Pragmatic and Observational Research

Dovepress

open access to scientific and medical research

Open Access Full Text Article

METHOD

Comparing Machine Learning and Advanced Methods with Traditional Methods to Generate Weights in Inverse Probability of Treatment Weighting: The INFORM Study

Doyoung Kwak[1], Yuanjie Liang[2], Xu Shi[3], Xi Tan[2]

[1]Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX, USA; [2]Novo Nordisk Inc, Plainsboro, NJ, USA; [3]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Correspondence: Xi Tan, Email mxtz@novonordisk.com

Purpose: Observational research provides valuable insights into treatments used in patient populations in real-world settings. However, confounding is likely to occur if there are differences in patient characteristics associated with both the exposure and outcome between the groups being evaluated. One approach to reduce confounding and facilitate unbiased comparisons is inverse probability of treatment weighting (IPTW) using propensity scores. Machine learning (ML) and entropy balancing can potentially be used in generating propensity scores for IPTW, but there is limited literature on this application. We aimed to assess the feasibility of applying these methods for reducing confounding in observational studies. These methods were assessed in a study comparing cardiovascular outcomes in adults with type 2 diabetes and established atherosclerotic cardiovascular disease taking once-weekly glucagon-like peptide-1 receptor agonists or dipeptidyl peptidase-4 inhibitors.

Methods: We applied advanced methods to generate the propensity scores compared to the original logistic regression method in terms of covariate balance. After calculating weights, a weighted Cox proportional hazards model was used to calculate the sample average treatment effect. Support Vector Classification, Support Vector Regression, XGBoost, and LightGBM were the ML models used. Entropy balancing was also performed on features identified in the original cardiovascular outcomes study.

Results: Accuracy (range: 0.71 to 0.73), area under the curve (0.77 to 0.79), precision (0.53 to 0.60), recall (0.66 to 0.68), and F1 score (0.60 to 0.64) were similar between all of the advanced propensity score methods and traditional logistic regression. Among ML models, only XGBoost achieved balance in all measured baseline characteristics between the two treatment groups, closely approximating the performance of the original logistic regression. Entropy balancing weights provided the best performance among all models in balancing baseline characteristics, achieving near perfect balancing.

Conclusion: Among the advanced methods examined, entropy balancing weights performed the best for optimizing balancing and can produce similar results compared to traditional logistic regression.

Keywords: propensity score, machine learning, entropy balancing, type 2 diabetes, glucagon-like peptide-1 receptor agonists

Introduction

Observational research offers valuable insights into patient populations exposed to different treatments in real-world settings. However, unlike in randomized clinical trials where confounding of intervention and control groups is reduced or eliminated, in real-world studies, differences in patient characteristics associated with both the exposure and outcome between the groups being evaluated can potentially lead to underestimating or overestimating the true effect of the exposure on the outcome being measured.[1,2] Commonly used methods for reducing confounding in observational studies include methods based on propensity scores such as propensity score matching and inverse probability of treatment weighting (IPTW).[1,3–5]

IPTW is an approach to controlling for confounding variables by using propensity scores to facilitate unbiased comparisons in observational studies.[1] Propensity scores are calculated as an individual's probability, or propensity, of

Received: 14 March 2024
Accepted: 19 September 2024
Published: 4 October 2024

being exposed or treated based on their characteristics, and can be generated using statistical models. After calculating propensity scores, the inverse of the score of being exposed/treated is calculated as a weight for each individual; this weight is then applied to the overall study population. When these weights are used in causal inference tasks, they effectively create a pseudo population of patients for which the demographic and clinical characteristics are more optimally balanced between the groups being compared.

Machine learning (ML), a subfield of artificial intelligence (AI), is the use of computer systems and algorithms to develop models that can complete complex tasks, including analyzing and identifying patterns in data. ML is increasingly viewed as a valuable tool in advancing healthcare and clinical research.[6–10] ML has been applied in real-world evidence (RWE) and health economics and outcomes research[11–13] including but not limited to extraction of patient information from unstructured electronic health records,[14,15] developing predictive models,[16–18] detecting medical conditions,[19–21] and in digital health interventions.[22] Further, ML has the potential to be used in generating propensity scores for IPTW or other methods of causal inference.[23–25] However, to our knowledge, there is limited literature on using ML for propensity scores for IPTW, including a recent observational real-world study of an ML-based propensity score approach (using Random Forest and Bayesian Additive Regression Trees), which generated similar findings to the traditional logistic regression-based propensity score approach; however, detailed comparisons between these approaches were not presented.[26]

Similar to IPTW, entropy balancing is a method that allows for the estimation of causal effects in observational studies.[27] It is a multivariate matching approach that adjusts for covariates instead of relying on a model to estimate propensity scores. This method identifies weights for the control sample to equalize the distribution of covariates across treatment and control samples. It has been used in various fields, including medical studies,[28–32] to apply causal findings from one population (a source population) to another (a target population) based on observed characteristics. This approach can be more effective in balancing covariates than traditional propensity score weighting methods because it operates directly on the covariate distributions as opposed to specification of a logistic function.[29,33,34] In addition to its flexibility, entropy balancing can generate better covariate balance and subsequently more precise effect estimates with large target populations.[34,35]

Despite the potential benefits of advanced methodologies for reducing confounding in observational studies, there is limited evidence comparing these methods with traditional methods in actual RWE research practice. The objective of this INFORM (Modernization of Real-World Research Methods) study is to establish the feasibility of applying ML methods and entropy balancing to reduce confounding in observational studies and compare them with a traditional method. These methods will be assessed in a study comparing cardiovascular outcomes in adults with type 2 diabetes (T2D) and established atherosclerotic cardiovascular disease (ASCVD) taking once-weekly (OW) glucagon-like peptide-1 receptor agonists (GLP-1 RA) or dipeptidyl peptidase-4 inhibitors (DPP-4i) in which a traditional method of propensity scoring and IPTW via logistic regression was used.[36] We consider the feasibility of these methods to be two-fold: (1) the propensity function does not need to be explicitly defined as it does with logistic regression, resulting in easier implementation, and (2) given their easier implementation, whether their use for propensity scoring results in equal or greater bias reduction as compared to logistic regression.

Methods

Study Design

The original T2D/ASCVD cardiovascular outcomes study[36] was an observational cohort study conducted between January 1, 2017, and September 30, 2021, using Optum's de-identified Clinformatics® Data Mart Database (CDM).[37] The primary objective of this study was to compare the time to ischemic stroke and myocardial infarction among US adults with T2D and ASCVD who initiated OW GLP-1RA vs DPP-4i. To reduce selection bias and confounding from observed covariates between the OW GLP-1RA and DPP-4i groups, propensity scores were estimated using a generalized linear model for binary outcome with logit function, ie, logistic regression, on pre-specified variables and ad hoc interaction effects informed by expert domain knowledge. The final study sample included 26,430 OW GLP-1RA users and 39,858 DPP-4i users before weighting. About 82 baseline patient characteristics were balanced, including sociodemographic factors, T2D and ASCVD history, healthcare utilization, healthcare costs, comorbidities, medication use, procedures, and proxy measures of overall health and diabetes severity. Weights were calculated for IPTW and applied in the calculation of weighted descriptive statistics, bivariate analyses, and multivariate analyses to assess covariate balance between groups.

After IPTW, baseline characteristics were well balanced between the two groups by using the conventional threshold of standardized mean difference (SMD) (<0.1). With adequate balancing through IPTW, a weighted Cox proportional hazards (PH) model was used to estimate the treatment effect (rate of ischemic stroke). Results showed that OW GLP-1RA users had 26% lower risk of ischemic stroke compared to DPP-4i users (hazard ratio [HR] and 95% confidence interval [95% CI] = 0.74 [0.63, 0.87], p<0.001); details of the study have been previously published.[36]

For this INFORM analysis, we reevaluated the comparative effectiveness on cardiovascular outcomes by applying advanced methods to generate the propensity scores and compared these methods to the previously used logistic regression method in terms of covariate balance, using the same analytic dataset, dependent variables, and independent variables. The new methods included ML models and entropy balancing as detailed later. Performance between new and old methods was compared using several measures as described later. Clinical outcomes in the original study were also compared using weighted Cox PH models. The sample average treatment effect (sATE) was estimated. This study was reviewed by the WCG Institutional Review Board. It was deemed exempt because the research utilizes retrospectively collected de-identified data for further analysis without further interaction with the human subjects.

Dataset

We used the CDM for the original cardiovascular outcomes study and the current INFORM study. CDM comprises administrative claims data for commercial healthcare and Medicare Advantage enrollees of large national managed care companies, across the United States. Their claims data includes verified, adjudicated, adjusted, and de-identified medical and pharmacy claims, as well as outpatient laboratory test results from large national laboratories.

ML Models

For this research, we utilized two types of ML models: Support Vector Machine (SVM) and Gradient Boosting. From these types, we have four different models: Support Vector Classification, Support Vector Regression (SVR, treated the outcome as continuous), XGBoost, and LightGBM. These models were chosen based on their proven effectiveness in various fields and their ability to handle complex, high-dimensional data. SVM is a set of supervised learning methods used for classification, regression, and outliers detection.[38] It is effective in high-dimensional spaces and is still effective in cases where the number of dimensions is greater than the number of samples. SVR is an extension of SVM that applies the SVM concepts to regression problems.[39] It has been proven to be an effective tool in real-value function estimation. While SVR is not typically used for binary outcomes, we chose to utilize this as a linear probability model and explore its potential in this specific context. Gradient boosting is an ML technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.[40] XGBoost[41] is a scalable ensemble technique that has been demonstrated to be a reliable and efficient ML model. LightGBM,[42] similar to XGBoost, is an ensemble learner of weak decision trees, but is focused on providing extremely fast training performance using selective sampling of high gradient instances. In Gradient Boosting frameworks, like XGBoost and LightGBM, loss functions set the learning objective and are optimized during model training. Any given loss function aims to minimize the error between predicted and actual outcomes, thus guiding the model's learning and determining its performance; however, the differences between loss functions can lead to differences in model performance for any given task. For this reason, we have chosen to evaluate XGBoost and LightGBM using two different loss functions: a mean squared error loss function ("XGBoost Error", "LightGBM Error") and a squared log loss function ("XGBoost Loss", "LightGBM Loss"). Mean squared loss is favored for its simplicity and ability to penalize discrepancies between predictions and real values across a broad spectrum of regression scenarios. On the other hand, squared log loss is advantageous for skewed target variable distributions, mitigating outlier impacts by focusing on logarithmic differences. These two objectives have been chosen and tested due to their demonstrated capacity to accurately capture complex data relationships. All ML models were trained on features that were identified in the original cardiovascular outcomes study. For each model, we conducted hyperparameter tuning using a subset of our data (approximately 5%), which enabled us to identify the optimal hyperparameters. Specifically, for SVM models, this tuning process involved fine-tuning the kernel function and regularization alongside additional parameters. For Gradient Boosting models, we adjusted the number of leaves/branches, depth limit, and learning rates, along with various other settings. This preparatory step allowed us to establish the best configurations before embarking on the main training phase with the entire dataset. The ML models were used to perform prediction tasks to create propensity scores. With the

generated propensity scores, weights were calculated and applied to the data. Thus, these models were used as an intermediary step for bias reduction when estimating the sATE and not for other purposes such as causal inference.

Entropy Balancing

In the INFORM study, entropy balancing was initially performed on features that had been identified in the original cardiovascular outcomes study. We also tried a "no interaction" approach. The primary distinction between the original and the no interaction approaches lies in the latter's use of all available features without specifying interaction terms, allowing models to autonomously determine feature significance without prior filtering.

Comparisons Between New Advanced Methods and Traditional Material and Methods

The performance of the ML models was compared between both the new advanced methods and the traditional logistic regression method in terms of classification accuracy, area under the curve (AUC), precision, recall, and F1 score. AUC measures the area under the Receiver Operating Characteristic (ROC) curve, with higher values indicating better model performance in distinguishing between treatment classes.[43] Precision measures the ratio of true positive predictions to the total number of positive predictions, with higher values indicating fewer false positives. Recall measures the ratio of true positive predictions to the total number of actual positives, with higher values indicating the model's effectiveness at identifying all positive instances. The F1 score is the harmonic mean of precision and recall, with higher values indicating a better balance between these two metrics.[44] All four metrics are percentages and, therefore, range from 0 to 1. The evaluation metrics were calculated based on the performance of the models on the test dataset, which constitutes 20% of the total data. These metrics were obtained after training the models on the training dataset, comprising the remaining 80% of the data. This process was done once; the metrics we report are from the single partition of test data, and therefore, 95% confidence intervals are not reported.

The distribution plots for propensity scores and sATE were produced for each method. The baseline characteristics and SMD[45–47] between the OW GLP-1RA group and DPP-4i group after IPTW were compared for each method of estimating propensity scores and for entropy balancing. SMD is the [difference between groups]/[SD of combined groups]. SMD <0.1 was considered a non-significant difference.[45–47]

Almost all variables were non-missing except race, BMI, region, and HbA1c. Missing data for these variables were treated as a separate "unknown" category. All analyses were conducted using R 4.2.2 (Supplementary Methods).

Results

Final Study Sample

This study included 26,430 OW GLP-1RA users and 39,858 DPP-4i users before propensity score weighting. Detailed patient characteristics can be found elsewhere.[36] Because no significant imbalance was observed in the sample sizes between OW GLP-1RA users and DPP-4i users, no imbalance correction was performed.

Predictive Performance of Propensity Score Models

Overall performance was similar between the advanced methods and traditional logistic regression. In Table 1 we compare predictive performance of the different propensity score models and see, in terms of classification accuracy, AUC, precision, recall, and F1 score, that all advanced ML methods perform similarly to logistic regression. For all models, classification accuracy is between 0.71 and 0.73, AUC is between 0.77 and 0.79, precision is between 0.53 and 0.60, recall is between 0.66 and 0.68, and F1 score is between 0.60 to 0.64. This parity evidences the utility of all models to adequately distinguish between the experimental groups. Among the ML models, XGBoost Error and XGBoost Log Loss models perform slightly better, and SVM regression and SVM classification seem to perform the worst. Further, to ensure that our models were not overfitting, we compared the evaluation metrics for the training dataset and the test dataset. The metrics did not show substantial differences, indicating that the models generalize well to unseen data.

Table 1 Predictive Performance Metrics for Different Propensity Score Models

Performance Metrics	Logistic Regression	SVM Regression	SVM Classification	XGBoost Error	XGBoost Log Loss	LightGBM Error	LightGBM Log Loss
Accuracy	0.72	0.71	0.71	0.73	0.73	0.72	0.72
AUC	0.79	0.77	0.77	0.79	0.79	0.79	0.79
Precision	0.58	0.55	0.53	0.60	0.60	0.58	0.57
Recall	0.67	0.66	0.68	0.68	0.68	0.68	0.67
F1 Score	0.62	0.60	0.60	0.63	0.64	0.63	0.62

Note: For Gradient-Boosting models, two learning objectives were tested: Error and Log Loss.
Abbreviations: AUC, area under the curve; RMSD, root mean squared deviation; SVM, support vector machine.

Balancing Results

The propensity score and sATE distributions were similar between XGBoost (Figure 1 C1 and C2), LightGBM (Figure 1 D1 and D2) and logistic regression (Figure 1 A1 and A2). However, the distribution of SVM regression (Figure 1 B1 and B2) has less separation between the two treatment groups. The ranking of ML methods in balancing of baseline characteristics from best to worst was XGBoost, LightGBM, and SVM. Among ML models, only XGBoost achieved balance in all measured baseline characteristics between the two groups, closely approximating the performance of the

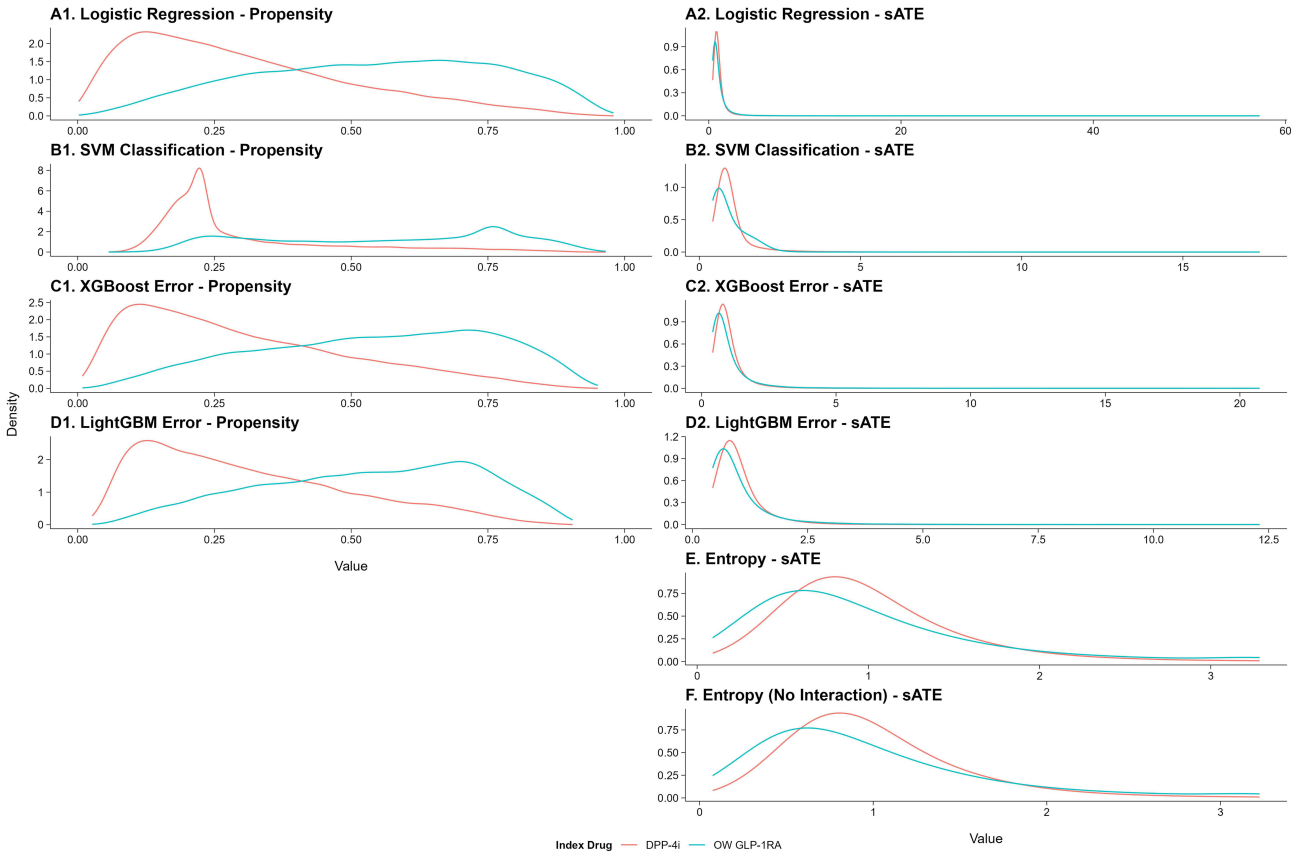

Figure 1 Distribution Plots for Propensity and sATE. Propensity graphs (**A1-D1**): These graphs display the distribution of the propensity score by the two index drugs. Models that are capable of distinguishing between the two groups will have minimal overlap concentrated near 0.5 with separate peaks at 0 and 1. sATE graphs (**A2-D2, E-F**): These graphs display the distribution of the sample average treatment effect weights by the two index drugs. Distributions that look different across models indicate a difference in sample balancing.
Notes: SVM Regression, XGBoost LogLoss, and LightGBM LogLoss results are similar to the results for SVM Classification, XGBoost Error, and LightGMB Error, respectively.
Abbreviations: DPP-4i, dipeptidyl peptidase-4 inhibitors; GLP-1RA, glucagon-like peptide-1 receptor agonists; sATE, sample average treatment effect; SVM, Support Vector Machine.

Table 2 Balancing Results (Key Variables), Standardized Mean Difference by ML Method

Variables	Baseline	Logistic Regression (reference)	Entropy Balance	SVM Regression	SVM Classification	XGBoost Error	XGBoost Log Loss	LightGBM Error	LightGBM Log Loss
Age group									
18–44	0.130	0.008	0.0002	0.014	0.029	0.021	0.017	0.039	0.048
45–64	0.442	0.032	0	0.059	0.112	0.034	0.047	0.060	0.081
65–79	0.079	0.015	0.0001	0.047	0.065	0.023	0.015	0.017	0.008
≥80	0.460	0.059	0	0.135	0.229	0.075	0.0796	0.104	0.119
Sex									
Female	0.035	0.010	0.0001	0.042	0.028	0.018	0.021	0.030	0.029
Male	0.035	0.010	0.0001	0.042	0.028	0.018	0.021	0.030	0.029
Race									
White	0.149	0.018	0	0.150	0.080	0.033	0.027	0.059	0.067
Black	0.005	0.002	0	0.041	0.016	0.0001	0.0003	0.006	0.014
Hispanic	0.115	0.016	0	0.103	0.055	0.025	0.017	0.041	0.042
Asian	0.148	0.024	0	0.104	0.097	0.058	0.049	0.084	0.088
Unknown	0.005	0.004	0.0002	0.007	0.023	0.018	0.011	0.018	0.018
Commercial insurance/Medicare									
Commercial	0.383	0.030	0	0.082	0.088	0.045	0.049	0.077	0.095
Medicare	0.383	0.030	0	0.082	0.088	0.045	0.049	0.077	0.095
Antidiabetic medications									
Metformin	0.044	0.001	0	0.006	0.020	0.010	0.004	0.026	0.038
Sulfonylureas	0.125	0.009	0	0.021	0.006	0.009	0.003	0.010	0.003
Thiazolidinediones	0.037	0.011	0	0.059	0.024	0.023	0.014	0.046	0.057
SGLT2 inhibitors	0.313	0.028	0.0001	0.174	0.074	0.044	0.044	0.072	0.086
Insulin	0.555	0.043	0	0.067	0.059	0.062	0.059	0.086	0.100
Cardiovascular disease									
Myocardial infarction	0.003	0.0003	0	0.006	0.012	0.012	0.006	0.016	0.013
Ischemic stroke	0.078	0.012	0.0001	0.069	0.060	0.027	0.022	0.052	0.054
Peripheral artery disease	0.125	0.012	0	0.051	0.049	0.021	0.016	0.034	0.039
Transient ischemic attack	0.061	0.013	0.01	0.040	0.039	0.012	0.009	0.024	0.027

Notes: See Supplementary Table 1 for the full set of balancing results.
Abbreviation: SGLT2, sodium-glucose cotransporter-2.

original logistic regression. The entropy balancing weights provided the best performance among all models in balancing baseline characteristics (Table 2 and Supplementary Table 1). This method achieved near perfect balancing, with SMDs for all baseline characteristics very close to zero.

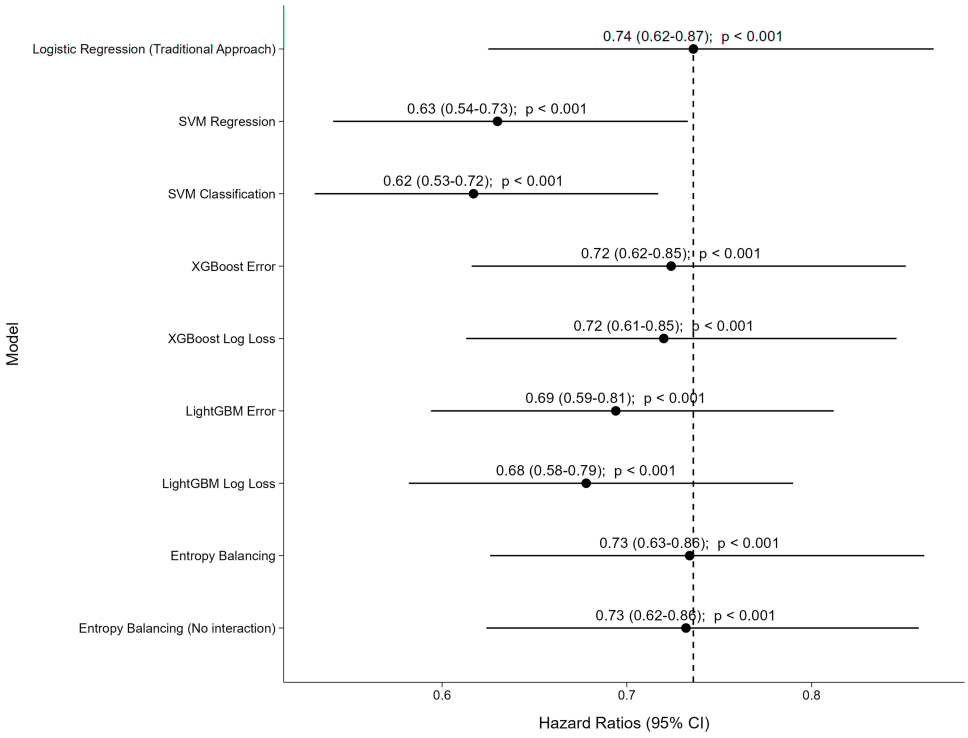

Figure 2 Forest Plots of Hazard Ratios of Ischemic Stroke Comparing OW GLP-1RA versus DPP-4i Using Different Weighting Methods.
Abbreviations: CI, confidence interval; DPP-4i, dipeptidyl peptidase-4 inhibitors; GLP-1RA, glucagon-like peptide-1 receptor agonists; OW, once-weekly; SVM, Support Vector Machine.

Additional Comparisons

Cox PH regression models showed significantly lower risk of ischemic stroke for the OW GLP-1RA users than for the DPP-4i users regardless of which method was used to estimate balancing weights. HRs from models using entropy balancing weights (with and without interaction) and XGBoost weights (XGBoost Error and XGBoost Log Loss) were very similar to the original logistic regression (Figure 2). The findings using SVM methods seem to have the largest difference in the HR magnitude from other methods, but the direction and statistical significance are the same as other methods (Figure 2).

Discussion

The INFORM study compared ML models and entropy balancing with a traditional logistic regression model to generate weights for IPTW. We utilized a large observational secondary data study, the cardiovascular outcomes study, as an example case to facilitate the comparison in real research practice. Overall, we found some ML models like XGBoost can achieve comparable performance and balancing compared to traditional logistic regression with manual/human/expert variable selection. In addition, entropy balancing appears to be the most suitable approach for optimizing weights, ie, achieving a highly balanced distribution. The key potential advantage of these advanced methods over the traditional approach, especially entropy balancing, is that they could minimize the need for expert-driven feature engineering and selection, particularly incorporation of the interaction terms into the model.

In our comparison of ML models, the performance of SVM was not the most favorable for this task, contrary to its reputation for effectiveness in binary classification.[48] Although SVM excels when faced with simpler feature spaces, its performance tends to diminish in higher dimensional feature spaces.[49] Although kernel functions are applied to address these complexities, their efficacy diminishes in highly complex feature spaces.[50] The propensity value distribution plot of the SVM models are different compared to all other models or the reference model, which can be attributed to the inherent nature of SVM in binary classification tasks. Unlike other algorithms that aim for a more symmetric separation between classes, SVM is designed to find the hyperplane that maximally separates the two classes while minimizing the

margin violations. This results in a unique characteristic where SVM attempts to position the hyperplane in a way that as few data points as possible are located close to it – that is, the propensity values will be denser at the extremes (0, 1) as compared to the other methods. Despite this, we deemed it worthwhile to explore SVM's utility in our study, considering that the relationships between the collected features and the prediction task might involve less intricate interconnections, making SVM well-suited for such scenarios.

On the other hand, our expectations for Gradient Boosting models, specifically XGBoost and LightGBM, were met with promising results. Given the high-dimensional nature of our data, Gradient Boosting was anticipated to perform well due to its ability to handle complex relationships within features. Both XGBoost and LightGBM demonstrated strong performance; however, XGBoost outperformed LightGBM. This could be attributed to the latter's emphasis on fast training performance rather than optimizing predictive accuracy. Nevertheless, our findings show that ML models perform similarly to traditional logistic regression and therefore, suggest a notable potential for ML models, particularly Gradient Boosting algorithms, in generating weights for IPTW in observational studies.

Entropy balancing, comparatively, shows the most promise, with near-perfect balancing between groups. The fact that performance remains high without the need to identify the true propensity function makes this technique ideal for real-world applications where domain knowledge is not available. In this study, the increase in balance did not result in a meaningfully different estimate as compared with the original logistic model; however, in studies where the logistic model is not correctly fully specified, we would expect this method to reduce bias in estimation of the true effect of the exposure on the outcome being measured. However, entropy balancing is not without its cautions. Because entropy balancing is an optimization problem, it is known that covariate balance, in some instances, may be achieved by assigning extreme weights to a small set of observations, in effect nearly removing the observations.[29,51] This is similar to instances of a propensity model where scores approach 0 or 1, but in entropy balancing, because we are estimating the weights directly, we cannot adjust for extreme results of this kind. Further, depending on the balance constraints (eg, covariates, distributional moments), entropy balancing may fail to converge where an otherwise simpler propensity model would succeed.[51] Future research could focus on improvements in covariate balance and estimation of the ATE for such an approach.

Examining the implications of our study underscores the continued utility of ML models for propensity scoring tasks, even though their performance, in terms of covariate balance, did not consistently surpass the traditional logistic regression model. Notably, ML models, devoid of explicit interaction knowledge, demonstrated comparable covariate balance to the logistic regression. It is essential to recognize that XGBoost emerged as a robust method for generating weights for use in IPTW, exhibiting superior performance compared to other ML models and closely approaching the effectiveness of the logistic model; exploring more suitable Gradient Boosting models for this task could yield even better results. The advantage of the traditional logistic regression remains in its familiarity and interpretability as long as interaction terms are held to a minimum. The traditional approach necessitated a manual, expert-driven process for feature engineering. This approach required domain knowledge to identify relevant variables that influence the outcome, a process that, while effective, introduced an inherent level of subjectivity and potential bias into the selection process. This complexity underscores the evolving landscape of the methodology, where ML models present a promising alternative. We expect it to autonomously identify and utilize the most predictive features from the dataset without explicit human intervention in the feature selection process. By inputting the entirety of the available data, these models leveraged algorithms capable of assessing the importance and relevance of each variable in relation to the predictive objective.

This study has several limitations. First, INFORM inherited some limitations from the original cardiovascular outcomes study, including the inability to establish causality, potential measurement errors (eg, misclassified billing codes), missing data, and unavailability of information (eg, date and cause of death). Second, as with the original study, INFORM used CDM data, including US commercial and Medicare Advantage enrollees; caution is needed to generalize the results to other populations. Third, our ML models focused on predicting the treatment administered to each patient, and subsequent IPTW were generated based on these predictions. While this approach provided valuable insights, an alternative and potentially more suitable task for ML models could involve the direct generation of weights optimized to minimize SMD values for better covariate balancing. Despite this limitation, our study revealed that ML models, including XGBoost, can achieve comparable covariate balance

compared to traditional logistic regression. Fourth, these advanced models are used as an intermediary step for bias reduction when estimating the sATE, but they may not be able to fully eliminate bias to establish causality. Lastly, we did not validate these methods using simulation, which warrants future research.

Conclusion

Among the advanced methods we explored, entropy balancing weights performed the best for optimizing balancing and can produce similar results compared to the traditional logistic regression. It could also minimize the expert-driven feature engineering and selection, especially efforts to incorporate the interaction terms into the model.

Abbreviations

AI, artificial intelligence; ASCVD, atherosclerotic cardiovascular disease; AUC, area under the curve; CDM, Optum's de-identified Clinformatics® Data Mart Database; CI, confidence interval; DPP-4i, dipeptidyl peptidase-4 inhibitors; GLP-1RA, glucagon-like peptide-1 receptor agonists; HR, hazard ratio; IPTW, inverse probability of treatment weighting; ML, machine learning; OW, once-weekly; PH, proportional hazards; RWE, real-world evidence; sATE, sample average treatment effect; SD, standard deviation; SMD, standardized mean difference; SVM, Support Vector Machine; SVR, Support Vector Regression; T2D, type 2 diabetes; US, United States.

Data Sharing Statement

Optum's de-identified Clinformatics® Data Mart Database was commercially licensed from the data vendor. Restrictions apply to the availability of these data, which were used under license of this study.

Acknowledgments

The authors thank Rebecca Hahn, MPH, of KJT Group, Inc. Rochester, NY, for medical writing support, which was funded by Novo Nordisk Inc. The authors also wish to thank Chris Claeys, MS, of KJT Group, Inc. for his review and contributions to the manuscript drafts.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study was funded by Novo Nordisk Inc. Novo Nordisk employees contributed to the design and conduct of the study, and writing and approving the paper.

Disclosure

X.T. and Y.L. are employees of Novo Nordisk.

D.K. is a PhD student at Texas A&M University and worked as a summer intern at Novo Nordisk during the summer of 2023.

X.S. has nothing to disclose.

References

1. Chesnaye NC, Stel VS, Tripepi G, et al. An introduction to inverse probability of treatment weighting in observational research. *Clin Kidney J*. 2022;15(1):14–20. doi:10.1093/ckj/sfab158
2. Jager KJ, Zoccali C, Macleod A, Dekker FW. Confounding: what it is and how to deal with it. *Kidney Int*. 2008;73(3):256–260. doi:10.1038/sj.ki.5002650
3. Braga LH, Farrokhyar F, Bhandari M. Confounding: what is it and how do we deal with it? *Can J Surg Apr*. 2012;55(2):132–138. doi:10.1503/cjs.036311

4. Kahlert J, Gribsholt SB, Gammelager H, Dekkers OM, Luta G. Control of confounding in the analysis phase - an overview for clinicians. *Clin Epidemiol*. 2017;9:195–204. doi:10.2147/clep.S129886
5. Pourhoseingholi MA, Baghestani AR, Vahedi M. How to control confounding effects by statistical analysis. *Gastroenterol Hepatol Bed Bench*. 2012;5(2):79–83.
6. Javaid M, Haleem A, Pratap Singh R, Suman R, Rab S. Significance of machine learning in healthcare: features, pillars and applications. *Inte J Intellig Net*. 2022;3:58–73. doi:10.1016/j.ijin.2022.05.002
7. Habehh H, Gohel S. Machine learning in healthcare. *Curr Genomics*. 2021;22(4):291–300. doi:10.2174/1389202922666210705124359
8. Weissler EH, Naumann T, Andersson T, et al. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*. 2021;22(1):537. doi:10.1186/s13063-021-05489-x
9. An Q, Rahman S, Zhou J, Kang JJ. A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges. *Sensors*. 2023;23(9). doi:10.3390/s23094178
10. Ericson O, Hjelmgren J, Sjövall F, Söderberg J, Persson I. The potential cost and cost-effectiveness impact of using a machine learning algorithm for early detection of sepsis in intensive care units in Sweden. *J Health Econ Outcomes Res*. 2022;9(1):101–110. doi:10.36469/jheor.2022.33951
11. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med Res Met* 2022;22 (1):287. doi:10.1186/s12874-022-01768-6
12. Lee W, Schwartz N, Bansal A, et al. A scoping review of the use of machine learning in health economics and outcomes research: part 2-data from nonwearables. *Value Health*. 2022;25(12):2053–2061. doi:10.1016/j.jval.2022.07.011
13. Padula WV, Kreif N, Vanness DJ, et al. Machine learning methods in health economics and outcomes research—The PALISADE checklist: a Good Practices Report of an ISPOR Task Force. *Value Health*. 2022;25(7):1063–1080. doi:10.1016/j.jval.2022.03.022
14. Adamson B, Waskom M, Blarre A, et al. Approach to machine learning for extraction of real-world data variables from electronic health records. Hypothesis and Theory. *Front Pharm*. doi:10.3389/fphar.2023.1180962
15. Benedum CM, Sondhi A, Fidyk E, et al. Replication of real-world evidence in oncology using electronic health record data extracted by machine learning. *Cancers*. 2023;15(6). doi:10.3390/cancers15061853
16. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Meth* 2019;19(1):64. doi:10.1186/s12874-019-0681-4
17. Hill NR, Ayoubkhani D, McEwan P, et al. Predicting atrial fibrillation in primary care using machine learning. *PLoS One*. 2019;14(11):e0224582. doi:10.1371/journal.pone.0224582
18. Ru B, Tan X, Liu Y, et al. Comparison of machine learning algorithms for predicting hospital readmissions and worsening heart failure events in patients with heart failure with reduced ejection fraction: modeling study. *JMIR Form Res*. 2023;7(41775). doi:10.2196/41775
19. Tsang C, Huda A, Norman M, et al. Detecting transthyretin amyloid cardiomyopathy (ATTR-CM) using machine learning: an evaluation of the performance of an algorithm in a UK setting. *BMJ Open*. 2023;13(10):e070028. doi:10.1136/bmjopen-2022-070028
20. Hill NR, Groves L, Dickerson C, et al. Identification of undiagnosed atrial fibrillation using a machine learning risk prediction algorithm and diagnostic testing (PULsE-AI) in primary care: cost-effectiveness of a screening strategy evaluated in a randomized controlled trial in England. *J Med Econ*. 2022;25(1):974–983. doi:10.1080/13696998.2022.2102355
21. Wang Z, Chen X, Tan X, et al. Using deep learning to identify high-risk patients with heart failure with reduced ejection fraction. *J Health Econ Outcomes Res*. 2021;8(2):6–13. doi:10.36469/jheor.2021.25753
22. Triantafyllidis AK, Tsanas A. Applications of machine learning in real-life digital health interventions: review of the literature. *J Med Internet Res*. 2019;21(4):e12286. doi:10.2196/12286
23. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am J Epidemiol*. 2017;185(1):65–73. doi:10.1093/aje/kww165
24. Crown WH. Real-world evidence, causal inference, and machine learning. *Value Health*. 2019;22(5):587–592. doi:10.1016/j.jval.2019.03.001
25. Rostami M, Saarela O. Normalized augmented inverse probability weighting with neural network predictions. *Entropy*. 2022;24(2). doi:10.3390/e24020179
26. Costello MJ, Li Y, Zhu Y, et al. Using conventional and machine learning propensity score methods to examine the effectiveness of 12-step group involvement following inpatient addiction treatment. *Drug Alcohol Depend*. 2021;227:108943. doi:10.1016/j.drugalcdep.2021.108943
27. Hainmueller J. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Political Anal*. 2012;20(1):25–46. doi:10.1093/pan/mpr025
28. Larrain N, Groene O. Improving the evaluation of an integrated healthcare system using entropy balancing: population health improvements in Gesundes Kinzigtal. *SSM Popul Health*. 2023;22:101371. doi:10.1016/j.ssmph.2023.101371
29. Parish WJ, Keyes V, Beadles C, Kandilov A. Using entropy balancing to strengthen an observational cohort study design: lessons learned from an evaluation of a complex multi-state federal demonstration. *Health Servi Outc Res Meth*. 2018;18(1):17–46. doi:10.1007/s10742-017-0174-z
30. Hwang U, Dresden SM, Vargas-Torres C, et al. Association of a geriatric emergency department innovation program with cost outcomes among Medicare beneficiaries. *JAMA Network Open*. 2021;4(3):e2037334–e2037334. doi:10.1001/jamanetworkopen.2020.37334
31. Yu G, Bian Y, Gamalo M. Power priors with entropy balancing weights in data augmentation of partially controlled randomized trials. *J Biopharl Stati*. 2022;32(1):4–20. doi:10.1080/10543406.2021.2021226
32. Ricci C, Kauffmann EF, Pagnanelli M, et al. Minimally invasive versus open radical antegrade modular pancreatosplenectomy for pancreatic ductal adenocarcinoma: an entropy balancing analysis. *HPB*. doi:10.1016/j.hpb.2023.09.013
33. Matschinger H, Heider D, König -H-H. A comparison of matching and weighting methods for causal inference based on routine health insurance data, or: what to do if an RCT is impossible. *Gesundheitswesen*. 2020;82(02):S139–S150. doi:10.1055/a-1009-6634
34. Tübbicke S. Entropy balancing for continuous treatments. *J Econom Meth*. 2022;11(1):71–89. doi:10.1515/jem-2021-0002
35. Josey KP, Berkowitz SA, Ghosh D, Raghavan S. Transporting experimental results with entropy balancing. *Stat Med*. 2021;40(19):4310–4326. doi:10.1002/sim.9031
36. Tan X, Liang Y, Rajpura JR, et al. Once-weekly glucagon-like peptide-1 receptor agonists vs dipeptidyl peptidase-4 inhibitors: cardiovascular effects in people with diabetes and cardiovascular disease. *Cardiovasc Diabetol*. 2023;22(1):319. doi:10.1186/s12933-023-02051-8
37. Optum Clinformatics® Data Mart. https://www.optum.com/content/dam/optum/resources/productSheets/Clinformatics_for_Data_Mart.pdf. Accessed December 7, 2023.

38. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. *IEEE Intelligent Sys Appli*. 1998;13(4):18–28. doi:10.1109/5254.708428
39. Awad M, Khanna R. Support Vector Regression. In: Awad M, Khanna R, editors. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. 2015:67–80.
40. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. *Artif Intell Rev*. 2021;54(3):1937–1967. doi:10.1007/s10462-020-09896-5
41. Chen T, Guestrin C XGBoost: a scalable tree boosting system. In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785–794.
42. Sai MJ, Chettri P, Panigrahi R, Garg A, Bhoi AK, Barsocchi P. An ensemble of Light Gradient Boosting Machine and Adaptive Boosting for prediction of type-2 diabetes. *Int J Comput Intell Syst*. 2023;16(1):14. doi:10.1007/s44196-023-00184-y
43. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36. doi:10.1148/radiology.143.1.7063747
44. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. *Cam Univ Press*. 2008;155–157.
45. Austin PC. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharma Drug Saf*. 2008;17(12):1202–1217. doi:10.1002/pds.1673
46. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*. 2011;46(3):399–424. doi:10.1080/00273171.2011.568786
47. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med*. 2015;34(28):3661–3679. doi:10.1002/sim.6607
48. Melgani F, Bruzzone L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans Geoscien Remote Sensin*. 2004;42(8):1778–1790. doi:10.1109/TGRS.2004.831865
49. Zhang S, Hossain M, Hassan M, Bailey J, Ramamohanarao K Feature Weighted SVMs Using Receiver Operating Characteristics. In: *Proceedings of the SIAM International Conference on Data Mining*. 2009:497–508.
50. Kamath U, Shehu A, Jong KD Using evolutionary computation to improve SVM classification. In: *Proceedings of the IEEE Congress on Evolutionary Computation*. 2010:1–8.
51. McMullin J, Schonberger B. When good balance goes bad: a discussion of common pitfalls when using entropy balancing. *J Fin Rep*. 2022;7(1):167–196. doi:10.2308/JFR-2021-007