Understanding the role of different proxy selection methods in High-Dimensional Propensity Score Framework - A Plasmode simulation study

Dr. Mohammad Ehsanul Karim MSc, PhD[1,2*] and Yang Lei[3]

[1*]School of Population and Public Health, University of British Columbia, Vancouver, Canada, V6T 1Z3, BC, 2206 East Mall.
[2]St. Paul's Hospital, Vancouver, Canada, V6Z 1Y6, BC, 588 - 1081 Burrard Street.
[3]Department of Statistics, University of British Columbia, Vancouver, Canada, V6T 1Z4, BC, Room 3182 Earth Sciences Building, 2207 Main Mall.

*Corresponding author(s). E-mail(s): ehsan.karim@ubc.ca;

Abstract

Purpose: We aims to evaluate various proxy selection methods within the context of high-dimensional propensity score (hdPS) analysis. The study focuses on assessing the performance of these methods, including alternative variable selection approaches, in selecting proxy variables for confounding adjustment compared to the traditional hdPS method that is rooted in Bross formula. The goal is to understand better the performance of these alternative methods in estimating treatment effects, and identify scenarios in which they may perform better. **Methods:** Using data from three cycles of the National Health and Nutrition Examination Survey (NHANES) spanning 2013-2018, we motivated the study by examining the association between obesity and diabetes. A plasmode simulation framework based on this data was employed to mimic real-world data structures. Simulations were conducted under three scenarios: Frequent Exposure and Outcome, Rare Exposure and Frequent Outcome, and Frequent Exposure and Rare Outcome. The performance of a variety of proxy selection methods—including tree-based methods, LASSO-based methods, and the Genetic Algorithm (GA)—was evaluated across these scenarios using standard simulation metrics. **Results:** XGBoost consistently demonstrated the lowest MSE and high coverage, making it a reliable method overall, although it did not always exhibit the lowest bias. In contrast, GA consistently showed the highest bias and MSE, with lower coverage and greater variability, indicating its unsuitability for accurate effect estimation. The kitchen sink model, Bross-based hdPS, and Hybrid hdPS methods performed moderately well, with low bias and moderate MSE, but varied in coverage across scenarios. Scenario-specific trends revealed that rare outcome scenarios yielded lower MSE and better precision, while rare exposure scenarios were associated with higher bias and MSE. **Conclusion:** Aside from GA, the performance of most other methods in your study appears to be comparable in each scenario, though with some variations depending on specific metrics. The findings underscore the importance of selecting appropriate methods for hdPS analysis based on the characteristics of the data, particularly the prevalence of exposure and outcome. While XGBoost excelled in overall accuracy and precision, the choice of method should be tailored to the specific epidemiological goal to optimize bias, coverage, and precision in estimating treatment effects.

Keywords: Machine learning, Propensity score, Deep learning, Causal inference

JEL Classification: C18

MSC Classification: 92D30 , 62P10

1 Background

need to write what hdps is, and why alternatives are considered

Introduction of hdPS: The use of medico-administrative datasets in research is frequently criticized due to the potential for the lack of key confounders. Estimating treatment effects utilizing such data sources may be subject to residual confounding. To address this issue, the high-dimensional propensity score (hdPS) algorithm was introduced in 2009. The hdPS method systematically incorporates some information that are associated with the unmeasured or mismeasured confounders as proxies, with the goal of reducing residual confounding bias.

Significance of alternatives: Since the introduction of hdPS algorithm, numerous extensions, rooted in machine learning and semi-parametric techniques, have been developed in order to improve the selection of proxy variables. The study aims to compare and evaluate the hdPS methods alongside the alternative proxy selection approaches to identify the ones that provide superior confounder adjustment, reduce bias, and enhance the precision of treatment effect estimation.

Aim: The aim of this research is to systematically evaluate and compare different proxy selection methods within the context of high-dimensional propensity score (hdPS) analysis. Specifically, the study focuses on assessing how these methods, including alternative variable selection approaches, perform in selecting proxy variables for confounding adjustment compared to the traditional Bross method within the hdPS framework. We seek to determine whether these alternative methods offer superior performance in estimating treatment effects compared to the default Bross formula.

2 Methods

Data and Simulation

Motivating Example: To explore the relationship between obesity and the risk of diabetes, we revisited this association using data from three cycles of the National Health and Nutrition Examination Survey (NHANES) covering the years 2013-2014, 2015-2016, and 2017-2018 [1]. This analysis was informed by a thorough review of the existing literature [2–5]. To identify relevant covariates,

we constructed a causal diagram based on established causal inference principles [6]. The covariates included in our analysis were carefully selected and categorized into Demographic, Behavioral, Health History, Access-related, and Laboratory variables [1]. While most of these variables were binary or categorical, the Laboratory variables were continuous.

Plasmode simulation: To rigorously assess the performance of the methods under consideration, we employed a plasmode simulation framework, which is particularly well-suited for reflecting real-world data structures and complexities [7]. This approach was modeled after the analytic dataset derived from NHANES and involved resampling from the observed covariates and exposure information (i.e., obesity) without altering them. By mirroring key aspects of an actual epidemiological study, this simulation framework offers a significant advantage over traditional Monte Carlo simulations, which often rely on idealized assumptions.

Simulation scenarios under consideration: Our plasmode simulation was conducted over 500 iterations. For the base simulation scenario, we set the prevalence of exposure (obesity) and the event rate (diabetes) at 30%, with a true odds ratio (OR) parameter of 1, corresponding to a risk difference (RD) of 0. Each simulated dataset had a sample size of 3,000 participants. The description of other scenarios under consideration is provided in Table 1.

Table 1: Overview of Plasmode Simulation Scenarios Reflecting Varying Exposure and Outcome Prevalences Based on National Health and Nutrition Examination Survey (NHANES) Data Cycles (2013-2018)

Plasmode Simulation Scenario	Exposure Prevalence	Outcome Prevalence	True Odds Ratio	Sample Size
(i) Frequent Exposure and Outcome (Base)	30%	30%	1	3,000
(ii) Rare Exposure and Frequent Outcome	5%	30%	1	3,000
(iii) Frequent Exposure and Rare Outcome	30%	5%	1	3,000

True Data Generating Mechanism Used in Plasmode Simulation: The primary goal of this study is to evaluate various variable selection methods under realistic conditions. To achieve this, we formulated the outcome data based on a specific model specification that incorporates both exposure and covariates, including investigator-specified and proxy variables. The model specification consists of three key components (See Appendices §A and B for further details):

1. *Investigator-Specified Covariates*: We retained the original investigator-specified covariates, which were either binary or categorical, reflecting how real-world studies typically operate.
2. *Transformation of Laboratory Variables*: In real-world studies, it is common for analysts to lack precise knowledge of the true model specification. To simulate this uncertainty, we transformed the

continuous laboratory variables using complex functions such as logarithmic, exponential, square root, polynomial transformations, and interactions. This reflects the challenges analysts face in correctly specifying models when dealing with continuous data.

3. *Inclusion of Proxy Variables*: Real-world studies often deal with unmeasured confounding, which researchers attempt to mitigate by adding proxy variables. However, when a large number of proxies are added, some may act as noise variables, contributing little to the analysis. To simulate this, we selected only those binary proxy covariates (referred to as recurrence covariates in hdPS terminology) that had a relative risk (RR) of less than 0.8 or greater than 1.2 concerning the outcome. Out of 143 **142** proxy covariates, 94 met this criterion and were included in calculating a simple comorbidity burden measure. The remaining 49 **49? or 48?** covariates were excluded from this calculation and considered noise. This comorbidity burden measure was then incorporated into our model specification for generating the plasmode data.

Performance Measures: From this simulation, we derived several performance metrics to evaluate the effectiveness of the methods under consideration: (1) bias, (2) average model standard error (SE; the average of estimated SEs obtained from a model over repeated samples), (3) empirical SE (the standard deviation of estimated treatment effects across repeated samples), (4) mean squared error (MSE), (5) coverage probability of 95% confidence intervals, (6) bias-corrected coverage, and (7) Zip plot [8, 9].

Estimators under consideration

check if the description of the methods look okay to you. Make sure the description matches with the analysis codes.

All looks good!

The comparison between the data generation process and the analysis process reveals two key differences: (i) The data generation used transformed laboratory variables, whereas the analysis was conducted using only the original laboratory variables. (ii) The data generation employed a simple sum of selected proxy variables (sum of 94 proxy covariates), while the analysis included all proxy variables (143 **142** binary proxies), with 49**49? or 48?** of these acting as noise variables. These differences help us assess how the proxy variable selection methods handle model misspecification and the presence of noise variables.

1. **Kitchen sink model**: This is a base model for comparison, where no variable selection approaches were used. All investigator-selected features and all proxy variables were used to model [10].

2. **Bross formula**: The Bross formula is a statistical method used to calculate the bias introduced by not adjusting for a covariate [11]. In hdPS analysis, this formula was originally applied to each proxy variable to measure and rank the potential bias if the covariate were not adjusted for. In our analysis, the 100 proxies with the highest bias rankings are selected for further modeling [12, 13].
3. **Least Absolute Shrinkage and Selection Operator (LASSO)**: LASSO is a variable selection technique that limits the number of variables by adding a penalty term to the regression model. Cross-validation (CV) is used in LASSO to identify variables with non-zero coefficients in the best model by optimizing the penalty value [10, 14, 15].
4. **Hybrid of hdPS and LASSO**: Instead of relying solely on LASSO for variable selection, a hybrid approach combines the Bross formula and LASSO. First, hdPS variables are selected using the hdPS algorithm (e.g., the top 100), and then LASSO is applied to further refine the selection [10, 14].
5. **Elastic Net**: Elastic Net is an extension of LASSO that includes an additional penalty term to handle multicollinearity by grouping correlated features and selecting the most representative ones [10].
6. **Random Forest**: The Random Forest (RF) algorithm is an ensemble learning method that constructs multiple decision trees to perform classification [16]. It calculates the importance of each proxy variable based on the decrease in impurity or Gini importance, providing a ranking of the proxies. The top 100 variables from this ranking are manually selected for further modeling [15].
7. **XGBoost**: XGBoost is a gradient boosting algorithm used to optimize machine learning models [17]. It builds decision trees that make splits based on maximum impurity reduction, and it assigns an importance score to each proxy variable by calculating the mean decrease in impurity [18].
8. **Stepwise**: Stepwise selection is a progressive feature selection method that can proceed in two directions—forward or backward—based on the maximum adjusted R-squared. We have implemented two versions: (a) Forward selection (FS) starts with an initial model (e.g., including all investigator-selected features) and adds proxies to the model one at a time. (b) Backward elimination (BE) starts with a full model (e.g., all investigator-selected features and all proxy variables) and removes features one at a time based on their contribution to the model.
9. **Genetic algorithm (GA)**: GA is an evolutionary algorithm inspired by the theory of natural selection [19]. It operates by evolving offspring from a population of the fittest individuals over

several generations, evaluating and selecting the best combination of features or variables that maximize prediction accuracy.

3 Results

Can you check the app in the doc folder of the repo, and compare of the results description look accurate to you?

Comments and edits in brown and blue color.

The results for each method under the different scenarios are summarized below. See Figures 1 and 2 for an overview of the performance in terms of bias and coverage, respectively.

(i) **Frequent Exposure and Outcome (base) scenario**:

1. *Bias*: The kitchen sink model, which includes all variables without selection, exhibited the smallest bias (0.0002). GA shows the highest bias (0.0287), indicating a substantial deviation from the true effect. Among the other methods, Bross-based hdPS (-0.0001), Hybrid hdPS (0.0016), and Elastic Net (0.0036) demonstrated low bias. XGBoost (0.0074) and Random Forest (RF) (0.0034) still had slightly higher bias.

It seems that Bross-based hdPS (-0.0001) instead of the kitchen sink model (0.0002) exhibited the smallest bias.

The Bross-based hdPS model exhibited the smallest bias (-0.0001). GA shows the highest bias (0.0287), indicating a substantial deviation from the true effect. Among the other methods, the kitchen sink model (0.0002), Hybrid hdPS (0.0016), and Elastic Net (0.0036) demonstrated low bias. XGBoost (0.0074) and Random Forest (RF) (0.0034) still had slightly higher bias.

2. *Coverage*: The coverage for most methods was high, with Hybrid hdPS, Forward Selection, Backward Elimination, LASSO, and Elastic Net achieving values around 98%, indicating well-calibrated confidence intervals. However, GA had significantly lower coverage (83.8%), indicating that its confidence intervals might be too narrow or biased, potentially missing the true effect. After applying bias elimination, GA's coverage improved to 96%, which is better but still lower than other methods.
3. *Mean Squared Error (MSE)*: XGBoost achieved the lowest MSE (0.0006), reaffirming it as the most accurate method overall. GA maintained the highest MSE (0.0016), reflecting its higher bias and variability. The kitchen sink model (0.0009), Bross-based hdPS (0.0008), Hybrid hdPS (0.0008), and Elastic Net (0.0009) all had relatively similar and moderate MSE values.

4. *Standard Error (SE)*: XGBoost exhibited the lowest Empirical SE (0.0229), indicating high precision in its estimates. The kitchen sink model had the highest Empirical SE (0.0305), suggesting greater variability. Other methods, including GA (0.0274), Hybrid hdPS (0.0278), and Bross-based hdPS (0.0287), showed moderate variability. LASSO (0.0299) and Elastic Net (0.0294) had slightly higher variability. The Model-based SE followed a similar pattern, with XGBoost (0.0268) showing the lowest variability and the kitchen sink model (0.0333) the highest, indicating less precision in its estimates. See Appendix §C for further details.

(ii) Rare Exposure and Frequent Outcome:

1. *Bias*: The kitchen sink model showed a relatively low bias (0.0025), while GA continued to exhibit the highest bias (0.0408), indicating a significant deviation from the true effect. XGBoost had a bias of 0.0259, which, while still higher than some other methods, was lower than GA. Other methods, such as Bross-based hdPS (0.0035), Hybrid hdPS (0.0049), and Elastic Net (0.0053), demonstrated moderate bias. Random Forest (RF) (0.0127) and Forward Selection (0.0108) had slightly higher bias but remained within an acceptable range.
2. *Coverage*: Coverage levels remained high for most methods, with XGBoost achieving the highest coverage at 96.2%, indicating well-calibrated confidence intervals despite its higher bias. The GA method had lower coverage (92.2%), suggesting that its confidence intervals might be narrower, potentially missing the true effect. Other methods such as RF, Forward Selection, Backward Elimination, and Hybrid hdPS maintained coverage values around 94-95%, suggesting adequate interval calibration. Bias-eliminated coverage for GA improved to 94.2%, but it was still slightly lower than other methods.
3. *Mean Squared Error (MSE)*: Hybrid hdPS and XGBoost both demonstrated the lowest MSE (0.0032), indicating that these methods were the most accurate in this scenario. The GA method had a higher MSE (0.0043), reflecting its substantial bias and variability. The kitchen sink model also had an MSE of 0.0043, similar to GA, while other methods like Bross-based hdPS (0.0035), RF (0.0039), and Elastic Net (0.0036) exhibited moderate MSE values, indicating reasonable accuracy.

It seems that forward selection model instead of Hybrid hdPS and XGBoost demonstrated the lowest MSE (0.003)

The forward selection algorithm exhibited the lowest MSE (0.0030), indicating that this method was the most accurate in this scenario. In addition, the hybrid hdPS and XGBoost also demonstrated

relatively low MSE (0.0032). The GA method had a higher MSE (0.0043), reflecting its substantial bias and variability. The kitchen sink model also had an MSE of 0.0043, similar to GA, while other methods like Bross-based hdPS (0.0035), RF (0.0039), and Elastic Net (0.0036) exhibited moderate MSE values, indicating reasonable accuracy.

4. *Standard Error (SE)*: The lowest Empirical SE was observed with XGBoost (0.0507) and GA (0.0510), reflecting high precision despite their higher bias. The kitchen sink model had the highest Empirical SE (0.0656), indicating greater variability. Hybrid hdPS (0.0564), Bross-based hdPS (0.0595), and RF (0.0609) showed moderate variability. Forward Selection (0.0537) and Backward Elimination (0.0576) had lower variability compared to the kitchen sink model. In terms of Model-based SE, XGBoost (0.0531) and GA (0.0533) continued to show low variability, while the kitchen sink model had the highest Model-based SE (0.0623), indicating less precision in its estimates.

(iii) Frequent Exposure and Rare Outcome:

1. *Bias*: In this scenario, the kitchen sink model exhibited a moderate negative bias (-0.0093), similar to the Bross-based hdPS method (-0.0088). GA showed a significantly higher bias (0.0362), indicating a substantial deviation from the true effect. Among other methods, XGBoost demonstrated the lowest bias (-0.0061), while methods like Hybrid hdPS (-0.0082), Forward Selection (-0.0070), and Backward Elimination (-0.0070) had slightly higher but still moderate biases. Elastic Net and LASSO both had biases of -0.0079, reflecting slightly larger deviations compared to XGBoost but still within acceptable limits.
2. *Coverage*: Most methods achieved good coverage, with XGBoost, RF, and Forward Selection each achieving a coverage rate of 95.4%, indicating well-calibrated confidence intervals. The GA method, however, had slightly lower coverage (91.8%), indicating that its confidence intervals might be narrower, potentially excluding the true effect. Bross-based hdPS and the kitchen sink model had slightly lower coverage values of 93.8% and 93.4%, respectively. After accounting for bias, the bias-eliminated coverage for most methods, except GA, remained high, with values ranging from 98.4% to 99.0%, indicating that most methods effectively adjusted for bias in their coverage estimates. GA's bias-eliminated coverage was lower at 93.4%, reflecting its higher inherent bias.
3. *Mean Squared Error (MSE)*: XGBoost exhibited the lowest MSE (0.0003), indicating it as the most accurate method overall in this scenario. GA had the highest MSE (0.0040), reflecting its substantial bias and variability. The kitchen sink model (0.0005), Bross-based hdPS (0.0005), and

other methods like Hybrid hdPS (0.0004) and Elastic Net (0.0005) all had relatively similar MSE values, indicating moderate accuracy.

4. *Standard Error (SE)*: The lowest Empirical SE was observed with XGBoost (0.0152), reflecting high precision in its estimates. The GA method exhibited the highest Empirical SE (0.0523), indicating greater variability and less precision. Methods like Hybrid hdPS (0.0184), Forward Selection (0.0187), and Elastic Net (0.0203) showed moderate variability, while Bross-based hdPS (0.0206) and the kitchen sink model (0.0212) had slightly higher variability. In terms of Model-based SE, XGBoost (0.0179) again showed the lowest variability, consistent with its low Empirical SE, indicating that it provided the most stable estimates. The kitchen sink model had a slightly higher Model-based SE (0.0219), indicating less precision in its estimates.

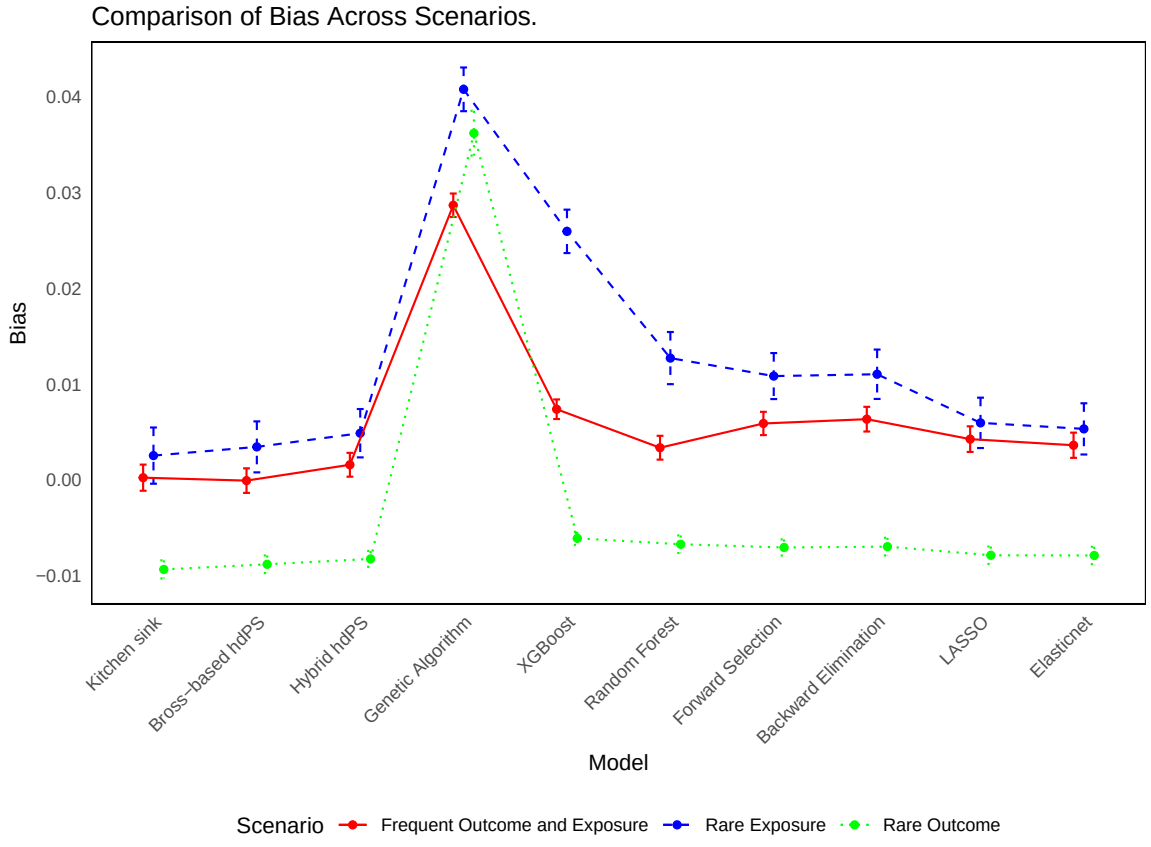

Fig. 1: Comparison of Bias Across Different Methods in hdPS Analysis

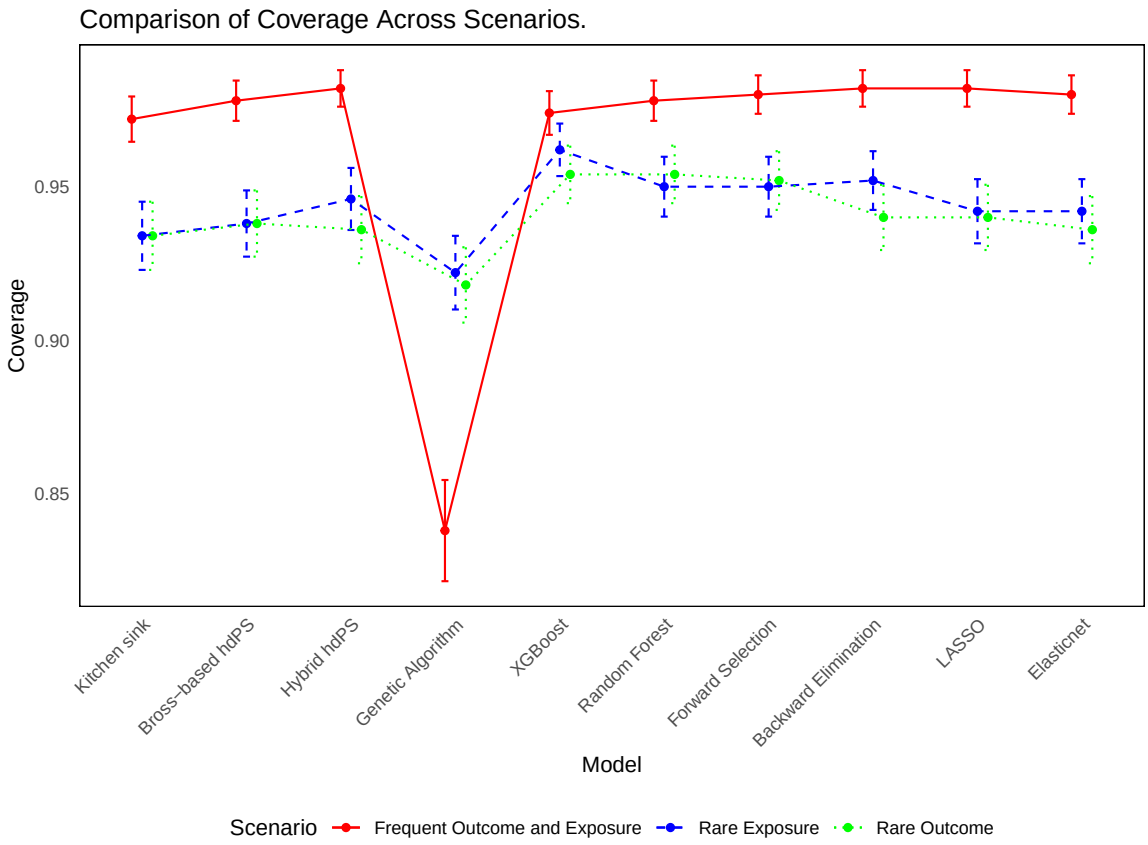

Fig. 2: Comparison of Coverage Probability Across Different Methods in hdPS Analysis

4 Real-world analysis

Here we include full data analysis (with some summary results like exposure and outcome prevalence, and sample size) and report OR and RD. Also mention how many proxies were chosen (add in the picture of RD and OR; side by side for each method, ordered my magnitude of RD), and how many were in common with hdPS (add table).

Summary results: The dataset comprises 7585 individuals. Among these, the prevalence of the exposure is 48.8% or 0.488 , while the prevalence of the outcome is 23.7% or 0.237.

See Figure 3 for the results from analyzing the NHANES (2013-2018) dataset.

I found the results in the table here is a little bit different from the value I got from the real data analysis. Also I found the cap of the figure not printing properly on my computer. I have to add double backslash before the pct symbol %. If on other computers the double backslash causes some error maybe deleting it would help? See the updated plot "comparison-plot-updated" and its code below:

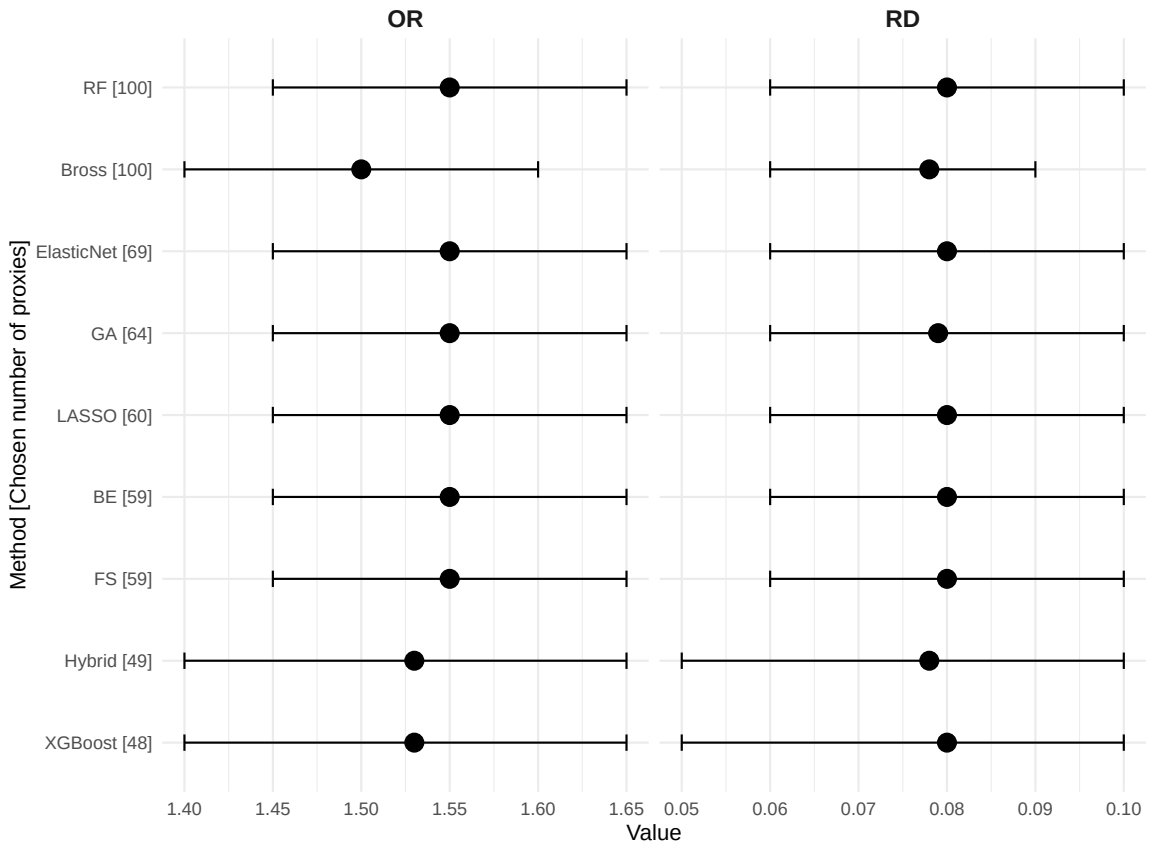

Fig. 3: Figure presenting a comparison of Risk Differences (RD) and Odds Ratios (OR) with 95% confidence intervals for different methods used to evaluate the association between obesity and diabetes risk. The analysis is based on data from the National Health and Nutrition Examination Survey (NHANES) for the years 2013-2018. Methods are arranged by the number of variables used in the models.

See Figure 4 for the updated results from analyzing the NHANES (2013-2018) dataset. The methods are arranged according to the number of selected proxy variables. Among all variable selection algorithms, Random Forest (RF) and XGBoost demonstrate the highest odds ratios (ORs), with values of 1.588 and 1.564, respectively. The ORs for the remaining methods cluster around 1.523. Additionally, with the exception of RF and the Bross formula hdPS, a general pattern emerges where methods selecting a larger number of proxy variables yield higher ORs. For risk difference (RD), RF and XGBoost also exhibit the highest values, 0.085 and 0.082, respectively, while the remaining methods converge around 0.077. The trend observed in the OR results appears to persist in the RD analysis.

Table 2 presents a pairwise comparison of the number of proxy features shared between different variable selection methods used in the analysis. Each cell in the table indicates the count of common proxy variables selected by the method in the corresponding row and column. The diagonal cells,

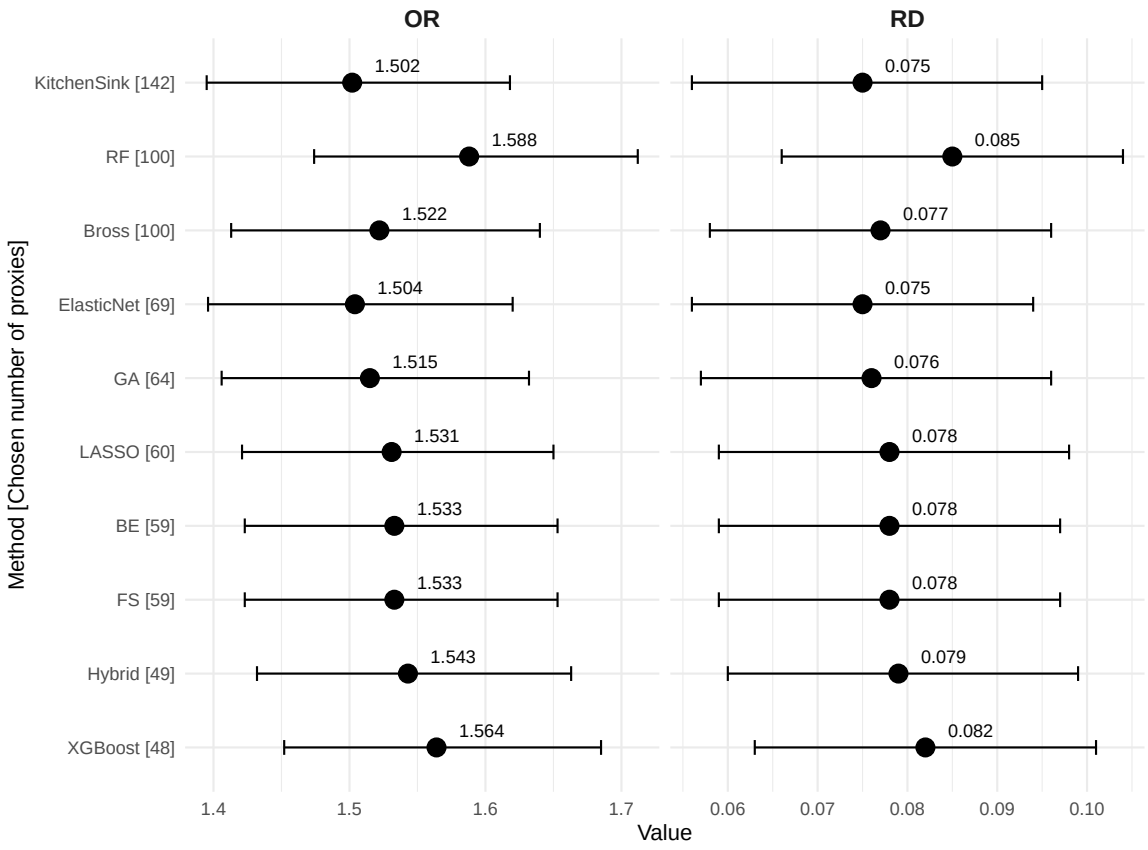

Fig. 4: Figure presenting a comparison of Risk Differences (RD) and Odds Ratios (OR) with 95% confidence intervals for different methods used to evaluate the association between obesity and diabetes risk. The analysis is based on data from the National Health and Nutrition Examination Survey (NHANES) for the years 2013-2018. Methods are arranged by the number of variables used in the models.

where the row and column methods are the same, represent the total number of proxy variables selected exclusively by each method.

need to think about how to best present this proxy in common table.

Updated the pairwise comparison table with adding the value for kitchen sink (ks). Though as the ks model contains all variables, adding the values for it would not add anything useful, I think it is weird to omit a method in the table without saying anything. Also here I changed the order/arrange of the methods, to match to the order/arrange of methods in the Methods section.

Table 3 presents a pairwise comparison of the number of proxy features shared between different variable selection methods used in the analysis. Each cell in the table indicates the count of common proxy variables selected by the method in the corresponding row and column. The diagonal cells, where the row and column methods are the same, represent the total number of proxy variables selected exclusively by each method. The first column displays the number of proxy variables shared

Table 2: Comparison of variable overlap of selected proxies across different methods used to evaluate the association between obesity and diabetes

	Bross	Hybrid	LASSO	ElasticNet	GA	XGBoost	RF	FS	BE
Bross formula	100								
Hybrid (Bross and LASSO)	49	49							
LASSO	47	47	60						
Elastic Net	54	48	60	69					
Genetic algorithm (GA)	44	28	36	40	64				
XGBoost	38	24	28	30	25	48			
Random Forest (RF)	72	37	42	50	36	48	100		
Forward selection (FS)	45	41	51	54	35	25	43	59	
Backward elimination (BE)	45	41	51	54	35	25	43	59	59

between each method and the kitchen sink (KS) model. Given that the KS model includes all proxy variables, the values in the first column are identical to those in the diagonal cells for each row, which also presents the total number of proxy variables selected by the method in the corresponding row.

Table 3: Updated Comparison of variable overlap of selected proxies across different methods used to evaluate the association between obesity and diabetes

	KS	Bross	Hybrid	LASSO	EN	RF	XGB	FS	BE	GA
Kitchen sink (KS)	142									
Bross formula	100	100								
Hybrid (Bross and LASSO)	49	49	49							
LASSO	60	47	47	60						
Elastic Net (EN)	69	54	48	60	69					
Random Forest (RF)	100	71	38	46	52	100				
XGBoost (XGB)	48	38	24	28	30	37	48			
Forward selection (FS)	59	45	41	51	54	45	25	59		
Backward elimination (BE)	59	45	41	51	54	45	25	59	59	
Genetic algorithm (GA)	64	44	28	36	40	49	25	35	35	64

Table 4 presents the comparison of common proxy counts and rates between different methods with the bross formula hdPS. Column 1 indicates the number of proxy variables selected by each variable selection method. Column 2 shows the number of common features selected by each method and the Bross formula method. Column 3 demonstrates the percentage of the common features in the total number of features of each method.

From Table 4, it can be observed that, with the exception of the bross hdPS and random forest (RF) models, where the number of proxies was manually set to 100, and the kitchen sink (KS) model, which inherently includes all proxies, the number of proxy variables selected by all other models ranges from 49 to 69. The hybird method of bross and lasso hdPS, by definition, performs lasso proxy selection on the 100 proxy variables selected by bross, resulting in an inherent common rate of 1.00.

For the remaining methods, the common rates are clustered around 0.74, with XGBoost exhibiting the highest common rate of 0.792 and GA presenting the lowest common rate of 0.688.

Table 4: Comparison of the count and percentage of proxy variables selected by each methods in common with that by the Bross formula hdPS

Method	Total Count	Common Count	Rate in Common
Kitchen sink	142	-	-
Bross formula	100	-	-
Hybrid (Bross and LASSO)	49	49	1.000
LASSO	60	47	0.783
Elastic Net	69	54	0.783
Random Forest (RF)	100	71	0.710
XGBoost	48	38	0.792
Forward selection (FS)	59	45	0.763
Backward elimination (BE)	59	45	0.763
Genetic algorithm (GA)	64	44	0.688

Computing time:

Report computing time for the Real-world analysis for each method. (add ordered table or figure)

Table 5 presents the computing time for each method. All methods, aside from RF and GA, exhibit relatively fast computing times. Although RF and GA are slower, their computational demands still remain within reasonable time limits.

Table 5: Computing time for the real-world analysis for each algorithm

Method	Computing Time
Kitchen Sink	2.942
Bross formula	2.280
Hybrid (Bross and LASSO)	2.517
LASSO	2.540
Elastic Net	2.814
Random Forest (RF)	379.485
XGBoost	1.444
Forward selection (FS)	7.417
Backward elimination (BE)	12.794
Genetic algorithm (GA)	105.697

5 Discussion

5.1 Summary of the simulation findings

Comparison of methods:

Across the three scenarios—Frequent Exposure and Outcome, Rare Exposure and Frequent Outcome, and Frequent Exposure and Rare Outcome—XGBoost consistently exhibited the lowest Mean

Squared Error (MSE) and high coverage, making it one of the most reliable methods overall. However, XGBoost did not always have the lowest bias; in fact, methods such as the kitchen sink model, Bross-based hdPS, and Hybrid hdPS often showed lower bias. The GA, in contrast, consistently exhibited the highest bias and MSE, along with lower coverage and greater variability, indicating it was less reliable for accurate effect estimation. The kitchen sink model, while demonstrating low bias and reasonable MSE, generally had higher variability and lower precision compared to other methods like XGBoost. Methods such as Bross-based hdPS, Hybrid hdPS, and Elastic Net performed moderately well across all scenarios, balancing bias, coverage, and MSE, but they did not outperform XGBoost in overall accuracy and precision.

Comparison of scenarios:

Scenarios with rare exposure tended to produce higher bias, particularly for methods like GA and XGBoost, while frequent outcomes generally led to lower bias. Overcoverage was common in the frequent exposure and outcome scenario, while the other scenarios tended to have more balanced or slightly under-coverage. Frequent exposure scenarios had higher relative error, indicating more difficulty in precisely estimating effects compared to rare exposure scenarios, which had lower relative error. Rare outcome scenarios tended to have the lowest MSE, indicating more precise effect estimates under these conditions, whereas rare exposure scenarios had higher MSE, reflecting the challenges of estimating effects when exposure is uncommon.

5.2 Contextualizing the literature

need to write what hdps offers in the literature; what additional selection criteria was used so far

5.3 Data analysis findings

summarize real data analysis finding in 1 para

5.4 Future Direction

TMLE

5.5 Conclusion

In conclusion, this analysis highlights the importance of carefully selecting appropriate methods for hdPS analysis based on the specific characteristics of the data, particularly the prevalence of exposure and outcome. While XGBoost consistently demonstrated strong performance in terms of

MSE and coverage, it did not always have the lowest bias, suggesting that it may be most suitable when precision is prioritized over bias minimization. GA, despite its potential, showed significant limitations with consistently high bias and MSE, making it less reliable for accurate effect estimation. The kitchen sink model, Bross-based hdPS, and Hybrid hdPS methods provided a balanced approach, often delivering low bias and moderate MSE, but with variability in coverage depending on the scenario. Scenario-specific trends revealed that rare outcomes generally yielded lower MSE and better precision, while rare exposures were associated with higher bias and MSE, emphasizing the challenges of accurately estimating effects in such contexts. Ultimately, the findings underscore the need to tailor method selection to the epidemiological scenario at hand, ensuring that the chosen approach aligns with the specific goals and challenges of the analysis.

List of abbreviations

- hdPS: High-dimensional Propensity Score
- NHANES: National Health and Nutrition Examination Survey
- OR: Odds Ratio
- RD: Risk Difference
- SE: Standard Error
- MSE: Mean Squared Error
- KS: Kitchen Sink
- LASSO: Least Absolute Shrinkage and Selection Operator
- ElasticNet / EN: A regularized regression method that combines LASSO and Ridge regression
- RF: Random Forest
- XGBoost: Extreme Gradient Boosting
- FS: Forward Selection
- BE: Backward Elimination
- GA: Genetic Algorithm
- CV: Cross-Validation
- RR: Relative Risk

Declarations

Ethics approval and consent to participate

The analysis conducted on secondary and de-identified data is exempt from research ethics approval requirements. Ethics for this study was covered by item 7.10.3 in University of British Columbia's Policy #89: Research and Other Studies Involving Human Subjects 19 and Article 2.2 in of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS2).

Consent for publication

Availability of data and materials

Competing interests

Over the past three years, MEK has received consulting fees from Biogen Inc. for consulting unrelated to this current work. MEK was previously supported by the Michael Smith Foundation for Health Research Scholar award.

Funding

This work was supported by MEK's Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants and Discovery Accelerator Supplements.

Authors' contributions

MEK: Conceptualization, Writing – Original Draft, Review & Editing YL: Formal Analysis, Review & Editing

Acknowledgements

Not applicable.

References

[1] Karim ME. High-dimensional propensity score and its machine learning extensions in residual confounding control. The American Statistician. 2024;(1):1–38.

[2] Saydah S, Bullard KM, Cheng Y, Ali MK, Gregg EW, Geiss L, et al. Trends in cardiovascular disease risk factors by obesity level in adults in the United States, NHANES 1999-2010. Obesity. 2014;22(8):1888–1895.

[3] Liu J, Hay J, Faught BE, et al. The association of sleep disorder, obesity status, and diabetes mellitus among US adults—The NHANES 2009-2010 survey results. International journal of endocrinology. 2013;2013.

[4] Kabadi SM, Lee BK, Liu L. Joint effects of obesity and vitamin D insufficiency on insulin resistance and type 2 diabetes: results from the NHANES 2001–2006. Diabetes care. 2012;35(10):2048–2054.

[5] Ostchega Y, Hughes JP, Terry A, Fakhouri TH, Miller I. Abdominal obesity, body mass index, and hypertension in US adults: NHANES 2007–2010. American journal of hypertension. 2012;25(12):1271–1278.

[6] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;p. 37–48.

[7] Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Computational statistics and data analysis. 2014;72:219–226.

[8] Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in medicine. 2019;38(11):2074–2102.

[9] White IR, Pham TM, Quartagno M, Morris TP. How to check a simulation study. International Journal of Epidemiology. 2023;p. dyad134.

[10] Karim ME, Pang M, Platt RW. Can we train machine learning methods to outperform the high-dimensional propensity score algorithm? Epidemiology. 2018;29(2):191–198.

[11] Bross ID. Spurious effects from an extraneous variable. Journal of chronic diseases. 1966;19(6):637–647.

[12] 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.