

Article

Non-Invasive Sociodemographic Cues for Type 2 Diabetes Risk Stratification in South Africa

Reza Fahimi ^{1,*}, Abdulaziz Alrubayyi ^{1,2}, Habibullah Muhammad Kamal ¹, Ibrahim Khalil Ja'afar ¹,
Mohanad Alhothify ¹, Thomas M. Barber ¹ and Olalekan A. Uthman ¹

¹ Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK; abdulaziz.alrubayyi@warwick.ac.uk (A.A.); habeeb.muhammad-kamal@warwick.ac.uk (H.M.K.); ibrahim.ja-afar@warwick.ac.uk (I.K.J.); mochanad.alhothify@warwick.ac.uk (M.A.); t.barber@warwick.ac.uk (T.M.B.); olalekan.uthman@warwick.ac.uk (O.A.U.)

² College of Applied Medical Science, Jazan University, Jazan 45142, Saudi Arabia

* Correspondence: reza.fahimi@warwick.ac.uk

Abstract

Background/Objectives: Type 2 diabetes mellitus (T2DM) is a growing burden in low- and middle-income countries, and non-invasive risk tools have been proposed where laboratory screening is not feasible. Fast-and-frugal decision trees (FFDTs) are transparent, non-compensatory rules that make sensitivity–specificity trade-offs explicit. We asked how far five routinely collected sociodemographic cues can go for classifying self-reported diabetes status and benchmarked FFDTs against the simplest possible alternatives. **Methods:** A cross-sectional secondary analysis of the 2016 South Africa Demographic and Health Survey Adult Health recode (N = 10,292; 459 self-reported cases, 4.46%) was carried out. Five binary cues were derived: age category, household wealth, employment, education and sex. Data were split 70/30 into development and a locked hold-out set. Three purpose-specific FFDTs were selected by 5 × 10-fold cross-validation within the development data only, frozen, and then evaluated once on the hold-out set. Comparators were a pre-specified age-only rule and logistic regression using the same five predictors at a development-derived Youden threshold. We report survey-weighted estimates, internal–external cross-validation across nine provinces, external validation in 12 further DHS, subgroup performance, and a log–log analysis of error against training sample size. **Results:** On the hold-out set the balanced screening FFDT achieved sensitivity 0.920 (95% CI 0.862–0.959), specificity 0.585 (0.567–0.602) and balanced accuracy 0.752 (0.727–0.776). The pre-specified age-only rule achieved balanced accuracy 0.751 (0.724–0.775), while logistic regression at the development-derived Youden threshold achieved 0.766 (0.735–0.797). Across 50 independent splits, the case-finding FFDT was arithmetically identical to the age-only rule in every split; the balanced FFDT exceeded it by a median of 0.001 (95% range –0.004 to +0.025); the referral-minimisation FFDT was worse by a median of 0.124. Performance was stable across provinces (balanced accuracy 0.731–0.801) but degraded across countries (0.508–0.724), with referral rate ranging from 0.225 to 0.806. Error scaled with training size as $\beta = -0.044$ (95% CI –0.063 to –0.025). **Conclusions:** Transparent decision trees built from non-invasive sociodemographic cues classify self-reported diabetes status no better than a single question about age, do not transport reliably across countries, and are unlikely to be materially improved by simply increasing the training sample size. Their value lies in auditability rather than accuracy. Progress in low-burden risk stratification for T2DM in these settings will require cues that carry a biological rather than health system signal.



Academic Editor: Peter Clifton

Received: 24 June 2026

Revised: 29 July 2026

Accepted: 3 August 2026

Published: 7 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: type 2 diabetes; risk stratification; fast-and-frugal decision trees; decision trees; model benchmarking; external validation; South Africa; demographic and health surveys; low- and middle-income countries; diabetes epidemiology

1. Introduction

Type 2 diabetes mellitus (T2DM) is a major and growing cause of premature morbidity and mortality worldwide, with the largest absolute increases in disease burden projected for sub-Saharan Africa and other low- and middle-income countries (LMICs) [1–3]. Despite well-defined diagnostic criteria, a substantial proportion of cases remain undiagnosed in these settings: health system evaluations consistently document major gaps in diagnosis, treatment, and glycaemic control, and treatment coverage is frequently inadequate even in nationally representative samples [4–6]. Laboratory-based screening is difficult to operationalise at scale in resource-constrained environments, where supply chains, equipment, and follow-up capacity may all be limited [4–6]. Risk assessment methods that use routinely available, non-invasive variables have therefore been proposed as practical complements to confirmatory testing and as instruments for targeting early intervention [7,8]. According to current estimates, hundreds of millions of adults have diabetes, and by the middle of the century, there will likely be a significant increase in the number of people with the disease. These trends are accompanied by high and increasing societal costs related to treatment, complications, lost productivity, and disability [1–3,9]. T2DM's significant correlation with cardiovascular morbidity and mortality, as well as microvascular problems that significantly lower quality of life and raise care needs, exacerbates its clinical and public health effects [9].

Numerous non-invasive diabetes risk scores have been developed, most incorporating anthropometric, demographic, and lifestyle variables. However, systematic evidence synthesis has documented high model multiplicity, inconsistent performance, and frequent miscalibration when scores are transferred between populations with differing baseline risk, case-mix, and measurement practices [8,10–12]. Machine learning (ML) approaches can achieve stronger discrimination under some conditions but typically require larger feature sets, substantial computational resources, and careful management of overfitting and dataset shift [13–16]. Translating apparent performance in development data into dependable real-world impact remains difficult; opacity, auditability concerns, and governance challenges frequently constrain adoption in clinical and public health settings [15,17–19]. Population-level risk classification and prioritisation continue to be crucial for programme design and resource allocation, as evidenced by nationally representative data showing that full treatment coverage is frequently inadequate in LMIC settings [6].

In response, reporting and evaluation frameworks for clinical prediction models have increasingly emphasised transparency, reproducibility, and risk-of-bias assessment. The TRIPOD statement and its elaboration formalised requirements for prediction model development and validation studies [20,21]; TRIPOD + AI extended these requirements to ML-based models, with particular attention to deployment-relevant evaluation [22]. Complementary tools such as PROBAST provide structured approaches to assessing bias and applicability [23,24], and calibration has been identified as a critical but frequently overlooked dimension of model performance [11]. These requirements have strengthened the rationale for model classes that are interpretable by construction, require no post hoc explanation methods, and remain implementable in settings without computational infrastructure, characteristics that are particularly relevant for community-level screening in LMICs [4–6,19]. As a result, risk assessment methods based on routinely obtained, non-

invasive variables have been frequently suggested as useful supplements to confirmatory testing and as instruments for focusing on early intervention and prevention [7,8].

Fast-and-frugal decision trees (FFDTs) are sequential decision heuristics that enforce structural simplicity: at each node, a single binary question is asked, and the answer either assigns an immediate classification or passes the individual to the next node [25]. Because each cue can trigger an early exit, the model is non-compensatory and fully specified by a small number of if-then rules that can be depicted in a single figure, implemented on paper, and audited without computational tools [25,26]. FFDTs make sensitivity-specificity trade-offs explicit at the point of model selection rather than post hoc, enabling purpose-specific models to be constructed for distinct clinical and public health objectives, including case-finding, population screening, and referral minimisation, each with a transparently defined error profile [25,27]. In population survey settings, where outcome prevalence is typically low and class imbalance inflates naïve accuracy, balanced accuracy (BAcc) provides a more appropriate performance summary than overall accuracy [28–30]. Non-invasive sociodemographic variables available in standardised surveys such as the Demographic and Health Surveys (DHS) are well suited to this architecture, offering coverage across LMICs without requiring laboratory inputs [4–6]. However, systematic evidence synthesis has brought to light the multiplicity of models, inconsistent performance, and the necessity of rigorous validation and selection when transferring scores between groups [8]. When models created in one epidemiological environment are used in another, where baseline risk, case-mix, measurement regimes, and health system paths vary, these issues are more pertinent [10–12].

In this study we developed three purpose-specific FFDTs for classifying self-reported diabetes status in a South African population survey, using five non-invasive sociodemographic cues, and then asked how much they add over the simplest available alternatives. We aimed to: (i) identify cues that provide discrimination in this survey context; (ii) construct FFDT configurations representing distinct operating points on the sensitivity-specificity spectrum; (iii) benchmark those configurations against a pre-specified age-only rule and against logistic regression using the same predictors; (iv) assess transportability across provinces and across countries; (v) characterise how classification error scales with training sample size. The intended contribution is not to replace biochemical diagnosis, but to establish what routinely collected sociodemographic information can and cannot achieve for low-burden risk stratification in resource-constrained settings.

2. Materials and Methods

This study was conducted as a cross-sectional secondary analysis and reported in accordance with STROBE guidance for observational studies and TRIPOD + AI guidance for prediction and classification model reporting [22]. The analytical objective was to develop interpretable, point-of-care-compatible classification rules for self-reported diabetes status using routinely available, non-invasive sociodemographic cues. Three purpose-specific fast-and-frugal decision trees were constructed to encode distinct sensitivity-specificity trade-offs for case-finding, balanced screening, and referral minimisation [22].

2.1. Study Design and Dataset Description

We conducted a cross-sectional secondary analysis of anonymised, individual-level data from the 2016 South Africa Demographic and Health Survey, using the Adult Health recode for women and men merged with the household member recode. Of 10,336 Adult Health respondents, 10,292 had a non-missing response to the diabetes question and complete data on all five cues; these formed the analytic sample (459 self-reported cases,

4.46%). The individual was the unit of analysis. All analyses used de-identified data. Participant flow is shown in Figure 1.

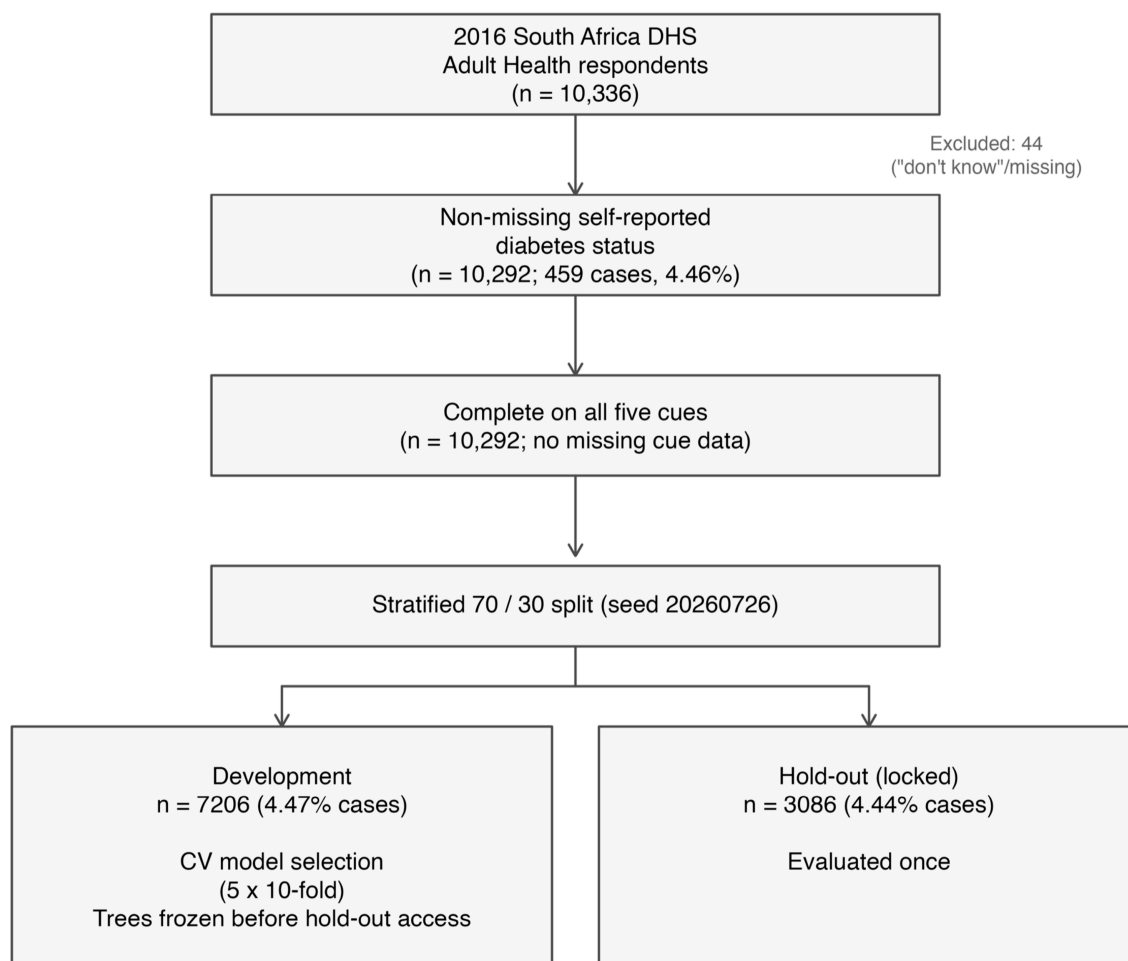


Figure 1. Participant flow from Adult Health respondents to the development and locked hold-out sets.

2.2. Outcome Definition

The outcome was a self-reported diagnosis of diabetes by a health professional, taken from the 2016 SADHS Adult Health recode: variable S1413F (“Doctor, nurse diagnosed with: diabetes”) for women and SM1108F for men, each coded 0 = No, 1 = Yes, 8 = Don’t know. Respondents coded 1 were classified “Diabetes” and those coded 0 “No Diabetes”; those coded 8 or with a missing response were excluded ($n = 44$ of 10,336). The models therefore classify self-reported, previously diagnosed diabetes rather than biomarker-confirmed disease. Performance metrics are interpreted as agreement with this dataset-defined label and not as diagnostic accuracy or as detection of undiagnosed diabetes.

2.3. Candidate Predictors and Cue Engineering

Candidate predictors comprised five routinely collected individual-level sociodemographic variables available in the 2016 South Africa Demographic and Health Survey: age category, household wealth index, employment status, educational attainment, and sex. Each variable was operationalised as a binary cue suited to the FFDT architecture; for example, “middle-aged or older adult,” “richer or richest wealth quintile,” “any recorded work activity,” “primary schooling or below,” and “female sex.” Binary operationalisation is central to the FFDT methodology: it constrains model complexity, produces auditable decision paths, and ensures each node poses a single answerable question [25,29,31,32].

We calculated cue-level sensitivity, specificity, and balanced accuracy ($B\text{Acc} = [\text{sensitivity} + \text{specificity}]/2$) for each candidate cue to quantify its marginal discriminative capacity. Operating points were visualised in a ROC space using the FFTrees “cue accuracies” function. $B\text{Acc}$ was used to rank cue thresholds because it weights sensitivity and specificity equally and remains informative under class imbalance. Cues with higher $B\text{Acc}$ and interpretable thresholds were prioritised for early placement in candidate trees, consistent with the principle that fast-and-frugal decision-making should examine the most informative cues first [25,29,31,32].

2.4. Data Preparation, Class Imbalance, and Missingness

Before model fitting, we audited variable types and standardised categorical predictors to uniform factor levels to ensure stable cue thresholds across training and testing subsets. The analytic sample had a self-reported diabetes prevalence of 4.46%, indicating substantial class imbalance. Under such conditions, overall accuracy is misleading: a naïve classifier that assigns “No Diabetes” to everyone would achieve approximately 96% accuracy with zero sensitivity. We therefore prioritised sensitivity, specificity, and $B\text{Acc}$ over raw accuracy in model selection and reporting [29].

Rather than optimising against a single global metric, we explicitly accounted for the asymmetric costs of false negatives and false positives for each intended application, following current guidance on imbalanced binary classification [29,30]. Missing data handling was applied uniformly across all candidate trees to ensure that any performance differences reflected model structure rather than variation in sample composition [29,30].

2.5. FFDT Model Development

We built FFDTs in R using the FFTrees package, which provides tools for constructing, visualising, and evaluating fast-and-frugal decision trees [25]. Structurally, FFDTs process each individual sequentially: at each node, a single binary question is asked, and the answer either assigns an immediate classification (“Diabetes” or “No Diabetes”) or passes the individual to the next node. No information from earlier nodes can override a subsequent decision, a property termed non-compensatory processing, which limits the number of cues examined per individual whilst maintaining discriminatory capacity [25,29,31,32]. This architecture is grounded in the bounded rationality literature on lexicographic and one-reason decision-making [31,32].

Candidate trees were generated using the FFTrees “fan” algorithm, which evaluates alternative cue orderings and exit configurations; full construction details are provided in Supplementary Methods S1 [25].

The analytic dataset was split into a development set (70%; $n = 7206$) and a locked hold-out set (30%; $n = 3086$), stratified on the outcome, under a fixed random seed. All model development and selection used the development data only.

Three purpose-specific configurations were pre-specified by the relative weight placed on sensitivity in the FFTrees weighted accuracy criterion: a case-finding configuration ($\text{sens.w} = 0.75$), a balanced screening configuration ($\text{sens.w} = 0.50$) and a referral-minimisation configuration ($\text{sens.w} = 0.25$). For each configuration, candidate trees were generated by the FFTrees fan algorithm, and the winning tree was chosen by mean weighted accuracy across 5×10 -fold cross-validation within the development data. The winning configuration was then refitted on the full development set and frozen; its decision rules were exported programmatically before the hold-out set was accessed. Each frozen tree was applied to the hold-out set once. No hold-out information was used at any point in model selection.

Because the choice of development and hold-out partition proved consequential, the entire pipeline, comprising split, cross-validated selection, freezing and hold-out evaluation, was repeated across 50 independent random partitions, and the distribution of the difference in balanced accuracy against the age-only rule is reported.

2.6. Strategy for Selecting Three Purpose-Specific FFDTs

Three FFDT configurations were retained to represent distinct operating profiles rather than because they were expected to outperform benchmark models. The case-finding configuration gave greater weight to sensitivity, the balanced screening configuration weighted sensitivity and specificity equally, and the referral-minimisation configuration gave greater weight to specificity. These objectives were implemented through the FFTrees weighted accuracy criterion using sensitivity weights of 0.75, 0.50, and 0.25, respectively [25,27,33].

For each configuration, model selection was performed within the development set only, using 5×10 -fold cross-validation. The selected configuration was then refitted on the full development set, frozen, and then applied once to the locked hold-out set. No hold-out data were used during model selection. The three configurations were therefore interpreted as transparent, purpose-specific classification rules with different error profiles and were subsequently benchmarked against an age-only rule and logistic regression using the same predictors [27,33].

2.7. Model Evaluation and Statistical Reporting

For each frozen FFDT we summarised hold-out performance using confusion matrices and derived metrics with 95% confidence intervals: exact binomial (Clopper–Pearson) intervals for sensitivity, specificity, positive and negative predictive value, and referral rate; the Simel log method for likelihood ratios; the inverted predictive value interval for number needed to test; and a 2000-resample bootstrap for balanced accuracy, which has no closed form. Differences in balanced accuracy between rules were estimated by paired bootstrap on the hold-out set. Subgroup performance of the balanced screening FFDT was evaluated on the locked hold-out set by sex, age category, household wealth quintile, educational attainment, and employment status, using sensitivity, specificity, positive predictive value, and referral rate. For all three FFDT configurations, between-group gaps in sensitivity, specificity, and referral rate were calculated as the maximum minus the minimum value across levels of each characteristic; strata with fewer than 40 participants or fewer than eight self-reported cases were omitted. Balanced accuracy was the primary scalar summary because it is robust to outcome prevalence under class imbalance.

2.8. Comparators and Benchmarking

Two benchmarks were pre-specified. The first was an age-only rule classifying all respondents aged 40 years or over as “Diabetes”. The second was a multivariable logistic regression using the same five predictors, fitted on the development set. The classification threshold was selected in the development data by maximising Youden’s index and was then applied unchanged to the locked hold-out set [34].

Established non-invasive risk scores, including the African Diabetes Risk Score, FIND-RISC and the Cambridge Risk Score, could not be computed because their required combinations of anthropometric measurements and family history variables were not all available for the full Adult Health analytic sample. As a partial substitute we tested whether measured anthropometry contributes at all by adding a body mass index category as a sixth cue in the subsample with measured height and weight.

2.9. Survey Design, Missing Data and Validation

Survey-weighted performance was estimated using the Adult Health sample weights (SWEIGHT for women, SMWEIGHT for men), with primary sampling units and sampling strata specified and Taylor-linearised confidence intervals. DHS does not publish a combined adult health weight, so pooled male-female weighted estimates should be interpreted with that caveat.

The five cues had no missing values among respondents with a non-missing outcome, so the complete-case analysis is the full analysis and no imputation was performed.

Transportability was assessed at two levels. Internal–external cross-validation was performed by leave-one-province-out resampling: the full development pipeline was repeated on eight provinces and evaluated on the ninth, for all nine provinces in turn. External validation was performed by applying the frozen South African balanced screening FFDT, without refitting or threshold adjustment, to 12 further DHS, using the same self-reported outcome definition.

2.10. Error Scaling

To characterise the relationship between training sample size and classification error, development subsamples were drawn at ten fractions of the development set ($n = 360$ to 7206) with 30 resamples at each fraction. The full selection pipeline was repeated within each subsample, and every resulting model was evaluated on the same locked hold-out set. Error was defined as 1 minus balanced accuracy, and the scaling relationship $\ln(\text{error}) = \alpha + \beta \times \ln(n)$ was estimated by least squares.

Throughout, performance interpretation emphasised the clinical and public health meaning of false-negative and false-positive rates in relation to each tree's intended use (case-finding, screening, or referral minimisation), consistent with recommendations to align evaluation criteria with decision-relevant consequences rather than relying solely on summary discrimination indices [27,30,33].

2.11. Software and Reproducibility

All analyses were conducted in R version 4.5.2 (R Foundation for Statistical Computing, Vienna, Austria). FFDT construction and model evaluation were implemented using the FFDtrees package, version 2.1.0 [25]. The revised balanced screening tree diagram was regenerated in R from the decision rules exported from the frozen model object. Analysis scripts applied cue encoding, data splitting, and performance evaluation consistently across all candidate trees and the three selected FFDTs, supporting reproducible comparison under a common data-generating procedure [33].

2.12. Ethics

Analyses used anonymised secondary data from the 2016 South Africa Demographic and Health Survey. The DHS programme obtains ethical approvals and informed consent for the original survey data collection. No new data were collected; no identifiable participant information was accessed; no attempt was made to re-identify individuals. Additional ethics approval was therefore not required for this secondary analysis. All results are reported in aggregate form. Data are available through the DHS programme subject to registration and approval of a data access request; analysis code is available from the corresponding author on reasonable request.

3. Results

3.1. Sociodemographic Cue Performance

The analytic sample comprised 10,292 individuals, of whom 459 (4.46%) reported a previous diagnosis of diabetes. The locked hold-out set comprised 3086 individuals with 137 self-reported cases (4.44%). The five candidate cues were age category, household wealth, employment status, educational attainment, and sex. Age category provided the strongest marginal discrimination. The remaining sociodemographic cues were less informative, consistent with the subsequent finding that the multivariable FFDT configurations did not reliably outperform the age-only rule.

3.2. Reproducibility of Candidate Tree Selection

Because the FFTrees development and testing partition is stochastic, the identity and performance of candidate trees depend on the partition. Across 200 random 75/25 partitions, testing set sensitivity for the case-finding configuration ranged from 0.693 to 0.948 and specificity from 0.573 to 0.736; for the referral-minimisation configuration, specificity ranged from 0.119 to 0.922 (Supplementary Table S1, Figure S1). The number of candidate trees generated varied between six and eight. Candidate tree indices are therefore not stable identifiers, and all results below use the locked design described in the Methods section. Under the locked design, apparent, cross-validated, and hold-out performance agreed closely for the case-finding and balanced screening configurations, while the referral-minimisation configuration showed a larger decline on the hold-out set (Table 1).

Table 1. Apparent, cross-validated and hold-out balanced accuracy.

Model	Apparent (Development)	Cross-Validated	Hold-Out
Case-finding FFDT	0.749	0.749	0.751
Balanced screening FFDT	0.747	0.745	0.752
Referral-minimisation FFDT	0.644	0.644	0.603

Apparent performance is measured on the development set used to fit the model; cross-validated performance is the mean across five repeats of 10-fold cross-validation within the development set; hold-out performance is the single evaluation on the locked hold-out set. Agreement was close for the case-finding and balanced screening configurations; the referral-minimisation configuration showed a larger decline on the hold-out set.

3.3. High-Sensitivity FFDT (Case-Finding Model)

Decision rules were generated directly from the frozen model objects. Complete machine-generated rules and node definitions for all configurations are provided in Supplementary Tables S5 and S6.

The case-finding configuration returned a tree with three non-discriminating nodes: two test membership conditions that cannot be satisfied and one that is always satisfied. Its behaviour therefore reduces to the age cue alone. On the locked hold-out set, it produced a confusion matrix identical to the age-only rule (TP 125, FP 1210, TN 1739, FN 12). We report it on that basis rather than as a four-cue model.

3.4. Balanced FFDT (Screening; Primary Model)

The balanced screening FFDT was retained as the primary model. Its rule was generated from the frozen model object: respondents aged ≥ 40 years are classified as “Diabetes”; among younger respondents, those outside the richer or richest wealth quintiles are classified as “No Diabetes”; among the remainder, those with primary or no education are

classified as “Diabetes”, and all others are classified as “No Diabetes” (Figure 2). On the locked hold-out set, the model achieved sensitivity 0.920 (95% CI 0.862–0.959), specificity 0.585 (0.567–0.602), and balanced accuracy 0.752 (0.727–0.776). It referred 43.7% of respondents, with a positive predictive value of 0.093 and a number needed to test of 10.7. The pre-specified age-only rule achieved balanced accuracy 0.751 (0.724–0.775), while logistic regression using the same five predictors achieved 0.766 (0.735–0.797) at the Youden threshold (Table 2; Figure 3). The balanced screening configuration was used for the province-level and external validation analyses.

Hold-out set $n = 3086$; 137 self-reported diabetes cases (4.44%). Values are estimates (95% confidence intervals). TP, true positives; FP, false positives; TN, true negatives; FN, false negatives. The case-finding FFDT and the age-only rule produce identical classifications for every individual.

Panel (b) uses the same models and hold-out set as Panel (a). PPV, positive predictive value; NPV, negative predictive value; LR+, positive likelihood ratio; LR–, negative likelihood ratio; NNT, number needed to test. Low predictive values reflect the 4.44% prevalence of self-reported diabetes in this sample.

Differences against the age-only benchmark were small (Table 3). Across 50 independent development/hold-out partitions, the case-finding FFDT was identical to the age-only rule in every partition (median difference 0.0000), while the balanced screening FFDT showed only a negligible median improvement of 0.0012 (95% range –0.0039 to 0.0250), favouring the tree in 33 of 50 partitions. The referral-minimisation FFDT was consistently inferior, with a median difference of –0.1242 (95% range –0.1719 to –0.0799) across all 50 partitions (Figure 4; Supplementary Table S7). Age-only balanced accuracy was stable across the partitions (median 0.751, range 0.733–0.771).

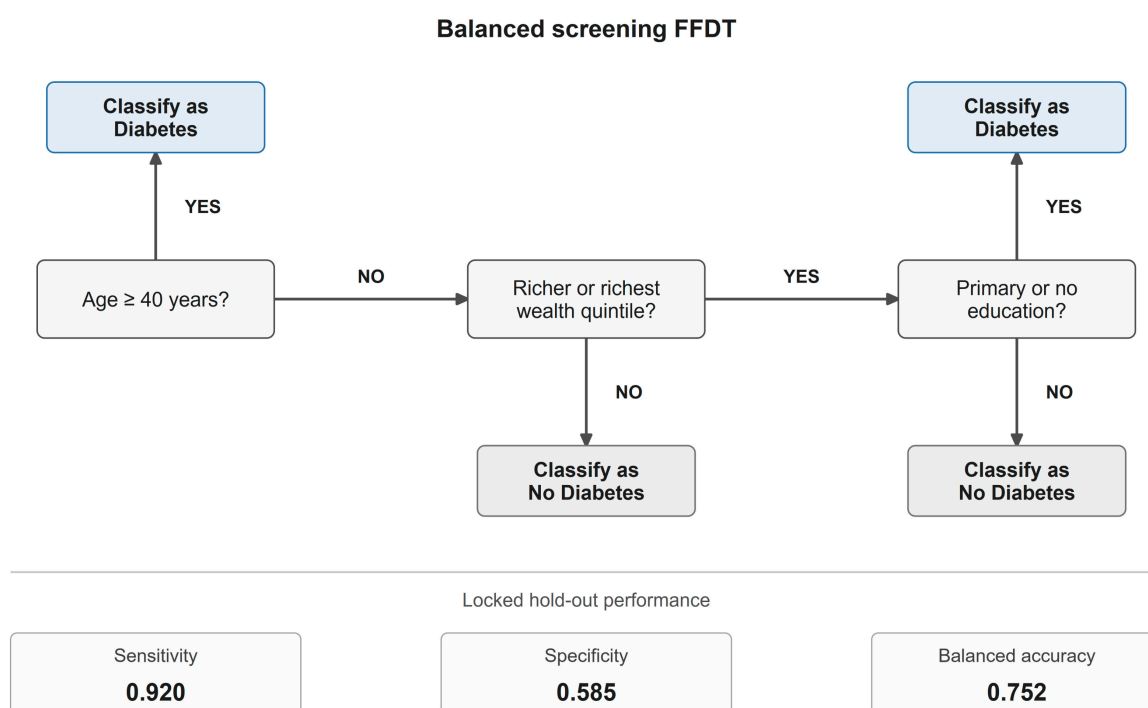


Figure 2. Balanced screening fast-and-frugal decision tree (FFDT) for classification of self-reported, previously diagnosed diabetes status. The tree was regenerated from the frozen model object. Respondents aged ≥ 40 years are classified as “Diabetes”. Among respondents aged < 40 years, those outside the richer or richest wealth quintiles are classified as “No Diabetes”; the remainder are classified as “Diabetes” if they have primary or no education and as “No Diabetes” otherwise. Hold-out performance was sensitivity 0.920, specificity 0.585, and balanced accuracy 0.752.

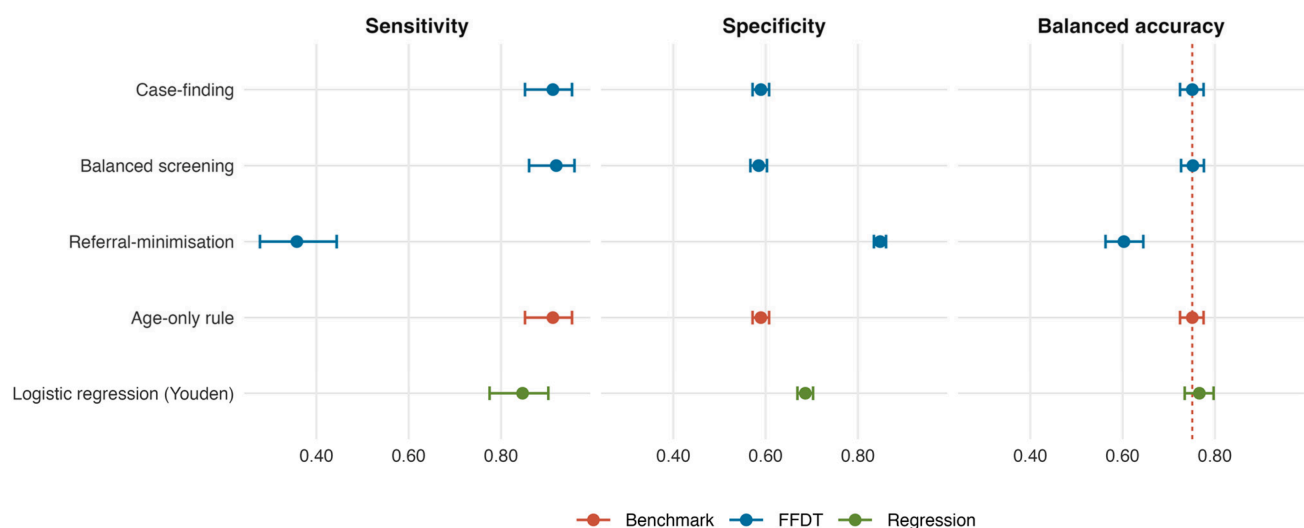
Table 2. Hold-out performance of all models and benchmarks.

(a) Classification performance.							
Model	TP	FP	TN	FN	Sensitivity	Specificity	Balanced Accuracy
Case-finding FFDT	125	1210	1739	12	0.91 (0.85–0.95)	0.59 (0.57–0.61)	0.75 (0.72–0.78)
Balanced screening FFDT	126	1224	1725	11	0.92 (0.86–0.96)	0.58 (0.57–0.60)	0.75 (0.73–0.78)
Referral-minimisation FFDT	49	448	2501	88	0.36 (0.28–0.44)	0.85 (0.83–0.86)	0.60 (0.56–0.64)
Age-only rule	125	1210	1739	12	0.91 (0.85–0.95)	0.59 (0.57–0.61)	0.75 (0.72–0.78)
Logistic regression (Youden threshold)	116	926	2023	21	0.85 (0.78–0.90)	0.69 (0.67–0.70)	0.77 (0.73–0.80)

(b) Predictive values, likelihood ratios, and workload.							
Model	PPV	NPV	LR–	LR+	Referral Rate	NNT	
Case-finding FFDT	0.09 (0.08–0.11)	0.99 (0.99–1.00)	0.15 (0.09–0.26)	2.22 (2.08–2.38)	0.43 (0.42–0.45)	10.7 (9.0–12.7)	
Balanced screening FFDT	0.09 (0.08–0.11)	0.99 (0.99–1.00)	0.14 (0.08–0.24)	2.22 (2.08–2.37)	0.44 (0.42–0.46)	10.7 (9.1–12.8)	
Referral-minimisation FFDT	0.10 (0.07–0.13)	0.97 (0.96–0.97)	0.76 (0.67–0.86)	2.35 (1.85–2.99)	0.16 (0.15–0.17)	10.1 (7.8–13.5)	
Age-only rule	0.09 (0.08–0.11)	0.99 (0.99–1.00)	0.15 (0.09–0.26)	2.22 (2.08–2.38)	0.43 (0.42–0.45)	10.7 (9.0–12.7)	
Logistic regression (Youden threshold)	0.11 (0.09–0.13)	0.99 (0.98–0.99)	0.22 (0.15–0.33)	2.70 (2.47–2.95)	0.34 (0.32–0.35)	9.0 (7.6–10.8)	

Hold-out performance against benchmark rules

n = 3086; 137 self-reported cases (4.44%). Dashed line: age-only balanced accuracy



Bars are 95% confidence intervals (exact binomial; bootstrap for balanced accuracy).

Figure 3. Hold-out performance of the three fast-and-frugal decision tree configurations, the age-only rule, and logistic regression at the development-derived Youden threshold. Sensitivity, specificity, and balanced accuracy are shown with 95% confidence intervals; the dashed line marks the balanced accuracy of the age-only rule.**Table 3.** Difference in balanced accuracy versus the age-only rule.

Model	Difference in Balanced Accuracy	95% CI
Case-finding FFDT	0.0000	0.0000 to 0.0000
Balanced screening FFDT	0.0013	−0.0034 to 0.0105
Referral-minimisation FFDT	−0.1482	−0.1886 to −0.1043
Logistic regression (Youden threshold)	0.0153	−0.0072 to 0.0354

Note: Paired bootstrap, 2000 resamples, on the locked hold-out set.

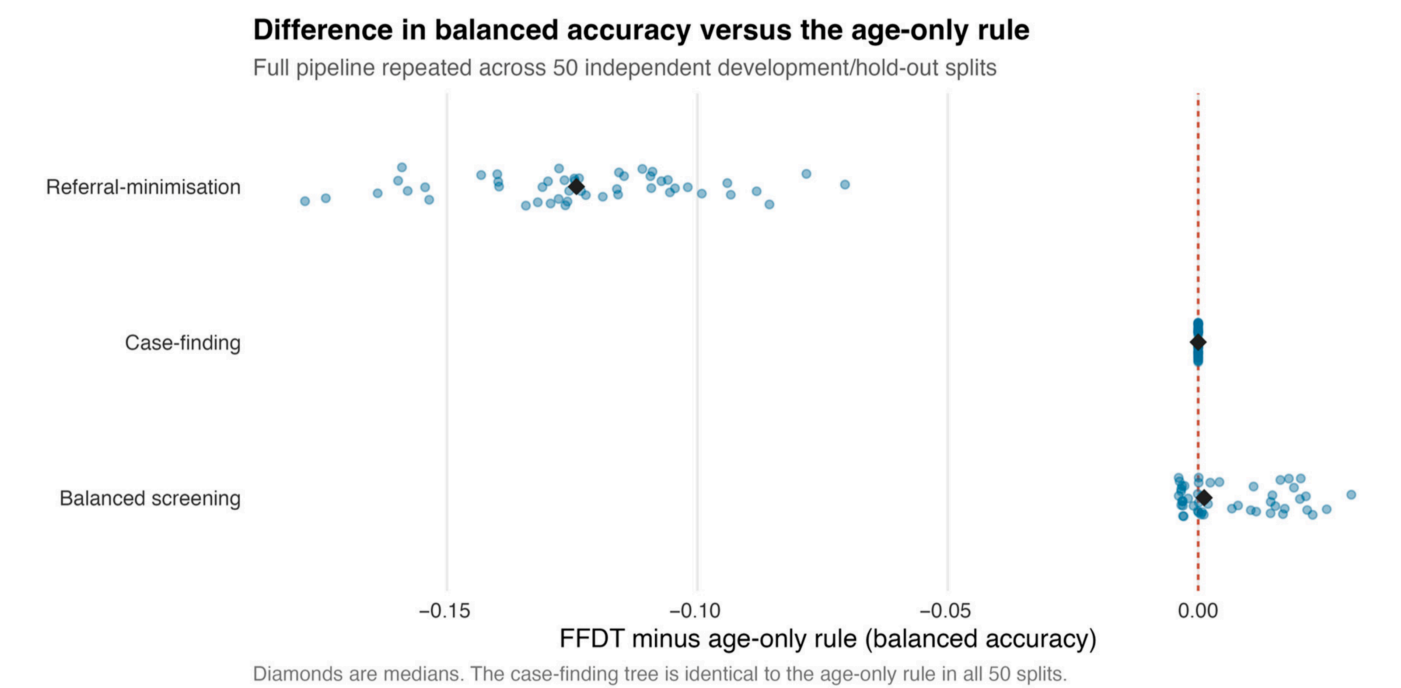


Figure 4. Difference in balanced accuracy between each fast-and-frugal decision tree configuration and the age-only rule across 50 independent development/hold-out splits. The diamonds indicate median differences. The case-finding configuration was identical to the age-only rule in all 50 splits.

Adding a measured body mass index category as a sixth cue changed balanced accuracy by +0.003 (95% CI −0.009 to +0.014) relative to the age-only rule in the subsample with anthropometry. Adding four community-level cues produced a tree numerically identical to the five-cue model. Survey-weighted estimates were close to unweighted estimates (Supplementary Table S2).

Under leave-one-province-out internal–external cross-validation, performance was consistent across all nine provinces: balanced accuracy ranged from 0.731 (Limpopo) to 0.801 (KwaZulu-Natal), the median was 0.740, and sensitivity was from 0.814 to 0.939 (Table 4, Figure S2). Positive predictive value varied from 0.049 to 0.158, tracking provincial prevalence.

Table 4. Internal–external cross-validation, leave-one-province-out.

Province	<i>n</i>	Prevalence	Sensitivity	Specificity	Balanced Accuracy	PPV	Referral Rate
Western Cape	752	0.084	0.937	0.544	0.740	0.158	0.496
Eastern Cape	1345	0.064	0.814	0.658	0.736	0.140	0.372
Northern Cape	881	0.048	0.881	0.591	0.736	0.097	0.431
Free State	1029	0.049	0.880	0.634	0.757	0.109	0.391
Kwazulu-Natal	1553	0.050	0.885	0.718	0.801	0.142	0.312
North West	1085	0.024	0.923	0.556	0.740	0.049	0.455
Gauteng	1025	0.032	0.939	0.606	0.773	0.073	0.412
Mpumalanga	1213	0.035	0.905	0.636	0.770	0.082	0.383
Limpopo	1409	0.028	0.897	0.565	0.731	0.055	0.448

Note: The balanced screening configuration was redeveloped on eight provinces and validated on the ninth, for each province in turn.

When the frozen balanced screening FFDT was applied unchanged to 12 further DHS, performance degraded substantially, and the error profile inverted (Table 5, Figure 5). Balanced accuracy fell from 0.752 in the South African hold-out set to between 0.508 (Timor-Leste) and 0.724 (Albania). Sensitivity, 0.920 in South Africa, ranged from 0.437 to 0.964,

while the referral rate ranged from 0.225 (Tajikistan) to 0.806 (Haiti), a more than 3-fold difference in workload generated by the same fixed algorithm.

Table 5. External validation in 12 further DHS.

Country	<i>n</i>	Prevalence	Sensitivity	Specificity	Balanced Accuracy	PPV	Referral Rate
Albania	21,142	0.016	0.964	0.485	0.724	0.029	0.522
The Gambia	10,806	0.009	0.674	0.731	0.702	0.022	0.272
India	824,838	0.017	0.601	0.742	0.672	0.039	0.264
Lesotho	8473	0.010	0.609	0.730	0.670	0.023	0.273
Kenya	31,354	0.009	0.603	0.731	0.667	0.019	0.272
Zambia	13,683	0.004	0.545	0.762	0.654	0.009	0.239
Tajikistan	10,718	0.025	0.437	0.781	0.609	0.049	0.225
Namibia	3473	0.043	0.906	0.244	0.575	0.051	0.762
Philippines	1308	0.042	0.491	0.653	0.572	0.058	0.353
Haiti	2537	0.028	0.915	0.197	0.556	0.032	0.806
Benin	11,363	0.011	0.527	0.548	0.538	0.013	0.452
Timor-Leste	4264	0.023	0.615	0.402	0.508	0.023	0.598

Note: Frozen South African balanced screening FFDT applied without refitting or threshold adjustment. Outcome is self-reported diabetes in every survey. South African hold-out reference: balanced accuracy 0.752, referral rate 0.437.

External validation: frozen South African trees applied unchanged

Balanced screening FFDT, 12 further DHS surveys. Dashed line: South African hold-out value

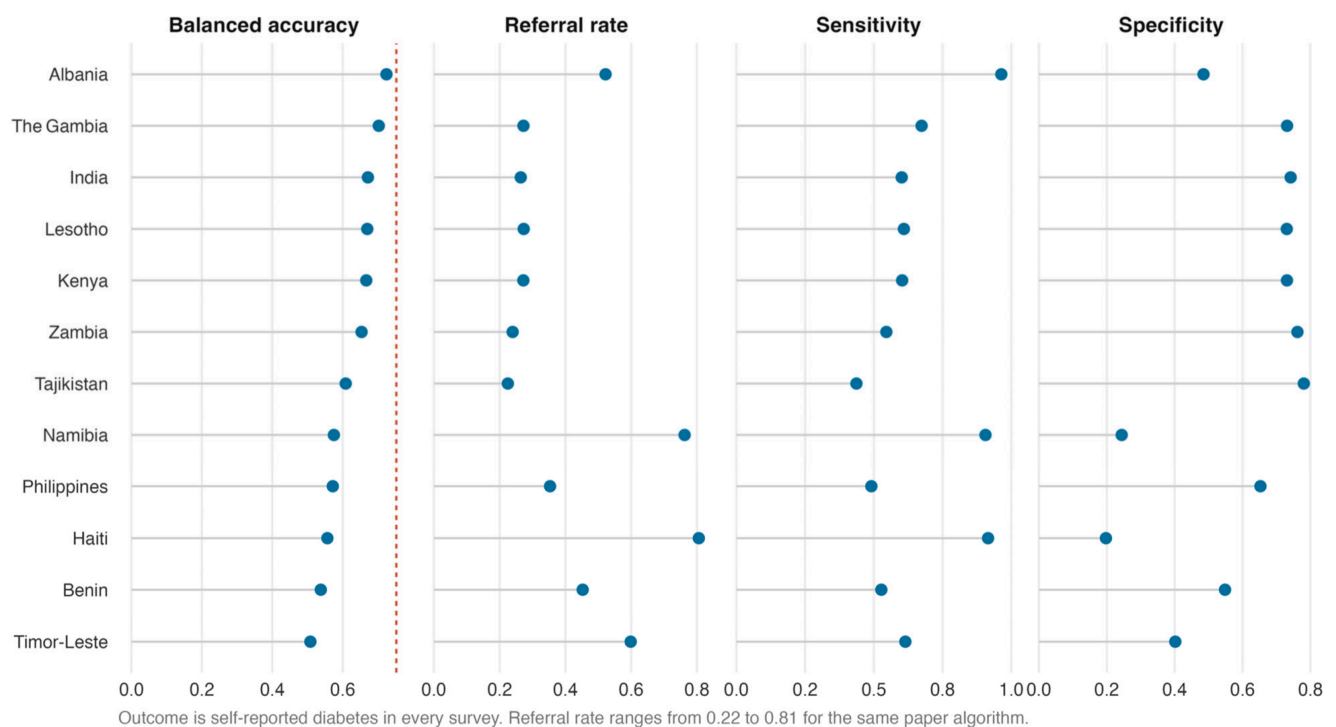


Figure 5. External validation of the frozen balanced screening tree across 12 further DHS, ordered by balanced accuracy. The dashed line marks the South African hold-out value.

3.5. High-Specificity FFDT (Referral-Minimisation Model)

The referral-minimisation FFDT was retained to represent a high-specificity operating profile rather than a rule-out model. Its rule was generated from the frozen model object: respondents with no education are classified as “Diabetes”; among the remaining respondents, those not in the richest wealth group are classified as “No Diabetes”; among the remaining respondents, those who are middle-aged or older are classified as “Diabetes”, and all others are classified as “No Diabetes”. On the locked hold-out set, this configuration

achieved sensitivity 0.358, specificity 0.848, and BAcc 0.603. Because it missed 88 of 137 self-reported cases and showed the largest between-group performance gaps, it is reported as a negative finding and is not recommended for use.

Error scaled weakly with training sample size. Fitting $\ln(\text{error}) = \alpha + \beta \times \ln(n)$ across development subsamples from $n = 360$ to $n = 7206$ gave $\beta = -0.044$ (95% CI -0.063 to -0.025) and $R^2 = 0.78$ (Figure 6, Supplementary Table S8). A 20-fold increase in training data reduced error by approximately 3.5 percentage points, and the curve was flat beyond approximately $n = 1500$.

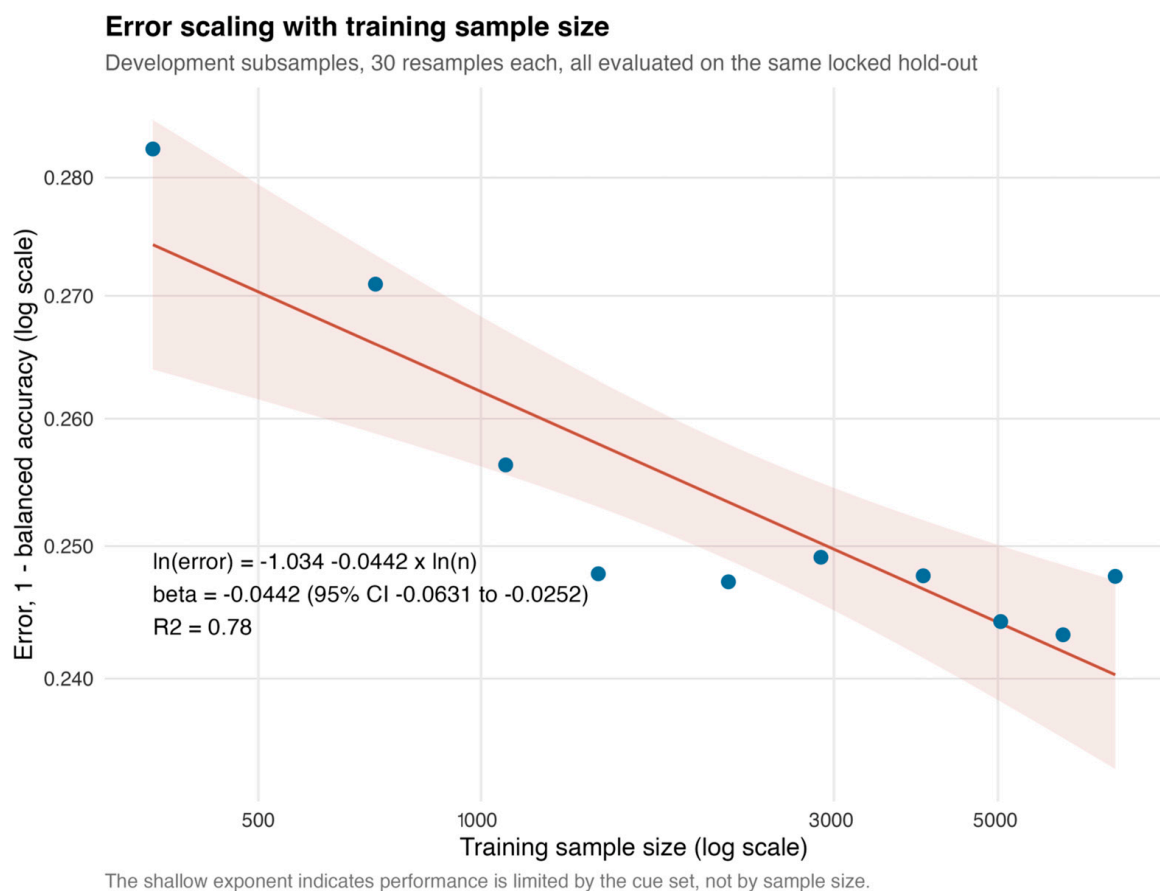


Figure 6. Error scaling with training sample size on log–log axes, with fitted line and 95% confidence band.

Performance differences by sex were small for all configurations (sensitivity gap 0.10 or less; Supplementary Tables S3 and S4). Differences by age category were large, particularly for the case-finding and balanced screening configurations, in which age is the first decision cue. The referral-minimisation configuration showed the largest disparities, with a sensitivity gap of 0.939 across education strata and 0.830 across wealth quintiles, indicating that it systematically fails to identify cases in specific socioeconomic groups.

3.6. Comparative Performance and Error Profiles

Table 2 summarises hold-out performance for the three FFDT configurations and benchmark models. The case-finding FFDT and the age-only rule produced identical classifications. The balanced screening FFDT showed nearly identical balanced accuracy to the age-only rule, while logistic regression showed only a small and non-significant advantage. The referral-minimisation FFDT reduced referral workload but at the cost of substantially lower sensitivity. These results support the revised interpretation that

the value of FFDTs in this setting is transparency and auditability rather than improved classification accuracy.

4. Discussion

In a nationally representative South African survey, transparent decision trees built from five routinely collected sociodemographic cues classified self-reported diabetes status with a balanced accuracy of approximately 0.75. That figure is, however, indistinguishable from what is achieved by a single question about age and from what is achieved by logistic regression using the same predictors. Across 50 independent development and hold-out partitions, one of our three configurations was arithmetically identical to the age-only rule; one matched it; one was consistently worse. We therefore do not claim that fast-and-frugal decision trees provide superior classification in this setting.

Two further results explain why. First, error scaled with training sample size with an exponent of only -0.044 , so a 20-fold increase in data reduces error by about three percentage points; performance is limited by the information content of the cue set rather than by sample size. Consistent with this, neither four community-level cues nor a measured body mass index category improved on the age-only rule. Second, performance was stable across South African provinces but degraded sharply across countries, and the referral rate generated by the same fixed rule varied more than 3-fold between surveys. A rule whose workload implications change this much between settings is not transportable in any operationally meaningful sense.

Taken together, these findings suggest that non-invasive sociodemographic cues in this context largely encode age together with health system contact, rather than biological risk. This is a substantive constraint on the literature that has repeatedly proposed such cues as a low-cost substitute for laboratory screening in resource-constrained settings.

4.1. The Reference Standard and Why HbA1c Was Not Used

Our outcome is self-reported, previously diagnosed diabetes rather than biomarker-confirmed disease, and this constrains interpretation in a specific direction: individuals with undiagnosed diabetes are labelled as non-cases, so the models are trained partly to identify who has had contact with health services. The socioeconomic gradient we observe should be read with that in mind.

The 2016 SADHS measured HbA1c in a subsample, and we considered using it. We retained self-report for three reasons. First, a single reference standard was required across nine provinces and 12 countries, and HbA1c is available only in the South African survey; mixing standards would have confounded transportability with measurement. Second, HbA1c thresholds may not perform consistently across populations of African descent: a scoping review reported underdiagnosis relative to oral glucose tolerance testing in studies of African participants, and work in Black South African patients documented substantial discordance between HbA1c-derived and directly measured average glucose [35,36]. Assay performance itself appears robust, including in the presence of the sickle cell trait [37], and HbA1c has performed comparably to fasting glucose in South African risk score validation [38,39]. Third, in this particular file the HbA1c distribution was shifted upward (median 6.1%; 88% of adults $\geq 5.7\%$), and the $\geq 6.5\%$ threshold yielded a survey-weighted prevalence of 21.3%, compared with previously published estimates of 11.4% and 14.6% from the same survey [40,41]. Because this discrepancy could not be resolved, validation against a biomarker-confirmed outcome remains an essential next step.

4.2. Interpretation of Main Findings

Age category, household wealth, and educational attainment appeared in the effective FFDT rules, consistent with the well-documented epidemiology of T2DM in sub-Saharan Africa, where risk is shaped by structural deprivation, rapid urbanisation, socioeconomic transition, and individual-level characteristics [42,43]. However, the revised benchmarking analyses showed that the case-finding configuration reduced exactly to an age-only rule, the balanced screening FFDT produced almost the same balanced accuracy as the age-only benchmark, and the referral-minimisation configuration reduced referral workload but missed a large proportion of self-reported cases. These results show that the three FFDT configurations encoded different error profiles, but they do not support a claim of improved classification over simpler alternatives. Their main contribution is therefore transparency: the rules make the operating assumptions and error trade-offs visible, which is useful for audit and implementation planning even when classification performance is limited [27,44,45].

The predominance of age, together with the limited incremental contribution of the remaining sociodemographic cues, is consistent with the ecological-rationality perspective underlying FFDTs, which holds that simple heuristics perform well when they exploit stable structure in the environment. In this setting, age captured most of the available classification signal, plausibly reflecting the strong age-patterning of T2DM risk and the broader cardiometabolic consequences of the nutrition transition in urbanising South African populations [26,27,46]. The resulting rules remained shallow, transparent, and non-compensatory, requiring inspection of only a small number of cues per individual and thereby reducing cognitive and data-collection burden compared with multivariable regression or machine learning models [25,26]. However, this structural simplicity did not translate into meaningful improvement over the age-only benchmark.

4.3. Relationship to Existing Diabetes Risk Tools

These findings contribute to the growing literature on non-invasive diabetes risk assessment in African contexts. A recent systematic review identified few risk tools that were developed and validated specifically for sub-Saharan African populations; imported tools from high-income settings frequently underperform [45]. External validation studies in African ancestry cohorts have found acceptable discrimination but poor calibration, with systematic over- or under-estimation of risk that is only partially corrected through recalibration [38,47]. Recent South African evidence further demonstrates that several established non-invasive risk scores fail to identify individuals at very high risk, highlighting the limitations of off-the-shelf algorithms in this setting [48].

FFDTs still offer a complementary strategy in this context, but mainly as transparent benchmarking tools rather than as superior screening instruments. Unlike many existing tools that generate a numerical risk score or require laboratory inputs, the rules described here can be implemented as simple paper algorithms or verbal checklists and explicitly encode the operational context for which each rule is intended. However, the case-finding configuration added nothing beyond age alone, and the referral-minimisation configuration is not recommended for use because of its low sensitivity and subgroup performance gaps. This pattern supports the need for resource-driven prediction models that account explicitly for input cost, availability, and the information content of candidate cues [49].

4.4. Methodological Implications: Heuristics Versus Complex Models

This study demonstrates that interpretable heuristic models need not be simplistic. FFDTs are constructed through a principled search over cue orderings and exit structures, subject to the constraint that each node uses a single binary question and is optimised

against a defined performance target such as BAcc [25,26,46]. Simulation and empirical evidence suggest that FFDTs can match or outperform logistic regression and more complex machine learning models in many real-world settings, particularly when sample sizes are modest and predictor noise is high [26,32,46]. Our results are consistent with this literature: even at a disease prevalence of approximately 4%, a small set of sociodemographic cues arranged in a non-compensatory tree achieved BAcc in the mid-70% range on held-out data.

We acknowledge that FFDTs do not produce individual risk probabilities and cannot be calibrated in the conventional sense. This simplifies communication and implementation (“follow this path to assign a classification”) but shifts the basis for evaluating model adequacy from risk calibration to classification performance and error trade-offs. Most multivariable diabetes risk scores, by contrast, generate estimated probabilities that can be assessed using Brier scores, calibration slopes, and calibration plots [50,51]. Exploring hybrid approaches, for example, using FFDTs to stratify individuals into broad risk bands that feed into calibrated probability estimates from more complex models, or extending FFDT methodology to include basic probability bins, represents an important direction for future work.

4.5. Calibration and the Need for External Validation

This study adhered to TRIPOD + AI reporting guidelines in its documentation of data processing, model construction, and internal testing [22]. Nevertheless, a thorough evaluation of probability calibration and prospective validation against biomarker-confirmed outcomes remains outside the scope of the current analysis and represents an important limitation. Evidence from multiple clinical domains consistently shows that prediction models with good apparent discrimination in development data are frequently miscalibrated when applied to new populations, producing systematic over- or under-estimation of risk in specific subgroups [50–52]. External validations of diabetes risk scores in African and African ancestry cohorts have documented exactly this pattern: reasonable discrimination (AUC 0.70–0.80) but clinically significant calibration drift requiring population-specific recalibration, particularly when HbA1c rather than fasting glucose is used as the reference standard [38,39,47].

Although FFDTs do not generate individual risk probabilities, group-level calibration remains important. At 4.4% prevalence, a tree that classifies a large proportion of the population as positive will generate predominantly false positives, potentially overwhelming confirmatory testing capacity. Conversely, an overly conservative referral-minimisation tree may provide false reassurance to a non-trivial number of undiagnosed individuals. Future work should therefore evaluate the calibration-in-the-large of each FFDT across age, sex, and socioeconomic strata; conduct prospective validation in independent primary care or community datasets with biomarker-confirmed outcomes using an established validation design [51–53]; and assess predictive values, decision-curve net benefit, and workload implications under realistic screening pathways.

4.6. Strengths and Limitations

This study has several strengths. First, it draws on a large population-based dataset from South Africa, the country at the epicentre of the projected diabetes burden increase in sub-Saharan Africa [42,43]. Second, exclusive use of non-invasive, routinely available sociodemographic indicators maximises potential scalability to primary care and community screening settings where biochemical testing is unavailable. Third, rather than seeking a single optimal model, we explicitly evaluated three pre-specified FFDT operating profiles to compare the consequences of different sensitivity–specificity trade-offs. Fourth, performance was reported on a held-out evaluation partition to reduce optimistic bias.

Fifth, explicit specification of cue thresholds and tree architectures supports reproducibility and aligns with TRIPOD + AI requirements for model transparency and transportability [22,25,51].

Several limitations should be noted. The outcome is self-reported previous diagnosis, not biomarker-confirmed diabetes; given that a substantial proportion of diabetes in South Africa is undiagnosed, individuals with undetected disease are misclassified as non-cases, and our estimates should not be interpreted as diagnostic accuracy or as detection of undiagnosed disease. The design is cross-sectional and the cues are not to be interpreted causally. Established non-invasive risk scores could not be computed because the required anthropometric and family history variables were not all available for the full Adult Health analytic sample, so our benchmarking is against an age-only rule and logistic regression rather than against published instruments. External validation used the same self-report outcome in every country, so cross-country degradation reflects transportability of the rule structure under a self-report label rather than of diagnostic performance. DHS does not publish a combined adult health sampling weight, so pooled male-female weighted estimates carry the usual caveat. Finally, we did not assess calibration in the conventional sense, because fast-and-frugal decision trees return classifications rather than probabilities.

4.7. Transparency, Reproducibility, and TRIPOD + AI

The TRIPOD and TRIPOD + AI frameworks emphasise complete reporting of model development, validation, and performance metrics, and require the availability of code and model specifications to support independent review and practical application [22,51]. This study advances that agenda in three respects. First, FFDTs are inherently transparent: the complete model specification consists of a small number of if-then rules, can be depicted in a single figure, and requires no computational infrastructure to implement. Second, we have documented data preprocessing, partitioning, and performance reporting in alignment with TRIPOD + AI requirements, including explicit treatment of class imbalance, missingness, and intended use cases. Third, by constructing distinct trees for case-finding, screening, and referral minimisation, we define the decision-analytic context in which each model is intended to operate, a dimension frequently absent from the prediction modelling literature [22,52]. Future work should include openly archived analysis code, prospective biomarker-based validation, and further assessment of subgroup performance and implementation impact across socioeconomic and demographic strata.

4.8. Future Directions and Impact

Several future research directions follow naturally from this work. Prospective validation studies in primary care and community settings are needed to assess how well these transparent rules perform when used by non-specialist health workers, and to measure their effect on detection rates, referral patterns, and health system workload. Research on adaptive implementation could examine how local disease prevalence, resource constraints, and confirmatory testing availability should inform the use of age-only and balanced screening rules; the referral-minimisation configuration is retained only as a negative finding and is not recommended for use. Methodologically, investigating the robustness of tree structures across time, regions, and alternative diabetes definitions (e.g., HbA1c-based criteria) will be important, given evidence that diagnostic criteria and biomarker choice significantly affect measured prevalence and model performance in African populations [39,54]. The trees could also serve as front-end triage instruments within calibrated risk-stratification pipelines.

5. Conclusions

Transparent fast-and-frugal decision trees built from five non-invasive sociodemographic cues classified self-reported diabetes status in a South African national survey with a balanced accuracy of approximately 0.75, essentially the same as a single question about age and close to the performance of logistic regression using the same predictors. Performance was stable across provinces but did not transport across countries, and error scaling indicates that additional data would not materially improve it.

Among the candidate cues, age category had the highest individual balanced accuracy. The case-finding configuration reduced to the age-only rule, the balanced model achieved the strongest FFDT performance but was not meaningfully better than age alone, and the referral-minimisation configuration reduced referral workload but missed many self-reported cases. These findings reinforce that model utility is context-dependent: the choice between the age-only and balanced screening rules depends on available diagnostic capacity, public health objectives, and tolerance for false positives, whereas the referral-minimisation configuration is not recommended because its lower referral workload was achieved at the cost of missing many self-reported cases.

The contribution of this model class in such settings therefore lies in transparency and auditability rather than in accuracy: the rules can be written on a card, applied without computation, and checked by anyone. Whether these cues can identify undiagnosed diabetes cannot be established from the present self-reported outcome. Future work should prioritise biologically informative cues, such as anthropometric or point-of-care biochemical measures, and validate them against a biomarker-confirmed reference standard.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/diabetology7080151/s1>, Supplementary Methods S1: Fast-and-frugal decision tree construction; Supplementary Methods S2: TRIPOD + AI reporting checklist; Supplementary Methods S3: PROBAST risk-of-bias self-assessment; Supplementary Methods S4: Derivation of cues from DHS variables; Table S1: Instability of the originally submitted design; Table S2: Survey-weighted hold-out performance; Table S3: Subgroup performance of the balanced screening FFDT; Table S4: Between-group performance gaps; Table S5: Machine-generated decision rules; Table S6: Node definitions of the frozen models; Table S7: Difference in balanced accuracy versus the age-only rule across 50 partitions; Table S8: Error against training sample size; Figure S1: Instability of the originally submitted design; Figure S2: Internal-external cross-validation by province.

Author Contributions: Conceptualisation, R.F. and O.A.U.; Methodology, R.F. and O.A.U.; Software, R.F.; Formal Analysis, R.F.; Investigation, R.F.; Data Curation, R.F.; Writing—Original Draft Preparation, R.F.; Writing—Review and Editing, A.A., H.M.K., I.K.J., M.A., T.M.B. and O.A.U.; Visualisation, R.F.; Supervision, T.M.B. and O.A.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study used anonymised secondary data from the 2016 South Africa Demographic and Health Survey. The DHS programme obtained ethical approvals for the original survey data collection. No new data were collected; no identifiable participant information was accessed; no attempt was made to re-identify individuals. Additional institutional ethics approval was not required for this secondary analysis. All results are reported in aggregate form.

Informed Consent Statement: Informed consent was obtained by the DHS programme for the original survey data collection. This secondary analysis used anonymised, de-identified data.

Data Availability Statement: Individual-level DHS data are available from the DHS programme subject to registration and approval of a data access request. The authors do not have permission to redistribute individual-level DHS data. The analysis code and derived outputs supporting the

findings of this study are available from the corresponding author on reasonable request and are reproducible from a single seeded script.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sun, H.; Saeedi, P.; Karuranga, S.; Pinkepank, M.; Ogurtsova, K.; Duncan, B.B.; Stein, C.; Basit, A.; Chan, J.C.; Mbanya, J.C.; et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res. Clin. Pract.* **2022**, *183*, 109119. [[CrossRef](#)] [[PubMed](#)]
2. Saeedi, P.; Petersohn, I.; Salpea, P.; Malanda, B.; Karuranga, S.; Unwin, N.; Colagiuri, S.; Guariguata, L.; Motala, A.A.; Ogurtsova, K.; et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes Res. Clin. Pract.* **2019**, *157*, 107843. [[CrossRef](#)] [[PubMed](#)]
3. Bommer, C.; Heesemann, E.; Sagalova, V.; Manne-Goehler, J.; Atun, R.; Bärnighausen, T.; Vollmer, S. The global economic burden of diabetes in adults aged 20–79 years: A cost-of-illness study. *Lancet Diabetes Endocrinol.* **2017**, *5*, 423–430. [[CrossRef](#)]
4. Manne-Goehler, J.; Geldsetzer, P.; Agoudavi, K.; Andall-Brereton, G.; Aryal, K.K.; Bicaba, B.W.; Bovet, P.; Brian, G.; Dorobantu, M.; Gathecha, G.; et al. Health system performance for people with diabetes in 28 low-and middle-income countries: A cross-sectional study of nationally representative surveys. *PLoS Med.* **2019**, *16*, e1002751. [[CrossRef](#)] [[PubMed](#)]
5. Atun, R.; Davies, J.I.; Gale, E.A.M.; Bärnighausen, T.; Beran, D.; Kengne, A.P.; Levitt, N.S.; Mangugu, F.W.; Nyirenda, M.J.; Ogle, G.D.; et al. Diabetes in sub-Saharan Africa: From clinical care to health policy. *Lancet Diabetes Endocrinol.* **2017**, *5*, 622–667. [[CrossRef](#)] [[PubMed](#)]
6. Flood, D.; Seiglie, J.A.; Dunn, M.; Tschida, S.; Theilmann, M.; Marcus, M.E.; Brian, G.; Norov, B.; Mayige, M.T.; Gurung, M.S.; et al. The state of diabetes treatment coverage in 55 low-income and middle-income countries: A cross-sectional study of nationally representative, individual-level data in 680 102 adults. *Lancet Healthy Longev.* **2021**, *2*, e340–e351. [[CrossRef](#)] [[PubMed](#)]
7. Lindstrom, J.; Tuomilehto, J. The diabetes risk score: A practical tool to predict type 2 diabetes risk. *Diabetes Care* **2003**, *26*, 725–731. [[CrossRef](#)] [[PubMed](#)]
8. Noble, D.; Mathur, R.; Dent, T.; Meads, C.; Greenhalgh, T. Risk models and scores for type 2 diabetes: Systematic review. *BMJ* **2011**, *343*, d7163. [[CrossRef](#)] [[PubMed](#)]
9. Rawshani, A.; Rawshani, A.; Franzén, S.; Sattar, N.; Eliasson, B.; Svensson, A.M.; Zethelius, B.; Miftaraj, M.; McGuire, D.K.; Rosengren, A.; et al. Risk factors, mortality, and cardiovascular outcomes in patients with type 2 diabetes. *N. Engl. J. Med.* **2018**, *379*, 633–644. [[CrossRef](#)] [[PubMed](#)]
10. Steyerberg, E.W.; Vickers, A.J.; Cook, N.R.; Gerds, T.A.; Gonen, M.; Obuchowski, N.; Pencina, M.J.; Kattan, M.W. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* **2010**, *21*, 128–138. [[CrossRef](#)] [[PubMed](#)]
11. Van Calster, B.; McLernon, D.J.; van Smeden, M.; Wynants, L.; Steyerberg, E.W.; Bossuyt, P.; Collins, G.S.; MacAskill, P.; Moons, K.G.M.; Vickers, A.J. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **2019**, *17*, 230. [[CrossRef](#)] [[PubMed](#)]
12. Pencina, M.J.; D’Agostino, R.B., Sr.; D’Agostino, R.B., Jr.; Vasan, R.S. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat. Med.* **2008**, *27*, 157–172. [[PubMed](#)]
13. Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. Machine learning and data mining methods in diabetes research. *Comput. Struct. Biotechnol. J.* **2017**, *15*, 104–116. [[CrossRef](#)] [[PubMed](#)]
14. Beam, A.L.; Kohane, I.S. Big data and machine learning in health care. *JAMA* **2018**, *319*, 1317–1318. [[CrossRef](#)] [[PubMed](#)]
15. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. [[CrossRef](#)] [[PubMed](#)]
16. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nat. Med.* **2019**, *25*, 24–29. [[CrossRef](#)] [[PubMed](#)]
17. Kelly, C.J.; Karthikesalingam, A.; Suleyman, M.; Corrado, G.; King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **2019**, *17*, 195. [[CrossRef](#)] [[PubMed](#)]
18. Subbaswamy, A.; Saria, S. From development to deployment: Dataset shift, causality, and shift-stable models in health AI. *Biostatistics* **2020**, *21*, 345–352. [[PubMed](#)]
19. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [[CrossRef](#)] [[PubMed](#)]
20. Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G.M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *J. Br. Surg.* **2015**, *102*, 148–158. [[CrossRef](#)] [[PubMed](#)]
21. Moons, K.G.M.; Altman, D.G.; Reitsma, J.B.; Ioannidis, J.P.A.; Macaskill, P.; Steyerberg, E.W.; Vickers, A.J.; Ransohoff, D.F.; Collins, G.S. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and elaboration. *Ann. Intern. Med.* **2015**, *162*, W1–W73. [[CrossRef](#)] [[PubMed](#)]

22. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; Altman, D.G.; et al. TRIPOD+ AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. [[CrossRef](#)] [[PubMed](#)]
23. Wolff, R.F.; Moons, K.G.M.; Riley, R.D.; Whiting, P.F.; Westwood, M.; Collins, G.S.; Reitsma, J.B.; Kleijnen, J.; Mallett, S.; PROBAST Group. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern. Med.* **2019**, *170*, 51–58. [[CrossRef](#)] [[PubMed](#)]
24. Moons, K.G.M.; Wolff, R.F.; Riley, R.D.; Whiting, P.F.; Westwood, M.; Collins, G.S.; Reitsma, J.B.; Kleijnen, J.; Mallett, S. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann. Intern. Med.* **2019**, *170*, W1–W33. [[CrossRef](#)] [[PubMed](#)]
25. Phillips, N.D.; Neth, H.; Woike, J.K.; Gaissmaier, W. FFTrees: A toolbox to create, visualize, and evaluate fast-and-frugal decision trees. *Judgm. Decis. Mak.* **2017**, *12*, 344–368. [[CrossRef](#)]
26. Gigerenzer, G.; Gaissmaier, W. Heuristic decision making. *Annu. Rev. Psychol.* **2011**, *62*, 451–482. [[CrossRef](#)] [[PubMed](#)]
27. Djulbegovic, B.; Hozo, I.; Dale, W. Transforming clinical practice guidelines and clinical pathways into fast-and-frugal decision trees to improve clinical care strategies. *J. Eval. Clin. Pract.* **2018**, *24*, 1247–1254. [[CrossRef](#)] [[PubMed](#)]
28. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
29. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [[CrossRef](#)]
30. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [[CrossRef](#)] [[PubMed](#)]
31. Gigerenzer, G.; Goldstein, D.G. Reasoning the fast and frugal way: Models of bounded rationality. *Psychol. Rev.* **1996**, *103*, 650. [[CrossRef](#)] [[PubMed](#)]
32. Luan, S.; Schooler, L.J.; Gigerenzer, G. A signal-detection analysis of fast-and-frugal decision trees. *Psychol. Rev.* **2011**, *118*, 316. [[CrossRef](#)] [[PubMed](#)]
33. Jaeschke, R.; Guyatt, G.H.; Sackett, D.L.; Hayward, R.S.A.; Cook, D.J.; Gerstein, H.C. Users' guides to the medical literature: III. How to use an article about a diagnostic test B. What are the results and will they help me in caring for my patients? *JAMA* **1994**, *271*, 703–707. [[CrossRef](#)] [[PubMed](#)]
34. Schisterman, E.F.; Perkins, N.J.; Liu, A.; Bondell, H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology* **2005**, *16*, 73–81. [[CrossRef](#)] [[PubMed](#)]
35. Khosla, L.; Bhat, S.; Fullington, L.A.; Horlyck-Romanovsky, M.F. HbA1c performance in African descent populations in the United States with normal glucose tolerance, prediabetes, or diabetes: A scoping review. *Prev. Chronic Dis.* **2021**, *18*, 200365. [[CrossRef](#)] [[PubMed](#)]
36. Nkoana, M.K.; Khine, A.A. Correlation between self-monitored mean blood glucose and average plasma glucose estimated from glycated haemoglobin in patients attending the diabetes clinic at Dr George Mukhari Academic Hospital, Pretoria, South Africa. *S. Afr. Med. J.* **2020**, *110*, 416–421. [[CrossRef](#)] [[PubMed](#)]
37. Balungi, P.A.; Niwaha, A.J.; Nice, R.; Rodgers, L.; Mubiru, N.; Mukasa, R.; Jones, A.; Hattersley, A.; Shields, B.; Nyirenda, M.; et al. Impact of haemoglobin variants on the diagnostic sensitivity of glycated haemoglobin (HbA1c) assay methodologies in sub-Saharan Africa: A laboratory-based method validation study. *Pan Afr. Med. J.* **2024**, *48*, 10. [[CrossRef](#)] [[PubMed](#)]
38. Royce, N.; Cronjé, H.T.; Kengne, A.P.; Kruger, H.S.; Dolman-Macleod, R.C.; Pieters, M. HbA1c comparable to fasting glucose in the external validation of the African Diabetes Risk Score and other established risk prediction models in Black South Africans. *BMC Endocr. Disord.* **2024**, *24*, 213. [[CrossRef](#)] [[PubMed](#)]
39. Hird, T.R.; Pirie, F.J.; Esterhuizen, T.M.; O'Leary, B.; McCarthy, M.I.; Young, E.H.; Sandhu, M.S.; Motala, A.A. Burden of diabetes and first evidence for the utility of HbA1c for diagnosis and detection of diabetes in urban black South Africans: The Durban diabetes study. *PLoS ONE* **2016**, *11*, e0161966. [[CrossRef](#)] [[PubMed](#)]
40. Kamerman, P. Underdiagnosis of hypertension and diabetes mellitus in South Africa. *S. Afr. Med. J.* **2022**, *112*, 53–60. [[CrossRef](#)] [[PubMed](#)]
41. Dawson, A.Z.; Walker, R.J.; Campbell, J.A.; Williams, J.S.; Egede, L.E. Prevalence and sociodemographic correlates of diabetes among adults in Namibia and South Africa. *J. Natl. Med. Assoc.* **2022**, *113*, 636–644. [[CrossRef](#)] [[PubMed](#)]
42. Motala, A.A.; Mbanya, J.C.; Ramaiya, K.; Pirie, F.J.; Ekoru, K. Type 2 diabetes mellitus in sub-Saharan Africa: Challenges and opportunities. *Nat. Rev. Endocrinol.* **2022**, *18*, 219–229. [[CrossRef](#)] [[PubMed](#)]
43. Wang, Q.-Y.; Mensah, E.; Li, Z.-C.; Wang, D.-G.; Zhang, C.-W.; Baako, B.M.; Zha, L.; Kong, X. African regional and national burden of diabetes mellitus and its attributable risk factors from 1990 to 2021: Results from the global burden of disease study 2021. *Front. Endocrinol.* **2025**, *16*, 1643999. [[CrossRef](#)] [[PubMed](#)]
44. Damen, J.A.A.G.; Hooft, L.; Schuit, E.; Debray, T.P.A.; Collins, G.S.; Tzoulaki, I.; Lassale, C.M.; Siontis, G.C.M.; Chiochia, V.; Roberts, C.; et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ* **2016**, *353*, i2416. [[CrossRef](#)] [[PubMed](#)]

45. Ekure, E.; Ovenseri-Ogbomo, G.; Osuagwu, U.L.; Agho, K.E.; Ekpenyong, B.N.; Ogbuehi, K.C.; Ndep, A.O.; Okonji, P.; Mashige, K.P.; Naidoo, K.S. A systematic review of diabetes risk assessment tools in sub-Saharan Africa. *Int. J. Diabetes Dev. Ctries.* **2022**, *42*, 380–393. [[CrossRef](#)]
46. Martignon, L.; Erickson, T.; Viale, R. Transparent, simple and robust fast-and-frugal decision trees and their construction. *Front. Hum. Dyn.* **2022**, *4*, 790033. [[CrossRef](#)]
47. Osei-Yeboah, J.; Kengne, A.-P.; Owusu-Dabo, E.; Schulze, M.B.; Meeks, K.A.; Klipstein-Grobusch, K.; Smeeth, L.; Bahendeka, S.; Beune, E.; van Charante, E.P.M.; et al. Validation of prevalent diabetes risk scores based on non-invasively measured predictors in Ghanaian migrant and non-migrant populations—The RODAM study. *Public Health Pract.* **2023**, *6*, 100453. [[CrossRef](#)] [[PubMed](#)]
48. Wentzel, A.; Patterson, A.C.; Karera, M.G.D.; Waldman, Z.C.; Schenk, B.R.; DuBose, C.W.; Sumner, A.E.; Horlyck-Romanovsky, M.F. Non-invasive type 2 diabetes risk scores do not identify diabetes when the cause is β -cell failure: The Africans in America study. *Front. Public Health* **2022**, *10*, 941086. [[CrossRef](#)] [[PubMed](#)]
49. Gurudas, S.; Nugawela, M.; Prevost, A.T.; Sathish, T.; Mathur, R.; Rafferty, J.M.; Blighe, K.; Rajalakshmi, R.; Mohan, A.R.; Saravanan, J.; et al. Development and validation of resource-driven risk prediction models for incident chronic kidney disease in type 2 diabetes. *Sci. Rep.* **2021**, *11*, 13654. [[CrossRef](#)] [[PubMed](#)]
50. Abbasi, A.; Peelen, L.M.; Corpeleijn, E.; van der Schouw, Y.T.; Stolk, R.P.; Spijkerman, A.M.W.; van der A, D.L.; Moons, K.G.M.; Navis, G.; Bakker, S.J.L.; et al. Prediction models for risk of developing type 2 diabetes: Systematic literature search and independent external validation study. *BMJ* **2012**, *345*, e5900. [[CrossRef](#)] [[PubMed](#)]
51. Collins, G.S.; de Groot, J.A.; Dutton, S.; Omar, O.; Shanyinde, M.; Tajar, A.; Voysey, M.; Wharton, R.; Yu, L.-M.; Moons, K.G.M.; et al. External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Med. Res. Methodol.* **2014**, *14*, 40. [[CrossRef](#)] [[PubMed](#)]
52. Debray, T.P.; Damen, J.A.A.G.; Snell, K.I.E.; Ensor, J.; Hooft, L.; Reitsma, J.B.; Riley, R.D.; Moons, K.G.M. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* **2017**, *356*, i6460. [[CrossRef](#)] [[PubMed](#)]
53. Riley, R.D.; Archer, L.; Snell, K.I.E.; Ensor, J.; Dhiman, P.; Martin, G.P.; Bonnett, L.J.; Collins, G.S. Evaluation of clinical prediction models (part 2): How to undertake an external validation study. *BMJ* **2024**, *384*, e074820. [[CrossRef](#)] [[PubMed](#)]
54. NCD Risk Factor Collaboration (NCD-RisC). Global variation in diabetes diagnosis and prevalence based on fasting glucose and hemoglobin A1c. *Nat. Med.* **2023**, *29*, 2885–2901. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.