Understanding the role of different proxy selection methods in High-Dimensional Propensity Score Analysis

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: **Methods: Results: Conclusion:**

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

JEL Classification: C18

MSC Classification: 92D30 , 62P10

1 Background

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:

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 proxy covariates, 94 met this criterion and were included in calculating a simple comorbidity burden measure. The remaining 49 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

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 binary proxies), with 49 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. **Elasticnet**: 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

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). In contrast, the Genetic Algorithm (GA) showed the highest bias (0.0287). Among the other methods, Bross (-0.0001), Hybrid (0.0016), and Elasticnet (0.0036) demonstrated low bias, indicating that these methods provide estimates closer to the true effect. XGBoost (0.0074) and Random Forest (RF) (0.0034) had slightly higher bias but remained within acceptable limits.
2. *Coverage*: Most methods, including Hybrid, Forward, Backward, LASSO, and Elasticnet, achieved high coverage values around 98%, indicating well-calibrated confidence intervals. However, the GA method had significantly lower coverage (83.8%), suggesting that its confidence intervals might be too narrow or biased, potentially missing the true effect. Given the bias in the GA method's results, we also calculated bias-eliminated coverage. This adjustment improved GA's coverage to 96%, but it still remained lower than that of the other methods.
3. *Mean Squared Error (MSE)*: XGBoost achieved the lowest MSE (0.0006), indicating it as the most accurate method overall. Conversely, the GA method had the highest MSE (0.0016), reflecting its higher bias and variability. The kitchen sink model (0.0009), Bross (0.0008), Hybrid (0.0008), and Elasticnet (0.0009) all had relatively similar and moderate MSE values.
4. *Standard Error (SE)*: The lowest Empirical SE was observed with XGBoost (0.0229), indicating high precision in its estimates. The kitchen sink model had the highest Empirical SE (0.0305), suggesting greater variability. Other methods such as GA (0.0274), Hybrid (0.0278), and Bross (0.0287) exhibited moderate variability, while LASSO (0.0299) and Elasticnet (0.0294) had slightly

higher variability. Similarly, XGBoost also showed the lowest Model-based SE (0.0268), consistent with its low Empirical SE and indicating high precision. The kitchen sink model had the highest Model-based SE (0.0333), suggesting less precision in its estimates.

(ii) Rare Exposure and Frequent Outcome:

1. *Bias*: In this scenario, the kitchen sink model showed a relatively low bias (0.0025), but the Genetic Algorithm (GA) had the highest bias (0.0408), indicating a substantial deviation from the true effect. XGBoost (0.0259) also exhibited higher bias compared to other methods. On the other hand, Bross (0.0035), Hybrid (0.0049), and Elasticnet (0.0053) demonstrated moderate bias, while Random Forest (RF) (0.0127) and Forward Selection (0.0108) had slightly higher bias but remained within an acceptable range.
2. *Coverage*: Most methods maintained high coverage levels, with XGBoost showing the highest coverage (96.2%), suggesting that its confidence intervals were well-calibrated despite the higher bias. The GA method, however, had slightly lower coverage (92.2%), indicating that its confidence intervals might be narrower, potentially excluding the true effect. Other methods such as RF, Forward, Backward, and Hybrid had coverage values around 95%, suggesting adequate interval calibration. To account for the bias in GA, bias-eliminated coverage was calculated, which improved the coverage for GA to 94.2%, still slightly lower than other methods.
3. *Mean Squared Error (MSE)*: XGBoost and Hybrid methods both demonstrated the lowest MSE (0.0032), indicating that these methods were the most accurate in this scenario. The GA method, with a higher MSE (0.0043), reflected its substantial bias and variability. The kitchen sink model had an MSE of 0.0043, similar to GA, while other methods such as Bross, RF, and Elasticnet 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 the high precision of these methods despite the higher bias. The kitchen sink model exhibited the highest Empirical SE (0.0656), indicating greater variability. Methods like Hybrid (0.0564), Bross (0.0595), and RF (0.0609) showed moderate variability, while Elasticnet (0.0597) and LASSO (0.0587) had slightly higher variability. In terms of Model-based SE, XGBoost (0.0531) and GA (0.0533) also showed low variability, while the kitchen sink model had the highest Model-based SE (0.0623), suggesting 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 method (-0.0088). The Genetic Algorithm (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 (-0.0082), Forward (-0.0070), and Backward (-0.0070) had slightly higher but still moderate biases. Elasticnet 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, Random Forest (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 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 (0.0005), and other methods like Hybrid (0.0004) and Elasticnet (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 (0.0184), Forward (0.0187), and Elasticnet (0.0203) showed moderate variability, while Bross (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.

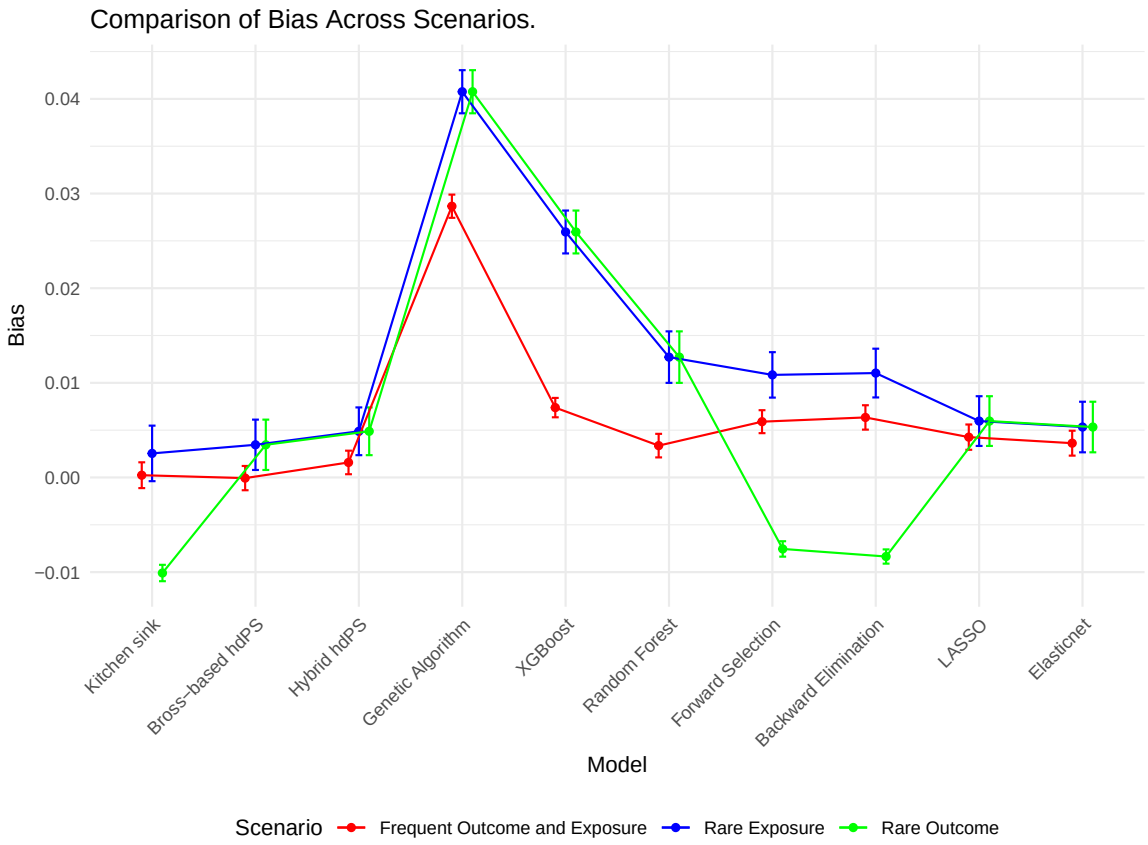

Fig. 1: Comparison of Bias Across Different Methods in hdPS Analysis

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).

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

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

Computing time:

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

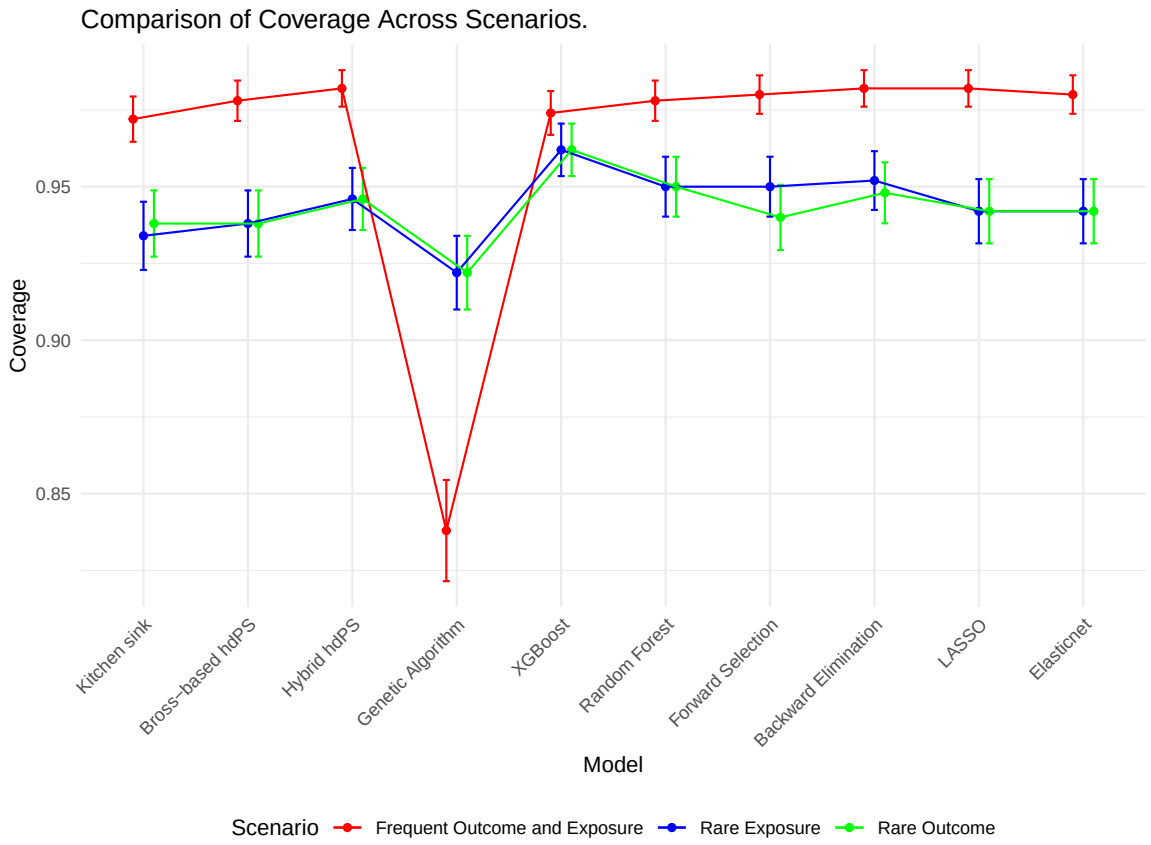

Fig. 2: Comparison of Coverage Probability Across Different Methods in hdPS Analysis

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						
Elasticnet	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

5 Discussion

XGBoost consistently outperformed other methods across all scenarios, demonstrating low bias, high coverage, and the lowest MSE. The GA, on the other hand, consistently underperformed with higher bias, lower coverage, and higher MSE, indicating less reliable estimates. The Bross method and the kitchen sink model generally provided moderate performance with low to moderate bias and

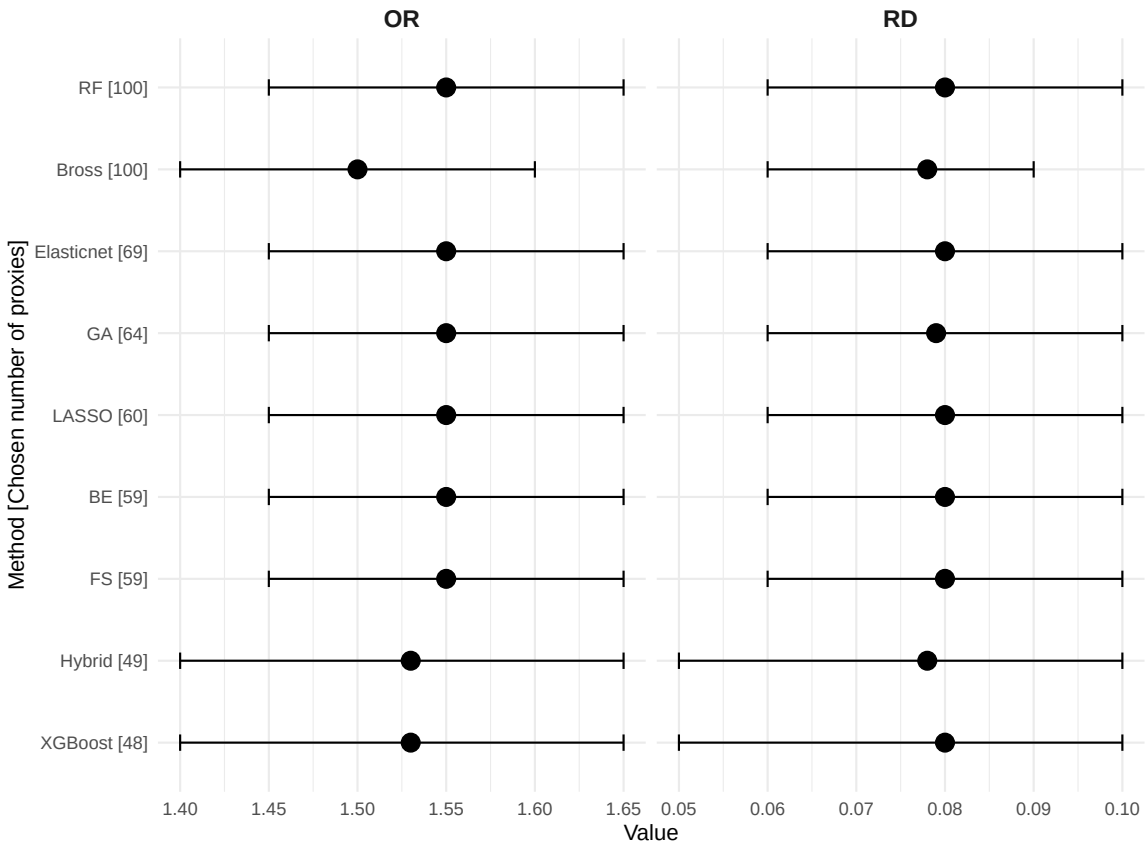

Fig. 3: Figure presenting a comparison of Risk Differences (RD) and Odds Ratios (OR) with 95% confidence intervals for different methods used to evaluate the association between obesity and diabetes risk. The analysis is based on data from the National Health and Nutrition Examination Survey (NHANES)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

reasonable coverage, but they were less accurate than XGBoost. Hybrid, Elasticnet, and LASSO methods showed competitive performance with moderate bias, good coverage, and acceptable MSE, making them reliable alternatives, though slightly less optimal than XGBoost.

Contextualizing the literature:

Summary of the simulation findings:

Data analysis findings:

Future Direction:

Conclusion:

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
- LASSO: Least Absolute Shrinkage and Selection Operator
- GA: Genetic Algorithm
- RF: Random Forest
- XGBoost: Extreme Gradient Boosting
- FS: Forward Selection
- BE: Backward Elimination
- CV: Cross-Validation
- RR: Relative Risk
- Elasticnet: A regularized regression method that combines LASSO and Ridge regression

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.