Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Calibrated Risk Stratification Model for Impaired Fasting Glycemia in the Ugandan Population: A Nested Cross-Validation and Threshold Optimization Approach

Bashir Ssuna

sbn144@gmail.com

Makerere University Hospital

Hannah Kibuuka

Makerere University Walter Reed Project

Maiya G Block Ngaybe

University of Arizona

Ivan Mufumba

University of California, Berkeley

Allan Omalla

[1]Makerere University Hospital

Raymond Bernard Kihumuro

Makerere University Walter Reed Project

Silver Bahendeka

Uganda Martyrs University

Research Article

Keywords: machine learning, impaired fasting glycemia, diabetes, risk prediction, Uganda, LASSO

Posted Date: August 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10639053/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

The burden of type 2 diabetes is rapidly increasing in sub-Saharan Africa, and impaired fasting glycemia (IFG) is a reversible pre-disease state. In Uganda, the prevalence of hyperglycemia has doubled in the past decade, with more than 60% of cases undiagnosed, which points to an urgent need for cost-effective risk stratification tools.

Methods

We trained machine learning algorithms using secondary data from the nationally representative STEPS survey. We used case-cohort sampling: all IFG cases (n = 287) plus a 1:4 sub-cohort of controls (n = 1,053). Preprocessing included exclusion of participants with missing outcomes, multiple imputation (MICE, five imputations), and standardization of continuous features. Nine candidate models, including LASSO, random forest, and gradient boosting, were trained and evaluated under 5-fold nested cross-validation. Discrimination (ROC AUC), calibration, precision-recall AUC, and F1-score guided model selection. LASSO was selected as the final model for its balance of discrimination, calibration, and interpretability. Threshold moving used the G-mean to balance sensitivity and specificity. Out-of-fold predictions defined cutoffs for low-risk (NPV > = 0.90) and high-risk (PPV > = 0.40, Sn > = 0.70) strata, with an intermediate moderate-risk band.

Results

A total of 1,340 participants were included. The mean age was 39 years, with balanced representation across sex, regions, and residence types. The final calibrated LASSO model achieved an out-of-fold ROC AUC of 0.68 [95%CI: 0.64–0.71] and PR AUC of 0.40 [95%CI: 0.34–0.46], with Sn = 60%, Sp = 69%, PPV = 34% and NPV = 86%. Selected features included age, BMI, SBP, DBP, cholesterol, waist/hip circumference, heart rate, sedentary time, and categorical covariates (education, region, tobacco use, physical activity). Using the calibrated thresholds, 1% of participants were classified as low risk, 76% as moderate risk, and 23% as high risk. The high-risk band achieved a PPV of 0.41 and an IFG rate of 41%.

Conclusions

This calibrated, threshold-based risk stratification model provides a contextually relevant, non-invasive tool for identifying Ugandans at high risk of IFG. Integration into primary care could enable efficient allocation of confirmatory testing and preventive interventions. Further validation in prospective cohorts and implementation research are recommended to support scale-up.

Background

Diabetes mellitus (DM) is among the fastest-growing global non-communicable diseases (NCD). The number of people living with DM is estimated to reach 850 million by 2050, an increase of nearly 45% from current levels[1]. As many as 4 in 10 people living with DM are unaware that they have the condition[1]. Over 80% of these cases occur in low- and middle-income countries (LMICs), regions often least equipped to manage these chronic diseases[1]. Individuals with prediabetes states of impaired glucose tolerance (IGT) or impaired fasting glycemia (IFG) represent a population at higher risk of overt DM [2]. Current estimates place the age-adjusted prevalence of IGT at 9.1% for adults aged 20-79, while IFG prevalence stands at 5.8%; these numbers are projected to increase to 10.0% for IGT and 6.5% for IFG by 2045, mostly in LMICs[3]. Over 90% of these cases are type 2 diabetes[1], driven by an interplay of socio-economic, demographic, environmental, and genetic factors, including urbanization, aging populations, and rising rates of overweight and obesity[1]. Recent work argues for reclassifying diabetes into more precise subtypes rather than the traditional binary split, which may sharpen how prediabetes states like IFG are understood going forward[4-6].

In Uganda, the first national assessment from the 2014 STEPS survey reported an overall diabetes prevalence of 1.4% and an IFG prevalence of 2.0% among adults aged 18-69 years[7]. In 2023, another national STEPS survey found that the prevalence increased to 3.3%[8,9]. This trend is mirrored by a rise in a key metabolic risk factor, with the prevalence of overweight and obesity increasing from 19.3% to 24.1% over the same period[8]. The 2014 survey revealed significantly higher prevalence of diabetes in urban areas (2.7%) compared to rural areas (1.0%), while 48.9% of participants with DM were unaware of their condition[7]. The proportion of participants unaware of their DM status increased to 61.2% in 2023[10], meaning this large, undiagnosed population is at high risk of developing debilitating and costly complications even before entering the healthcare system.

Accessing and delivering effective diabetes care in Uganda is a complex process with challenges at every level, from the individual patient to the national health system. Such barriers include financial hardship and frequent stockouts of essential medicines and supplies in public facilities, forcing patients to purchase them from private outlets, which is often unaffordable[11,12]. Beyond finances, a lack of awareness about the asymptomatic nature of early diabetes and the importance of routine screening contributes to poor health-seeking behaviors [13-15]. Many individuals only seek care once symptoms have become severe, by which point complications may have already developed[16-18]. Frequent and prolonged stockouts of fundamental supplies like blood glucose test strips are reported across numerous studies and settings in Uganda and neighbouring East African countries, from rural health centers to regional referral hospitals[12,19]. This calls for the development of machine learning (ML) models that promote efficient utilization of the limited screening tests on patients at the highest risk of developing IFG.

A key component of an effective public health response to the DM epidemic is the ability to identify high-risk individuals early, before the onset of overt disease or its complications. In resource-constrained

settings like Uganda, this requires cost-effective screening strategies like non-invasive ML risk scores that offer a promising triage to stratify the population and target more expensive blood tests to those most in need. However, the performance of these tools, largely developed in Western populations, is highly variable and often suboptimal in populations of African descent. The aim of this study, therefore, is to create an ML data-driven, resource-sensitive IFG screening tool that supports early identification of at-risk individuals and informs targeted prevention strategies in clinical and community settings.

Methods

Study design and area

We trained machine learning algorithms using secondary data from the nationally (Uganda) representative 2023 World Health Organization (WHO) STEPS survey[9], and the data dictionary describing the variable names is presented in Table 1. This study is reported in accordance with the STROBE guideline for cross-sectional studies (Additional file 1) and the TRIPOD+AI guideline for prediction model studies using regression or machine learning methods [34] (Additional file 2).

Data cleaning and preparation

Data cleaning was done in R. Missingness was examined separately for numeric and categorical variables, and all participants with missingness in the outcome variable (b5 = fasting blood sugar level, mmol/L) were excluded from the analysis. To address missing data, multiple imputation using chained equations (MICE package, version 3.18.0) was performed with five imputations and 50 iterations. Predictive mean matching (PMM) was used for continuous variables and logistic regression imputation for categorical variables. The identifier (ID variable) and outcome were excluded from the predictor matrix. Strip plots and density plots were used to visually inspect the plausibility of the imputations. The first completed dataset was extracted for analysis and to illustrate the distribution of imputed values (Figure 1).

Case-cohort sampling strategy

We implemented a case-cohort design to efficiently develop and evaluate the prognostic model. All participants with impaired fasting glycemia (IFG = 1; n = 287) were included, and a random sub-cohort of non-cases (IFG = 0) was sampled from the 3,268 participants without IFG. This approach balances focus on the groups that matter most, i.e., those with and without the outcome, so the analysis runs faster and with fewer computational resources while remaining representative of the full population. To determine the optimal sub-cohort fraction, we conducted simulations across fractions of 15%, 25%, 35%, and 50% of the full cohort. For each fraction, we combined the sampled sub-cohort controls with all IFG cases and fit a logistic regression model to estimate discrimination between individuals with and without IFG. We used the area under the ROC curve (AUC) as the performance metric and obtained 95% confidence intervals through bootstrap resampling, allowing us to compare the precision of the AUC across sampling fractions.

We therefore selected a sub-cohort fraction of approximately one third, yielding 1,053 controls for 287 cases (a realized fraction of 32.2% of the 3,268 non-cases; approximately a 1:4 case-to-control ratio). Our simulation showed that this fraction produced a stable AUC estimate of approximately 0.72 with a narrow 95% CI width of 0.07, while larger fractions added little gain in precision but increased the computational burden (Figure 2).

Model development and risk stratification

We began with a model discovery phase in which several candidate machine learning classifiers were trained and evaluated, including logistic regression with LASSO penalty, random forest, balanced random forest, support vector machine, gradient boosting (XGBoost and LightGBM), multi-layer perceptron, Gaussian naive Bayes, and logistic regression. Each model was assessed using nested cross-validation to minimize overfitting, with discrimination (AUC), calibration, precision-recall area under the curve (AUPRC), sensitivity, specificity, and F1-score reported (Table 2). Following this comparison, LASSO logistic regression was selected as the final model due to its balance of discrimination, calibration, and parsimony. During feature selection, both clinical and behavioral predictors were considered; continuous variables (e.g., age, BMI, blood pressure, waist/hip circumference, heart rate) were standardized, and categorical variables (e.g., education, region, tobacco use, physical activity) were encoded. LASSO regularization[20] shrank non-informative features toward zero, retaining only those with stable predictive contribution across folds.

The final modeling pipeline integrated preprocessing (scaling, encoding), feature selection, and classification within a reproducible framework. Out-of-fold (OOF) predictions from nested cross-validation were used to calibrate probabilities and evaluate performance across the entire dataset. Threshold moving was performed to optimize the balance between sensitivity, specificity, and predictive values, so the model could be adapted for different screening priorities. To enable clinical risk stratification, probability cutoffs were defined at clinically meaningful operating points: (i) a low-risk cutoff to maximize negative predictive value (NPV >= 0.90), (ii) a high-risk cutoff to maximize positive predictive value (PPV >= 0.40) while maintaining adequate sensitivity, and (iii) an intermediate band for moderate risk. These thresholds classified participants into low, medium, and high risk groups, aligning model outputs with practical resource allocation for targeted screening (Figures 4-6).

Results

Description of study participants

A total of 1,340 participants were included in this analysis. The mean and standard deviation of age was 39.4 +/- 13.7 years, and 833 (62.2%) were female. In this case-cohort, 287 (21.4%) of the participants had IFG (Table 3 and Table 4).

Correlation matrix of the numerical features

The correlation matrix of numeric predictors showed that most variables were only weakly correlated (|r| < 0.3), suggesting low redundancy across features. SBP and DBP were strongly correlated (r = 0.82), and waist and hip circumferences also demonstrated a strong positive correlation (r = 0.82) (Figure 3).

Feature selection in the final model

Of the 20 candidate predictors, LASSO retained 17 with non-zero coefficients. Living in the Northern region and higher educational attainment were the strongest protective signals, while current alcohol use, waist circumference, diastolic blood pressure, resting heart rate, and age were the strongest risk signals (Figure 4).

Diagnostic performance of the final IFG prediction model

The performance of the final calibrated model was evaluated on out-of-fold predictions. At the operating point that balanced sensitivity and specificity (threshold 0.21), sensitivity was 0.60 and specificity was 0.69. The negative predictive value was high at 0.86, indicating a strong ability to rule out disease, while the positive predictive value was modest at 0.34. Overall discrimination was fair, with a ROC AUC of 0.68 (95% CI: 0.64-0.71), while the precision-recall AUC was 0.40 (95% CI: 0.34-0.46) (Table 5, Figures 5 and 6). Applying the calibrated thresholds, 18 participants (1.3%) were classified as low risk, 1,019 (76.0%) as moderate risk, and 303 (22.6%) as high risk; the high-risk band contained 124 of the 287 cases and had an IFG rate of 41%.

Discussion

We built a LASSO logistic regression model to flag Ugandan adults likely to have impaired fasting glycemia, using the 2023 WHO STEPS survey. On out-of-fold predictions the model reached a ROC AUC of 0.68 (95% CI: 0.64-0.71), with sensitivity 0.60, specificity 0.69, and NPV 0.86. IFG produces no symptoms; a screening tool has to find people with no complaint and point them toward a test they would otherwise skip [21]. All nine models clustered between 0.65 and 0.68 (Table 2); LASSO tied with random forest and was chosen for readable coefficients over performance.

Several established scores predict type 2 diabetes rather than IFG, an easier target with a stronger metabolic signal, so the figures below are context, not a like-for-like benchmark. FINDRISC reports an AUC of 0.85 to 0.87 for incident T2DM in its original Finnish cohorts [22], but fell to 0.63 (sensitivity 0.48, specificity 0.73) for undiagnosed type 2 diabetes in Botswana [23], below our model's performance in a population of closer ancestry. QDiabetes, from UK primary care, reports 0.83 to 0.85 [24], and AUSDRISK reports 0.78 in its derivation cohort, with validation cohorts ranging from 0.66 to 0.79 [25]. Our 0.68 is lower, consistent with a milder phenotype in a different population (Table 6).

The high NPV matters most in a Ugandan clinic, where test strips are often out of stock [26], so the question is who should get the next strip. A person rated low risk has about an 86 percent chance of being free of IFG, though our low-risk cutoff cleared only 18 of 1,340 people, so that group is tiny. The

real value is the high-risk band: 303 people had a 41 percent rate, about double the sample average, so testing them first roughly doubles the yield. Directing confirmatory testing toward those most likely to benefit is a standard triage principle in resource-limited diabetes screening [27].

LASSO kept 17 of 20 candidate features. Northern region and higher education were the strongest protective signals; alcohol use, waist circumference, diastolic pressure, heart rate, and age carried the most risk, broadly matching FINDRISC, AUSDRISK, and QDiabetes on core anthropometric measures [22,24,25]. Family history, a staple of Western scores, was deliberately excluded from STEPS [9]. We instead kept heart rate and sedentary time, which rarely appear in Western scores but track cardiometabolic risk [28,29].

To change practice, the score would need to sit inside routine primary care. Uganda's Ministry of Health has pushed task-shifting [30], and a score using only age, blood pressure, waist circumference, and simple questions needs no laboratory; high-risk referral and moderate-risk counselling can both help delay progression [31,32]. A cholesterol-free version was trained alongside the full model but not separately validated. HIV and antenatal clinics already see many of the same adults [33], so an IFG check there would need little new infrastructure. The survey is nationally representative [9], and nested cross-validation kept model selection separate from performance estimation, with every reported number from out-of-fold predictions.

Several limitations affect how these numbers should be read. The design is cross-sectional, identifying prevalent IFG rather than forecasting incident cases as QDiabetes does [24]. The label was not restricted to the WHO range (6.1-6.9 mmol/L), so roughly a third of the 287 cases may actually have diabetes. Only the first of five imputed datasets was used, and sampling was unweighted, so probabilities reflect the 21 percent analytic prevalence rather than the roughly 5 percent population rate. Cholesterol still needs a blood draw, and no external validation exists yet; fold-level AUC ranged from 0.58 to 0.73 across five folds (~57 cases each).

Conclusions

In short, this model sorts Ugandan adults into IFG risk bands from measurements a clinic already collects. It cannot replace a glucose test, but it can help decide who gets one first. External validation on an independent Ugandan cohort, followed by prospective testing of whether high-risk flagging speeds up diagnosis, is the priority next step.

Abbreviations

AUC: area under the curve; BMI: body mass index; CI: confidence interval; CV: cross-validation; DBP: diastolic blood pressure; DM: diabetes mellitus; IFG: impaired fasting glycemia; IGT: impaired glucose tolerance; LASSO: least absolute shrinkage and selection operator; LMIC: low- and middle-income

country; MICE: multiple imputation by chained equations; ML: machine learning; NCD: non-communicable disease; NPV: negative predictive value; OOF: out-of-fold; PPV: positive predictive value; PR: precision-recall; ROC: receiver operating characteristic; SBP: systolic blood pressure; Sn: sensitivity; Sp: specificity; STEPS: STEPwise approach to NCD risk factor surveillance; T2DM: type 2 diabetes mellitus; WHO: World Health Organization.

Declarations

Ethics approval and consent to participate

The 2023 Uganda STEPS survey was approved by the Research and Ethics Committee (REC) of Nsambya St. Francis Hospital, Kampala, Uganda (Approval #: SFHN-2022-43)[8]. All participants in the original survey provided informed consent. This manuscript is a secondary analysis of the same STEPS survey data conducted by co-authors of the original survey, using the original survey's ethics approval; no additional ethical review was required.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request, and from the custodians of the Uganda 2023 WHO STEPS survey[9]. The analysis code (data preprocessing, model training, evaluation, and figure generation) is available at https://github.com/bashirssuna/Impaired-Fasting-Glucose-Risk-Model- under an MIT license.

Competing interests

The authors declare that they have no competing interests.

Funding

This was an independent secondary analysis with no external funding.

Authors' contributions

BS was involved in proposal development and writing, data analysis, and manuscript writing. HK contributed to manuscript review. MGBN contributed to manuscript writing, results interpretation, and manuscript review. IM contributed to manuscript writing, results interpretation, and manuscript review. AO contributed to manuscript writing, results interpretation, and manuscript review. BRK contributed to

manuscript writing, results interpretation, and manuscript review. SB contributed to data, manuscript writing, results interpretation, and review. All authors read and approved the final manuscript.

Acknowledgements

We thank the team of the WHO Uganda STEPS survey and the team from Makerere University Walter Reed Program (MUWRP).

References

1. International Diabetes Federation. IDF Diabetes Atlas, 11th ed. Brussels: International Diabetes Federation; 2025.
2. Nathan DM, Davidson MB, DeFronzo RA, Heine RJ, Henry RR, Pratley R, et al. Impaired fasting glucose and impaired glucose tolerance: implications for care. Diabetes Care. 2007;30(3):753-9.
3. Rooney MR, Fang M, Ogurtsova K, Ozkan B, Echouffo-Tcheugui JB, Boyko EJ, et al. Global prevalence of prediabetes. Diabetes Care. 2023;46(7):1388-1394.
4. Narayan KMV, Jagannathan R, Ridderstrale M. Managing type 2 diabetes needs a paradigm change. Lancet Diabetes Endocrinol. 2023;11(8):534-536.
5. Jagannathan R, Staimez LR, Narayan KMV. Reuniting diabetes through the islet coordinate framework. Lancet Diabetes Endocrinol. 2026 (in press). doi:10.1016/S2213-8587(26)00029-X.
6. Jagannathan R, Staimez LR, Narayan KMV. Type 2 diabetes subtypes classification: a global reckoning with heterogeneity. Lancet Diabetes Endocrinol. 2025;13(11):903-905.
7. Bahendeka S, Wesonga R, Mutungi G, Muwonge J, Neema S, Guwatudde D. Prevalence and correlates of diabetes mellitus in Uganda: a population-based national survey. Trop Med Int Health. 2016;21(3):405-16.
8. Kusolo R, Mutungi GN, Mbuliro M, Kajjura R, Wesonga R, Bahendeka SK, et al. Changes in the prevalence of the common risk factors for non-communicable diseases in Uganda between 2014 and 2023. PLOS Glob Public Health. 2025;5(4):e0003755.
9. Kusolo R, Mutungi GN, Mbuliro M, Kajjura R, Wesonga R, Bahendeka SK, et al. Uganda 2023 NCD risk factor STEPS survey data [dataset]. 2025. doi:10.1371/journal.pgph.0003755.s002.
10. International Diabetes Federation. Uganda diabetes country report 2000-2050. Diabetes Atlas; 2025.
11. Bukenya D, Van Hout MC, Shayo EH, Kitabye I, Junior BM, Kasidi JR, et al. Integrated healthcare services for HIV, diabetes mellitus and hypertension in selected health facilities in Kampala and Wakiso districts, Uganda: a qualitative methods study. PLOS Glob Public Health. 2022;2(2):e0000084.
12. Birabwa C, Bwambale MF, Waiswa P, Mayega RW. Quality and barriers of outpatient diabetes care in rural health facilities in Uganda: a mixed methods study. BMC Health Serv Res. 2019;19:706.
13. Musoke D, Boynton P, Butler C, Musoke MB. Health seeking behaviour and challenges in utilising health facilities in Wakiso district, Uganda. Afr Health Sci. 2014;14(4):1046-1055.

14. Brubacher LJ, Berrang-Ford L, Clark SN, Patterson K, Lwasa S, Namanya D, et al. Place, displacement, and health-seeking behaviour among the Ugandan Batwa: a qualitative study. PLOS Glob Public Health. 2024;4(6):e0003321.
15. Nabagereka H, Nalubuga B. Individual determinants of health-seeking behavior among the youth aged 14-18 years attending Kasozi Health Centre III, Wakiso district: a cross-sectional study. J World Health Res. 2025;2(1):1-8.
16. Kiconco R, Kalyesubula R, Sanyu N, Agaba BB, Ondari EN, Tusubira D, et al. Early nephropathy screening as a health seeking behavior among type 2 diabetes mellitus patients: a phenomenological study in southwestern Uganda. J Multidiscip Healthc. 2025;18:5875-5892.
17. Kasujja FX, Nuwaha F, Daivadanam M, Kiguli J, Etajak S, Mayega RW. Understanding the diagnostic delays and pathways for diabetes in eastern Uganda: a qualitative study. PLoS One. 2021;16(4):e0250421.
18. Abindu V, Hope D, Aleni M, Andru M, Ayiasi RM, Afayo V, et al. Missed diagnosis of gestational diabetes mellitus due to selective screening: evidence from a cross-sectional study in the West Nile sub-region, Uganda. Diabetes Metab Syndr Obes. 2024;17:1309-1319.
19. Chona EZ, Kayange LF, Iseselo MK. Barriers and facilitators to satisfaction with diabetes care: the perspectives of patients attending public diabetic clinics in Dar es Salaam, Tanzania. PLoS One. 2024;19(5):e0302858.
20. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol. 1996;58(1):267-288.
21. Tabak AG, Herder C, Rathmann W, Brunner EJ, Kivimaki M. Prediabetes: a high-risk state for diabetes development. Lancet. 2012;379(9833):2279-90.
22. Lindstrom J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care. 2003;26(3):725-731.
23. Omech B, Mwita JC, Tshikuka JG, Tsima B, Nkomazna O, Amone-P'Olak K. Validity of the Finnish Diabetes Risk Score for detecting undiagnosed type 2 diabetes among general medical outpatients in Botswana. J Diabetes Res. 2016;2016:4968350.
24. Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. BMJ. 2009;338:b880.
25. Chen L, Magliano DJ, Balkau B, Colagiuri S, Zimmet PZ, Tonkin AM, et al. AUSDRISK: an Australian type 2 diabetes risk assessment tool based on demographic, lifestyle and simple anthropometric measures. Med J Aust. 2010;192(4):197-202.
26. Katende D, Mutungi G, Baisley K, Biraro S, Ikoona E, Peck R, et al. Readiness of Ugandan health services for the management of outpatients with chronic diseases. Trop Med Int Health. 2015;20(10):1385-95.
27. Roglic G, editor. Global report on diabetes. Geneva: World Health Organization; 2016.
28. Fox K, Borer JS, Camm AJ, Danchin N, Ferrari R, Lopez Sendon JL, et al. Resting heart rate in cardiovascular disease. J Am Coll Cardiol. 2007;50(9):823-30.

29. Thorp AA, Owen N, Neuhaus M, Dunstan DW. Sedentary behaviors and subsequent health outcomes in adults: a systematic review of longitudinal studies, 1996-2011. Am J Prev Med. 2011;41(2):207-15.
30. Ministry of Health, Uganda. Health Sector Development Plan 2015/16-2019/20. Kampala: Ministry of Health; 2015.
31. 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.
32. Pan XR, Li GW, Hu YH, Wang JX, Yang WY, An ZX, et al. Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance: the Da Qing IGT and Diabetes Study. Diabetes Care. 1997;20(4):537-44.
33. Kintu A, Sando D, Okello S, Mutungi G, Guwatudde D, Menzies NA, et al. Integrating care for non-communicable diseases into routine HIV services: key considerations for policy design in sub-Saharan Africa. J Int AIDS Soc. 2020;23(Suppl 1):e25508.
34. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.

Tables

Table 1. Data dictionary of the 2023 WHO STEPS survey data

Feature	Label	Codes
age	Age in years	
b8	Total cholesterol (mmol/L)	
bmi	Body Mass Index (kg/m2)	
c1	Sex	1=Male; 2=Female
c5	Highest level of education	
d1	Days per week eating fruits	
d2	Servings of fruit on one of those days	
d3	Days per week eating vegetables	
d4	Servings of vegetables on one of those days	
DBP	Diastolic Blood Pressure (mmHg)	
district_2020	District of residence (2020)	
HR	Heart Rate (beats per minute)	
m14	Waist circumference (cm)	
m15	Hip circumference (cm)	
p1	Work involves vigorous-intensity activity	1=Yes; 2=No
p10	Any vigorous-intensity sports/fitness/recreation	1=Yes; 2=No
p13	Any moderate-intensity leisure activities	1=Yes; 2=No
p16a	Hours spent sitting or reclining per day	
p16b	Days per week of physical activity	
p4	Work involves moderate-intensity activity	1=Yes; 2=No
p7	Walk or cycle to and from places	1=Yes; 2=No
region	Region of residence	1=Central; 2=Eastern; 3=Northern; 4=Western
SBP	Systolic Blood Pressure (mmHg)	
status	Place of residence	1=Urban; 2=Rural
t1	Currently smoke any tobacco product	1=Yes; 2=No
t12	Currently use smokeless tobacco products	1=Yes; 2=No

t8	Currently use other tobacco product	1=Yes; 2=No
village	Village	

Table 2. Performance metrics during model selection (5-fold cross-validation)

Model	Sn mean	Sp mean	PPV mean	F1 mean	ROC AUC mean
LASSO	0.71	0.56	0.31	0.43	0.68
Random Forest	0.70	0.54	0.30	0.42	0.68
Balanced Random Forest	0.71	0.53	0.29	0.41	0.67
LightGBM	0.71	0.49	0.28	0.40	0.66
XGBoost	0.71	0.50	0.28	0.40	0.65
SVM	0.71	0.50	0.28	0.40	0.65
Gaussian Naive Bayes	0.74	0.45	0.28	0.40	0.65
MLP	0.71	0.47	0.27	0.39	0.63
Logistic Regression	0.71	0.37	0.23	0.35	0.56

Table 3. Summary statistics of the numerical features according to the outcome

Feature	IFG Mean (SD)	Normal FG Mean (SD)	IFG Range	Normal FG Range
DBP	86.0 (12.6)	80.9 (11.0)	57.0-153.7	51.0-147.7
HR	80.1 (11.9)	76.7 (12.1)	50.0-113.3	44.0-127.0
SBP	129.1 (21.0)	123.0 (17.2)	90.0-216.3	85.7-223.0
age	43.4 (13.2)	38.4 (13.7)	18.0-69.0	18.0-69.0
b8	3.5 (0.9)	3.1 (0.8)	2.6-6.1	2.6-10.4
BMI	25.6 (9.1)	22.7 (5.8)	13.1-122.8	12.6-67.4
d1	3.2 (2.4)	3.3 (2.3)	0.0-7.0	0.0-7.0
d2	2.3 (2.5)	2.6 (2.8)	1.0-20.0	1.0-20.0
d3	3.7 (2.3)	3.6 (2.3)	0.0-7.0	0.0-7.0
d4	2.0 (1.5)	2.2 (2.3)	1.0-17.0	1.0-20.0
m14	86.6 (14.1)	80.1 (10.2)	57.5-132.0	30.0-135.0
m15	97.7 (14.0)	92.2 (10.4)	68.0-159.2	45.0-160.5
p16a	3.2 (2.4)	2.8 (2.1)	0.0-13.0	0.0-12.0
p16b	6.0 (12.5)	6.3 (12.7)	0.0-50.0	0.0-50.0

Table 4. Summary statistics of the categorical features according to the outcome

Feature	Level	IFG n (%)	Normal FG n (%)
Sex (c1)	Male	83 (28.9)	424 (40.3)
	Female	204 (71.1)	629 (59.7)
Education (c5)	None	184 (64.1)	591 (56.1)
	Primary	60 (20.9)	282 (26.8)
	Secondary	31 (10.8)	141 (13.4)
	Tertiary	12 (4.2)	39 (3.7)
Vigorous work (p1)	Yes	76 (26.5)	379 (36.0)
	No	211 (73.5)	674 (64.0)
Vigorous sports (p10)	Yes	27 (9.4)	146 (13.9)
	No	260 (90.6)	907 (86.1)
Vigorous leisure (p13)	Yes	37 (12.9)	193 (18.3)
	No	250 (87.1)	860 (81.7)
Moderate work (p4)	Yes	256 (89.2)	952 (90.4)
	No	31 (10.8)	101 (9.6)
Walking/cycling (p7)	Yes	256 (89.2)	972 (92.3)
	No	31 (10.8)	81 (7.7)
Region	Central	90 (31.4)	260 (24.7)
	Eastern	87 (30.3)	294 (27.9)
	Northern	36 (12.5)	254 (24.1)
	Western	74 (25.8)	245 (23.3)
Residence (status)	Urban	67 (23.3)	238 (22.6)
	Rural	220 (76.7)	815 (77.4)
Currently smoking (t1)	Yes	18 (6.3)	81 (7.7)
	No	269 (93.7)	972 (92.3)
Smokeless tobacco (t12)	Yes	4 (1.4)	37 (3.5)
	No	283 (98.6)	1016 (96.5)
History of smoking (t8)	Yes	28 (9.8)	83 (7.9)
	No	259 (90.2)	970 (92.1)

Table 5. Performance metrics of the final calibrated LASSO model (out-of-fold evaluation)

Metric	Estimate	95% CI
Sensitivity	0.60	-
Specificity	0.69	-
PPV	0.34	-
NPV	0.86	-
F1	0.44	-
ROC AUC	0.68	0.64-0.71
PR AUC	0.40	0.34-0.46
Brier score	0.155	0.143-0.168

Out-of-fold predictions from nested 5-fold cross-validation (N=1340; 287 cases). 95% CIs from 2,000 bootstrap resamples. Sensitivity, specificity, PPV, NPV and F1 are reported at the G-mean optimal threshold (approximately 0.21; TP=171, FP=326, FN=116, TN=727). Isotonic calibration did not change discrimination (uncalibrated OOF AUC 0.68).

Table 6. Comparison of the final LASSO model with other risk scores in the literature

Study / Score	Country / Cohort (N)	Outcome	Model family	Validation	Reported AUC
This study (IFG-LASSO)	Uganda case-cohort (n=1340)	IFG	LASSO-logistic + isotonic calibration	Nested CV (5x), OOF	0.68; Sn 0.60, Sp 0.69
FINDRISC	Finland primary care	Undiagnosed / future T2DM	Points score (logistic)	Multiple external	0.85-0.87
QDiabetes	UK CPRD (millions)	Incident T2DM	Cox / Poisson	Extensive external	0.83-0.85
AUSDRISK	Australia population	Undiagnosed / 5-yr T2DM	Points score (logistic)	External (Asia-Pacific)	0.66-0.81

IFG and T2DM are different endpoints. T2DM carries a stronger metabolic signal, so the AUCs above are not directly comparable to the IFG endpoint used in this study. Our threshold strategy: low risk < 0.05, high risk >= 0.28, tunable to resource levels.

Figures

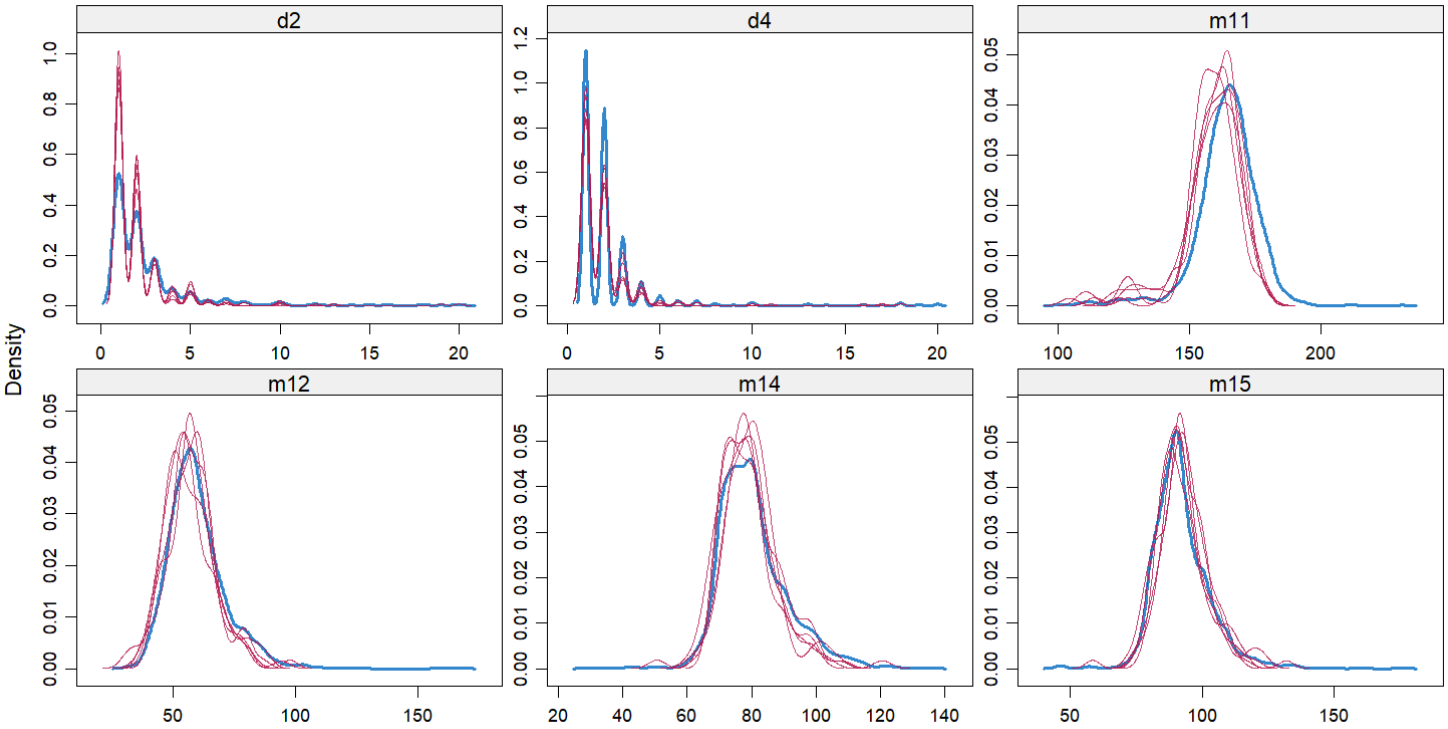

Figure 1

Density plots showing data imputation.

Density plots comparing observed (blue) and multiply-imputed (red) values for variables with missing data, generated from the first of five completed MICE datasets.

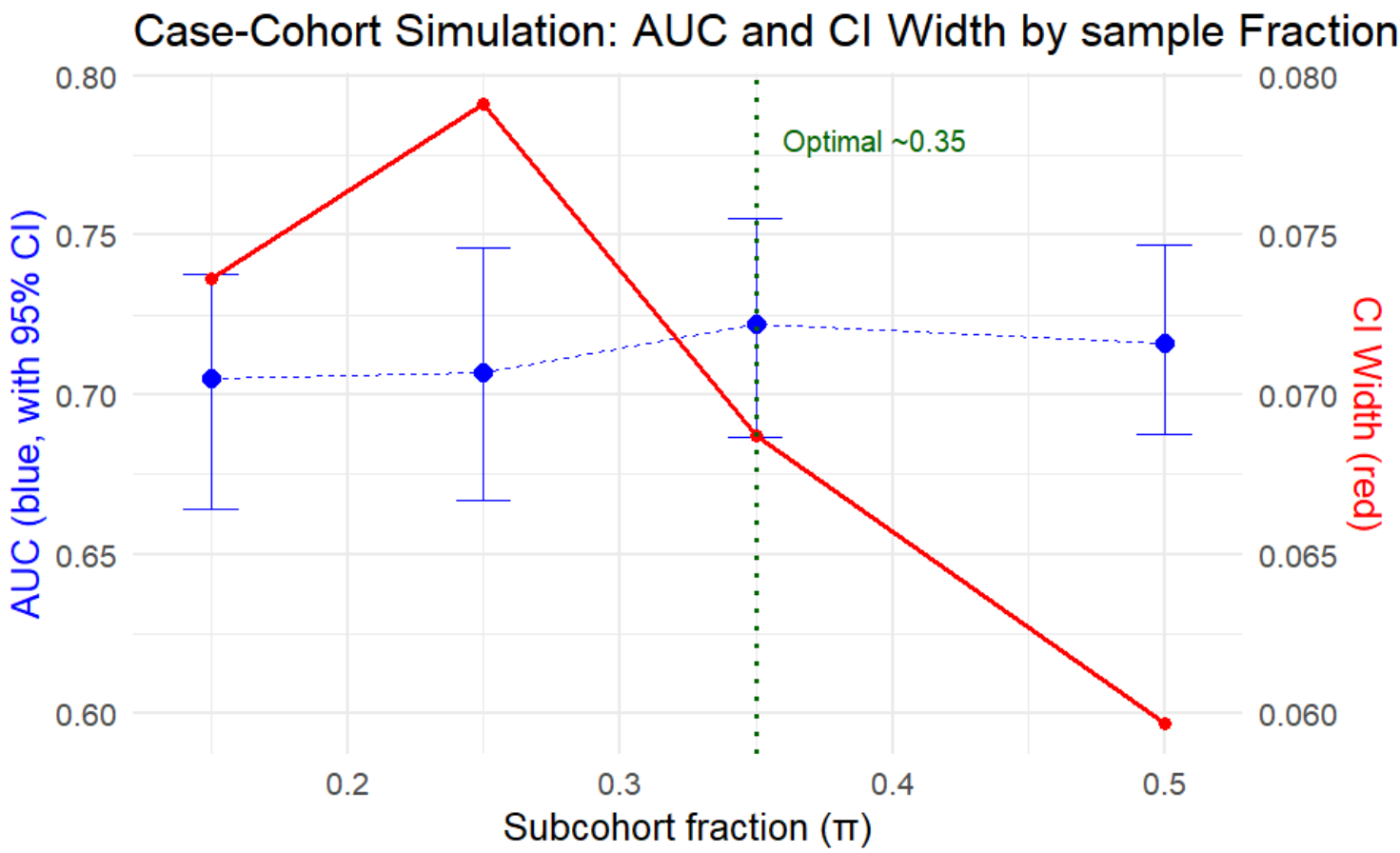

Figure 2

Simulation of the case-cohort sampling fraction.

AUC (blue, left axis) and 95% bootstrap confidence interval width (red, right axis) plotted across candidate sub-cohort sampling fractions (15%, 25%, 32%, 50%). The realized fraction of 32.2% was selected.

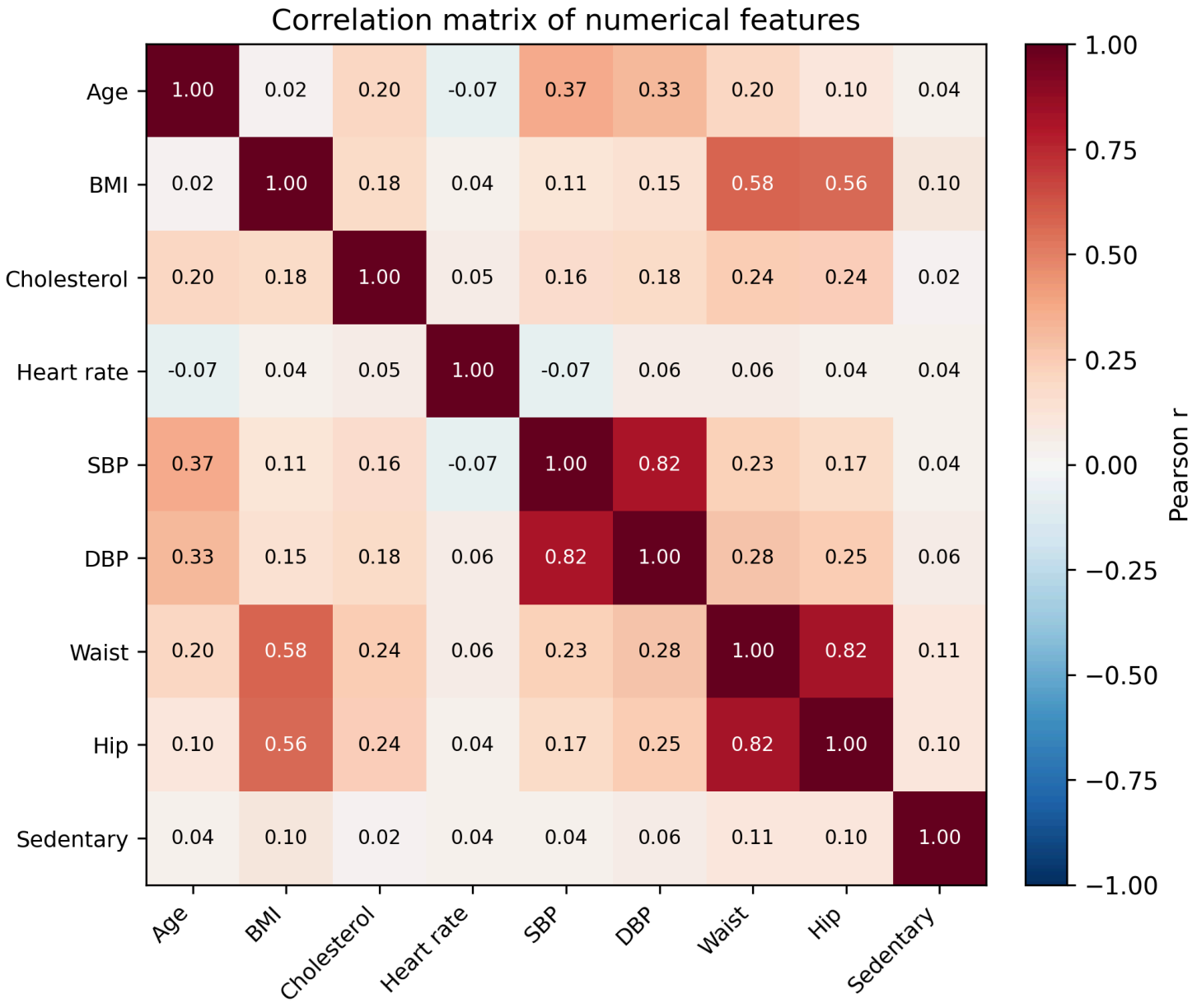

Figure 3

Correlation matrix of the numerical features.

Pearson correlation coefficients among numeric predictors (age, BMI, total cholesterol, heart rate, systolic and diastolic blood pressure, waist and hip circumference, sedentary hours), N=1,340.

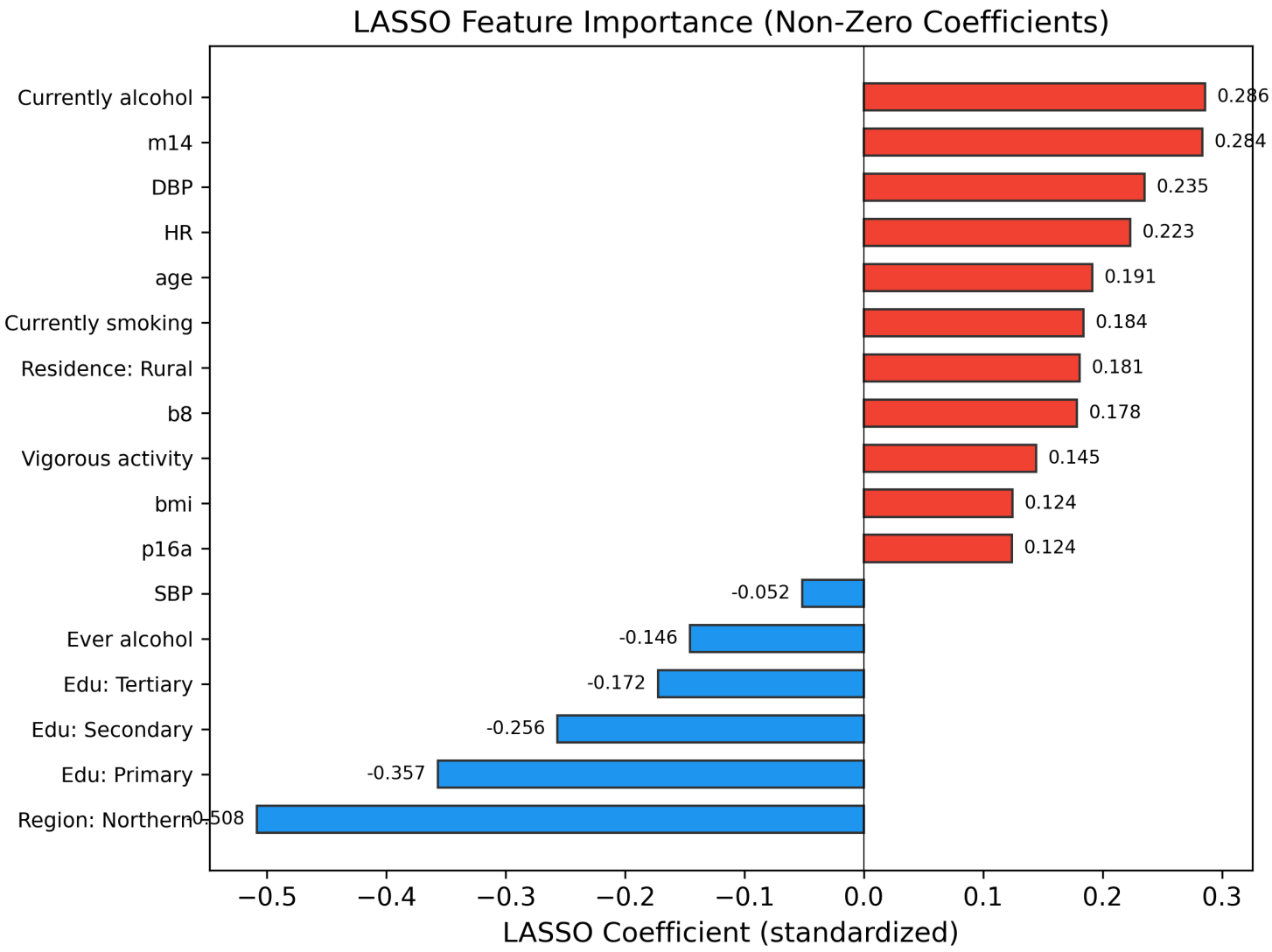

Figure 4

Feature importance from the final LASSO model.

Non-zero standardized coefficients from the LASSO model fit on the full dataset; negative values indicate a protective association with IFG, positive values indicate an increased-risk association.

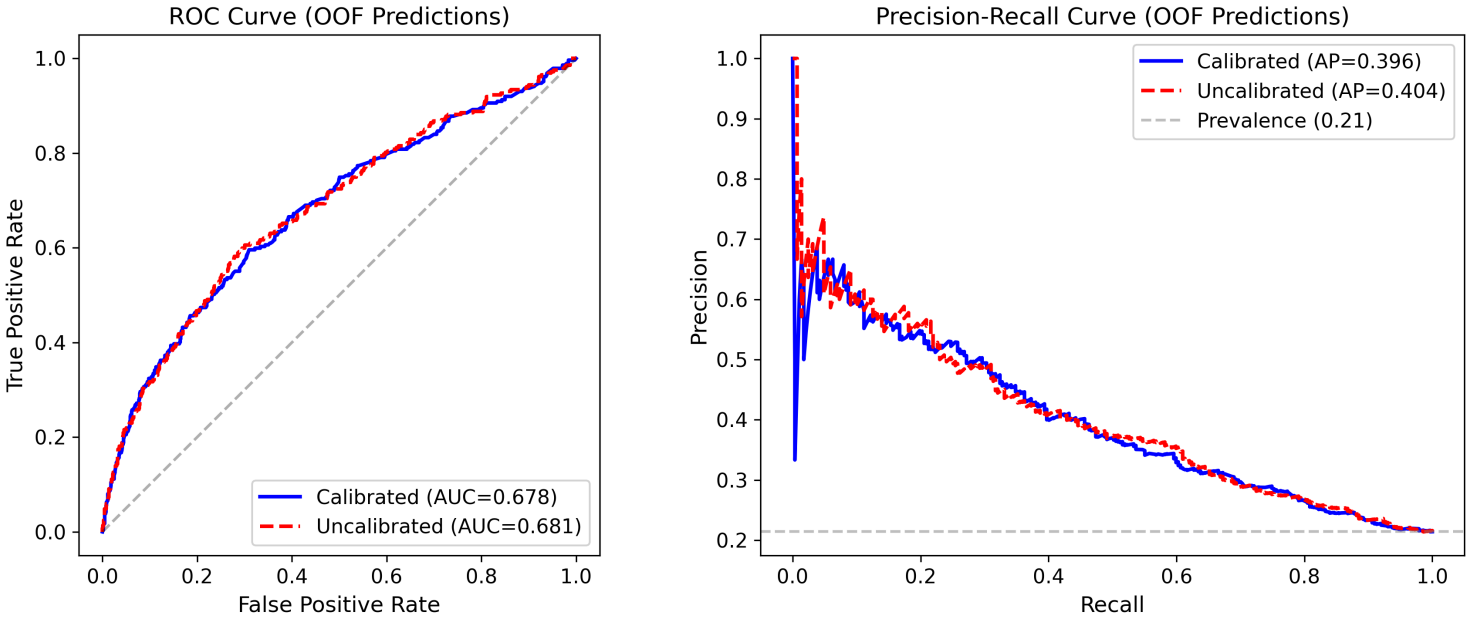

Figure 5

Out-of-fold ROC and precision-recall curves.

Left: receiver operating characteristic curve (calibrated vs. uncalibrated). Right: precision-recall curve (calibrated vs. uncalibrated). Both are computed from concatenated out-of-fold predictions of the final calibrated LASSO model across the 5 outer cross-validation folds.

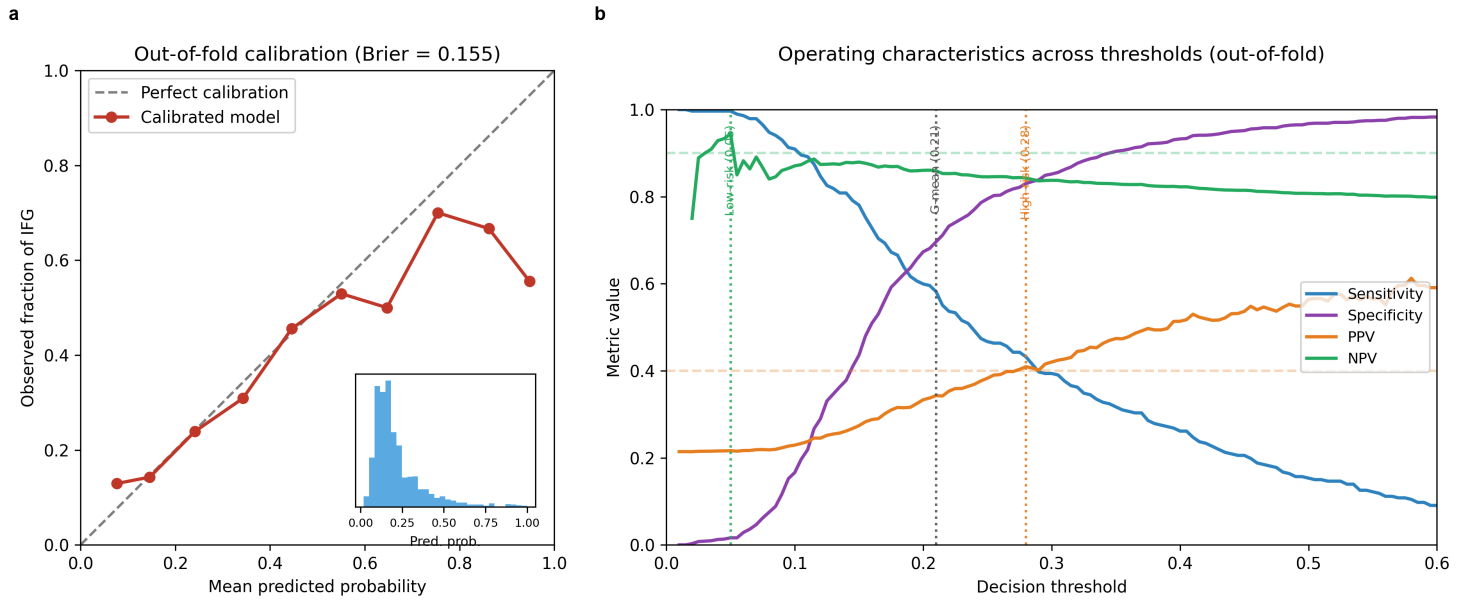

Figure 6

Out-of-fold calibration and threshold operating characteristics.

(a) Calibration curve (Brier score = 0.155) comparing mean predicted probability to observed fraction of IFG across 10 bins; inset shows the distribution of predicted probabilities. (b) Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) across decision thresholds, with the low-risk (0.05), balanced G-mean (0.21), and high-risk (0.28) cutoffs marked; dashed horizontal lines show the NPV >= 0.90 and PPV >= 0.40 targets.

Supplementary Files

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

- Listofadditionalfiles.docx