



Diabetes Risk Prediction under Severe Class Imbalance: Comparing Firth Logistic Regression and SMOTEN-Based Logistic Regression

Lintang Dwi Laga Pertiwi¹ and A'yunin Sofro^{*2}

¹*Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Negeri Surabaya, Indonesia*

²*Department of Actuarial Science, Faculty of Mathematics and Natural Science, Universitas Negeri Surabaya, Indonesia*

Abstract

Diabetes mellitus is a major noncommunicable disease, and reliable early-detection prediction can support Sustainable Development Goal (SDG) Target 3.4. This study compares standard logistic regression, Firth's Penalized Logistic Regression, and SMOTEN-based logistic regression for diabetes risk prediction under severe class imbalance using Wave 5 of the Indonesia Family Life Survey (IFLS5). The analysis included 115 respondents, comprising 8 diabetes cases (6.96%), with all predictors represented as categorical variables. Model performance was evaluated using 50 repeated stratified 75/25 train-test splits, with SMOTEN applied only to training data to prevent leakage. Standard logistic regression failed to converge across all partitions. Firth's Penalized Logistic Regression identified waist circumference as significantly associated with diabetes status (odds ratio = 10.69, 95% CI: 1.56–73.14) and achieved mean accuracy of 89.17%, sensitivity of 8.00%, specificity of 95.19%, and AUC of 0.692. SMOTEN-based logistic regression achieved accuracy of 86.90%, sensitivity of 12.00%, specificity of 92.44%, and AUC of 0.632. Brier scores were 0.089 and 0.098, respectively, compared with 0.064 for the prevalence benchmark, indicating poorer overall probabilistic prediction performance. These findings illustrate the challenges of rare-event prediction under severe class imbalance and highlight the need for larger samples and further validation before clinical application.

Keywords: Diabetes Risk Prediction; Class Imbalance; Firth Logistic Regression; SDG 3; Disaster Risk Mitigation.

Copyright © 2026 by Authors, Published by CAUCHY Group. This is an open access article under the CC BY-SA License (<https://creativecommons.org/licenses/by-sa/4.0>)

1. Introduction

Diabetes mellitus remains a major noncommunicable-disease burden worldwide, and in Indonesia its prevalence among adults has reached double digits, affecting millions [1, 2], with projections suggesting continued increases toward 2045 [3]. As a catastrophic disease, diabetes represents a non-natural disaster-related health burden that requires risk mitigation through prevention and early detection. Data-driven diabetes risk prediction can support such mitigation efforts by identifying individuals at higher risk and thereby contributing to SDG 3, Target 3.4, which seeks to reduce premature mortality from noncommunicable diseases [4].

^{*}Corresponding author. E-mail: ayuninsofro@unesa.ac.id

Ariyanto et al. [5] modeled diabetes risk in prospective East Java athletes using logistic and probit regression, favoring the logistic model (AIC = 129.85 vs. 130.35, mother's occupation and financial status as significant predictors). Despite 80.53% overall accuracy, the model correctly identified only 9 of 28 diabetic cases (sensitivity $\approx 32\%$) on an imbalanced sample (74.64% non-diabetic), illustrating how high accuracy can mask poor minority-class detection; the identified socio-economic predictors also align with evidence linking childhood socio-economic conditions to later type 2 diabetes risk [6]. As the present study draws on a different population and data source, this earlier result is treated only as a contextual benchmark, not a controlled comparison.

Under severe imbalance, conventional maximum-likelihood logistic regression can be dominated by the majority class and become unstable or separated when events are sparse [7]. Firth's Penalized Logistic Regression addresses this via bias reduction, yielding finite, stable estimates [8] and has recently been applied to sparse, imbalanced binary outcomes [9, 10], directly relevant here given only 8 diabetic cases among 115 respondents, though it can still distort predicted probabilities under severe imbalance [11]. An alternative, data-side strategy is SMOTE, which synthesizes minority-class observations and has been used in several diabetes-risk studies [12–15]; however, standard SMOTE interpolates continuous features and is not directly suited to the fully categorical predictors used here, and resampling can itself bias predicted probabilities [16]. Because Firth's estimation-side correction and SMOTE-type resampling tackle imbalance through fundamentally different mechanisms, contrasting them clarifies the trade-offs between these two general strategies rather than establishing either as superior.

Firth's method, SMOTE, and standard logistic regression are each well established individually. The contribution of this study is an empirical, leakage-free comparison of an estimation-based approach (Firth) against a resampling-based approach (SMOTEN, adapted to purely categorical data) on a real, severely imbalanced dataset (6.96% prevalence, 8 events). The study also demonstrates, through a convergence-failure diagnostic and a repeated-holdout validation design, two pitfalls common in small-sample imbalanced classification: undiagnosed separation and evaluation-data leakage.

Accordingly, this study aims to (1) compare standard, Firth-penalized, and SMOTEN-based logistic regression for diabetes-risk prediction under severe imbalance using leakage-free repeated-holdout validation; (2) identify predictors associated with diabetes risk, distinguishing this descriptive aim from predictive performance; and (3) characterize the resulting sensitivity–specificity–precision trade-offs and their uncertainty given the very small event count. The analysis uses IFLS5's general-population sample, unlike the prospective-athlete population in [5], and includes only 8 diabetic cases among 115 respondents (6.96%). This substantially more severe imbalance means that the findings should be read as illustrative of rare-event prediction challenges rather than as evidence for a validated clinical tool.

2. Methods

This section describes the data and variable construction used for diabetes risk prediction. The statistical models include baseline logistic regression, Firth's Penalized Logistic Regression, and SMOTEN-based logistic regression. Repeated train-test validation and performance measures were used to evaluate predictive performance.

2.1. Data Source and Construction

The data used to build the prediction models were obtained from Wave 5 of the Indonesia Family Life Survey (IFLS5, 2014–2015), a nationally representative longitudinal household survey administered by the RAND Corporation in collaboration with partner institutions in Indonesia [17]. The study variables were constructed from three survey modules: Book K (household roster, for age, gender, and parental employment status), Book US (physical measurements, for height, weight, and waist circumference), and the Dried Blood Spot (DBS) module (for HbA1c values). The modules were merged based on each individual's unique identifier (*pidlink*), and complete

response and anthropometric data were required, resulting in a final analytic sample of 115 respondents.

Within this sample of 115, parental employment data were incomplete for a subset of respondents (father's job available for $n = 67$; mother's job available for $n = 98$), because these variables could only be constructed for respondents whose corresponding parent was listed as a co-resident household member in Book K. Rather than further restricting the analytic sample by excluding these respondents (which would reduce the already very small number of diabetes cases), missingness on father's job and mother's job was retained and handled explicitly through missing-indicator encoding, described in Section 2.2. This missingness pattern is most plausibly related to household structure (i.e., whether a parent co-resides with the respondent) rather than to diabetes status itself, and is discussed further as a limitation in Section 4.

2.2. Variable Definitions

The response variable (Y) to be predicted is diabetes status, categorized as binary based on HbA1c values following the WHO/ADA diagnostic standard ($Y = 1$ if $\text{HbA1c} \geq 6.5\%$; $Y = 0$ otherwise). Nine predictor variables were categorized following the structure of the reference study [5], as shown in Table 1. These definitions were applied consistently throughout the modeling and validation procedures described in the subsequent subsections.

Table 1: Definition of predictor variables

Variable	Definition
Height (X_1)	0 = < 170 cm; 1 = ≥ 170 cm
Weight (X_2)	0 = < 60 kg; 1 = ≥ 60 kg
BMI (X_3)	0 = < 25 kg/m ² ; 1 = ≥ 25 kg/m ²
Waist (X_4)	0 = < 85 cm; 1 = ≥ 85 cm
Age (X_5)	0 = below sample median; 1 = above sample median
Gender (X_6)	0 = female; 1 = male
Father's Job (X_7)	0 = not working; 1 = working; missing-indicator $X_{7,NA} = 1$ if unavailable
Mother's Job (X_8)	0 = not working; 1 = working; missing-indicator $X_{8,NA} = 1$ if unavailable
Finance (X_9)	0 = < Rp3 million/month; 1 = Rp3–6 million/month; 2 = > Rp6 million/month (not observed)

The BMI cutoff of 25 kg/m² follows the WHO overweight threshold. The waist-circumference cutoff of 85 cm was retained from the reference study [5] solely to preserve direct comparability with that study's variable definitions, and not because it represents a validated sex-specific clinical threshold for the present sample: the International Diabetes Federation's ethnicity-specific abdominal-obesity thresholds for Asian populations are 90 cm for men and 80 cm for women, so a single 85 cm cutoff applied to both sexes does not correspond exactly to either of these values. A single, non-sex-specific cutoff was used here because the available data module did not support constructing sex-specific cutoffs without further reducing the number of usable cases; this simplification may attenuate the true waist–diabetes association and is noted as a limitation in Section 4. All continuous anthropometric measures (height, weight, BMI, waist circumference) and age were dichotomized to preserve direct comparability with the reference study's variable definitions; we acknowledge that dichotomization discards within-category information and can itself contribute to sparse-data problems when the resulting binary predictors are highly correlated (Section 3.1). A continuous-predictor sensitivity analysis was not feasible within the current data-construction pipeline, because the processed analytic file used for modeling retains only the categorized indicators rather than the original continuous measurements; this is noted explicitly as a limitation rather than omitted silently, and is a priority for the immediate next iteration of this work.

Variables X_7 and X_8 (father's and mother's job) are proxies for employment status (working/not working), which differ from the formal/informal employment categories used in the reference study due to limitations in the coverage of the data module used. Missing values on X_7

and X_8 (Section 2.1) were encoded using a standard missing-indicator approach. Each variable was set to 0 for missing cases, and a corresponding binary indicator ($X_{7,NA}$, $X_{8,NA}$) was added to the predictor set, allowing the models to use all 115 respondents while flagging that these values were imputed rather than directly observed.

The finance variable (X_9) was defined with three intended categories, but no respondent in the final analytic sample fell into the highest income category (>Rp6 million/month). Consequently, only two effective categories are present in this dataset, and only one contrast (0 vs. 1, i.e., X_{9_1} in the coefficient tables) could be estimated. Because the highest category is empty, the finance variable is effectively binary in this analysis, and its coefficient should not be interpreted as reflecting a dose-response gradient across three income levels.

2.3. Binary Logistic Regression (Baseline Prediction Model)

Binary logistic regression was used as the baseline model to predict the probability of diabetes occurrence. For the i -th individual, let $\pi_i = P(Y_i = 1 \mid \mathbf{x}_i)$ denote the predicted probability that the individual has diabetes. Following the binary logistic regression formulation in the reference study [5], the predicted probability is defined as follows:

$$\pi_i = \frac{\exp\left(\beta_0 + \sum_{j=1}^p \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^p \beta_j x_{ij}\right)}, \quad (1)$$

where p is the number of predictor variables. The logit transformation provides an equivalent representation of the predicted probability in linear form. Corresponding to Eq. (1), the logit transformation is given as follows:

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}. \quad (2)$$

In Eq. (2), β_0 is the intercept, β_j is the regression coefficient of the j -th predictor, and x_{ij} represents the value of the j -th predictor for the i -th individual. The model parameters were estimated using Maximum Likelihood Estimation (MLE), solved numerically using the Newton–Raphson iterative algorithm. This estimation procedure was used for the standard logistic regression model before applying the penalized and resampling-based approaches described below.

2.4. Testing of Overall Model Significance

In addition to the partial (Wald) test of individual predictors, the overall significance of each fitted model was assessed with a simultaneous likelihood-ratio (G) test, following the same framework used in the reference study [5]. This test evaluates whether the fitted predictors jointly provide evidence against the intercept-only model. The null and alternative hypotheses for this test are given in Equations (3)–(4).

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0, \quad (3)$$

$$H_1 : \text{at least one } \beta_j \neq 0, \quad j = 1, 2, \dots, p. \quad (4)$$

The likelihood-ratio test compares the maximized likelihood of the full model with that of the intercept-only model. Its test statistic measures the improvement in model fit obtained by including the predictor variables. The likelihood-ratio test statistic is defined as follows:

$$G = -2[\ell(\beta_0) - \ell(\beta)] = 2[\ell(\beta) - \ell(\beta_0)]. \quad (5)$$

In Eq. (5), $\ell(\beta_0)$ is the maximized log-likelihood of the null (intercept-only) model, whereas $\ell(\beta)$ is the maximized log-likelihood of the full model. Under H_0 in Eq. (3), the statistic G

approximately follows a Chi-square distribution with p degrees of freedom. The null hypothesis is rejected if G exceeds the corresponding critical value or if the associated p -value is below α .

For the baseline model, this test could not be computed because estimation failed to converge, as discussed in Section 3.2. The standard Chi-square reference distribution for the statistic in Eq. (5) is derived under unpenalized likelihood theory. Therefore, for the Firth-penalized model, this simultaneous test is interpreted as an approximate descriptive complement to the penalized-likelihood-based inference discussed in Section 2.5.

2.5. Firth's Penalized Logistic Regression

Firth's Penalized Logistic Regression was employed to reduce the bias of maximum likelihood estimation (MLE) in small or imbalanced datasets, thereby producing finite and more stable parameter estimates even under data separation. The method introduces a bias-reducing adjustment to the score function based on Jeffreys' invariant prior [8]. This approach was therefore used as an estimation-based alternative when standard maximum likelihood logistic regression failed to converge.

The adjusted score function for the j -th parameter incorporates the bias-reducing adjustment described above. It consists of the usual logistic score term and an additional leverage-based correction. The resulting adjusted score function is given as follows:

$$U_j^*(\beta) = \sum_i x_{ij}(y_i - \pi_i) + \frac{1}{2} \sum_i h_i(1 - 2\pi_i)x_{ij}. \quad (6)$$

In Eq. (6), h_i denotes the i -th diagonal element of the hat matrix. The hat matrix is defined in Eq. (7) below. This matrix provides the leverage terms required in the bias-reducing adjustment of the Firth score function.

$$\mathbf{H} = \mathbf{W}^{1/2} \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{1/2}, \quad (7)$$

The hat matrix depends on a diagonal weight matrix that reflects the fitted probabilities. The weight assigned to each observation is determined by its fitted probability of diabetes. The diagonal weight matrix is given in Eq. (8) as follows:

$$\mathbf{W} = \text{diag}\{\pi_i(1 - \pi_i)\}, \quad (8)$$

with π_i denoting the fitted probability of diabetes for the i -th observation. The hat matrix in Eq. (7) therefore depends on the current fitted weights $\mathbf{W}$ from Eq. (8) and is recomputed at each iteration of the Newton–Raphson algorithm described below. This iterative updating allows the bias-reducing adjustment to reflect the current fitted model at each step of the estimation procedure.

Parameter estimates were obtained iteratively via the Newton–Raphson algorithm by solving the adjusted score equations in Eq. (6) until convergence [8]. The significance of each predictor was assessed under $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ using the Wald statistic. This procedure provided the coefficient-level inference reported for the full-sample Firth model.

$$z_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}. \quad (9)$$

In Eq. (9), $\hat{\beta}_j$ denotes the estimated coefficient and $SE(\hat{\beta}_j)$ denotes its standard error. A predictor was considered statistically significant at $\alpha = 0.05$. The odds ratio for the j -th predictor was calculated using the exponential transformation of its estimated coefficient, as follows:

$$OR_j = \exp(\hat{\beta}_j). \quad (10)$$

Thus, Eq. (10) expresses the multiplicative change in the odds of diabetes associated with a one-unit increase in the corresponding predictor, holding the other predictors constant. Approximate

95% confidence intervals were calculated for the odds ratios to quantify uncertainty around the estimated effects. The confidence intervals are given as follows:

$$CI_{95\%} = \exp \left(\hat{\beta}_j \pm 1.96 \times SE(\hat{\beta}_j) \right). \quad (11)$$

The confidence intervals in Eq. (11) are Wald-type intervals and should therefore be interpreted as approximate. Profile penalized-likelihood confidence intervals would provide a more exact alternative, particularly given the very small number of diabetes cases in the present sample. The use of approximate intervals is therefore acknowledged as an additional source of uncertainty in interpreting the estimated odds ratios.

2.6. SMOTEN-Based Logistic Regression for Categorical Predictors

The Synthetic Minority Over-sampling Technique (SMOTE) [18] was originally developed to address class imbalance by generating synthetic minority-class observations through linear interpolation. The method constructs new minority observations using information from existing minority-class cases and their nearest neighbors. Its original formulation is therefore most naturally suited to continuous predictor spaces.

For a minority-class observation x_i and one of its k nearest minority-class neighbors x_{nn} , the synthetic observation is generated by interpolation. The interpolation factor λ determines the position of the synthetic observation between the selected minority-class observation and its neighbor. The resulting synthetic observation is defined as follows:

$$x_{\text{new}} = x_i + \lambda (x_{nn} - x_i), \quad \lambda \in [0, 1]. \quad (12)$$

As shown in Eq. (12), standard SMOTE generates a new observation through continuous interpolation. This approach presumes continuous or mixed-type features and is not directly meaningful for purely categorical or binary predictors, since a linear combination of two binary values need not itself be a valid category. Therefore, an alternative oversampling method specifically designed for categorical features is required for the present analysis.

Because all predictor variables used in this study are categorical (Table 1), standard SMOTE was not used. Instead, oversampling was performed using SMOTEN (SMOTE for Nominal features), a variant designed specifically for datasets in which every feature is categorical. SMOTEN identifies the nearest neighbors of a minority-class observation using a value-difference metric appropriate for nominal data and generates a synthetic observation by selecting the modal category for each feature. This preserves the categorical nature of the predictors.

Oversampling was implemented using the `imbalanced-learn` Python package (v0.14.2), using the `SMOTEN` class. The number of nearest neighbors was set adaptively according to Eq. (13). This adaptive choice ensured that the neighborhood size remained valid across the repeated training partitions.

$$k = \min(5, n_{\text{minority,train}} - 1). \quad (13)$$

Eq. (13) ensures that the number of nearest neighbors remains valid when a training partition contains very few minority-class cases, as few as six cases in the present study. For each repetition, the SMOTEN procedure used a random seed equal to the repetition index, from 0 to 49, to make the oversampling step reproducible across repetitions. This seed specification applies to the SMOTEN operation and does not by itself imply that all random components of the repeated train-test procedure were generated from the same seed.

To avoid data leakage, oversampling was applied exclusively within the training partition of each repeated train-test split described in Section 2.7. It was never applied to the test partition or to the full dataset before splitting. The resulting oversampled training set was then used to fit a standard binary logistic regression model, which was evaluated on the corresponding untouched, non-oversampled test partition.

2.7. Validation Procedure and Evaluation Criteria

Given the very small number of minority-class cases (8 of 115), a single train-test split would allocate as few as 2 positive cases to the test set, making point estimates highly sensitive to which specific cases happen to fall into the test partition. To obtain stable estimates of predictive performance while retaining a genuine held-out test-set design, model evaluation used a repeated stratified train-test split procedure. In each of 50 repetitions, the 115 observations were randomly partitioned into a 75% training set ($n = 86$, including 6 diabetes cases) and a 25% test set ($n = 29$, including 2 diabetes cases), stratified to preserve the diabetes prevalence in both partitions. Within each repetition, the baseline, Firth, and SMOTEN-based logistic models were fitted using only the training partition, including, for the SMOTEN-based model, the oversampling step described in Section 2.6, and evaluated exclusively on the corresponding held-out test partition; no observation used for evaluation in a given repetition was ever used for training in that same repetition.

Firth's Penalized Logistic Regression was also fitted once on the complete sample ($n = 115$) for the purpose of interpreting predictor associations (Section 3.3, Table 2). This full-sample model is explicitly distinct from, and is not used for, the repeated-holdout predictive-performance evaluation reported in Sections 3.3–3.5. This distinction reflects a broader point emphasized throughout this study: identifying a statistically significant predictor, an associational question addressed by the full-sample model, is conceptually distinct from establishing that a fitted model generalizes to unseen data, a predictive question addressed only by the repeated-holdout evaluation. Statistical significance of a coefficient does not by itself imply predictive value, and conversely, a variable can contribute to prediction without reaching conventional significance.

Because the standard 0.5 probability threshold is not necessarily appropriate for classification under severe imbalance, each model was evaluated at two thresholds. The first was the conventional default threshold of 0.5, while the second was an approximately optimal threshold selected separately for each of the 50 repetitions. The latter threshold was determined by maximizing the Youden index, defined in Eq. (14) below.

The Youden index combines sensitivity and specificity into a single threshold-selection criterion. It is used here to identify a threshold that balances the two measures on the training partition. The Youden index is defined as follows:

$$J = \text{sensitivity} + \text{specificity} - 1, \quad (14)$$

As shown in Eq. (14), the Youden index combines sensitivity and specificity into a single criterion. The threshold value that maximizes J on the training partition was selected as the repetition-specific optimal threshold. The threshold was selected using the training partition only and was then applied unmodified to the corresponding held-out test partition, so test-set observations were never used in threshold selection.

Performance metrics, including accuracy, sensitivity, specificity, precision, and the area under the receiver operating characteristic curve (AUC), were computed on each held-out test partition and then summarized across the 50 repetitions by their mean, standard deviation (SD), and 95% percentile interval. The percentile interval was defined by the 2.5th and 97.5th percentiles across repetitions. These intervals were reported to convey the substantial uncertainty inherent in evaluating an 8-event outcome on small test partitions.

In addition to these threshold-dependent metrics, confusion matrices were aggregated by summing true positive, false positive, true negative, and false negative counts across all 50 test partitions. This provided a single, stable, and interpretable summary of classification behavior, as discussed in Section 3.5, alongside the threshold-independent AUC. The aggregated counts were therefore reported as a complementary summary rather than as an additional independent validation sample.

Predictive-probability accuracy, defined as the agreement between predicted probabilities and observed outcome frequencies, was assessed using the Brier score. The Brier score jointly

reflects a model's calibration (whether predicted probabilities match observed frequencies) and its resolution (how sharply the model separates cases from non-cases). The Brier score is defined as follows:

$$BS = \frac{1}{n} \sum_i (\hat{\pi}_i - y_i)^2, \quad (15)$$

The Brier score in Eq. (15) was computed on each held-out test partition and then averaged across repetitions. As a reference point, the Brier score of a trivial model that always predicts the training-set prevalence was also computed. This benchmark provides a simple reference for assessing whether the fitted models improve probabilistic prediction beyond a prevalence-only prediction.

3. Results and Discussion

This section presents the results of diabetes risk prediction under severe class imbalance. The analysis covers predictor distribution, baseline model convergence, and the association and prediction results of Firth's Penalized Logistic Regression. SMOTEN-based results are then compared with Firth's model using threshold-dependent and probabilistic performance measures, followed by a contextual comparison with the reference study.

3.1. Descriptive Statistics and Predictor Redundancy

Of the 115 respondents, 8 (6.96%) were classified as diabetic based on HbA1c, and 107 (93.04%) were not (Figure 1), a substantially more severe imbalance than in the reference study (25.36%) [5]. This distribution establishes the rare-event setting underlying the subsequent model-fitting and validation results. The small number of diabetes cases also limits the precision with which classification performance can be estimated across the repeated test partitions.

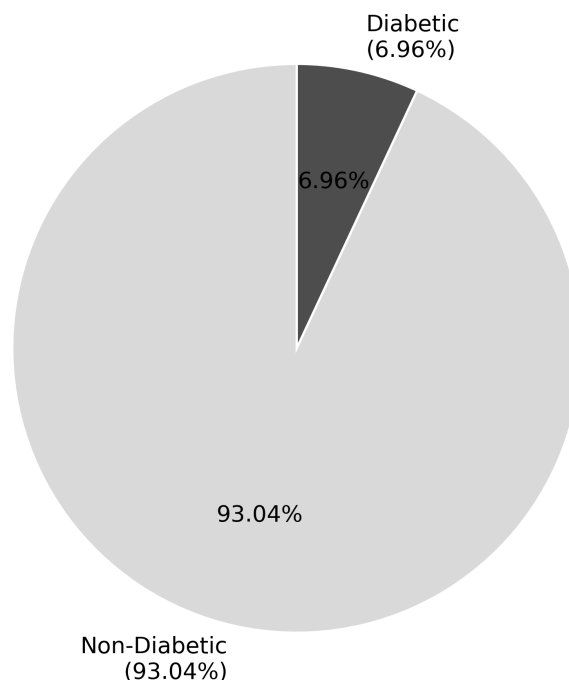


Fig. 1: Proportion of respondents with and without diabetes based on HbA1c.

Figure 2 shows diabetes prevalence within each category of the anthropometric and demographic predictors; consistent with the Firth model results in Section 3.3, the largest contrast is observed for waist circumference (1.3% among respondents with waist < 85 cm vs. 19.4% among

those with waist ≥ 85 cm). The observed contrast provides descriptive motivation for examining waist circumference in the subsequent regression analysis. However, these category-level prevalences do not account for the other predictors included simultaneously in the fitted model.

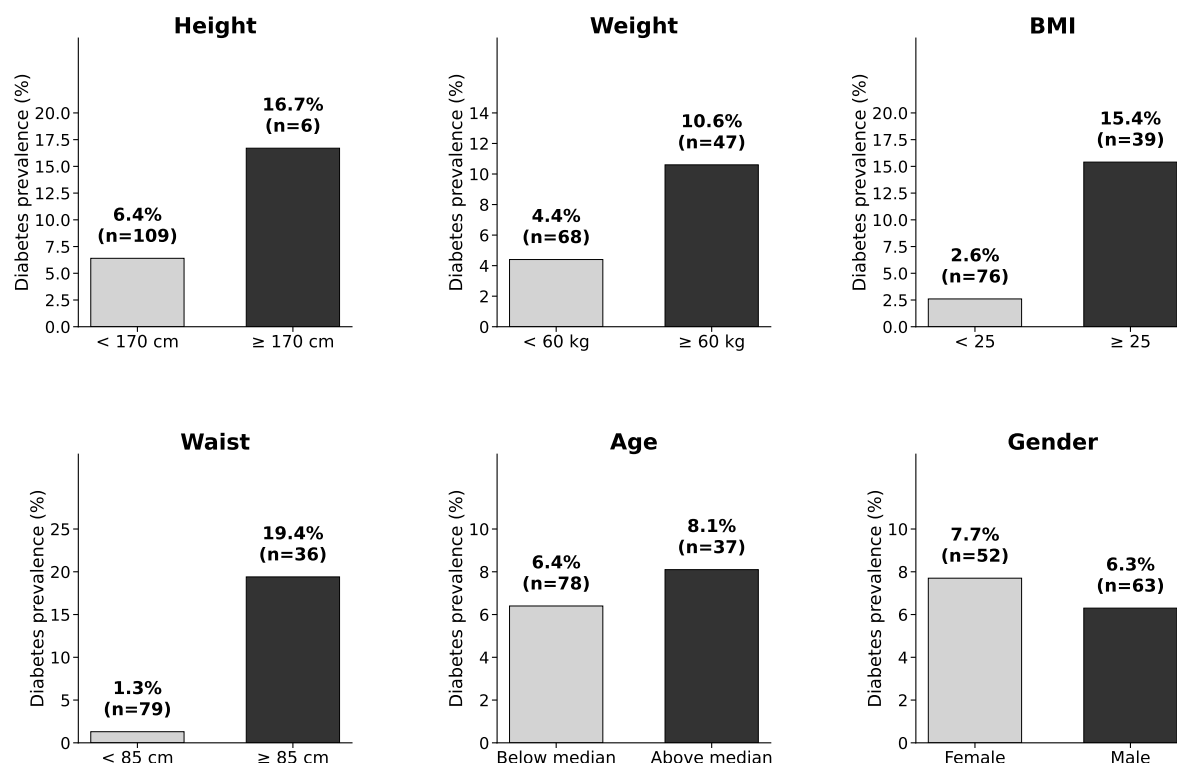


Fig. 2: Diabetes prevalence across categories of anthropometric and demographic predictors. Category boundaries follow Table 1 (e.g., < 170 cm and ≥ 170 cm for height).

Because height, weight, BMI, and waist circumference are conceptually related, and BMI is mathematically derived from height and weight, their simultaneous inclusion raises the possibility of multicollinearity, which can further destabilize estimation when only 8 events are available. Pairwise correlations among these four variables ranged from 0.55 to 0.64 (weight-BMI: $r = 0.64$; BMI-waist: $r = 0.59$; weight-waist: $r = 0.55$; height was weakly correlated with the others, $|r| < 0.05$). Variance inflation factors (VIFs) for all predictors, computed from the full-sample design matrix, were below 3 for all anthropometric variables and for age, gender, and finance; two variables exceeded conventional concern thresholds: father's job (VIF = 5.79) and its missing-indicator (VIF = 5.55), reflecting structural collinearity between a variable and its own missingness indicator. This relationship arises because a respondent's job status and whether it is observed are related whenever missingness is associated with household composition, as noted in Section 2.1. These VIFs indicate that redundancy among the anthropometric predictors is moderate rather than severe, but combined with only 8 available events, even moderate collinearity can meaningfully widen standard errors and contribute to the wide confidence intervals reported in Section 3.3. We retained all four anthropometric variables to preserve comparability with the reference study's variable set, while flagging this as a limitation rather than resolving it by ad hoc predictor removal.

3.2. Baseline Model Results

Estimation of the standard logistic regression model (unpenalized maximum likelihood) failed to converge on the full sample (a singular-matrix error, accompanied by overflow in the exponential function during iteration). To confirm that this failure was not a technical artifact, two additional diagnostic checks were performed: (1) removing predictors with zero variance from the model, and (2) simplifying the model to the nine core predictors without missing-value indicators. Both

diagnostics produced the same convergence failure. The same failure also occurred within the training partition of all 50 repeated train-test splits used for the leakage-free evaluation described in Section 2.7, confirming that this is a structural property of the data rather than a property of the particular full-sample fit.

A low events-per-variable (EPV) ratio does not, by itself, prove separation; a low EPV can also simply produce large standard errors and unstable but finite estimates. What distinguishes this case as genuine (quasi-)complete separation is the specific pattern observed during estimation: the log-likelihood did not plateau but continued to increase toward the iteration limit, at least one coefficient diverged toward an extreme value with a correspondingly inflated standard error, and the fitted probabilities for several observations approached the boundary values of 0 or 1 rather than stabilizing at intermediate values—the standard signature of separation in logistic regression, under which maximum likelihood estimates may fail to exist as finite values [19], rather than of ordinary small-sample imprecision. This pattern persisted consistently across the full-sample fit, both simplified-model diagnostics, and all 50 repeated training partitions.

According to the commonly cited guideline of Peduzzi et al. [20], stable logistic MLE estimation ideally requires approximately 10 events per predictor variable. The EPV in this study (fewer than 1 event per predictor, both in the full sample and within each 86-observation training partition) is far below this guideline and is consistent with, though not sufficient on its own to establish, the separation diagnosed above. This convergence failure is therefore treated as a substantive empirical finding rather than a computational error, motivating the use of Firth's Penalized Logistic Regression and SMOTEN-based logistic regression evaluated below.

3.3. Firth's Penalized Logistic Regression: Association and Prediction Results

For interpretability of predictor associations, Firth's Penalized Logistic Regression was fitted once on the full sample ($n = 115$). Table 2 reports the resulting coefficients, Wald test statistics, and odds ratios (OR) with approximate 95% confidence intervals. This full-sample fit is used only for describing associations and is distinct from the repeated-holdout predictive evaluation described below in Section 2.7.

Table 2: Parameter estimates and odds ratios from Firth's Penalized Logistic Regression (full-sample fit, $n = 115$).

Variable	Coef.	SE	z	p -value	OR	95% CI (OR)
Constant	-5.9694	2.7518	-2.169	0.0301*	0.003	[0.000, 0.562]
Height	1.0181	1.3769	0.739	0.4596	2.77	[0.19, 41.13]
Weight	-1.7246	1.2382	-1.393	0.1637	0.18	[0.02, 2.02]
BMI	1.6016	1.1295	1.418	0.1562	4.96	[0.54, 45.40]
Waist	2.3692	0.9812	2.415	0.0158*	10.69	[1.56, 73.14]
Age	0.1058	0.7625	0.139	0.8896	1.11	[0.25, 4.95]
Gender	0.2586	0.7954	0.325	0.7451	1.30	[0.27, 6.16]
Father's Job	1.4440	2.1577	0.669	0.5034	4.24	[0.06, 290.94]
Father's Job NA ($X_{7,NA}$)	1.6133	2.0952	0.770	0.4413	5.02	[0.08, 304.84]
Mother's Job	0.9351	1.5415	0.607	0.5441	2.55	[0.12, 52.27]
Mother's Job NA ($X_{8,NA}$)	1.1225	1.7340	0.647	0.5174	3.07	[0.10, 91.95]
Finance	0.3125	1.9591	0.160	0.8733	1.37	[0.03, 63.58]

*Significant at $\alpha = 0.05$.

Based on the Wald test, $H_0 : \beta_j = 0$ is rejected only for waist circumference ($z = 2.415$, $p = 0.0158$), with an estimated odds ratio of 10.69 (95% CI: 1.56–73.14); the width of this interval, spanning roughly a 47-fold range, itself illustrates the substantial estimation uncertainty that follows from having only 8 events, and this finding should accordingly be treated as suggestive rather than conclusive. All remaining predictors, including mother's occupation and financial status (significant in the reference study [5]), were not statistically significant here ($p > 0.05$), each with wide confidence intervals reflecting low precision. The direction of the waist-circumference

association is consistent with the medical literature identifying waist circumference as a low-cost indicator of central obesity and a correlate of insulin resistance and metabolic syndrome [21], but given the small sample this study cannot be considered a confirmatory test of that relationship.

Separately, to evaluate genuine out-of-sample predictive performance, the Firth-penalized model was refit within the training partition of each of the 50 repeated stratified train-test splits described in Section 2.7 and evaluated on the corresponding held-out test partition. At the default 0.5 threshold, averaged across repetitions, the Firth-penalized model achieved a mean accuracy of 89.17% (SD 4.35%; 95% percentile interval: 79.31–93.10%), mean sensitivity of 8.00% (SD 18.52%; 95% percentile interval: 0–50.00%), mean specificity of 95.19% (SD 4.81%; 95% percentile interval: 85.19–100%), mean precision of 6.83%, and a mean AUC of 0.692 (SD 0.177; 95% percentile interval: 0.365–0.922). The aggregated confusion matrix across all 50 test partitions (1,450 held-out predictions in total) was: 1,285 true negatives, 65 false positives, 92 false negatives, and 8 true positives. At the Youden-optimal threshold (selected on training data only; mean optimal threshold = 0.231), mean sensitivity rose to 26.00% at the cost of mean specificity falling to 81.78% (aggregated confusion matrix: 1,104 true negatives, 246 false positives, 74 false negatives, and 26 true positives); AUC is unchanged, since it is threshold-independent. These leakage-free estimates are both considerably lower and far more variable than a single in-sample fit would suggest, reflecting the genuine difficulty of detecting an average of 2 positive cases per 29-observation test partition using a model trained on as few as 6 positive cases.

3.4. SMOTEN-Based Logistic Regression: Prediction Results

Within each of the 50 repeated training partitions, SMOTEN increased the minority-class proportion from approximately 7% to 50% using the categorical-appropriate oversampling procedure described in Section 2.6. The corresponding held-out test partition retained the original class distribution and was never modified. Thus, the reported SMOTEN performance reflects evaluation on untouched observations rather than on oversampled or otherwise resampled test data.

At the default 0.5 threshold, averaged across repetitions, the SMOTEN-based model achieved a mean accuracy of 86.90% (SD 6.19%), mean sensitivity of 12.00% (SD 21.57%), mean specificity of 92.44% (SD 6.90%), mean precision of 8.41%, and a mean AUC of 0.632 (SD 0.139; 95% percentile interval: 0.363–0.813). These results indicate that the model detected relatively few diabetes cases at the conventional threshold despite maintaining high specificity. The aggregated confusion matrix (1,450 held-out predictions) was: 1,248 true negatives, 102 false positives, 88 false negatives, and 12 true positives.

At the Youden-optimal threshold (mean optimal threshold = 0.224), mean sensitivity rose to 32.00% with mean specificity of 79.26%. The aggregated confusion matrix was 1,070 true negatives, 280 false positives, 68 false negatives, and 32 true positives. Thus, the lower threshold increased detection of diabetes cases while also increasing the number of false-positive classifications.

Contrary to the substantially higher sensitivity obtained under in-sample (non-held-out) evaluation in earlier analyses of this dataset, the leakage-free SMOTEN-based sensitivity (12.00% at the 0.5 threshold; 32.00% at the optimal threshold) is much closer to, and at the 0.5 threshold slightly below, that of the Firth-penalized model. This large discrepancy between in-sample and out-of-sample estimates is itself an important finding: it demonstrates concretely how evaluating a resampling-based model on data that were also used (directly or indirectly) for oversampling and fitting can substantially overstate real predictive performance, reinforcing the necessity of the leakage-free repeated-holdout design adopted in this study. This pattern is consistent with prior methodological warnings that class-imbalance corrections such as SMOTE can distort apparent predictive performance when evaluation is not strictly separated from resampling [16], underscoring that reported sensitivity gains from resampling-based methods should be interpreted cautiously unless validated on genuinely held-out data.

3.5. Comparison of the Two Converged Models and Predictive-Probability Accuracy

Table 3 compares Firth and SMOTEN-based logistic regression under the leakage-free repeated-split evaluation, at both the default and Youden-optimal thresholds. The baseline model has no entries because it failed to converge in all 50 training partitions (Section 3.2). This comparison therefore focuses on the two models for which complete held-out predictions were available.

Table 3: Performance comparison (mean [SD]; 50 repeated 75/25 stratified splits).

Model	Threshold	Accuracy	Sensitivity	Specificity	Precision	AUC
Logistic (baseline)	–	Failed to converge*	–	–	–	–
Firth Penalized	0.5	0.892 [0.043]	0.080 [0.185]	0.952 [0.048]	0.068 [0.186]	0.692 [0.177]
Firth Penalized	Optimal	0.779 [0.056]	0.260 [0.272]	0.818 [0.065]	0.087 [0.097]	0.692 [0.177]
SMOTEN + Logistic	0.5	0.869 [0.062]	0.120 [0.216]	0.924 [0.069]	0.084 [0.188]	0.632 [0.139]
SMOTEN + Logistic	Optimal	0.760 [0.074]	0.320 [0.263]	0.793 [0.083]	0.098 [0.087]	0.632 [0.139]

*Convergence failure across every repeated training partition is a substantive finding (Section 3.2), not a computational error.

At the conventional 0.5 threshold, both converged models are highly conservative, correctly classifying the large majority of non-diabetic cases (specificity above 92%) while detecting very few diabetes cases (sensitivity 8–12%). At the Youden-optimal threshold, sensitivity roughly triples for both models (26% for Firth, 32% for SMOTEN) at a substantial cost to specificity (78–82%); precision remains low (under 10%) under either threshold for either model, reflecting the very low prevalence in the test partitions. Firth retains a modestly, though not dramatically, higher mean AUC than SMOTEN (0.692 vs. 0.632; overlapping 95% percentile intervals), indicating a slight, uncertain advantage in overall ranking ability. Notably, the difference between the two models at the default threshold is much smaller than reported in earlier, leakage-affected evaluations of this dataset, underscoring that the previously observed “SMOTE advantage” in sensitivity was, to a substantial extent, an artifact of evaluating the model on data involved in its own training.

Overall predictive-probability accuracy was assessed via the Brier score: Firth achieved a mean Brier score of 0.089 (SD 0.027) and SMOTEN 0.098 (SD 0.028), both computed on held-out test partitions. As a reference, a trivial model that always predicts the training-set prevalence achieved a Brier score of 0.064. Both fitted models therefore had higher (worse) Brier scores than this trivial reference, indicating poorer overall probabilistic prediction performance than a model that ignores the predictors entirely. Because the Brier score reflects both calibration and resolution jointly [22], this result alone does not establish that either model is miscalibrated: it is also consistent with adequate calibration combined with weak resolution, an interpretation plausible here given the very low sensitivity reported above. Disentangling these two components – for instance via a calibration plot or a formal calibration test – was not feasible with only 8 events in the full sample and 2 events per test partition, and is left as a direction for future work with a larger dataset. This finding is nonetheless reported directly, without minimization, as it is central to an honest assessment of what this dataset can and cannot support.

Figures 3 and 4 provide graphical summaries of the predictive-performance results. Figure 3 compares the mean threshold-dependent metrics, whereas Figure 4 presents the mean ROC curves. Together, these figures provide a visual complement to the numerical results reported in Table 3.

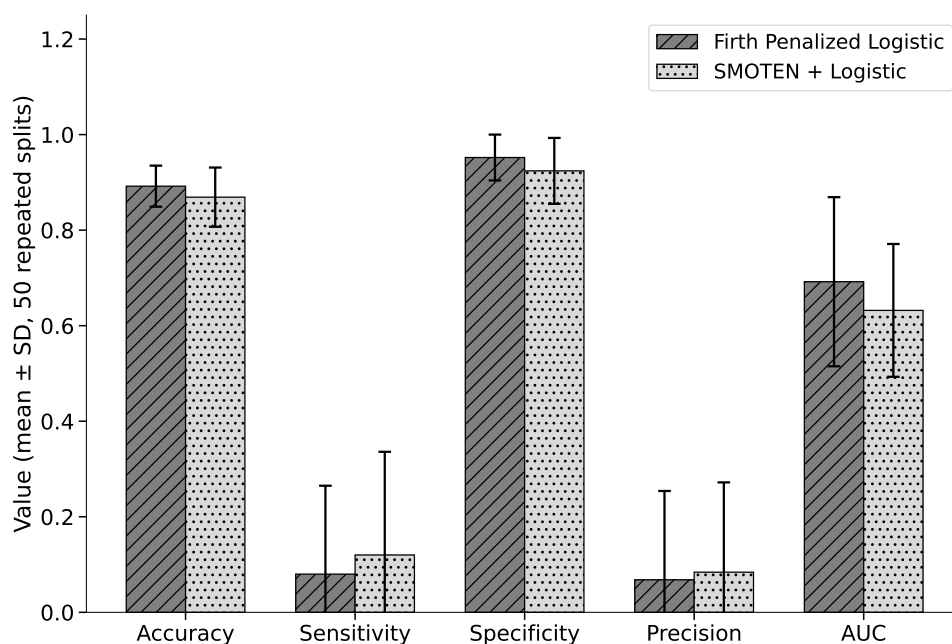


Fig. 3: Comparison of mean predictive performance metrics (accuracy, sensitivity, specificity, precision, and AUC) between Firth and SMOTEN-based logistic regression at the default classification threshold (0.5), averaged across the 50 repeated held-out test partitions.

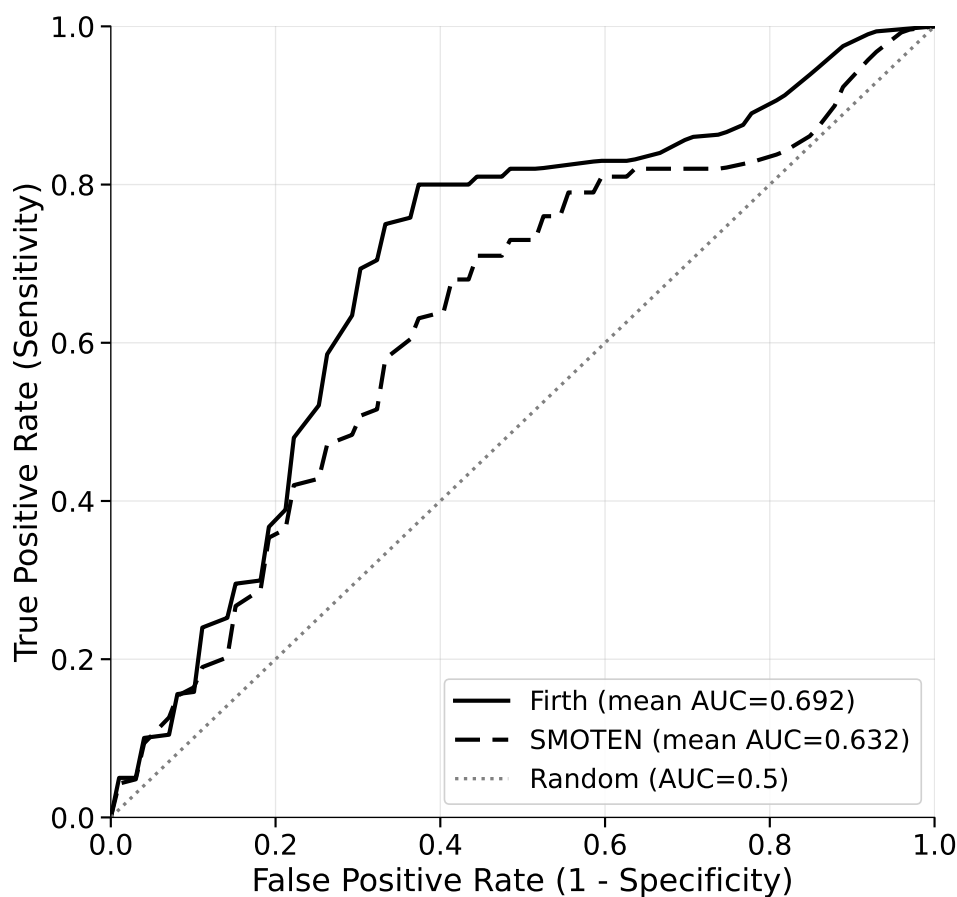


Fig. 4: Mean ROC curves of Firth and SMOTEN-based logistic regression, averaged across the 50 repeated held-out test partitions.

3.6. Contextual Comparison with the Reference Study

Ariyanto et al. [5] reported a standard logistic model, evaluated in-sample on a population of prospective athletes, with 80.53% accuracy and approximately 32% sensitivity. The present study's leakage-free, repeated-holdout evaluation on IFLS5 data yields a mean AUC of 0.692 for Firth and 0.632 for SMOTEN, with sensitivity at the Youden-optimal threshold of 26–32%. These sensitivity values are broadly similar in magnitude to the reference study's in-sample sensitivity, although the evaluation designs differ substantially.

Because the two studies differ in population (prospective athletes vs. general adult IFLS respondents), data source, sample size, and event count, this comparison should be read as contextual rather than as a controlled test of any specific method. In particular, the reference study's sensitivity was computed in-sample, whereas the present study's figures are held-out estimates. The two sets of results therefore should not be interpreted as measuring identical predictive-performance quantities.

This pattern is broadly consistent with other IFLS-based diabetes research. Tanoey and Becher [1], using earlier IFLS waves, found that HbA1c-based case ascertainment yields substantially higher and more clinically realistic diabetes prevalence than self-report alone. This supports the use of the biomarker-based response variable adopted here, independent of which classification method is subsequently applied.

4. Conclusion

This study compared standard logistic regression, Firth's Penalized Logistic Regression, and SMOTEN-based logistic regression for diabetes-risk prediction on severely imbalanced IFLS5 data ($n = 115$, 8 diabetes cases, 6.96% prevalence), under a leakage-free, repeated stratified train-test split validation design. The standard logistic model failed to converge in every one of 50 repeated training partitions, a finding confirmed through multiple diagnostic checks rather than inferred from a low events-per-variable ratio alone, underscoring the necessity of methods designed for rare-event and severely imbalanced conditions. Under leakage-free evaluation, Firth's Penalized Logistic Regression identified waist circumference as an associated predictor (odds ratio 10.69, 95% CI 1.56–73.14) and achieved a mean AUC of 0.692, while SMOTEN-based logistic regression using an oversampling method appropriate for this study's entirely categorical predictor set achieved a mean AUC of 0.632. Both models showed low sensitivity at the conventional 0.5 threshold and only moderate sensitivity gains at a data-driven optimal threshold, together with Brier scores worse than a trivial constant-prevalence prediction, indicating poorer overall probabilistic prediction performance rather than necessarily poor calibration specifically.

These findings should be interpreted with substantial caution. With only 8 observed events, this study cannot provide a validated clinical prediction tool; the wide confidence intervals throughout (e.g., a 47-fold range for the waist-circumference odds ratio) reflect substantial statistical uncertainty rather than a limitation of the estimation method alone. Additional limitations include the dichotomization of continuous anthropometric and age variables, which discards information and was not evaluated against a continuous-predictor alternative in this iteration of the analysis; moderate collinearity among the anthropometric predictors; missingness in parental employment data handled via missing-indicator encoding rather than a formal imputation model; the use of Wald-based rather than profile-penalized-likelihood confidence intervals for the Firth model; the lack of a formal calibration assessment separating calibration from resolution in the Brier score; and the absence of external validation on an independent sample.

Consequently, we moderate our earlier framing of this comparison as directly informing model choice for screening or clinical use. Instead, we present it as an illustration of the estimation and validation challenges inherent to rare-event, severely imbalanced diabetes-risk prediction, following reporting practices consistent with the TRIPOD+AI guidance for prediction-model

studies [22]. These findings nevertheless highlight the potential role of reliable diabetes-risk prediction in supporting risk mitigation and early detection efforts related to noncommunicable diseases, contributing to SDG 3, particularly Target 3.4. Future research should prioritize larger and, ideally, externally validated samples, more detailed and less collinear predictor sets (including sex-specific anthropometric cutoffs and, where feasible, continuous rather than dichotomized measures), formal missing-data modeling, formal calibration assessment, and profile-likelihood-based inference for the Firth model, before any recommendation regarding a preferred method or clinical threshold can be considered well supported.

CRedit Authorship Contribution Statement

Lintang Dwi Laga Pertiwi: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Investigation, Visualization, Writing – Original Draft Preparation. **A’yunin Sofro:** Conceptualization, Methodology, Validation, Supervision, Writing – Review & Editing.

Declaration of Generative AI and AI-Assisted Technologies

During the preparation of this manuscript, an AI-assisted tool was used for language editing, improving clarity and academic writing, statistical-analysis code drafting, and formatting assistance. The authors reviewed, verified, and take full responsibility for the content of the manuscript, including all reported statistical results. An AI-writing detection report with a score below 25% is provided as required by the journal.

Declaration of Competing Interest

The authors declare no competing interests.

Ethics Approval and Consent to Participate

This study used secondary, de-identified data from Wave 5 of the Indonesia Family Life Survey (IFLS5). Data collection for IFLS5, including informed consent procedures, was conducted under the ethical approval framework described by Strauss et al. [17]; the present secondary analysis involved no direct interaction with participants and no collection of new personal data by the authors.

Funding and Acknowledgments

This research received no external funding.

Data and Code Availability

The IFLS5 dataset used in this study is publicly available through the RAND Corporation, subject to its standard data-use terms. The analysis code used in this study is not publicly available but can be made available from the corresponding author upon reasonable request.

References

- [1] J. Tanoey and H. Becher. “Diabetes prevalence and risk factors of early-onset adult diabetes: results from the Indonesian family life survey”. In: *Global Health Action* 14.1 (2021), p. 2001144. DOI: [10.1080/16549716.2021.2001144](https://doi.org/10.1080/16549716.2021.2001144).
- [2] F. R. Muharram, J. B. Swannjo, R. R. Melbiarta, and S. Martini. “Trends of diabetes and pre-diabetes in Indonesia 2013–2023: a serial analysis of national health surveys”. In: *BMJ Open* 15.9 (2025), e098575. DOI: [10.1136/bmjopen-2024-098575](https://doi.org/10.1136/bmjopen-2024-098575).

- [3] M. Wahidin, A. Achadi, B. Besral, et al. "Projection of diabetes morbidity and mortality till 2045 in Indonesia based on risk factors and NCD prevention and control programs". In: *Scientific Reports* 14 (2024), p. 5424. DOI: [10.1038/s41598-024-54563-2](https://doi.org/10.1038/s41598-024-54563-2).
- [4] United Nations. *Goal 3: Ensure Healthy Lives and Promote Well-Being for All at All Ages*. United Nations Sustainable Development Goals. Accessed August 19, 2026. 2026. <https://sdgs.un.org/goals/goal3>.
- [5] D. Ariyanto, A. Sofro, A. N. Hanifah, J. B. Prihanto, D. A. Maulana, and R. W. Romadhonia. "Logistic and probit regression modeling to predict the opportunities of diabetes in prospective athletes". In: *BAREKENG: Jurnal Ilmu Matematika dan Terapan* 18.3 (2024), pp. 1391–1402. DOI: [10.30598/barekengvol18iss3pp1391-1402](https://doi.org/10.30598/barekengvol18iss3pp1391-1402).
- [6] T. Tamayo, C. Herder, and W. Rathmann. "Impact of early psychosocial factors (childhood socioeconomic factors and adversities) on future risk of type 2 diabetes, metabolic disturbances and obesity: a systematic review". In: *BMC Public Health* 10 (2010), p. 525. DOI: [10.1186/1471-2458-10-525](https://doi.org/10.1186/1471-2458-10-525).
- [7] H. He and E. A. Garcia. "Learning from imbalanced data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284. DOI: [10.1109/TKDE.2008.239](https://doi.org/10.1109/TKDE.2008.239).
- [8] D. Firth. "Bias reduction of maximum likelihood estimates". In: *Biometrika* 80.1 (1993), pp. 27–38. DOI: [10.1093/biomet/80.1.27](https://doi.org/10.1093/biomet/80.1.27).
- [9] S. Uno, H. Noma, and M. Gosho. "Firth-type penalized methods of the modified Poisson and least-squares regression analyses for binary outcomes". In: *Biometrical Journal* 66.7 (2024), e202400004. DOI: [10.1002/bimj.202400004](https://doi.org/10.1002/bimj.202400004).
- [10] S. Suhas, N. Manjunatha, C. N. Kumar, V. Benegal, G. N. Rao, M. Varghese, and G. Gururaj. "Firth's penalized logistic regression: A superior approach for analysis of data from India's National Mental Health Survey, 2016". In: *Indian Journal of Psychiatry* 65.12 (2023), pp. 1208–1213. DOI: [10.4103/indianjpsychiatry.indianjpsychiatry_827_23](https://doi.org/10.4103/indianjpsychiatry.indianjpsychiatry_827_23).
- [11] R. Puhr, G. Heinze, M. Nold, L. Lusa, and A. Geroldinger. "Firth's logistic regression with rare events: accurate effect estimates and predictions?" In: *Statistics in Medicine* 36 (2017), pp. 2302–2317. DOI: [10.1002/sim.7273](https://doi.org/10.1002/sim.7273).
- [12] S. Sadeghi, D. Khalili, A. Ramezankhani, M. A. Mansournia, and M. Parsaeian. "Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods". In: *BMC Medical Informatics and Decision Making* 22.1 (2022), p. 36. DOI: [10.1186/s12911-022-01775-z](https://doi.org/10.1186/s12911-022-01775-z).
- [13] M. Talebi Moghaddam, Y. Jahani, Z. Arefzadeh, A. Dehghan, M. Khaleghi, M. Sharafi, and G. Nikfar. "Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm". In: *BMC Medical Research Methodology* 24 (2024), p. 220. DOI: [10.1186/s12874-024-02341-z](https://doi.org/10.1186/s12874-024-02341-z).
- [14] P. Sampath, G. Elangovan, K. Ravichandran, V. Shanmuganathan, S. Pasupathi, T. Chakrabarti, P. Chakrabarti, and M. Margala. "Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique". In: *Scientific Reports* 14 (2024), p. 28984. DOI: [10.1038/s41598-024-78519-8](https://doi.org/10.1038/s41598-024-78519-8).
- [15] Y. Jang. "Feature-based ensemble modeling for addressing diabetes data imbalance using the SMOTE, RUS, and random forest methods: a prediction study". In: *Ewha Medical Journal* 48.2 (2025), e32. DOI: [10.12771/emj.2025.00353](https://doi.org/10.12771/emj.2025.00353).
- [16] R. van den Goorbergh, M. van Smeden, D. Timmerman, and B. Van Calster. "The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression". In: *Journal of the American Medical Informatics Association* 29.9 (2022), pp. 1525–1534. DOI: [10.1093/jamia/ocac093](https://doi.org/10.1093/jamia/ocac093).

- [17] J. Strauss, F. Witoelar, and B. Sikoki. *The Fifth Wave of the Indonesia Family Life Survey (IFLS5): Overview and Field Report*. Tech. rep. WR-1143/1-NIA/NICHD. RAND Corporation, 2016. DOI: [10.7249/WR1143.1](https://doi.org/10.7249/WR1143.1).
- [18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. “SMOTE: Synthetic Minority Over-sampling Technique”. In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357. DOI: [10.1613/jair.953](https://doi.org/10.1613/jair.953).
- [19] G. Heinze and M. Schemper. “A solution to the problem of separation in logistic regression”. In: *Statistics in Medicine* 21.16 (2002), pp. 2409–2419. DOI: [10.1002/sim.1047](https://doi.org/10.1002/sim.1047).
- [20] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. “A simulation study of the number of events per variable in logistic regression analysis”. In: *Journal of Clinical Epidemiology* 49.12 (1996), pp. 1373–1379. DOI: [10.1016/S0895-4356\(96\)00236-3](https://doi.org/10.1016/S0895-4356(96)00236-3).
- [21] J. I. Ramírez-Manent, A. Martínez Jover, C. Silveira Martinez, P. Tomás-Gil, P. Martí-Llitas, and A. A. López-González. “Waist circumference is an essential factor in predicting insulin resistance and early detection of metabolic syndrome in adults”. In: *Nutrients* 15.2 (2023), p. 257. DOI: [10.3390/nu15020257](https://doi.org/10.3390/nu15020257).
- [22] Gary S. Collins et al. “TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods”. In: *BMJ* 385 (2024), e078378. DOI: [10.1136/bmj-2023-078378](https://doi.org/10.1136/bmj-2023-078378).