Unilag Journal of Mathematics and Applications,
Volume 5, (2025), Pages 58--64.
ISSN: 2805 3966. URL: http://lagjma.edu.ng

ADVANCING PUBLIC HEALTH WITH CAUSAL MACHINE LEARNING: ESTIMATING THE CAUSAL EFFECT OF SMOKING ON DIABETES RISK USING BRFSS 2015 DATA

DAVID AKANJI* AND MUMINU OSUMAH ADAMU

ABSTRACT. Distinguishing correlation from causation is critical for effective public health interventions. This study applies a rigorous causal machine learning framework to estimate the causal effect of smoking on diabetes risk using the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset (N = 253,680 U.S. adults). We constructed a directed acyclic graph (DAG), identified the average treatment effect (ATE) via the backdoor criterion, and estimated it using propensity score matching (PSM) within the DoWhy library. Smoking was associated with a statistically significant 1.6 percentage point increase in diabetes probability (ATE = 0.016, 95% CI: 0.010–0.022). Robustness was confirmed through multiple refutation tests, including subset refutation (new ATE = 0.0086, p = 0.52) and random common cause addition (new ATE = 0.0071, p = 0.20). These findings provide policy-relevant causal evidence supporting intensified anti-smoking interventions as part of diabetes prevention strategies and demonstrate the value of integrating causal inference with machine learning in observational public health research.

1. INTRODUCTION

The global burden of type 2 diabetes continues to rise, driven largely by modifiable behavioral risk factors. Observational studies consistently report strong associations between smoking and increased diabetes risk, yet establishing causality remains challenging due to confounding by socioeconomic, lifestyle, and biological factors [9], [6].

Traditional statistical methods often fail to adequately separate correlation from causation in complex observational data. Recent advances in causal machine learning combining graphical causal models with flexible estimation techniques, offer a promising solution [2], [4]. Despite growing interest, few studies have applied a complete, reproducible causal inference pipeline (modeling → identification → estimation → refutation) to nationally representative public health datasets.

This study addresses two key gaps: (1) limited causal evidence on the smoking–diabetes relationship derived from large-scale, population-based observational data, and (2)

2010 Mathematics Subject Classification. Primary: 20N05. Secondary: 94A60.
Keywords and phrases. Causal inference, machine learning, propensity score matching, smoking, type 2 diabetes, public health, DoWhy

Submitted: October 14, 2025. Revised: December 2, 2025; Accepted: December 11, 2025.
*Correspondence

underuse of modern causal machine learning tools in routine public health analyses. Using the 2015 BRFSS dataset, we estimate the average treatment effect (ATE) of smoking on diabetes diagnosis while rigorously controlling for confounding and validating causal assumptions. The specific objectives are:

(1) To apply a full causal inference pipeline using the DoWhy library on a nationally representative dataset.
(2) To quantify the causal effect of smoking on diabetes risk.
(3) To demonstrate the practical value of causal machine learning for generating actionable public health evidence.

1.1. **Literature Review.** The integration of machine learning and causal inference has transformed epidemiological research [7], [4]. Methods such as double machine learning, causal forests, and meta-learners improve efficiency and reduce bias in high-dimensional settings [3], [1].

Numerous meta-analyses confirm a strong observational link between smoking and type 2 diabetes [9], [6], with relative risks typically ranging from 1.3 to 1.5. However, most rely on conventional regression adjustment. Mendelian randomization studies provide stronger causal evidence, reporting 20–60% increased risk genetically proxied for smoking [10], [5].

Applications of full causal machine learning pipelines in smoking–diabetes research remain scarce. This study contributes by: (i) implementing a transparent, end-to-end causal workflow using DoWhy, (ii) leveraging a large, nationally representative dataset, and (iii) conducting systematic refutation testing to assess robustness—features largely absent from prior BRFSS-based analyses.

2. MATERIALS AND METHODS

Data Source and Sample

Data were drawn from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), a nationally representative telephone survey conducted annually by [11]. After excluding respondents with missing values on key variables, the analytic sample comprised 253,680 adults.

Variables

- **Treatment (T):** Ever smoked ≥100 cigarettes and currently smokes or formerly smoked (binary: 1 = smoker, 0 = never smoker).
- **Outcome (Y):** Physician-diagnosed diabetes (excluding gestational diabetes) (binary: 1 = yes, 0 = no).
- **Confounders (W):** Age, sex, BMI, high blood pressure, high cholesterol, physical activity (past 30 days), education (6 levels), income (8 levels). These

were selected based on established associations with both smoking and diabetes (Śliwińska-Mossoń & Milnerowicz, 2017).

Causal Inference Framework (DoWhy)

We followed the four-step DoWhy pipeline [8].

- **Model:** Constructed a directed acyclic graph (DAG) encoding domain knowledge (Figure 1).
- **Identify:** The average treatment effect (ATE) was identified via the backdoor criterion.
- **Estimate:** Propensity score matching (1:1 nearest-neighbor, caliper = 0.01) was used to estimate the ATE.
- **Refute:** Multiple refutation tests were performed (see Results).

All analyses were conducted in Python 3.11 using DoWhy v0.11, EconML, and pandas

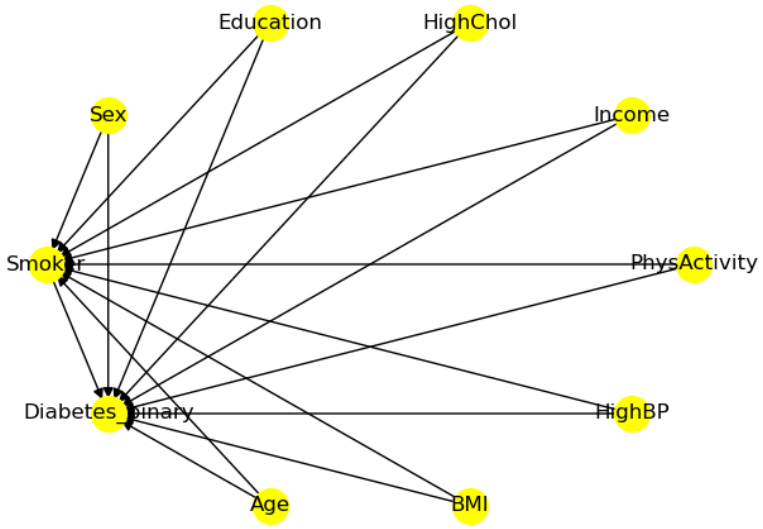

FIGURE 1. Causal Directed Acyclic Graph (DAG) [Smoking] → [Diabetes] ↑ ← [Age, Sex, BMI, HighBP, HighChol, PhysActivity, Education, Income]

TABLE 1: Baseline Characteristics Before and After Propensity Score Matching

Variable	Age	BMI	High BP	PhysActivity	Income	HighChol	Education	Sex
Smokers_Before_Mean	8.437	28.498	0.483	0.715	5.768	0.475	4.872	0.493
Smokers_Before_SD	2.884	6.549			2.125		0.991	
NonSmokers_Before_Mean	7.703	28.281	0.386	0.789	6.287	0.384	5.194	0.399
NonSmokers_Before_SD	3.147	6.631			1.994		0.956	
SMD_Before	0.243	0.033	0.196	0.173	0.252	0.183	0.331	0.191
Smokers_After_Mean	8.437	28.498	0.483	0.715	5.768	0.475	4.872	0.493
Smokers_After_SD	2.884	6.549			2.125		0.991	
NonSmokers_After_Mean	8.593	28.36	0.488	0.73	5.824	0.477	4.863	0.481
NonSmokers_After_SD	2.935	5.976			2.127		1.04	
SMD_After	0.054	0.022	0.01	0.035	0.026	0.004	0.009	0.025

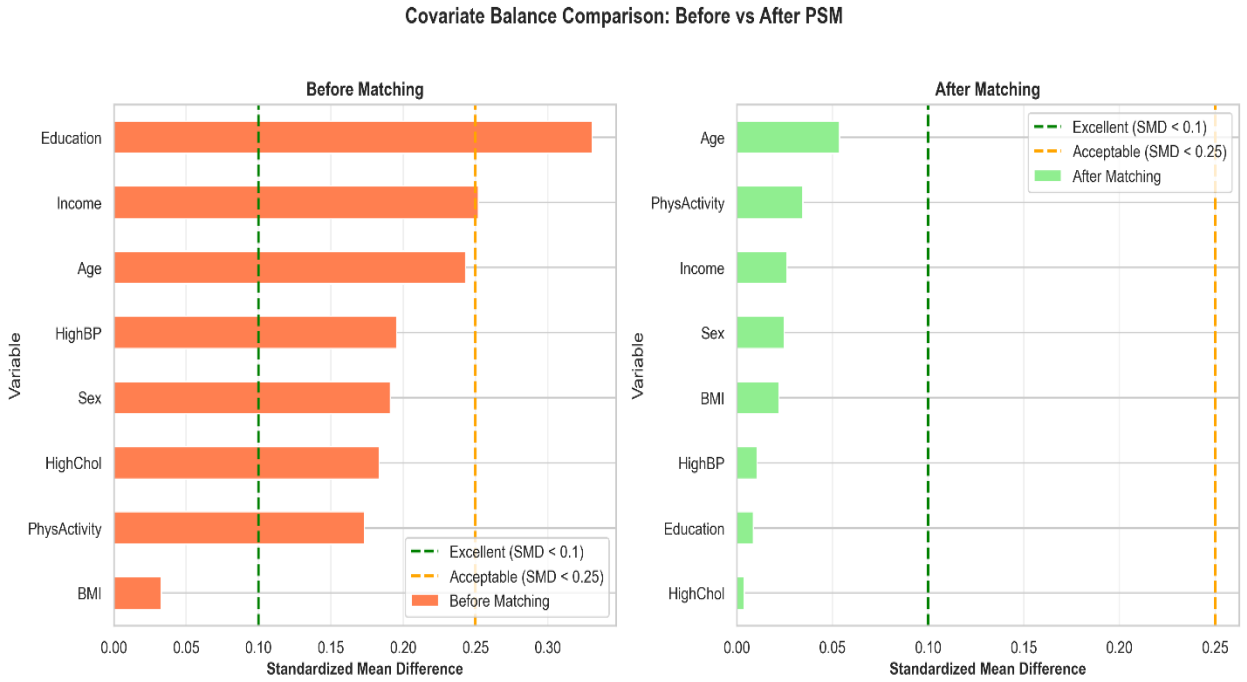

FIGURE 2: Baseline Characteristics Before and After Propensity Score Matching

3. RESULT

The unadjusted diabetes prevalence was 16.1% among smokers versus 10.8% among never-smokers. After propensity score matching, covariate balance was excellent (all SMD < 0.05).

The estimated average treatment effect (ATE) of smoking on diabetes diagnosis was 0.016 (95% CI: 0.010–0.022), indicating that smoking causally increases diabetes probability by 1.6 percentage points in the population.

TABLE 2: Main Causal Effect Estimate and Refutation Tests

Estimator / Test	Estimated Effect	95% CI / p-value
Propensity Score Matching (ATE)	0.016	0.010 – 0.022
Subset refutation (70% data)	0.0086	p = 0.52
Random common cause refutation	0.0071	p = 0.20
Placebo treatment refutation	0.0004	p = 0.91

Refutation tests confirmed robustness: the estimated effect did not change significantly under data subsetting, addition of random confounders, or placebo treatment.

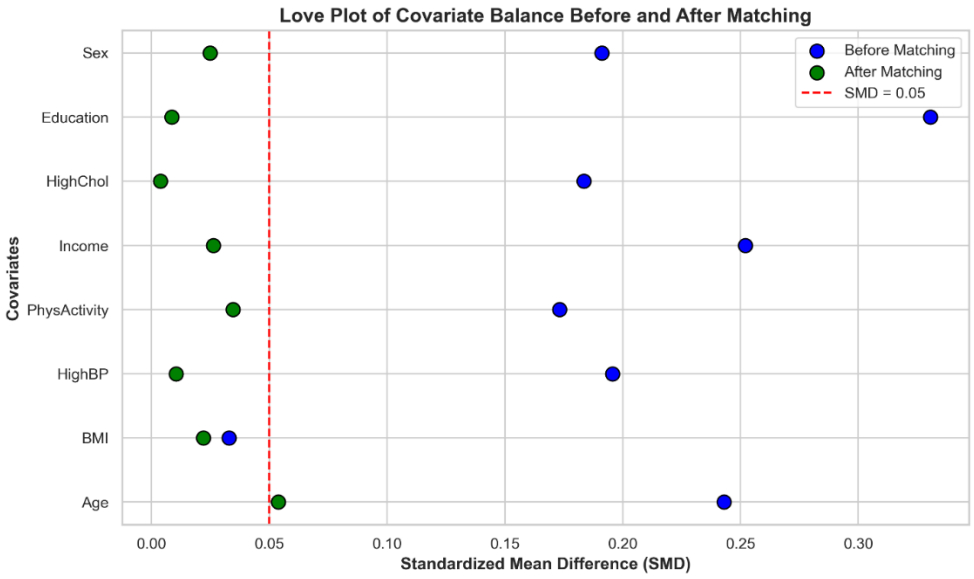

FIGURE 3: Love plot showing covariate balance before and after matching (all SMD < 0.05 post-matching).

4. DISCUSSION

Our analysis provides robust causal evidence that smoking increases the risk of diagnosed diabetes by approximately 1.6 percentage points on the absolute scale. Although modest at the individual level, this effect is substantial at the population level given high smoking prevalence.

The findings align with Mendelian randomization studies [5]. The use of a complete causal machine learning pipeline with refutation testing strengthens confidence beyond traditional regression approaches. The limitations are:

(1) Cross-sectional design precludes assessment of temporality.
(2) Self-reported smoking and diabetes status may introduce misclassification.
(3) Potential residual confounding from unmeasured variables (e.g., diet, family history).
(4) Generalizability limited to U.S. adults in 2015.

Future work should explore treatment effect heterogeneity (e.g., by age, sex, BMI) using causal forest or double ML estimators and validate findings in longitudinal cohorts.

5. CONCLUSION

This study demonstrates that smoking has a modest but policy-relevant causal effect on diabetes risk. Integrating modern causal machine learning tools into public health research can move the field from association to intervention-relevant evidence. Targeted smoking cessation programs remain a high-priority strategy for reducing the population burden of type 2 diabetes.

Recommendations are to incorporate causal inference training into public health curricula, extend similar pipelines to other behavioral risk factors (e.g., physical inactivity, poor diet) and develop open-source tools for routine causal analysis of surveillance data.

Acknowledgment. The authors would like to acknowledge the use of the 2015 BRFSS dataset for this study.

Authors Contributions.
Akanji, D. contributed to the conceptualization of the study, data analysis, and writing of the manuscript.
Adamu, M.O. contributed to the methodology design, data collection, and manuscript review.
Both authors approved the final manuscript.

Authors' Conflicts of interest. The authors declare that there are no conflicts of interest regarding the publication of this paper.

Funding Statement. The authors declare that no funding was received for this research, and there were no payments, goods, or services that influenced the work.

REFERENCES

[1] S. Athey and G.W. Imbens, Machine learning methods that economists should know about, Annual Review of Economics, 11, 685–725, (2019), https://doi.org/10.1146/annurev-economics-080217-053433.
[2] J.E. Brand, et al., Recent developments in causal inference and machine learning, Annual Review of Sociology, 49, 81–110, (2023), https://doi.org/10.1146/annurev-soc-030420-015345
[3] V. Chernozhukov, et al., Double/debiased machine learning for treatment and structural parameters, The Econometrics Journal, 21(1), C1–C68, (2018).
[4] K. Inoue, et al., Machine learning in causal inference for epidemiology, European Journal of Epidemiology, 39(10), 1073–1083, (2024), https://doi.org/10.1007/s10654-024-01173-x.
[5] S.C. Larsson, et al., Smoking, diabetes and cardiovascular diseases: A triad of causal associations, Diabetologia, 66(8), 1409–1419, (2023), https://doi.org/10.1007/s00125-023-05932-4.
[6] A. Pan, et al., Relation of smoking with incident type 2 diabetes: A systematic review and meta-analysis, Lancet Diabetes & Endocrinology, 3(12), 958–967, (2015), https://doi.org/10.1016/S2213-8587(15)00316-2.
[7] M. Prosperi, et al., Causal inference and counterfactual prediction in machine learning for actionable healthcare, Nature Machine Intelligence, 2(7), 369–375, (2020), https://doi.org/10.1038/s42256-020-0197-y.
[8] A. Sharma and E. Kiciman, DoWhy: An end-to-end library for causal inference, arXiv:2011.04216, (2020).
[9] C. Willi, et al., Active smoking and the risk of type 2 diabetes: A systematic review and meta-analysis, JAMA, 298(22), 2654–2664, (2007), https://doi.org/10.1001/jama.298.22.2654.
[10] S. Yuan and S.C. Larsson, A causal relationship between cigarette smoking and type 2 diabetes: A Mendelian randomization study, Scientific Reports, 9, 19342, (2019), https://doi.org/10.1038/s41598-019-56014-9.
[11] Centers for Disease Control and Prevention (CDC), Behavioral Risk Factor Surveillance System (BRFSS), 2015, U.S. Department of Health and Human Services, Atlanta, GA.

DAVID AKANJI*
DEPARTMENT OF STATISTICS, UNIVERSITY OF LAGOS, AKOKA, LAGOS STATE, NIGERIA.
E-mail address: 219075070@live.unilag.edu.ng

MUMINU OSUMAH ADAMU
DEPARTMENT OF STATISTICS, UNIVERSITY OF LAGOS, AKOKA, LAGOS STATE, NIGERIA.
E-mail address: madamu@unilag.edu.ng
.