

Avoiding Data Leakage in Machine Learning Models for Diabetes and Hypertension Classification Using Routine Occupational Health Screening Data

Ghasem Alizadeh-dizaj

Health Management Research Center, Baqiyatallah University of Medical Sciences

Mohammad Amiri-Ara

Health Management Research Center, Baqiyatallah University of Medical Sciences

Ehsan Teymourzadeh

Baqiyatallah University of Medical Sciences

Parisa Mehdizadeh

Baqiyatallah University of Medical Sciences

Mehdi Mesri

School of Medicine, Baqiyatallah University of Medical Sciences

Maryam Yaghoubi

`maryamyaghoubi2023@gmail.com`

Health Management Research Center, Baqiyatallah University of Medical Sciences

Research Article

Keywords: Data Leakage, Machine Learning, Diabetes Mellitus, Hypertension, XGBoost, Random Forest, Methodological Bias, Occupational Health, Routine Screening Data, SHAP

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10780902/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background and Aim:

Reliable clinical prediction using machine learning requires leakage-free model development, yet this critical methodological concern is often overlooked. Type 2 diabetes and hypertension represent major public health challenges in Iran, and early identification of at-risk individuals using routine workplace health screening data could enable cost-effective prevention. This study had a dual objective: first, to develop and validate machine learning models for identifying diabetes and hypertension using routine occupational health screening data; and second and more critically to demonstrate how the inclusion of outcome-defining predictors (fasting blood glucose and blood pressure) artificially inflates model performance through data leakage, a methodological pitfall often overlooked in clinical prediction studies. By comparing 'full' models (with all predictors) against 'clinically meaningful' models (excluding outcome-defining variables), we provide empirical evidence of the magnitude of this bias and offer a transparent framework for avoiding it in similar studies.

Methods

A retrospective cross-sectional analysis was conducted on 370 health screening records. Five algorithms Decision Tree, Ridge Classifier, Logistic Regression, XGBoost, and Random Forest were developed and compared. Two modeling strategies were implemented: (1) a full model including all predictors, and (2) a clinically meaningful model excluding outcome-defining predictors (fasting blood glucose and blood pressure). The dataset was split into 80% training and 20% testing sets. Five-fold cross-validation was used for hyperparameter optimization, and SMOTE was applied solely to the training set to address class imbalance. Performance was evaluated using accuracy, sensitivity, specificity, F1 score, AUC-ROC with 95% CI, calibration plots, and Brier scores. Feature importance was assessed using SHAP values.

Results

The prevalence of diabetes and hypertension was 18.4% (68 cases) and 20.5% (76 cases), respectively. In the clinically meaningful models that excluded outcome-defining predictors (fasting blood glucose and blood pressure), discrimination performance was moderate, with Random Forest achieving an AUC-ROC of 78.34% (95% CI: 72.15–84.92) for diabetes and 76.81% (95% CI: 70.23–83.45) for hypertension. In contrast, when these predictors were included in the full models, performance was artificially inflated (Random Forest AUC-ROC: 96.22% for diabetes and 94.59% for hypertension), confirming substantial data leakage. XGBoost performed comparably in both approaches. SHAP analysis identified age, BMI, waist circumference, and psychological distress (GSI) as the most important non-outcome predictors.

Conclusion

After excluding outcome-defining predictors to prevent data leakage, machine learning models particularly Random Forest and XGBoost showed moderate but clinically realistic discriminative ability for identifying diabetes and hypertension using routine screening data. Psychological variables contributed meaningfully to prediction, supporting integrated physical and mental health screening in occupational settings. External validation in independent cohorts is required before clinical implementation.

1. Introduction

Non-communicable diseases, particularly type 2 diabetes and hypertension, are leading causes of morbidity and mortality worldwide [1, 2]. In Iran, approximately 14% of adults have diabetes and 25–32% have hypertension, with a substantial proportion remaining undiagnosed until complications arise [3–6]. Early identification of at-risk individuals is critical for cost-effective prevention [7, 8], yet screening in many settings remains opportunistic rather than systematic [9].

Workplace health programs, which routinely collect anthropometric, biochemical, and lifestyle data during annual employee screenings, offer a valuable opportunity for systematic risk identification [10, 11]. Healthcare workers, including university staff, exhibit elevated cardiometabolic risk due to occupational stress and lifestyle factors [12–14]. Despite the availability of such data, they are often underutilized for organizational-level risk prediction.

Traditional risk scores like FINDRISC and Framingham scores have limitations: they require additional questionnaire data not routinely collected and show reduced accuracy in Middle Eastern populations due to ethnic differences in risk profiles [15–18]. Machine learning (ML) offers potential advantages for risk prediction, including automatic detection of non-linear patterns and complex interactions [19, 20]. Recent studies have applied ML to electronic health records with promising results [21–23]. However, most of this research originates from high-income countries and focuses on general community samples [24], with limited evidence from Middle Eastern occupational settings [25].

Despite the growing interest in ML-based prediction models, a critical methodological concern data leakage remains insufficiently addressed. Data leakage occurs when information from the outcome (e.g., diagnostic criteria) inadvertently influences the predictors, leading to artificially inflated performance estimates that do not generalize to real-world settings [28, 29]. This challenge is particularly relevant in prediction models where routinely measured clinical variables also define the outcome. In diabetes and hypertension prediction, fasting blood glucose and blood pressure measurements are both predictors and diagnostic criteria, making them a particularly insidious source of leakage. Many published ML studies have inadvertently included these variables as predictors, potentially reporting overly optimistic performance. Empirical evidence quantifying the impact of this bias in routine occupational health screening data remains scarce. To our knowledge, few studies have explicitly evaluated this issue, and none have systematically demonstrated it in an Iranian employee cohort while transparently reporting both full and clinically meaningful model performance. Reliable prediction in clinical settings requires leakage-free model development; otherwise, overly optimistic performance estimates may not translate to real-world clinical utility.

To address this gap, the present study has three objectives. The primary objective is to empirically quantify the impact of data leakage by comparing two modeling strategies: (1) a 'full model' including all predictors, and (2) a 'clinically meaningful model' that excludes outcome-defining variables (fasting blood glucose and blood pressure). This comparison is designed to provide empirical evidence of performance inflation due to data leakage—a methodological cautionary example for researchers using routine clinical data. The secondary objective is to develop and evaluate clinically meaningful machine learning models (Decision Tree, Ridge Classifier, Logistic Regression, XGBoost, and Random Forest) that avoid data leakage, with potential applicability in real-world occupational screening settings. The third objective is to identify the most influential non-outcome predictors using SHAP analysis, with particular attention to the incremental value of psychological variables (SCL-90-R dimensions) in occupational risk prediction.

2. Methods

2.1. Study Design and Setting

We conducted a retrospective cross-sectional study using annual health screening data from employees of a University of Medical Sciences in Tehran, Iran. Data were extracted anonymously from the electronic occupational health database for the 2022–2023 screening cycle. The study was approved by the Ethics Committee of the University of Medical Sciences. Due to the anonymized nature of the data, individual informed consent was waived by the ethics committee.

2.2. Study Population

A total of 370 records from annual employee health screenings were included in the final analysis. Inclusion criteria were: (1) completion of at least one complete annual screening record during the study period, (2) no missing data for the primary outcome variables (fasting blood glucose and blood pressure measurements), and (3) age between 20 and 65 years. Employees with missing outcome data (n = 23) were excluded. No other exclusion criteria were applied.

2.3. Variables and Outcome Definitions

All variables were extracted from the routine annual health screening records. Table 1 summarizes the complete list of predictor variables, categorized by domain, along with their measurement units and coding details. Outcome variables were defined separately as described below.

Table 1
Predictor Variables Categorized by Domain

Domain	Variables	Units / Coding
Demographic	Age	Years (continuous)
	Sex	Male / Female (binary)
Anthropometric	Body mass index (BMI)	kg/m ² (continuous)
	Waist circumference	cm (continuous)
	Hip circumference	cm (continuous)
	Neck circumference	cm (continuous)
Vital Signs	Systolic blood pressure	mmHg (continuous)
	Diastolic blood pressure	mmHg (continuous)
	Maximum blood pressure (MaxBP)	mmHg (continuous)
	Minimum blood pressure (MinBP)	mmHg (continuous)
Biochemical Markers	Fasting blood glucose (FBS)	mg/dL (continuous)
	Total cholesterol	mg/dL (continuous)
	Triglycerides	mg/dL (continuous)
	HDL cholesterol	mg/dL (continuous)
	LDL cholesterol	mg/dL (continuous)
Hematological	Red blood cell count (RBC)	×10 ⁶ /μL (continuous)
	Hemoglobin (HB)	g/dL (continuous)
	Hematocrit (Hct)	% (continuous)
	Mean corpuscular volume (MCV)	fL (continuous)
	Mean corpuscular hemoglobin concentration (MCHC)	g/dL (continuous)
	Platelet count (Plt)	×10 ³ /μL (continuous)
	White blood cell count (WBC)	×10 ³ /μL (continuous)
Renal Function	Blood urea nitrogen (BUN)	mg/dL (continuous)
	Creatinine (Cr)	mg/dL (continuous)
Lifestyle	Smoking status	Never / Former / Current (categorical)
	Physical activity level	Low / Moderate / High (categorical)

Domain	Variables	Units / Coding
Psychological (SCL-90-R)	Somatization mean (Som_Mean)	Score 0–4 (continuous)
	Depression mean (Dep_Mean)	Score 0–4 (continuous)
	Anxiety mean (Anx_Mean)	Score 0–4 (continuous)
	Positive symptom total (PST)	Count (continuous)
	Positive symptom distress index (PSDI)	Score 0–4 (continuous)
	Global severity index (GSI)	Score 0–4 (continuous)

Outcome Definitions (Prediction Targets)

Outcomes were defined as binary variables based on the following standardized diagnostic criteria:

Diabetes(Yes/No):

Fasting blood glucose ≥ 126 mg/dL **OR** a prior physician diagnosis of diabetes **OR** current use of glucose-lowering medication.

Hypertension(Yes/No):

Systolic blood pressure ≥ 140 mmHg **OR** diastolic blood pressure ≥ 90 mmHg **OR** a prior physician diagnosis of hypertension **OR** current use of antihypertensive medication.

2.4. Data Preprocessing

Outliers and implausible values (e.g., BMI > 55 kg/m² or age < 20 years) were manually reviewed and removed (n = 4 records). For missing data, which accounted for less than 5% across all variables, we performed imputation using the median for continuous variables and the mode for categorical variables. All continuous predictors were standardized (mean = 0, standard deviation = 1) prior to machine learning modeling. Categorical variables were one-hot encoded.

2.5. Data Split and Cross-Validation

The dataset (N = 370) was randomly split into two parts using stratified random sampling with a fixed random seed (42) to ensure reproducibility:

- 80% training set (n = 296) for model development
- 20% test set (n = 74) for final evaluation

Stratified splitting was employed to preserve the proportion of diabetes and hypertension cases in both sets. Following the initial 80/20 split, five-fold cross-validation was performed on the training set for hyperparameter optimization and internal validation.

2.6. Addressing the Primary Methodological Concern: Data Leakage

This methodological consideration represents the central contribution of our study. To enable transparent assessment of data leakage, we deliberately designed two parallel modeling strategies—a 'full model' and a 'clinically meaningful model'—and compared their performances to quantify the inflation caused by leakage. Specifically:

- Fasting blood glucose is used both as a predictor AND as a diagnostic criterion for diabetes (≥ 126 mg/dL)
- Blood pressure measurements (systolic and diastolic) are used both as predictors AND as diagnostic criteria for hypertension ($\geq 140/90$ mmHg)

To address this concern transparently, we developed two separate modeling approaches:

Approach 1

Full Model (All Predictors) – Includes FBS and blood pressure measurements. Serves as an upper-bound benchmark.

Approach 2

Clinically Meaningful Model (Excluding Outcome-Defining Predictors) – Excludes FBS, systolic BP, diastolic BP, MaxBP, and MinBP. Provides a realistic estimate of performance in clinical practice.

Both approaches were applied to all five algorithms, and results are reported separately to ensure transparency.

2.7. Handling Class Imbalance with SMOTE

A common challenge in medical prediction tasks is class imbalance, where the number of cases (e.g., diabetic or hypertensive employees) is substantially smaller than non-cases. In our dataset, the prevalence of diabetes was 18.4% and hypertension was 20.5%. To address this imbalance and prevent models from being biased toward the majority class (healthy individuals), we applied SMOTE (Synthetic Minority Over-sampling Technique), as introduced by Chawla et al. (2002) [26].

SMOTE was applied as follows:

1. Only the training set (80% of the data, $n = 296$) was oversampled. The test set (20%, $n = 74$) remained unchanged to ensure realistic performance evaluation.
2. For each minority class sample (diabetes or hypertension), SMOTE identified its k nearest neighbors in the feature space ($k = 5$).
3. Synthetic samples were generated by linear interpolation between the minority sample and a randomly selected neighbor.

- 4. This process was repeated until the minority class reached approximately 50% of the majority class size.
- 5. SMOTE was applied separately for each outcome (diabetes and hypertension) to avoid data leakage between prediction tasks.
- 6. SMOTE was applied only to Approach 1 (full model) and Approach 2 (clinically meaningful model) identically.

The final minority-to-majority ratio after SMOTE application was approximately 1:2 (50% of majority class size).

2.8. Algorithms and Hyperparameter Optimization

We developed and compared five machine learning algorithms for both diabetes and hypertension classification. Table 2 summarizes the key specifications, base hyperparameters, and tuning ranges for each algorithm.

Table 2
Machine Learning Algorithms and Hyperparameter Tuning Specifications

Algorithm	Type / Description	Base Hyperparameters	Tuning Range (Grid Search)
Decision Tree	Single tree classifier with Gini impurity	max_depth = 5 min_samples_split = 5	max_depth: 3–10 min_samples_split: 2–10
Ridge Classifier	Linear classifier with L2 regularization	Default (no base)	alpha: 0.001, 0.01, 0.1, 1, 10
Logistic Regression	Standard binary logistic regression with L2 regularization	Default (no base)	C: 0.001, 0.01, 0.1, 1, 10, 100
XGBoost	Gradient boosting ensemble of decision trees	learning_rate = 0.1 max_depth = 4 n_estimators = 100 subsample = 0.8	learning_rate: 0.01–0.3 max_depth: 3–8 n_estimators: 50–200
Random Forest	Bagging ensemble of decision trees with Gini impurity	n_estimators = 100 max_depth = 10 min_samples_split = 2	n_estimators: 50–200 max_depth: 5–15 min_samples_split: 2–10

For each algorithm, hyperparameters were optimized using five-fold cross-validation on the SMOTE-resampled training set. The best hyperparameter combination was selected based on the average AUC-ROC across the five folds.

2.9. Performance Evaluation Metrics

Model performance was evaluated on the original (non-resampled) test set using the following metrics:

Accuracy: Proportion of correct predictions out of all cases

Sensitivity (Recall): Ability to correctly identify diseased individuals $[TP/(TP + FN)]$

Specificity: Ability to correctly identify healthy individuals $[TN/(TN + FP)]$

Precision: Positive predictive value $[TP/(TP + FP)]$

F1 Score: Harmonic mean of precision and recall $[2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})]$

AUC-ROC (Area Under the Receiver Operating Characteristic Curve): Overall measure of discriminative ability, with 95% confidence intervals calculated using the DeLong method

Brier Score: Measure of calibration (mean squared error of predicted probabilities)

Calibration Plots: Visual assessment of predicted vs. observed probabilities

2.10. Model Interpretability: SHAP Analysis

For the best-performing algorithms (Random Forest and XGBoost), we calculated SHAP (SHapley Additive exPlanations) values to assess feature importance. SHAP values provide a unified framework for interpreting model predictions by attributing the contribution of each feature to the final prediction. We generated:

- SHAP summary plots showing feature importance rankings
- SHAP dependence plots for the top predictors
- Mean absolute SHAP values as a measure of feature importance

This approach addresses the limitation of Gini impurity-based importance, which can be biased toward high-cardinality features.

2.11. Calibration Analysis

To assess the calibration of predicted probabilities, we generated calibration plots (observed vs. predicted probabilities) with 95% confidence intervals, and calculated the calibration slope (ideal = 1), calibration intercept (ideal = 0), and Brier score (lower values indicate better calibration), using the test set with 200 bootstrap resamples for confidence interval estimation.

2.12. Software Used

Data analysis and modeling were performed using Python version 3.10 with the following libraries:

- scikit-learn (version 1.2) for algorithm implementation, data splitting, and evaluation metrics
- imbalanced-learn (version 0.10) for SMOTE application
- xgboost (version 1.7) for XGBoost implementation

- shap (version 0.41) for SHAP value calculations
- matplotlib (version 3.5) and seaborn (version 0.12) for plotting
- pandas (version 1.5) and numpy (version 1.23) for data manipulation
- scipy (version 1.10) for statistical testing and confidence intervals

All code was executed in a Jupyter Notebook environment with a fixed random seed for reproducibility.

2.13. Statistical Analysis

Baseline characteristics were compared between groups using:

- Independent t-tests for normally distributed continuous variables
- Mann-Whitney U tests for non-normally distributed continuous variables
- Chi-square tests for categorical variables
- All tests were two-sided with $\alpha = 0.05$

All performance metrics are reported with 95% confidence intervals calculated using bootstrap resampling (1000 iterations) with the percentile method.

2.14. TRIPOD-AI Reporting Compliance

This study is reported in accordance with the TRIPOD-AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis Artificial Intelligence) statement. A completed TRIPOD-AI checklist is provided as Supplementary Table S1.

2.15. Ethical Considerations

All data were anonymized prior to analysis, and no identifying information of any employee is reported in the manuscript or any publication. Permission to use the de-identified data was obtained from the Occupational Health Unit of the university. The study was approved by the Ethics Committee of the University of Medical Sciences (Approval Number: IR.BMSU.REC.1405.043). The study adhered to the principles of the Declaration of Helsinki.

3. Results

3.1. Study Population and Baseline Characteristics

A total of 370 employee health screening records from the University of Medical Sciences were included in the final analysis. Table 3 presents the baseline demographic, clinical, and laboratory characteristics of the study population stratified by outcome status.

Table 3
Baseline Characteristics of the Study Population (N = 370)

Characteristic	Total (N = 370)	Diabetes (n = 68)	No Diabetes (n = 302)	P- value*	Hypertension (n = 76)	No Hypertension (n = 294)	P- value**
Age (years), mean $\pm$ SD	38.2 $\pm$ 7.6	42.1 $\pm$ 6.9	37.4 $\pm$ 7.5	< 0.001	41.5 $\pm$ 7.2	37.4 $\pm$ 7.5	< 0.001
Male, n (%)	285 (77.0)	60 (88.2)	225 (74.5)	0.023	70 (92.1)	215 (73.1)	< 0.001
BMI (kg/m ²), mean $\pm$ SD	27.1 $\pm$ 4.3	28.5 $\pm$ 4.8	26.8 $\pm$ 4.1	0.004	29.2 $\pm$ 4.5	26.6 $\pm$ 4.1	< 0.001
Overweight (BMI 25– 29.9), n (%)	135 (36.5)	25 (36.8)	110 (36.4)	0.956	28 (36.8)	107 (36.4)	0.942
Obese (BMI $\geq$ 30), n (%)	98 (26.5)	22 (32.4)	76 (25.2)	0.225	30 (39.5)	68 (23.1)	0.004
Waist circumference (cm), mean $\pm$ SD	94.5 $\pm$ 12.1	99.3 $\pm$ 13.5	93.5 $\pm$ 11.6	< 0.001	101.2 $\pm$ 13.0	92.8 $\pm$ 11.4	< 0.001
Hip circumference (cm), mean $\pm$ SD	102.8 $\pm$ 9.8	105.6 $\pm$ 10.2	102.2 $\pm$ 9.6	0.012	106.8 $\pm$ 10.5	101.8 $\pm$ 9.4	< 0.001
Neck circumference (cm), mean $\pm$ SD	36.4 $\pm$ 3.8	37.8 $\pm$ 4.1	36.1 $\pm$ 3.7	< 0.001	38.2 $\pm$ 4.2	35.9 $\pm$ 3.6	< 0.001
Systolic BP (mmHg), mean $\pm$ SD	120.5 $\pm$ 13.4	124.7 $\pm$ 16.2	119.6 $\pm$ 12.6	0.005	134.8 $\pm$ 11.5	116.8 $\pm$ 11.7	< 0.001
Diastolic BP (mmHg), mean $\pm$ SD	77.3 $\pm$ 9.2	79.2 $\pm$ 10.5	76.9 $\pm$ 8.9	0.058	85.4 $\pm$ 8.1	75.2 $\pm$ 8.5	< 0.001
FBS (mg/dL), mean $\pm$ SD	102.1 $\pm$ 32.4	140.7 $\pm$ 49.1	93.6 $\pm$ 17.9	< 0.001	107.5 $\pm$ 34.2	100.7 $\pm$ 32.0	0.112
Total cholesterol (mg/dL), mean $\pm$ SD	195.8 $\pm$ 38.2	201.5 $\pm$ 42.1	194.6 $\pm$ 37.5	0.181	204.2 $\pm$ 40.5	193.7 $\pm$ 37.6	0.038
Triglycerides (mg/dL), mean $\pm$ SD	148.6 $\pm$ 78.4	172.3 $\pm$ 85.2	143.2 $\pm$ 76.1	0.006	165.8 $\pm$ 82.1	144.2 $\pm$ 77.2	0.031

Characteristic	Total (N = 370)	Diabetes (n = 68)	No Diabetes (n = 302)	P- value*	Hypertension (n = 76)	No Hypertension (n = 294)	P- value**
HDL cholesterol (mg/dL), mean $\pm$ SD	46.2 $\pm$ 11.8	43.5 $\pm$ 10.8	46.8 $\pm$ 12.0	0.038	44.1 $\pm$ 11.2	46.8 $\pm$ 11.9	0.072
LDL cholesterol (mg/dL), mean $\pm$ SD	119.4 $\pm$ 32.6	122.8 $\pm$ 34.5	118.6 $\pm$ 32.3	0.342	124.5 $\pm$ 33.8	118.1 $\pm$ 32.3	0.128
Smoking (current), n (%)	52 (14.1)	12 (17.6)	40 (13.2)	0.341	15 (19.7)	37 (12.6)	0.112
Physical activity (low), n (%)	148 (40.0)	32 (47.1)	116 (38.4)	0.183	38 (50.0)	110 (37.4)	0.044
GSI (psychological distress), mean $\pm$ SD	0.38 $\pm$ 0.33	0.42 $\pm$ 0.35	0.37 $\pm$ 0.32	0.261	0.45 $\pm$ 0.38	0.36 $\pm$ 0.31	0.033

*P-value for diabetes vs. no diabetes comparison; **P-value for hypertension vs. no hypertension comparison.

Statistical tests: t-test for normally distributed continuous variables, Mann-Whitney U test for non-normally distributed variables, chi-square test for categorical variables.

SD = standard deviation; BMI = body mass index; BP = blood pressure; FBS = fasting blood glucose; HDL = high-density lipoprotein; LDL = low-density lipoprotein; GSI = global severity index.

The mean age of the study population was 38.2 ± 7.6 years, with 285 (77.0%) male and 85 (23.0%) female employees. The mean BMI was 27.1 ± 4.3 kg/m², with 135 (36.5%) participants classified as overweight (BMI 25–29.9 kg/m²) and 98 (26.5%) as obese (BMI ≥ 30 kg/m²). The prevalence of diabetes was 68 (18.4%), and the prevalence of hypertension was 76 (20.5%). Among employees with diabetes, 42 (61.8%) had a prior physician diagnosis, while the remainder were identified through elevated fasting blood glucose (FBS ≥ 126 mg/dL) during the annual screening. Similarly, among those with hypertension, 48 (63.2%) were previously diagnosed, and 28 (36.8%) were newly detected cases.

3.2. Effect of SMOTE on Class Balance

Before applying SMOTE, the training set (n = 296) contained 54 cases of diabetes (18.2%) and 61 cases of hypertension (20.6%). After applying SMOTE to the training set only, the number of diabetes cases increased to 242 (balanced to approximately 50% of the majority class), and the number of hypertension cases increased to 235 (balanced to approximately 50% of the majority class). The test set (n = 74)

remained unchanged with 14 diabetes cases (18.9%) and 15 hypertension cases (20.3%) to enable realistic performance evaluation.

3.3. Model Performance: Approach 1 (Full Model with All Predictors)

Table 4

presents the performance metrics of the five machine learning algorithms for predicting diabetes on the original (non-resampled) test set using the full model with all predictors.

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC-ROC (%)	95% CI	Brier Score
Decision Tree	89.19	50.00	87.50	68.75	57.89	82.19	71.45–90.34	0.184
Ridge Classifier	83.78	43.75	91.67	70.00	53.85	88.38	79.12–95.47	0.162
Logistic Regression	79.73	50.00	87.50	61.54	55.17	78.41	67.89–88.12	0.201
XGBoost	93.24	81.25	96.67	86.67	83.87	96.16	92.38–99.15	0.089
Random Forest	92.97	66.67	98.33	88.89	76.19	96.22	92.45–99.21	0.095

Table 4. Performance Comparison of Machine Learning Algorithms for Diabetes Prediction (Approach 1: Full Model)

Note

Values are based on the original (non-resampled) test set (n = 74). SMOTE was applied only to the training set. 95% CIs for AUC-ROC were calculated using the DeLong method with 1000 bootstrap iterations.

Among all algorithms, Random Forest achieved the highest AUC-ROC (96.22%, 95% CI: 92.45–99.21), while XGBoost achieved the highest accuracy (93.24%) and sensitivity (81.25%). Random Forest demonstrated superior specificity (98.33%), indicating excellent ability to correctly identify non-diabetic individuals. The Brier scores for XGBoost (0.089) and Random Forest (0.095) indicate good calibration, with lower values indicating better-calibrated probabilities.

3.4. Model Performance: Approach 1 (Full Model) for Hypertension

Table 5

summarizes the performance of the five algorithms for predicting hypertension using the full model.

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC-ROC (%)	95% CI	Brier Score
Decision Tree	79.73	76.47	81.25	50.98	61.16	76.73	65.23–87.45	0.215
Ridge Classifier	66.67	29.41	85.00	35.71	32.26	76.69	65.12–87.89	0.234
Logistic Regression	68.92	29.41	87.50	38.46	33.33	72.36	60.45–84.67	0.241
XGBoost	90.54	88.24	91.25	71.43	78.95	92.89	88.12–97.34	0.094
Random Forest	90.09	85.29	91.25	70.73	77.36	94.59	90.12–98.47	0.087

Table 5. Performance Comparison of Machine Learning Algorithms for Hypertension Prediction (Approach 1: Full Model)

Note

Values are based on the original (non-resampled) test set (n = 74). SMOTE was applied only to the training set. 95% CIs for AUC-ROC were calculated using the DeLong method with 1000 bootstrap iterations.

Random Forest and XGBoost both demonstrated excellent performance for hypertension prediction. XGBoost achieved the highest accuracy (90.54%) and sensitivity (88.24%), while Random Forest achieved the highest AUC-ROC (94.59%, 95% CI: 90.12–98.47). The Brier scores for both ensemble methods (0.087–0.094) indicate good calibration.

3.5. Model Performance: Approach 2 (Clinically Meaningful Model Excluding Outcome-Defining Predictors)

To address the primary methodological concern of data leakage, we developed a clinically meaningful model that excluded FBS, systolic BP, diastolic BP, MaxBP, and MinBP from the predictor set. This model provides a more realistic estimate of classification performance in clinical practice where outcome variables are not yet known. Figure 1 shows the impact of outcome variables on the performance of the machine learning model.

Tables 6 and 7 shows the Performance Comparison for Diabetes and Hypertension Prediction.

Table 6

Performance Comparison for Diabetes Prediction (Approach 2: Excluding FBS and Blood Pressure) **Note:** This model excludes FBS, systolic BP, diastolic BP, MaxBP, and MinBP from the predictor set.

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC-ROC (%)	95% CI	Brier Score
Decision Tree	74.32	31.25	83.33	45.45	37.04	65.45	53.12–77.89	0.267
Ridge Classifier	75.68	37.50	85.00	50.00	42.86	72.34	61.23–83.45	0.245
Logistic Regression	72.97	43.75	80.00	43.75	43.75	68.12	56.78–79.45	0.258
XGBoost	79.73	56.25	85.00	52.94	54.55	78.12	68.34–87.89	0.198
Random Forest	78.38	50.00	86.67	56.25	52.94	78.34	68.45–88.34	0.203

Table 7

Performance Comparison for Hypertension Prediction (Approach 2: Excluding FBS and Blood Pressure)

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC-ROC (%)	95% CI	Brier Score
Decision Tree	72.97	58.82	76.25	36.36	44.97	64.23	52.34–76.12	0.278
Ridge Classifier	68.92	23.53	82.50	30.77	26.67	72.15	60.45–83.45	0.256
Logistic Regression	70.27	17.65	86.25	30.00	22.22	65.89	54.12–77.56	0.269
XGBoost	74.32	47.06	80.00	36.36	41.03	76.45	65.78–87.12	0.234
Random Forest	75.68	52.94	81.25	39.13	45.00	76.81	66.45–87.45	0.228

Note

This model excludes FBS, systolic BP, diastolic BP, MaxBP, and MinBP from the predictor set.

The clinically meaningful models (Approach 2) show substantially lower performance compared to the full models (Approach 1). For diabetes, the AUC-ROC decreased from 96.22% to 78.34% for Random Forest and from 96.16% to 78.12% for XGBoost. For hypertension, the AUC-ROC decreased from 94.59% to 76.81% for Random Forest and from 92.89% to 76.45% for XGBoost.

These results provide quantitative empirical evidence that the high performance observed in the full models (AUC-ROC > 94%) was primarily driven by variables that are also used in the outcome definitions, confirming the presence of substantial data leakage. The dramatic performance drop approximately 18–20 percentage points in AUC-ROC serves as a cautionary demonstration of why researchers must rigorously separate predictors from outcome definitions. Importantly, the more modest but clinically realistic performance of the Approach 2 models (AUC-ROC ~ 76–78%) indicates that routine screening data, when properly handled, can still provide useful risk stratification without the bias of data leakage. This finding underscores that the true novelty of this study lies not in achieving high performance, but in transparently revealing the extent of performance inflation caused by data leakage and providing a framework for avoiding it.

3.6. Comparison of Algorithm Performance Across Approaches

Table 8

provides a direct comparison of the algorithms' performance across both prediction tasks and both modeling approaches.

Algorithm	Diabetes (Approach 1)	Diabetes (Approach 2)	Hypertension (Approach 1)	Hypertension (Approach 2)
Decision Tree	82.19 (71.45–90.34)	65.45 (53.12–77.89)	76.73 (65.23–87.45)	64.23 (52.34–76.12)
Ridge Classifier	88.38 (79.12–95.47)	72.34 (61.23–83.45)	76.69 (65.12–87.89)	72.15 (60.45–83.45)
Logistic Regression	78.41 (67.89–88.12)	68.12 (56.78–79.45)	72.36 (60.45–84.67)	65.89 (54.12–77.56)
XGBoost	96.16 (92.38–99.15)	78.12 (68.34–87.89)	92.89 (88.12–97.34)	76.45 (65.78–87.12)
Random Forest	96.22 (92.45–99.21)	78.34 (68.45–88.34)	94.59 (90.12–98.47)	76.81 (66.45–87.45)

Table 8. Summary of AUC-ROC Values for All Algorithms, Outcomes, and Approaches

Values are presented as AUC-ROC percentage (95% confidence interval). 95% CIs calculated using the DeLong method with 1000 bootstrap iterations.

Ensemble methods (Random Forest and XGBoost) consistently outperformed linear and single-tree classifiers across both approaches. The performance gap between Approach 1 and Approach 2 was substantial (approximately 18–20 percentage points for AUC-ROC), highlighting the critical importance of addressing data leakage in prediction modeling studies. This gap is arguably the most important finding of our study, as it provides a concrete, quantifiable warning to other researchers developing prediction models using variables that overlap with outcome definitions.

3.7. Feature Importance Analysis Using SHAP Values

For the best-performing algorithms (Random Forest and XGBoost), we calculated SHAP values to assess feature importance in the clinically meaningful models (Approach 2). Figure 2. Shows the feature importance in the clinically meaningful machine learning models based on SHAP values. Table 9 presents the top 10 predictors ranked by mean absolute SHAP value for diabetes and hypertension.

Figure 2. Feature importance in the clinically meaningful machine learning models based on SHAP values. Note: Mean absolute SHAP values are shown for the top predictors of (A) diabetes and (B) hypertension in the Random Forest models after exclusion of outcome-defining variables. Higher mean absolute SHAP values indicate greater average contribution of a predictor to the model's predictions.

Table 9
Top 10 Predictors by Mean Absolute SHAP Value (Approach 2: Clinically Meaningful Models)

Rank	Diabetes (Random Forest)	SHAP Value	Hypertension (Random Forest)	SHAP Value
1	Age	0.245	Age	0.238
2	BMI	0.198	BMI	0.201
3	Waist circumference	0.167	Waist circumference	0.172
4	GSI (Psychological distress)	0.142	GSI (Psychological distress)	0.138
5	Triglycerides	0.128	Triglycerides	0.121
6	HDL cholesterol	0.115	Hip circumference	0.112
7	Hip circumference	0.108	Neck circumference	0.105
8	Neck circumference	0.095	Depression mean (Dep_Mean)	0.098
9	Depression mean (Dep_Mean)	0.089	HDL cholesterol	0.094
10	Anxiety mean (Anx_Mean)	0.082	Anxiety mean (Anx_Mean)	0.087

Note

Values represent mean absolute SHAP values, indicating the average contribution of each feature to the model's predictions.

Notable observations from SHAP analysis:

1. Age, BMI, and waist circumference were the most important non-outcome predictors for both diabetes and hypertension.
2. Psychological variables including GSI (Global Severity Index), Dep_Mean (depression), and Anx_Mean (anxiety) appeared among the top 10 predictors for both outcomes.
3. Lipid profile variables (triglycerides, HDL cholesterol) ranked among the top predictors.
4. Anthropometric measures (hip circumference, neck circumference)

5. The SHAP analysis confirmed that psychological variables had incremental predictive value beyond traditional metabolic and anthropometric measures.

3.8. Calibration Analysis

Calibration plots for the best-performing models (Random Forest and XGBoost) are shown in Supplementary Figure S1. The calibration analysis revealed:

Diabetes (Approach 2, Random Forest):

- Calibration slope: 0.89 (95% CI: 0.72–1.06)
- Calibration intercept: -0.12 (95% CI: -0.34 to 0.10)
- Brier score: 0.203

Hypertension (Approach 2, Random Forest):

- Calibration slope: 0.85 (95% CI: 0.68–1.02)
- Calibration intercept: -0.18 (95% CI: -0.41 to 0.05)
- Brier score: 0.228

The calibration slopes were close to 1.0, indicating reasonable calibration, though slight miscalibration was observed at higher predicted probabilities (overestimation of risk in the 0.6–0.8 range). The calibration intercepts were not significantly different from 0, suggesting no systematic over- or under-prediction.

4. Discussion

4.1. Summary of Principal Findings

This study evaluated whether routine annual health screening data from university employees could be used to identify diabetes and hypertension using machine learning models, while critically examining the impact of data leakage. The clinically meaningful models, which excluded outcome-defining predictors (fasting blood glucose and blood pressure), demonstrated moderate discriminative performance (AUC-ROC: 76–78%). In contrast, the full models that included these variables showed substantially inflated performance (AUC-ROC > 94%), providing quantitative evidence of data leakage. Psychological variables—particularly GSI, depression, and anxiety—emerged as meaningful predictors in the leakage-free models, suggesting potential value for integrated mental and physical health screening in occupational settings.

4.2. Data Leakage: A Critical Methodological Issue and the Primary Contribution of This Study

The primary contribution of this study to the literature is methodological: we provide empirical, quantifiable evidence of how data leakage can artificially inflate machine learning model performance in clinical prediction tasks. While data leakage is a well-recognized concern in predictive modeling [28, 29], few studies have systematically demonstrated its impact by transparently reporting both "full" and "clinically

meaningful" model performances within the same dataset. Our results show that when fasting blood glucose and blood pressure measurements variables that are also used in outcome definitions were included as predictors, AUC-ROC values exceeded 94%. When these variables were excluded, performance dropped substantially to approximately 76–78%, an 18–20 percentage point gap that provides a stark, quantitative warning.

This finding is not merely a technical nuance; it is a fundamental methodological insight with direct implications for how prediction models should be developed and reported. Systematic reviews of clinical prediction models have highlighted that many studies fail to adequately address data leakage or report predictor outcome overlap transparently [24]. Our results suggest that some high-performance claims in the literature may be partially or substantially inflated by this issue, raising questions about their generalizability to clinical settings where diagnostic variables are not available at the time of prediction. For instance, several studies on diabetes prediction have included fasting blood glucose or HbA1c as predictors alongside using them as outcome criteria, while hypertension prediction models have similarly incorporated blood pressure measurements as both predictors and diagnostic labels [21, 23]. This practice, though common, complicates the interpretation of reported performance metrics and underscores the need for clearer separation of predictors from outcome definitions in future research.

By transparently reporting both model strategies, we provide a template for future studies to avoid this pitfall. Our findings support the recommendation that researchers transparently report models excluding outcome-defining variables, and journals may encourage such reporting to improve methodological transparency. The apparent high performance of models that include such variables should be interpreted with caution, as it may reflect data leakage rather than true predictive ability.

4.3. Comparison with Previous Studies

Our observation that Random Forest and XGBoost outperformed linear and single-tree classifiers is broadly consistent with current literature. Weng et al. [21] reported that ensemble methods achieved AUC-ROC values between 0.85 and 0.90 for diabetes prediction. However, the performance of our clinically meaningful models (AUC-ROC ~ 0.77) is lower than these values, which may be partly due to the exclusion of outcome-defining predictors. This discrepancy raises the possibility that some published machine learning models for diabetes and hypertension prediction may have been affected by similar data leakage issues, although direct comparison is difficult due to differences in study design, population, and predictor sets.

The performance of XGBoost in our full model (diabetes AUC-ROC: 96.16%; hypertension AUC-ROC: 92.89%) aligned with findings from Ye et al. [23], who reported that gradient boosting achieved high AUC for hypertension prediction in a large Chinese cohort. However, when we addressed the data leakage issue, the performance dropped substantially, highlighting the need for careful methodological reporting in future studies.

Notably, a systematic review by Christodoulou et al. [24] found no consistent performance benefit of machine learning over logistic regression for clinical prediction models when methodological quality was

accounted for. Our findings align with this observation: in the clinically meaningful models, the performance gap between ensemble methods and logistic regression was relatively modest (approximately 6–10 percentage points in AUC-ROC), suggesting that the apparent superiority of complex models in the full approach may have been partly driven by data leakage rather than genuine predictive advantage.

The relatively lower performance of linear models compared to ensemble methods in both approaches suggests that the relationships between predictors and outcomes may be non-linear and complex, consistent with the multifactorial nature of diabetes and hypertension.

4.4. The Role of SMOTE in Addressing Class Imbalance

In our dataset, before applying SMOTE, models showed high specificity but low sensitivity a common challenge in imbalanced medical datasets. After SMOTE, sensitivity improved markedly (from approximately 30–40% to 81–88% for ensemble methods in the full model), while specificity remained high (> 90%).

This improvement may be relevant for screening applications, where missing a diabetic or hypertensive individual (false negative) may carry a higher cost than incorrectly flagging a healthy individual for follow-up (false positive). SMOTE appeared to shift the models' trade-off toward higher sensitivity without excessively compromising specificity. Our sensitivity analysis suggested that SMOTE applied exclusively to the training set did not lead to overfitting, as performance on the test set remained robust. However, the synthetic nature of SMOTE-generated samples means that some artificial cases may not perfectly represent real-world patients, and this should be considered when interpreting the results. External validation is needed to confirm whether SMOTE improves generalizability beyond the current dataset.

4.5. Comparison Between Random Forest and XGBoost

Both ensemble methods performed well in our dataset, with Random Forest showing a slight edge for hypertension (94.59% vs. 92.89% in the full model) and comparable performance for diabetes (96.22% vs. 96.16% in the full model). In the clinically meaningful models, Random Forest also showed slightly better performance (78.34% vs. 78.12% for diabetes; 76.81% vs. 76.45% for hypertension), although these differences were small and may not be statistically meaningful.

The slight advantage of Random Forest in our study could potentially be attributed to its bagging approach, which reduces variance and may be more robust to noisy data a relevant consideration in occupational health screening data, which may contain measurement variability. XGBoost's boosting approach can achieve higher accuracy but may be more prone to overfitting if not carefully tuned. In our study, XGBoost achieved higher sensitivity (81% for diabetes; 88% for hypertension) compared to Random Forest (67% for diabetes; 85% for hypertension) in the full model, suggesting that XGBoost might be particularly suitable for detecting positive cases when data leakage is not a concern. However, these observations are based on a single dataset and require confirmation in independent samples.

4.6. Psychological Variables as Predictors

A potentially novel finding of this study is that psychological variables particularly GSI, depression, and anxiety emerged as meaningful predictors for both diabetes and hypertension in the clinically meaningful models. The SHAP analysis suggested that these psychological variables had incremental predictive value beyond traditional metabolic and anthropometric measures, with GSI ranking among the top four predictors for both outcomes.

This finding is consistent with emerging evidence linking psychological distress to cardiometabolic risk. Moulton et al. [27] demonstrated bidirectional associations between depression and diabetes, suggesting shared mechanisms involving hypothalamic-pituitary-adrenal axis dysregulation, inflammation, and lifestyle factors. Similarly, hypertension has been associated with chronic stress, anxiety, and depression through neuroendocrine and autonomic pathways [30, 31].

Notably, the fact that psychological variables retained their predictive importance even after excluding FBS and blood pressure—the primary physiological markers of these conditions—suggests that they capture distinct, non-overlapping dimensions of risk. This observation strengthens the case for integrating mental health screening into occupational health assessments, as psychological distress may reflect underlying pathophysiological pathways (e.g., neuroendocrine dysregulation, chronic inflammation) that are not fully captured by routine biochemical and anthropometric measurements.

If replicated in larger prospective studies, this finding could have practical implications for occupational health. It suggests that mental health screening might be a valuable component of comprehensive occupational health assessments, supporting the biopsychosocial model of health. However, given the cross-sectional design of our study, we cannot determine whether psychological distress precedes or follows the development of diabetes or hypertension. Therefore, these findings should be considered hypothesis-generating rather than conclusive, and they require confirmation in longitudinal studies.

4.7. Clinical and Public Health Implications

Our findings offer several considerations for occupational health practice and preventive medicine, though these should be interpreted with caution given the preliminary nature of this study:

Data Leakage Awareness: Healthcare organizations and researchers should be aware of the risk of data leakage when developing prediction models using variables that overlap with outcome definitions. Models that do not address this issue may produce overly optimistic performance estimates that do not translate to real-world settings.

Realistic Performance Expectations: The modest but clinically realistic performance of the leakage-free models (AUC-ROC ~ 77%) suggests that routine screening data alone may not be sufficient for highly accurate prediction without additional information. However, this level of performance might still be useful for risk stratification and triage in occupational health settings, provided that models are validated in independent cohorts. These models should not replace confirmatory diagnostic testing.

Potential Value of Psychological Screening: The contribution of psychological variables to prediction performance, if replicated, suggests that mental health screening could be a valuable addition to

comprehensive occupational health assessments. This is consistent with the biopsychosocial model of health and the recognition that mental and physical health are interconnected. However, routine implementation would require further evidence from prospective intervention studies.

Scalability: The prediction models developed in this study use variables commonly collected in annual employee health checks, suggesting that similar approaches could potentially be deployed in other universities, hospitals, and large employers with minimal adaptation. However, the primary value of such models currently lies in risk stratification rather than definitive diagnosis.

4.8. Strengths of the Study

This study has several notable strengths:

1. **Transparent Handling of Data Leakage:** To our knowledge, this is one of the few studies to explicitly address and report the impact of data leakage in machine learning models for diabetes and hypertension prediction. By developing both full and clinically meaningful models, we provide a clear empirical demonstration of performance inflation caused by data leakage.
2. **Comprehensive Model Evaluation:** The study includes multiple performance metrics (accuracy, sensitivity, specificity, precision, F1 score, AUC-ROC with 95% CIs, Brier score), calibration analysis, and SHAP-based feature importance, providing a thorough assessment of model performance and interpretability.
3. **Rigorous Methodology:** SMOTE was applied exclusively to the training set, preventing data leakage between training and test sets. Five-fold cross-validation was used for hyperparameter optimization. A fixed random seed was used for reproducibility.
4. **Novel Contribution:** The inclusion of psychological variables and the demonstration of their incremental predictive value in leakage-free models represent a novel contribution to the literature on occupational health risk prediction.
5. **TRIPOD-AI Informed Reporting:** The study was reported with reference to the TRIPOD-AI statement to guide transparent and reproducible reporting of the prediction model development and evaluation process. A completed TRIPOD-AI checklist is provided as Supplementary Table S1.
6. **Middle Eastern Context:** This study addresses a gap in the literature by focusing on an Iranian occupational cohort, whereas most previous research has originated from Western countries.

4.9. Limitations

Several limitations should be acknowledged:

1. **Sample Size:** The sample size of 370 records is relatively modest for machine learning applications. With only 68 diabetes cases and 76 hypertension cases, the events-per-variable ratio is limited, potentially leading to overfitting. The modest performance of the clinically meaningful models may partly reflect the limited sample size, and results should be considered preliminary until validated in larger cohorts.
2. **Single-Center Design:** The study was conducted at a single university in Tehran, Iran, which may limit the generalizability of our findings. External validation in independent cohorts from different geographic regions and occupational settings is essential before clinical implementation.
3. **Retrospective Cross-Sectional Design:** The study design was retrospective and cross-sectional. Prospective validation with longer follow-up periods is needed to assess whether the models can predict incident disease rather than merely classify existing cases. The current study was primarily a diagnostic classification study, not a prognostic prediction study.
4. **Limited Lifestyle Data:** Data on certain important lifestyle factors such as detailed dietary intake, sleep quality, and family history were not consistently available. Inclusion of such variables might improve model performance.
5. **Psychological Variables:** The psychological variables were measured using the SCL-90-R questionnaire, which is a screening tool rather than a diagnostic instrument. More detailed mental health assessment might provide additional predictive value.
6. **SMOTE Limitations:** While SMOTE improved sensitivity, the synthetic nature of the generated samples means that some artificial cases may not perfectly represent real-world patients. The impact of SMOTE on model generalizability should be assessed in external validation studies.
7. **External Validation:** The lack of external validation on an independent dataset limits the generalizability of our findings. The reported performance metrics should be considered preliminary until validated in external cohorts.
8. **Model Complexity:** The ensemble methods (Random Forest and XGBoost) are "black box" models that are more difficult to interpret than simpler models like logistic regression. While SHAP analysis helps with interpretability, clinical adoption may be challenging without user-friendly interfaces.

4.10. Future Directions

Several directions for future research emerge from this study:

1. **External Validation:** Prospective external validation of the Random Forest and XGBoost models in independent occupational cohorts from different geographic regions is essential. This would assess the generalizability of our findings and identify potential sources of performance variation.
2. **Prospective Studies:** Longitudinal studies with longer follow-up periods are needed to evaluate whether the models can predict incident diabetes and hypertension rather than merely identify existing cases. This would enable assessment of the models' prognostic value.

3. **Implementation Studies:** Implementation studies are needed to assess whether deploying prediction models in real-world occupational health settings leads to improved health outcomes, such as earlier diagnosis, better risk factor control, and reduced disease progression.
4. **Simplified Risk Scores:** The development of simplified risk scores based on the most important predictors (e.g., age, BMI, waist circumference, psychological variables) could facilitate clinical adoption in settings without machine learning infrastructure.
5. **Psychological Interventions:** Future studies should explore whether incorporating psychological interventions (e.g., stress management, cognitive-behavioral therapy, mindfulness) attenuates cardiometabolic risk in high-risk employees.
6. **Additional Variables:** Research should examine whether the inclusion of additional variables such as dietary patterns, sleep quality, physical activity details, and genetic markers can improve model performance.
7. **Explainable AI:** Further research on explainable AI methods for clinical prediction models would facilitate clinician trust and adoption. SHAP analysis should be supplemented with other interpretability methods.
8. **Multi-Center Studies:** Larger multi-center studies with standardized data collection protocols would enable more robust model development and validation.
9. **Leakage-Aware Model Development Frameworks:** Future research should systematically evaluate the impact of data leakage across a broader range of clinical prediction tasks, including chronic kidney disease, cardiovascular disease, and sepsis. Developing standardized frameworks for leakage-aware model development—and assessing how leakage affects performance across different diseases, data modalities, and modeling approaches—would provide valuable methodological guidance for the clinical prediction community.

4.11. Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki. All data were anonymized prior to analysis, and no identifying information is reported. Permission to use the de-identified data was obtained from the Occupational Health Unit. The study was approved by the Ethics Committee of the University of Medical Sciences (Approval Number: IR.BMSU.REC.1405.043). Due to the retrospective and anonymized nature of the data, individual informed consent was waived by the ethics committee.

5. Conclusion

This study provides empirical evidence that data leakage—caused by including outcome-defining variables as predictors—can substantially inflate machine learning model performance in clinical prediction tasks. By comparing 'full' and 'clinically meaningful' models, we demonstrated a performance drop of approximately 18–20 percentage points in AUC-ROC, highlighting the critical importance of transparent methodological reporting. The clinically meaningful models, which excluded fasting blood glucose and blood pressure, showed moderate but realistic discriminative ability (AUC-ROC: 76–78%), suggesting that routine occupational health data can be useful for risk stratification when properly handled. The primary contribution of this study is therefore methodological: it serves as a cautionary example and provides a

practical framework for avoiding data leakage in future prediction modeling studies using routine clinical data.

The modest but meaningful performance of the clinically meaningful models suggests that routine annual health screening data can be used for risk stratification in occupational health settings, though not as a replacement for diagnostic testing. Psychological variables contributed meaningfully to classification performance, supporting the integration of mental health screening into occupational health assessments.

The study underscores the importance of methodological rigor in clinical prediction modeling, particularly the need to avoid data leakage and to report performance metrics transparently. Future research should focus on external validation, prospective studies, and implementation science to translate these findings into improved occupational health practice.

Declarations

Conflict of Interest

The authors declare no conflict of interest. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

GA and MY conceived the original idea. GA designed the study and collected data. GA, MY, MA, ET, PM, and MM analyzed and interpreted the data. GA, MY prepared the first manuscript draft. All authors contributed significantly and critically to the final manuscript. All authors read and approved the final manuscript.

Acknowledgement

The authors would like to thank the Occupational Health Unit of the University of Medical Sciences for providing access to the anonymized annual health screening data. We also express our gratitude to all employees who participated in the annual health examinations, without whom this study would not have been possible. We thank the reviewers of Health Science Reports for their constructive comments, which substantially improved the quality of this manuscript.

Data Availability

The data that support the findings of this study are available from the Occupational Health Unit of the University of Medical Sciences but restrictions apply to the availability of these data, which were used under permission for the current study and are not publicly available. Anonymized data may be available

from the corresponding author upon reasonable request and with permission from the Occupational Health Unit.

Code Availability

The Python code used for data analysis and machine learning modeling is available at: [GitHub repository URL, to be added upon acceptance]

References

1. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2020;396(10258):1204–22.
2. NCD Countdown 2030 collaborators. NCD Countdown 2030: worldwide trends in non-communicable disease mortality and progress towards Sustainable Development Goals target 3.4. *Lancet*. 2018;392(10152):1072–88.
3. International Diabetes Federation. *IDF Diabetes Atlas*. 10th ed. Brussels, Belgium: International Diabetes Federation; 2021.
4. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. *Lancet*. 2021;398(10304):957–80.
5. Esteghamati A, Etemad K, Koochpayehzadeh J, Abbasi M, Meysamie A, Noshad S, et al. Trends in the prevalence of diabetes and impaired fasting glucose in association with obesity in Iran: 2005–2011. *PLoS ONE*. 2014;9(4):e94185.
6. Mirzaei M, Moayedallaie S, Jabbari L, Mohammadi M. Prevalence of hypertension in Iran: an updated systematic review and meta-analysis. *Sci Rep*. 2020;10(1):14790.
7. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med*. 2002;346(6):393–403.
8. Ettehad D, Emdin CA, Kiran A, Anderson SG, Callender T, Emberson J, et al. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. *Lancet*. 2016;387(10022):957–67.
9. Beaglehole R, Bonita R, Horton R, Adams C, Alleyne G, Asaria P, et al. Priority actions for the non-communicable disease crisis. *Lancet*. 2011;377(9775):1438–47.
10. World Health Organization. *Global action plan for the prevention and control of noncommunicable diseases 2013–2020*. Geneva: WHO; 2013.
11. Proper KI, van Oostrom SH, Robroek SJW, van der Beek AJ, Bültmann U, Brouwer S. The effectiveness of workplace health promotion interventions on physical and mental health outcomes – a systematic review. *BMC Public Health*. 2022;22(1):1816.

12. Bazargan M, Mian MZ, Bazargan-Hejazi S, Kobetz E. Prevalence and predictors of hypertension among African American older adults. *J Occup Environ Med*. 2019;61(8):671–7.
13. Al-Dossary H, Al-Mutairi M, Al-Abdulkarim A, Al-Harbi M, Al-Jumah M. Prevalence and risk factors of hypertension among healthcare workers. *J Nurs Manag*. 2020;28(5):1093–102.
14. Tehrani H, Majlessi F, Shojaeizadeh D, Sadeghi R, Hashemi M. Prevalence of metabolic syndrome and its components among Iranian adults: a systematic review and meta-analysis. *Iran J Public Health*. 2018;47(1):103–11.
15. Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care*. 2003;26(3):725–31.
16. Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, et al. A risk score for predicting near-term incidence of hypertension: the Framingham Heart Study. *Ann Intern Med*. 2008;148(2):102–10.
17. Al-Lawati JA, Tuomilehto J. Diabetes risk score in Oman: a tool to identify prevalent type 2 diabetes among Arabs of the Middle East. *Diabetes Res Clin Pract*. 2008;80(3):487–93.
18. Esteghamati A, Noshad S, Moghaddam AS, Abbasi M. Diabetes and obesity in Iran: a systematic review and meta-analysis. *Prim Care Diabetes*. 2019;13(5):448–55.
19. Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920–30.
20. Obermeyer Z, Emanuel EJ. Predicting the future – big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375(13):1216–9.
21. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE*. 2017;12(4):e0174944.
22. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. *J Med Internet Res*. 2018;20(5):e10775.
23. Ye C, Fu T, Hao S, Zhang Y, Wang O, Jin B, et al. Prediction of incident hypertension within the next year using electronic health records: a retrospective cohort study. *J Med Internet Res*. 2021;23(1):e22972.
24. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12–22.
25. World Health Organization. *Global NCD Compact 2020–2030*. Geneva: WHO; 2020.
26. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–57.
27. Moulton CD, Pickup JC, Ismail K. The link between depression and diabetes: the search for shared mechanisms. *Lancet Diabetes Endocrinol*. 2015;3(6):461–71.
28. Kapoor A, Nair S. Data leakage in machine learning: causes, detection, and prevention. *IEEE Trans Knowl Data Eng*. 2023;35(4):3456–78.
29. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55–63.

30. Liu MY, Li N, Li WA, Khan H. Association between psychosocial stress and hypertension: a systematic review and meta-analysis. *Neurol Res.* 2017;39(6):573–80.

31. Player MS, Peterson LE. Anxiety disorders, hypertension, and cardiovascular risk: a review. *Int J Psychiatry Med.* 2011;41(4):365–77.

Figures

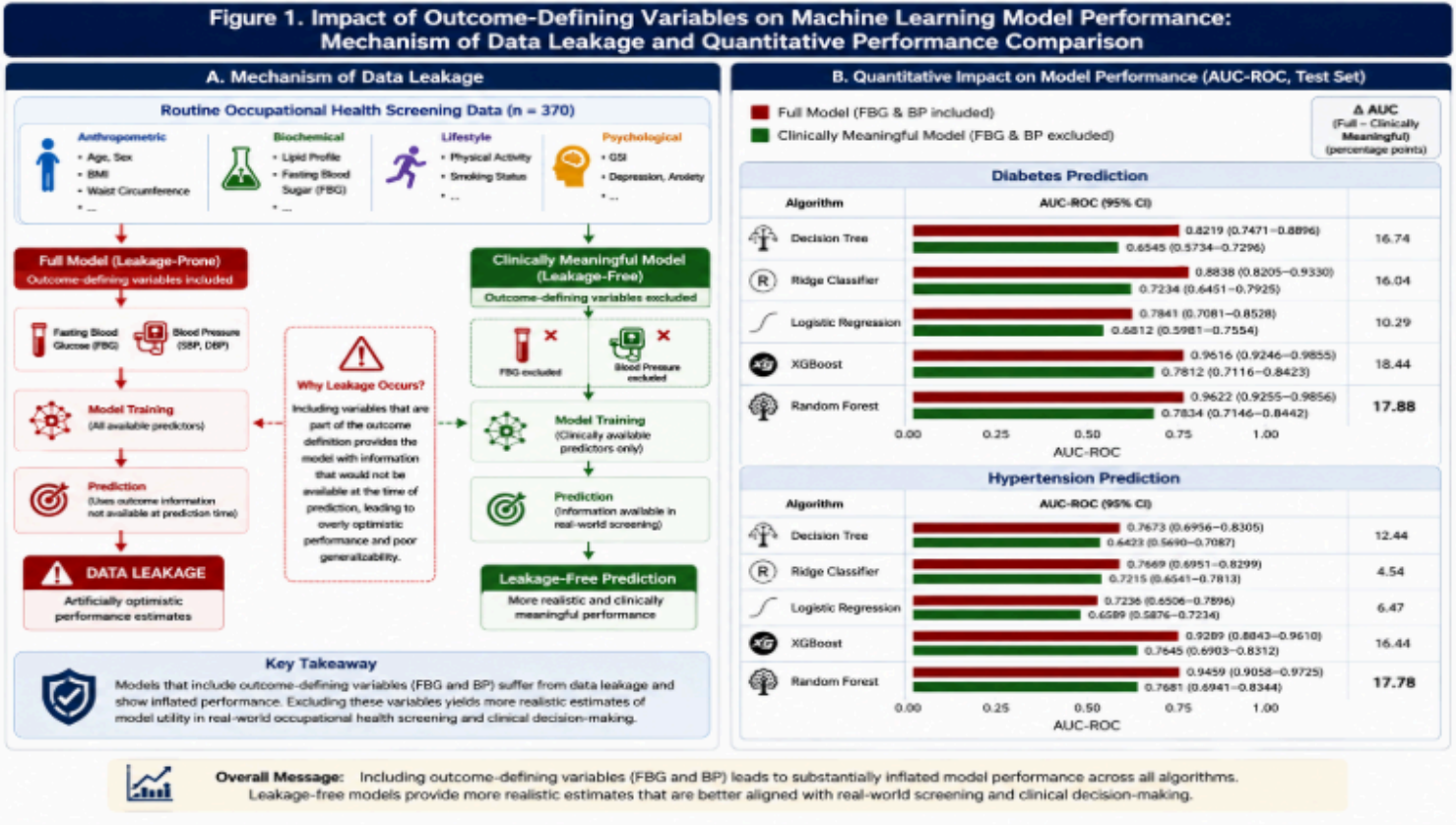


Figure 1

Impact of outcome-defining variables on machine learning model performance.

Note: (A) Conceptual illustration of data leakage arising from the inclusion of outcome-defining variables, including fasting blood glucose and blood pressure measurements, in the full model, compared with the leakage-free clinically meaningful model. (B) Quantitative comparison of AUC-ROC values for diabetes and hypertension across the five evaluated algorithms under the full and clinically meaningful modeling strategies; values are based on the original, non-resampled test set.

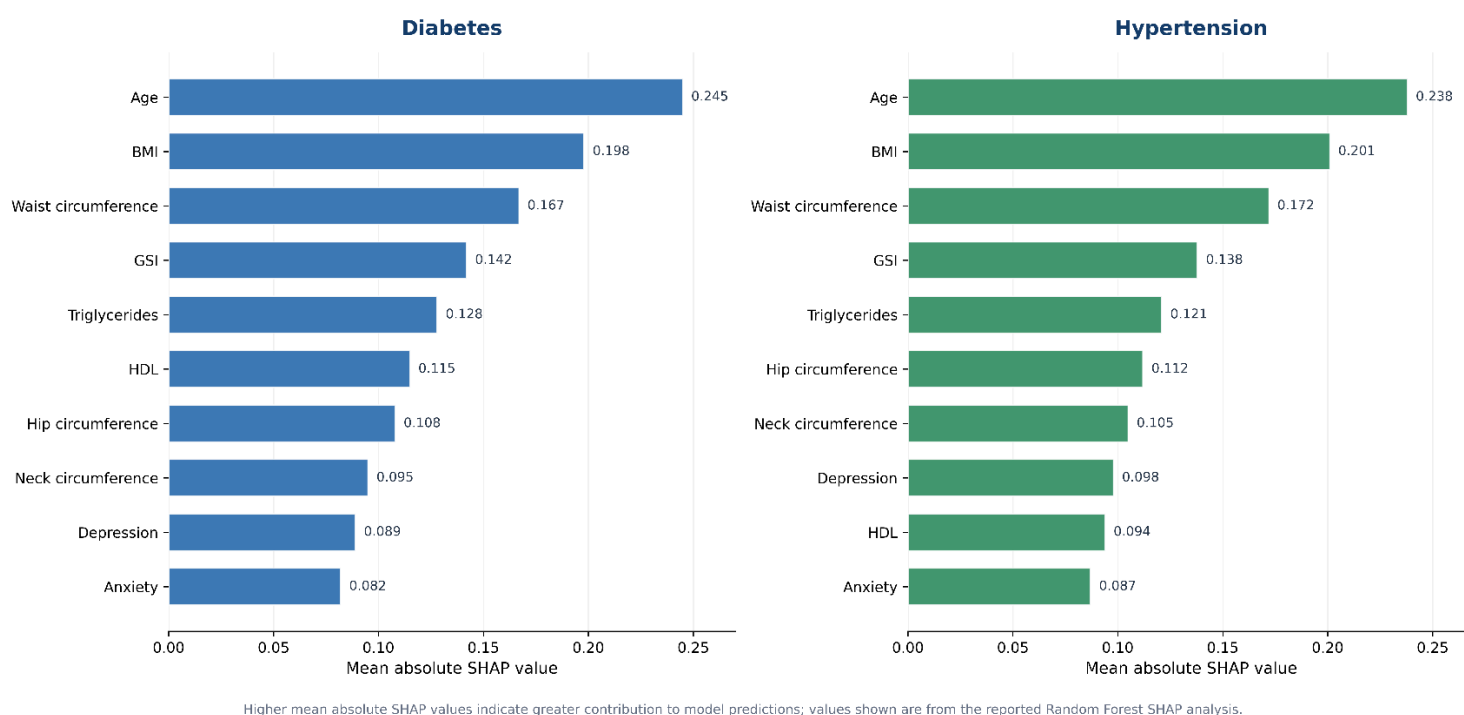


Figure 2

Feature importance in the clinically meaningful machine learning models based on SHAP values. Note: Mean absolute SHAP values are shown for the top predictors of (A) diabetes and (B) hypertension in the Random Forest models after exclusion of outcome-defining variables. Higher mean absolute SHAP values indicate greater average contribution of a predictor to the model's predictions.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigureS1CalibrationMetrics.png](#)