the Supplemental Appendix.

These studies were chosen because they 1) present challenges in terms of confounding due to selective prescribing (confounding by indication); and 2) address clinically relevant questions. For example, while RCTs suggest no association between PPI dose and GI bleeding complications,[19,20] our unadjusted analysis showed that high-dose PPI was associated with a higher risk of GI bleeding when compared to low-dose PPI. This is likely due to confounding where individuals who have a history of more severe gastrointestinal bleeding were channeled to higher doses of PPI. For the HTN Cohort, it is common for patients at high risk of hyperkalemia to be channeled away from ACE inhibitors. As a result, we observed an apparent

protective effect against hyperkalemia associated with ACE inhibitors when compared with beta-blockers, which is likely confounded.[21] We would expect effective confounding adjustment would move the estimates toward the null.[22] Finally, for the Analgesics Cohort, previous studies suggest that NSAIDs are nephrotoxic and can cause AKI.[23,24] Consequently, patients with a history of kidney diseases are more likely to be prescribed an opioid potentially leading to strong confounding by indication showing an apparent decreased risk of AKI associated with NSAIDs when compared to opioids. This confounded association was observed in our unadjusted analysis and an effective confounding adjustment would be expected to move the estimates toward the harmful effects associated with NSAIDs.

2.3. Generating Structured Features from Unstructured EHR Free Text Notes.

We applied 5 different unsupervised NLP approaches ranging from basic statistical features to contextual deep-learning representation features, to generate structured features from unstructured free-text notes. For all approaches, we considered unstructured information during the 365 days before cohort entry. Additional details and code for NLP generation are available on GitHub:[25]

- *bag-of-words or bag-of-n-grams*: An n-gram is a sequence of consecutive items (in this case, words), where “unigram” refers to a single word, “bigram” refers to 2 consecutive words, and so on. Each document was tokenized and processed into unigrams and bigrams. We excluded stop words, i.e., words that occur frequently but convey little semantic meaning, such as articles (e.g., “a”, “the”) and prepositions (e.g., “in”, “on”).
- *MTERMS*: An automatic tool that extracts clinical terms or concepts from EHRs. Using lexical and term pattern matching methods, MTERMS can extract medical problems, medications, adverse reactions, and allergies.[26]
- *Word Embeddings (GloVe)**: Word embeddings are commonly used to convert each word into a continuous vector that can capture the semantic similarity between words. With word embeddings, if two words have similar neighboring words, then their

embedding vectors will be close to each other, reflecting their similarity, even if the two words do not share morphological similarity.[27]

- *Contextual Word Embeddings (BioBERT)*: An advanced contextual word embeddings model that takes the context of a word into consideration when generating the embedding. Compared to GloVe word embeddings, contextual word embeddings can provide unique embeddings for the same word under different contexts, ensuring fine-grained semantic information for the word. The BioBERT model was trained with biomedical text, which makes it specifically tuned for tasks in the biomedical domain, enhancing its ability to understand and process medical terminology more accurately.[28]
- *Sentence Embeddings (BERT):* Similar to contextual word embeddings, sentence embeddings with BERT use contextual representation but represent the entire sentence rather than individual words. It takes the sentence as the basic unit to generate candidate features. For the two BERT approaches we used the KMeans algorithm (MiniBatchKMeans) to cluster all the embeddings and represent patients with a binary cluster feature.[29]

It is important to emphasize that a key factor for the scalability of using NLP tools to generate large numbers of structured features for real-time clinical decision support is that researchers can remain agnostic to the format and content of the processed information. The NLP methods described above were used to automatically identify concepts or patterns from EHR free-text notes which were then fed into a LASSO regression to model the treatment choice as a propensity score for high-dimensional proxy adjustment. We were not concerned about the specific clinical meaning of the extracted features but only their potential for confounding the specified causal analyses. In other words, the application of NLP tools in this setting does not require time-intensive training data creation and reference-standard annotation through manual chart review.

2.4. Data Screening, Propensity Score Development, and Confounding Adjustment

2.4.1. Data screening. For each empirical study, baseline covariates (features) included several dozen researcher-specified variables (which included demographic variables like age and sex), thousands of additional claims codes, and thousands of NLP-generated features from EHR. The researcher-specified variables were constructed using claims data only that were determined based on diagnostic and procedural codes along with all NDC drug codes. NLP-generated features only included EHR information in the year before cohort entry. We then screened for and excluded from the analysis claims codes and NLP-generated features with a prevalence < 0.01.

We then screened features that were strongly associated with treatment choice but not or only weakly with the outcome and thus behaved like instrumental variables (IVs). These were excluded since adjusting for IVs harms the properties of causal effect estimates.[30-32] We ranked each variable based on its marginal correlation with treatment and manually examined the top-ranked variables. Those of the top-ranked variables that were not deemed to be risk factors for the outcome based on clinical domain knowledge were excluded from the analysis. While not comprehensive, we were able to examine the strongest predictors of treatment to exclude instruments that are most likely to be harmful to causal analyses (Supplemental Appendix).

2.4.2. Propensity score models. We used Lasso regression to fit 8 different propensity score models, each containing one of the following covariate sets as candidate predictors:

1) Researcher specified variables only,
2) Researcher specified variables + automated feature extraction from claims codes,
3) Researcher specified variables + automated feature extraction from claims codes + EHR codes
4) Researcher-specified variables + automated feature extraction from claims codes + EHR codes + NLP features generated using Ngrams.
5) Researcher-specified variables + automated feature extraction from claims codes + EHR

codes + NLP features generated using MTERMS.

6) Researcher-specified variables + automated feature extraction from claims codes + EHR codes + NLP features generated using sentence embeddings (BERT).
7) Researcher-specified variables + automated feature extraction from claims codes + EHR codes + NLP features generated using word embeddings (BioBERT).
8) Researcher-specified variables + automated feature extraction from claims codes + EHR codes + NLP features generated using word embeddings (GloVe).

To avoid overfitting the Lasso PS models we randomly split the data into 10 equally sized non-overlapping groups, trained the Lasso model in 9 of the groups, which was then applied to the 10th group to assign predicted probabilities. This process was repeated 10 times for each model. Theory and simulations have shown that the use of cross-fitting reduces the impact of modeling spurious associations in the data when using data-adaptive algorithms for estimating the PS for causal inference.

We evaluated both the discrimination and calibration of each model by calculating the C-statistic (AUC) and negative log-likelihood (NLL) using 10-fold cross-validation. Note that high values for the C-statistic correspond with stronger discrimination while lower values for negative log-likelihood correspond with better calibration (more accurate predictions).

2.4.3. Propensity score analyses. We adjusted for PSs using the following weighting approaches:[33-35]

- Inverse Probability Treatment Weights (*IPTW*): $w = \frac{A}{ps} + \frac{(1-A)}{(1-ps)}$
- *Overlap Weights*: $w = A(1-ps) + (1-A)ps$
- *Matching Weights*: $w = A\left[\frac{\min(ps,\ 1-ps)}{ps}\right] + (1-A)\left[\frac{\min(ps,\ 1-ps)}{(1-ps)}\right]$

With *A* being the binary treatment choice and *ps* the estimated propensity score.

For each example study, we used a weighted Cox regression to estimate the hazard ratio

and 95% confidence intervals over a 6-month follow-up window with an intent-to-treat analysis. We assessed the covariate balance between treatment groups by calculating the standardized differences after PS weighting. For each PS weighted analysis, balance was assessed on variables within the covariate set that was available to be included as candidate predictors in the PS model, as well as all variables across all covariate sets even though they were not eligible to be included in a given model. The latter was done to also assess the balance achieved in NLP-derived features not included in the adjustment. Covariates with standardized differences < 0.1 were assumed to provide adequate confounding control.[36]

3. RESULTS

All results for PS analyses shown here used matching weights for covariate adjustment, results using other weights are shown in the Supplemental Appendix.

3.1. Predicting Treatment Choice and PS Overlap

The number of candidate predictors within each covariate set, the number of candidate predictors selected by each of the 8 LASSO PS model, and their prediction diagnostics are shown in **Table 2**.

Prediction diagnostics show that PS models that only included researcher-specified variables resulted in the poorest predictive performance in terms of the C-statistic (AUC) and NLL. Supplementing researcher-specified variables with large numbers of features from claims and NLP-generated EHR features improved predictive performance. Overall, PS models that included Covariate Set 4 as candidate predictors (researcher-specified variables + claims codes + NLP-generated features using ngrams) resulted in the best predictive performance for both the AUC and NLL in comparison with the other covariate sets (**Table 2**).

The improved predictive performance when supplementing researcher-specified variables with large numbers of claims codes and NLP-generated features is reflected in the propensity score distributions plotted across treatment groups (**Figure 1 and Supplemental**

Figures S1 an S2). As predictive performance improved, separation in the distribution of the propensity score across treatment increased. PS Model 4, which included Covariate Set 4 as candidate predictors, had the strongest predictive performance and resulted in the largest separation across treatment groups for each of the 3 studies (Plot d in **Figure 1 and Supplemental Figures S1 an S2**). As previous studies have shown, however, PS models with better discrimination (C-statistic) does not necessarily correspond with better confounding control.[37,38]

3.2. Covariate Balance

Figures 3-5 show the absolute standardized differences of each covariate across treatment groups for Studies 1-3, respectively, before and after matching weights were used for covariate adjustment. Orange dots represent variables that were not available as candidate predictors for the given PS model. For example, the black dots in Panel a) represent the candidate predictors available for Model 1 which was Covariate Set 1 (in this case only researcher-specified variables), while the orange dots in Panel a) represent all other variables.

All PS models performed well in terms of balancing the covariates that were available as candidate predictors for the given model (black dots), with absolute standardized differences being <0.1 for all models across all studies. However, in terms of balancing covariates that were not available as candidate predictors for the given model (orange dots), results show that the models that did not include NLP-generated EHR features as candidate predictors in the LASSO PS models (PS Models 1-3) resulted in large imbalances in the NLP features after PS weighting. This is reflected in Plots a) through c) in **Figures 3-5** with a large proportion of the orange dots having standardized differences >0.1. Among the models that included NLP-generated features (PS Models 4-8), Model 4 performed best across all studies in terms of balancing covariates that were not available as candidate predictors (orange dots) with absolute standardized differences being <0.1 for all covariates in studies 1 and 2, and <0.1 for all but 1

covariate in study 3. Overall patterns for covariate balance were similar when using overlap weights for covariate adjustment (**Supplemental Figures S3-S5**). However, when using inverse probability weighting, we found that it was difficult to consistently balance covariates for the majority of models across all studies, regardless of whether or not they were available as candidate predictors (**Supplemental Figures S6-S8**).

3.3. Treatment Effect Estimates

Table 2 also shows the estimated hazard ratios and 95% confidence intervals after using PS matching weights for covariate adjustment. The estimated treatment effects after PS-weighting moved the hazard ratios in the expected direction for each study based on previous trial and clinical evidence discussed above. For example, trial evidence suggests that there is a null relationship between PPI dose and GI bleeding (expected HR close to 1). The unadjusted estimate comparing hi- vs. low-dose PPI was strongly impacted by confounding by indication with the hazard ratio being 2.77 (2.05, 3.74). Adjustment for each covariate set resulted in substantial movement in the estimated treatment effect towards the null with adjustment for Covariate Set 4 (PS Model 4) resulting in an estimated effect with the 95% CI including the null (HR: 1.38; 95% CI: 0.95, 1.97).

For the HTN cohort comparing ACE inhibitors versus beta-blockers on the risk of acute kidney injury, effect estimates were impacted by strong confounding by indication with the unadjusted estimate showing a protective effect for individuals initiating ACE inhibitors with a HR of 0.59 (0.52, 0.69). Adjustment for each covariate set moved the estimated effect towards the null with the 95% CI of all 8 models including the null. Adjustment for Covariate Set 5 (PS Model 5) resulted in an estimated effect that was furthest from the unadjusted estimate with a HR of 1.01 (0.83, 1.23).

For the Analgesics cohort comparing the effect of opioids vs. NSAIDs on the risk of risk of renal failure, results in **Table 2** show strong confounding by indication with the unadjusted

estimate showing a protective effect for individuals initiating NSAIDs with a HR of 0.41 (0.34, 0.50). Similar to the PPI study, adjustment for each covariate set moved the estimated effect towards the null with adjustment for Covariate Set 4 resulting in an estimated effect that was furthest from the unadjusted estimate with a HR of 0.81 (0.64, 1.02).

4. DISCUSSION

We investigated the impact of supplementing administrative claims with NLP-generated features from EHR free-text notes when using various propensity score-based weighting methods for high-dimensional proxy adjustment. After fitting LASSO PS models, we found that supplementing prespecified variables and claims codes with additional NLP-generated features using n-grams performed best in terms of treatment prediction. Interestingly, the overall covariate balance in all 3 empirical studies also improved and moved treatment effect estimates furthest from the unadjusted estimates in a direction consistent with prior expectation in 2 of the 3 studies. In our experiments, NLP tools more sophisticated than n-grams, such as MTERMS, and word or sentence embeddings, did not further improve covariate balance or confounding adjustment.

The use of NLP technology to leverage information from unstructured free-text EHR notes to improve confounding adjustment in healthcare database studies is a growing area of work.[39-43] To our knowledge, this study is the first to apply and compare the scalability and performance of several alternative NLP tools for purposes of high-dimensional proxy adjustment. These findings indicate that large-scale NLP-enabled feature generation is quite feasible and may provide additional confounder information when conducting high-dimensional proxy adjustment. This may result in improved effect estimates, although differences in effect estimates were modest in our example studies, and it supports automated approaches to confounding control.

When controlling for a large number of variables, we found that weighting methods that down-weight the tails of the PS distribution (matching weights and overlap weights) were more robust in balancing covariates compared to inverse probability weighting. This finding was consistent across all 3 empirical studies and is also consistent with previous studies that have compared alternative weighting approaches.[44] Because people with extreme PS values represent those for whom the confounding may be refractory, PS weighting approaches that down-weight the tails of the PS distribution may also mitigate unmeasured confounding that is likely stronger in the tails of the PS distribution.[45] However, the tradeoff is that the target population for matching weights and overlap weights cannot be defined a priori and is dependent on the distribution of the estimated PS. This creates additional challenges when comparing effect estimates across different PS models as it can be unclear if movement in the estimated treatment effect is due to better confounding control or heterogeneity in the treatment effect caused by changing target populations. Ideally, we would target the same population across all analyses by weighting to the same population (e.g., using inverse probability weights); however, across all 3 empirical studies we were only able to adequately balance covariates when using methods that down-weight the tails of the PS when controlling for high-dimensional sets of variables. Consequently, we did not focus on results from analyses using inverse probability weights.

Some additional limitations deserve attention. First, other factors including random chance and hidden biases not captured through PS weighting could also contribute to unpredictable movement in estimated effects and a true ‘gold standard’ is never known with certainty. Consequently, as discussed previously, methodological comparisons of estimated treatment effects in real-world data examples are inherently limited. However, such comparisons are widely used and can be informative, particularly when consistent trends are observed across multiple studies increasing confidence in the robustness of findings.

Second, while we explored several NLP methods for generating structured features from free-text notes, there are many NLP tools that were not considered or customized to our EHR data due to resource constraints. Future work could explore additional NLP tools such as large language models, for extracting confounder information from EHR data, as well as training embeddings on our local EHR data for better representing features in EHRs. In addition, different ways of modeling the NLP-generated features may also improve the performance. For example, in this study we simply created dichotomous variables from the n-gram and other NLP output. Other options could consider modeling the NLP output as categorical or continuous features, such as using the term frequency and the reciprocal document frequency (tf-idf) of the n-gram output or the weights of embedding clusters. These approaches help integrate the fine-grained contribution weights of NLP-generated features. In this study, we chose to focus on generating binary features from the NLP output to simplify the functional form of the PS model and reduce the likelihood of model misspecification. While more flexible nonparametric models could potentially be used to model complex relationships between continuous/categorical features and treatment, both theory and simulations have shown that using flexible nonparametric models for PS estimation comes at a cost of slow convergence rates, which can harm the properties of causal estimators, particularly in high-dimensional models.[46-48] A thorough comparison of modeling more complex NLP-output with flexible nonparametric models is beyond our scope.

In conclusion, we found that off-the-shelf NLP tools can scale to generate large numbers of structured features from free-text EHR notes. We found that supplementing administrative claims with large numbers of NLP-generated EHR features improved overall covariate balance in PS-weighted analyses, but the observed impact on estimated treatment effects was incremental and more complex NLP tools had no advantage over simple n-grams.

FIGURES

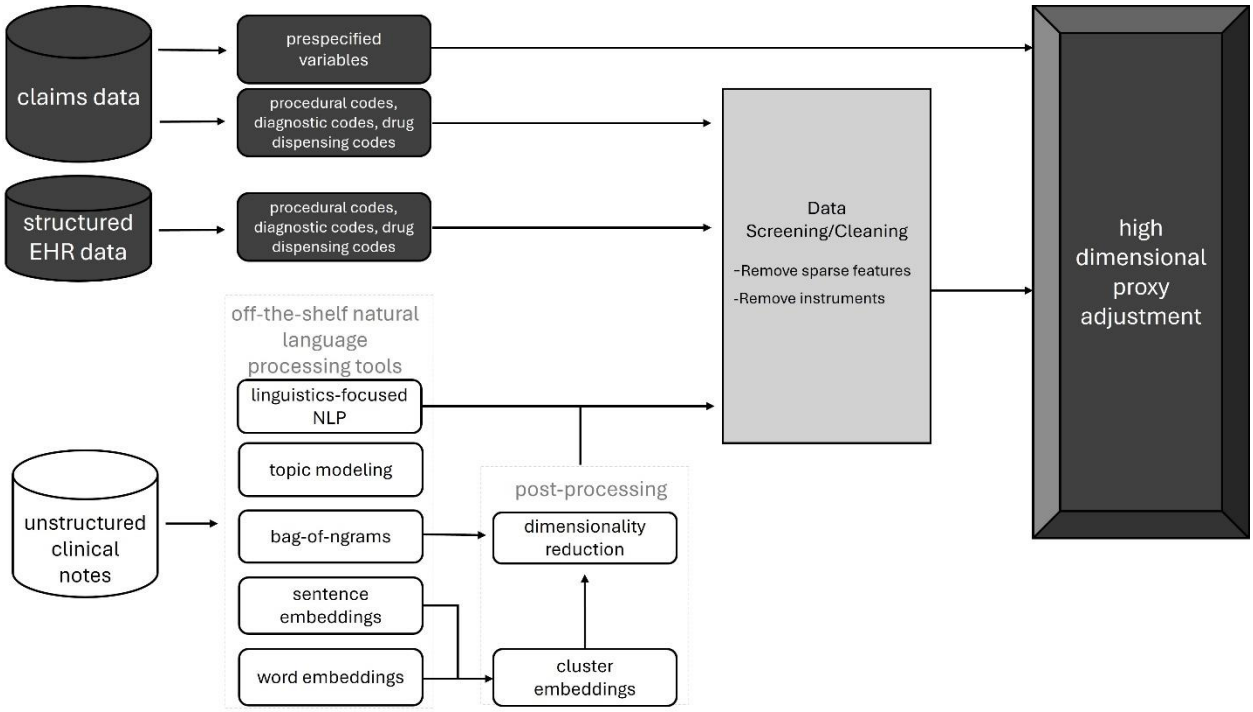

Figure 1. Analytic pipeline illustrating the process for feature engineering of structured and unstructured data for high-dimensional proxy confounder adjustment.

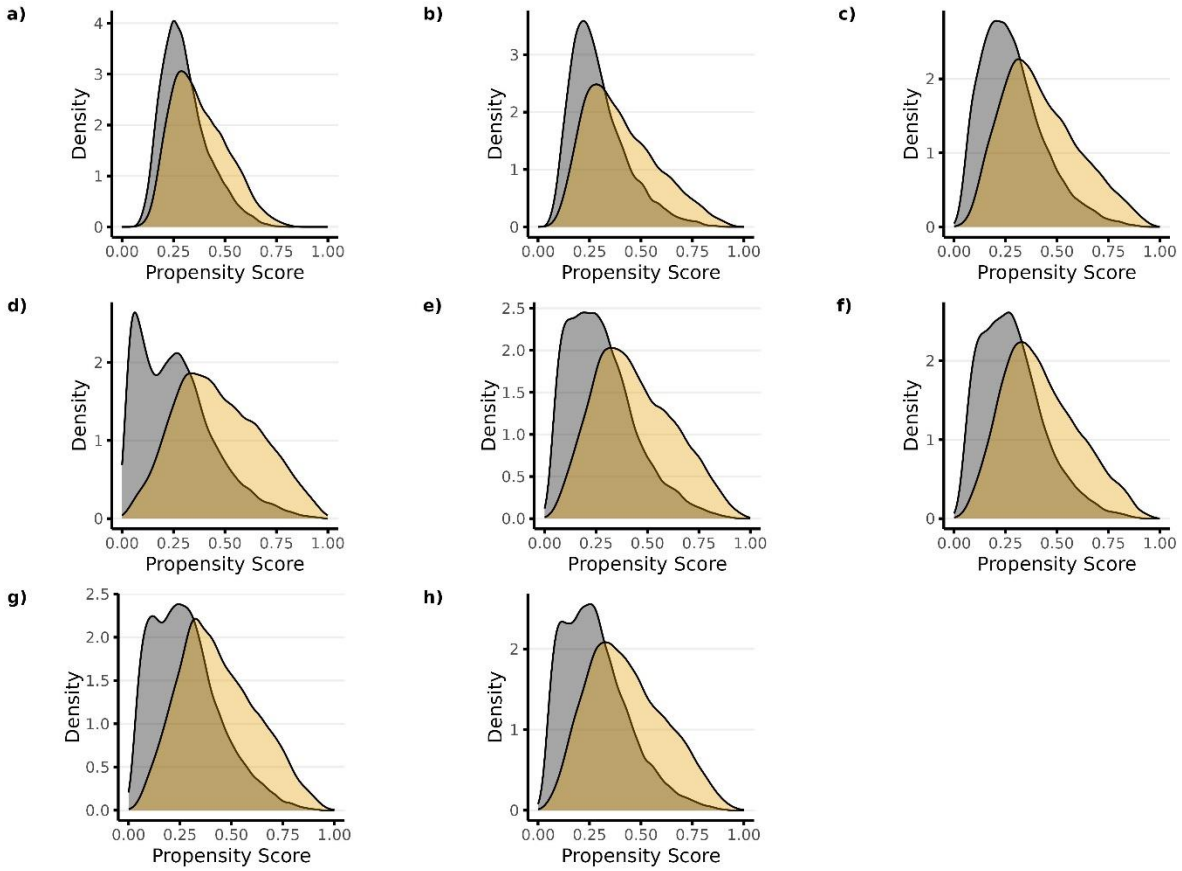

Figure 2. Propensity score distributions plotted across treatment groups for each propensity score model in Study 1 (high-dose PPI vs low-dose PPI). Panels a) through h) correspond to PS models 1-8, respectively. Candidate predictors for each model included the following: Model 1—researcher-specified variables only; Model 2—researcher-specified variables + claims codes; Model 3—research-specified variables + claims codes + EHR codes; Model 4—research specified variables + claims codes + EHR codes + NLP-generated features using ngrams; Model 5—researcher-specified variable + claims codes + EHR codes + NLP-generated features using mterms; Model 6—researcher-specified variable + claims codes + EHR codes + NLP-generated features using sentence embeddings; Model 7—researcher-specified variable + claims codes + EHR codes + NLP-generated features using word embeddings with bert; Model 8—researcher-specified variables + claims codes + EHR codes + NLP-generated features using word embeddings with glove.

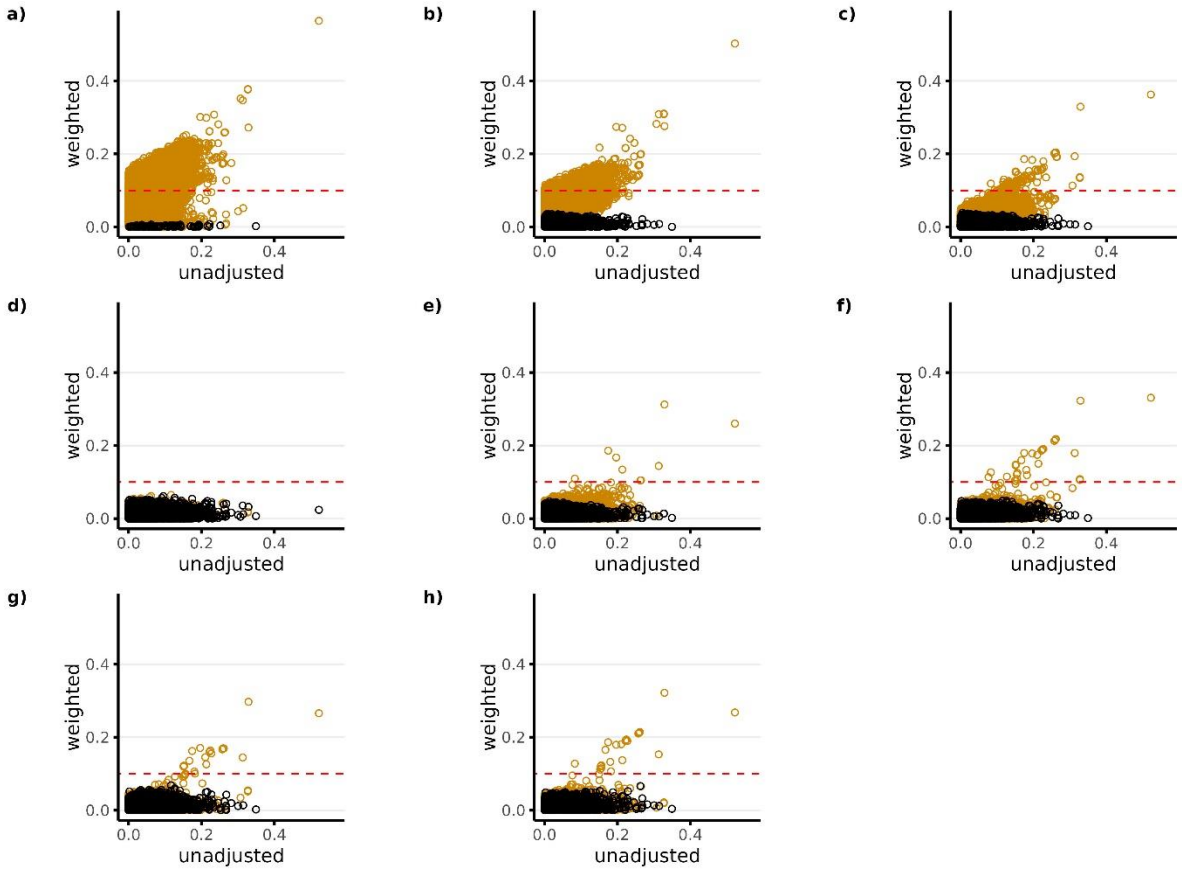

Figure 3. Standardized covariate difference across treatment groups before and after PS weighting (matching weights) for Study 1 (High-dose PPI vs Low-dose PPI). Panels a) through h) correspond to PS models 1-8, respectively. Black dots in each panel represent the candidate predictors that were available for the given PS model. Orange dots represent all other variables that were not available as candidate predictors for the given PS model. The candidate predictors available for each model (black dots) include the following: Model 1—researcher-specified variables only; Model 2—researcher-specified variables + claims codes; Model 3—research-specified variables + claims codes + EHR codes; Model 4—research specified variables + claims codes + EHR codes + NLP-generated features using ngrams; Model 5—researcher-specified variable + claims codes + EHR codes + NLP-generated features using mterms; Model 6—researcher-specified variable + claims codes + EHR codes + NLP-generated features using sentence embeddings; Model 7—researcher-specified variable + claims codes + EHR codes + NLP-generated features using word embeddings with bert; Model 8—researcher-specified variables + claims codes + EHR codes + NLP-generated features using word embeddings with glove. The red horizontal dotted line represents the largest level of balance after PS weighting that was considered adequate (<0.1).

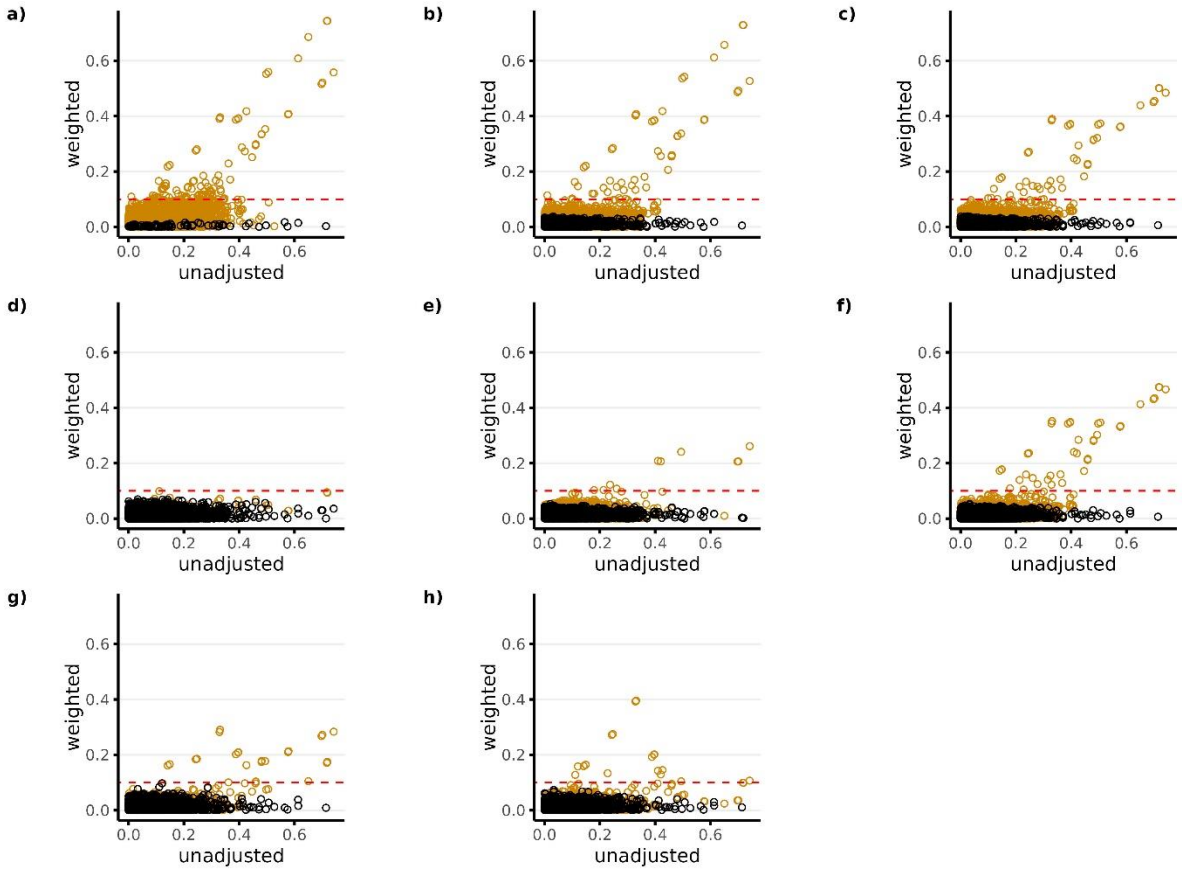

Figure 4. Standardized covariate difference across treatment groups before and after PS weighting (matching weights) for Study 2 (ACEIs vs Beta Blockers). Panels a) through h) correspond to PS models 1-8, respectively. Black dots in each panel represent the candidate predictors that were available for the given PS model. Orange dots represent all other variables that were not available as candidate predictors for the given PS model. The candidate predictors available for each model (black dots) include the following: Model 1—researcher-specified variables only; Model 2—researcher-specified variables + claims codes; Model 3—research-specified variables + claims codes + EHR codes; Model 4—research specified variables + claims codes + EHR codes + NLP-generated features using ngrams; Model 5—researcher-specified variable + claims codes + EHR codes + NLP-generated features using mterms; Model 6—researcher-specified variable + claims codes + EHR codes + NLP-generated features using sentence embeddings; Model 7—researcher-specified variable + claims codes + EHR codes + NLP-generated features using word embeddings with bert; Model 8—researcher-specified variables + claims codes + EHR codes + NLP-generated features using word embeddings with glove. The red horizontal dotted line represents the largest level of balance after PS weighting that was considered adequate (<0.1).

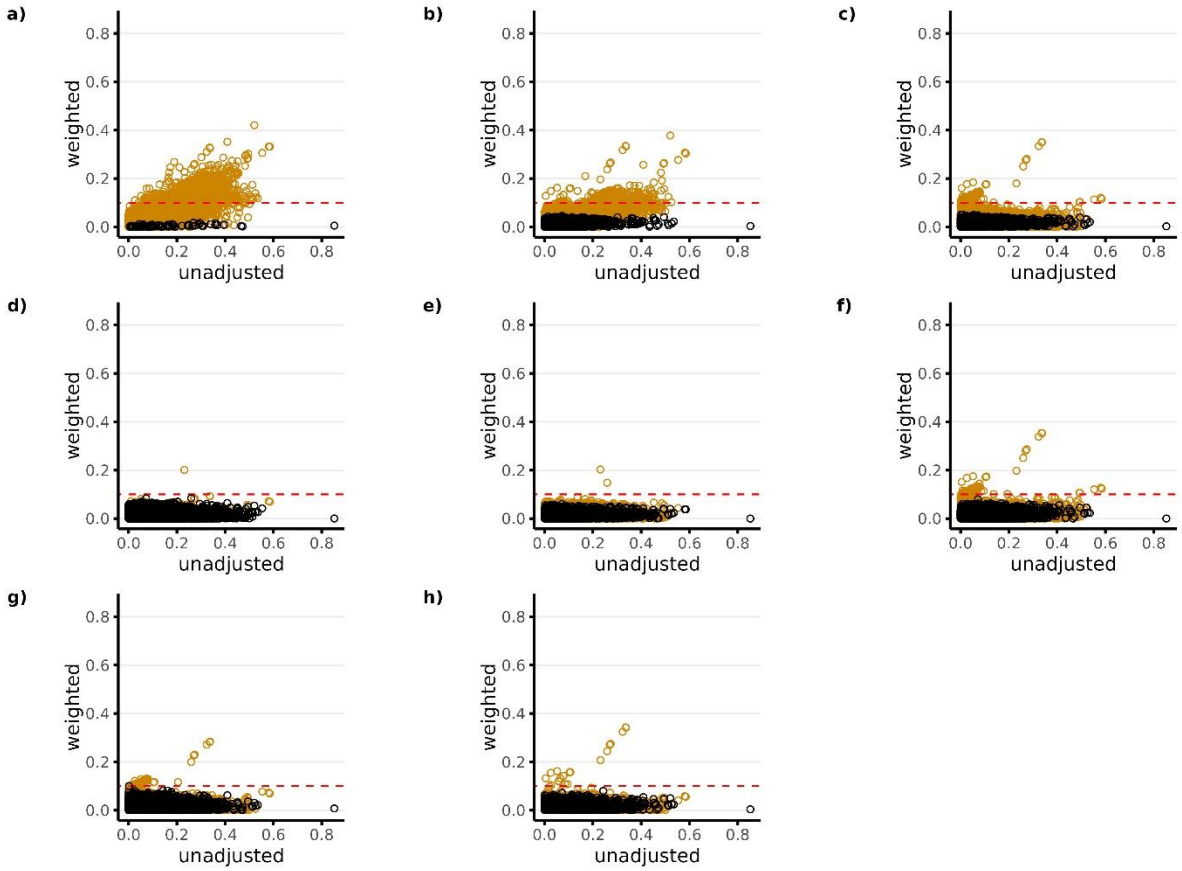

Figure 5. Standardized covariate difference across treatment groups before and after PS weighting (matching weights) for Study 3 (NSAIDs vs Opioids). Panels a) through h) correspond to PS models 1-8, respectively. Black dots in each panel represent the candidate predictors that were available for the given PS model. Orange dots represent all other variables that were not available as candidate predictors for the given PS model. The candidate predictors available for each model (black dots) include the following: Model 1—researcher-specified variables only; Model 2—researcher-specified variables + claims codes; Model 3—research-specified variables + claims codes + EHR codes; Model 4—research specified variables + claims codes + EHR codes + NLP-generated features using ngrams; Model 5—researcher-specified variable + claims codes + EHR codes + NLP-generated features using mterms; Model 6—researcher-specified variable + claims codes + EHR codes + NLP-generated features using sentence embeddings; Model 7—researcher-specified variable + claims codes + EHR codes + NLP-generated features using word embeddings with bert; Model 8—researcher-specified variables + claims codes + EHR codes + NLP-generated features using word embeddings with glove. The red horizontal dotted line represents the largest level of balance after PS weighting that was considered adequate (<0.1).

Table 1. Characteristics of study cohorts

Study	Example study	Treatment	Comparator	Outcome	N		
					Treatment (%)	Outcome (%)	Study Size
1.	**PPI Cohort**	High-dose PPI	Low-dose PPI	GI Bleeding	8,590 (33%)	175 (0.7%)	26,056
2.	**HTN Cohort**	ACEIs	Beta Blockers	Hyperkalemia	9,690 (36%)	675 (4.0%)	26,861
3.	**Analgesics Cohort**	NSAIDs	Opioids	AKI	5,750 (34%)	920 (3.4%)	16,990

Table 2. Prediction diagnostics and hazard ratio estimates when using matching weights for covariate adjustment

Example study	Model[a]	# Candidate Predictors	# Predictors Selected	AUC	NLL	HR	95% CI
PPI Cohort							
	Unadjusted	-----	-----	-----	-----	2.77	2.05, 3.74
	Model 1	84	84	0.66	0.60	1.80	1.31, 2.48
	Model 2	2,152	489	0.69	0.58	1.40	1.01, 1.95
	Model 3	3,060	523	0.72	0.57	1.43	1.03, 2.01
	Model 4	35,634	806	0.77	0.52	1.38	0.95, 1.97
	Model 5	6,870	554	0.74	0.55	1.43	1.02, 2.01
	Model 6	12,682	637	0.72	0.56	1.46	1.04, 2.05
	Model 7	13,022	809	0.74	0.55	1.40	1.00, 1.97
	Model 8	6,051	610	0.73	0.55	1.54	1.08, 2.16
HTN Cohort							
	Unadjusted	-----	-----	-----	-----	0.59	0.52, 0.69
	Model 1	82	82	0.80	0.51	0.93	0.79, 1.11
	Model 2	1,622	426	0.83	0.49	0.96	0.80, 1.15
	Model 3	2,278	535	0.86	0.45	0.99	0.83, 1.20
	Model 4	32,424	644	0.89	0.39	0.96	0.79, 1.16
	Model 5	5,475	663	0.89	0.39	1.01	0.83, 1.23
	Model 6	11,376	969	0.86	0.45	0.96	0.79, 1.16
	Model 7	12,176	847	0.88	0.42	0.99	0.82, 1.20
	Model 8	5,304	599	0.88	0.41	0.97	0.79, 1.17
Analgesic Cohort							
	Unadjusted	-----	-----	-----	-----	0.41	0.34, 0.50
	Model 1	77	77	0.78	0.52	0.76	0.61, 0.95
	Model 2	1,545	462	0.82	0.49	0.76	0.61, 0.96
	Model 3	2,029	454	0.84	0.46	0.79	0.63, 0.98
	Model 4	29,394	746	0.85	0.45	0.81	0.64, 1.02
	Model 5	4,905	512	0.85	0.45	0.78	0.62, 0.99
	Model 6	10,822	432	0.84	0.47	0.79	0.63, 0.99
	Model 7	11,881	1035	0.90	0.37	0.79	0.61, 1.03
	Model 8	4,937	655	0.84	0.46	0.80	0.64, 1.01

[a]**Model 1)** Researcher specified variables only, **Model 2)** Researcher specified variables + automated feature extraction from claims codes, **Model 3)** Researcher specified variables + automated feature extraction from claims codes + EHR codes, **Model 4)** Researcher-specified variables + automated feature extraction from claims codes + EHR codes + NLP features generated using Ngrams, **Model 5)** Researcher-specified variables + automated feature extraction from claims codes + EHR codes + NLP features generated using MTERMS, **Model 6)** Researcher-specified variables + automated feature extraction from claims codes + EHR codes + NLP features generated using sentence embeddings (BERT), **Model 7)** Researcher-specified variables + automated feature extraction from claims codes + EHR codes + NLP features generated using word embeddings (BioBERT), **Model 8)** Researcher-specified variables + automated feature extraction from claims codes + EHR codes + NLP features generated using word embeddings (GloVe).

REFERENCES

1. Corrigan-Curay J, Sacks L, Woodcock J. Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. *JAMA*. Sep 04 2018;320(9):867-868. doi:10.1001/jama.2018.10136
2. Streeter AJ, Lin NX, Crathorne L, et al. Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. *J Clin Epidemiol*. Jul 2017;87:23-34. doi:10.1016/j.jclinepi.2017.04.022
3. US Food and Drug Administration. Framework for FDA's real world evidence program. Updated December 2018. 2025. https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/RealWorldEvidence/UCM627769.pdf
4. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol*. Mar 2019;34(3):211-219. doi:10.1007/s10654-019-00494-6
5. Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clin Epidemiol*. 2018;10:771-788. doi:10.2147/CLEP.S166545
6. 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*. Jul 2009;20(4):512-22. doi:10.1097/EDE.0b013e3181a663cc
7. Wyss R, Yanover C, El-Hay T, et al. Machine learning for improving high-dimensional proxy confounder adjustment in healthcare database studies: An overview of the current literature. *Pharmacoepidemiol Drug Saf*. Sep 2022;31(9):932-943. doi:10.1002/pds.5500
8. Zhang L, Wang Y, Schuemie MJ, Blei DM, Hripcsak G. Adjusting for indirectly measured confounding using large-scale propensity score. *J Biomed Inform*. Oct 2022;134:104204. doi:10.1016/j.jbi.2022.104204
9. Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *Int J Epidemiol*. 12 01 2018;47(6):2005-2014. doi:10.1093/ije/dyy120
10. Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA. Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philos Trans A Math Phys Eng Sci*. Sep 13 2018;376(2128)doi:10.1098/rsta.2017.0356
11. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc Natl Acad Sci U S A*. 03 13 2018;115(11):2571-2577. doi:10.1073/pnas.1708282114
12. Guertin JR, Rahme E, LeLorier J. Performance of the high-dimensional propensity score in adjusting for unmeasured confounders. *Eur J Clin Pharmacol*. Dec 2016;72(12):1497-1505. doi:10.1007/s00228-016-2118-x
13. Guertin JR, Rahme E, Dormuth CR, LeLorier J. Head to head comparison of the propensity score and the high-dimensional propensity score matching methods. *BMC Med Res Methodol*. Feb 19 2016;16:22. doi:10.1186/s12874-016-0119-1
14. Wyss R, Plasek JM, Zhou L, et al. Scalable Feature Engineering from Electronic Free Text Notes to Supplement Confounding Adjustment of Claims-Based Pharmacoepidemiologic Studies. *Clin Pharmacol Ther*. Apr 2023;113(4):832-838. doi:10.1002/cpt.2826
15. Rassen J, Wahl P, Angelino E, Seltzer M, Rosenman M, Schneeweiss S. Automated use of electronic health record text data to improve validity in pharmacoepidemiology studies. *Pharmacoepidemiol Drug Saf*. 2013;22(S1):376.
16. Lin KJ, Singer DE, Glynn RJ, Murphy SN, Lii J, Schneeweiss S. Identifying Patients With High Data Completeness to Improve Validity of Comparative Effectiveness Research in Electronic Health Records Data. *Clin Pharmacol Ther*. May 2018;103(5):899-905. doi:10.1002/cpt.861

17. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol*. Nov 01 2003;158(9):915-20. doi:10.1093/aje/kwg231
18. Lund JL, Richardson DB, Stürmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. *Curr Epidemiol Rep*. Dec 2015;2(4):221-228. doi:10.1007/s40471-015-0053-5
19. Wu LC, Cao YF, Huang JH, Liao C, Gao F. High-dose vs low-dose proton pump inhibitors for upper gastrointestinal bleeding: a meta-analysis. *World J Gastroenterol*. May 28 2010;16(20):2558-65. doi:10.3748/wjg.v16.i20.2558
20. Yang M, He M, Zhao M, et al. Proton pump inhibitors for preventing non-steroidal anti-inflammatory drug induced gastrointestinal toxicity: a systematic review. *Curr Med Res Opin*. Jun 2017;33(6):973-980. doi:10.1080/03007995.2017.1281110
21. Tomson C, Tomlinson LA. Stopping RAS Inhibitors to Minimize AKI: More Harm than Good? *Clin J Am Soc Nephrol*. Apr 05 2019;14(4):617-619. doi:10.2215/CJN.14021118
22. Brar S, Ye F, James MT, et al. Association of Angiotensin-Converting Enzyme Inhibitor or Angiotensin Receptor Blocker Use With Outcomes After Acute Kidney Injury. *JAMA Intern Med*. Dec 01 2018;178(12):1681-1690. doi:10.1001/jamainternmed.2018.4749
23. Klomjit N, Ungprasert P. Acute kidney injury associated with non-steroidal anti-inflammatory drugs. *Eur J Intern Med*. Jul 2022;101:21-28. doi:10.1016/j.ejim.2022.05.003
24. Kim S, Joo KW. Electrolyte and Acid-base disturbances associated with non-steroidal anti-inflammatory drugs. *Electrolyte Blood Press*. Dec 2007;5(2):116-25. doi:10.5049/EBP.2007.5.2.116
25. Jie Y. Confounding Adjustment NLP. GitHub. https://github.com/jiesutd/Cofounding_adjustment_NLP
26. Zhou L, Plasek JM, Mahoney LM, et al. Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes. *AMIA Annu Symp Proc*. 2011;2011:1639-48.
27. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)2014. p. 1532-1543.
28. Lee J, Yoon W, Kim S, Kim D, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Feb 15 2020;36(4):1234-1240. doi:10.1093/bioinformatics/btz682
29. Turc I, Chang M, Lee K, Toutanova K. Well-read students learn better: On the importance of pre-training compact models. arXiv preprint arXiv:1908.089622019.
30. Wooldridge J. Should instrumental variables be used as matching variables? *Research in Economics*. 2016;70(2):232-237.
31. Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol*. Dec 01 2011;174(11):1213-22. doi:10.1093/aje/kwr364
32. Bhattacharya J, Vogt WB. Do instrumental variables belong in propensity scores? 2007:
33. Li F, Morgan KL, Zaslavsky MA. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*. 2018;113(521)
34. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *The international journal of biostatistics*. 2013;9(2):215-34. doi:10.1515/ijb-2012-0030
35. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*. Sep 15 2008;168(6):656-64. doi:10.1093/aje/kwn164
36. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. Nov 10 2009;28(25):3083-107. doi:10.1002/sim.3697

37. Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiol Drug Saf*. Mar 2011;20(3):317-20. doi:10.1002/pds.2074
38. 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*. Sep 15 2014;180(6):645-55. doi:10.1093/aje/kwu181
39. Feder A KK, Manzoor E, Pryzant R, Srdhar D, Wood-Doughty Z, et al. Causal inference in natural language processing: estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*. 2022;10:1138-1158.
40. Salmasian H, Freedberg DE, Friedman C. Deriving comorbidities from medical records using natural language processing. *J Am Med Inform Assoc*. Dec 2013;20(e2):e239-42. doi:10.1136/amiajnl-2013-001889
41. Zeng J, Gensheimer MF, Rubin DL, Athey S, Shachter RD. Uncovering interpretable potential confounders in electronic medical records. *Nat Commun*. Feb 23 2022;13(1):1014. doi:10.1038/s41467-022-28546-8
42. Malec SA, Wei P, Bernstam EV, Boyce RD, Cohen T. Using computable knowledge mined from the literature to elucidate confounders for EHR-based pharmacovigilance. *J Biomed Inform*. May 2021;117:103719. doi:10.1016/j.jbi.2021.103719
43. Afzal Z, Masclee GMC, Sturkenboom MCJM, Kors JA, Schuemie MJ. Generating and evaluating a propensity model using textual features from electronic medical records. *PLoS One*. 2019;14(3):e0212999. doi:10.1371/journal.pone.0212999
44. Li F, Thomas LE. Addressing Extreme Propensity Scores via the Overlap Weights. *Am J Epidemiol*. 01 01 2019;188(1):250-257. doi:10.1093/aje/kwy201
45. Stürmer T, Webster-Clark M, Lund JL, et al. Propensity Score Weighting and Trimming Strategies for Reducing Variance and Bias of Treatment Effect Estimates: A Simulation Study. *Am J Epidemiol*. Aug 01 2021;190(8):1659-1670. doi:10.1093/aje/kwab041
46. Naimi AI, Mishler AE, Kennedy EH. Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms. *Am J Epidemiol*. Jul 15 2021;doi:10.1093/aje/kwab201
47. Kennedy EH. Semiparametric theory and empirical processes in causal inference. *Statistical causal inferences and their applications in public health research*. Springer; 2016:141-167.
48. Zivich PN, Breskin A. Machine Learning for Causal Inference: On the Use of Cross-fit Estimators. *Epidemiology*. 05 01 2021;32(3):393-401. doi:10.1097/EDE.0000000000001332