Supplementary Materials – Table of Contents

Machine-learning prediction of patient delay among tuberculosis patients

I. Supplementary Methods

1. Study design, data sources and quality control

This study was designed as a secondary prediction-modeling analysis using routinely collected tuberculosis registration, clinical, health insurance, public-health management and geographic accessibility data. The objective was to develop and evaluate machine-learning models for predicting whether a newly registered tuberculosis patient would experience patient delay before first medical contact.

The original working dataset, analysis_df, contained 67,371 records and 217 variables. Data sources included tuberculosis registration records, symptom and first-visit dates, demographic information, household registration and migration information, occupation and ethnicity, patient source, symptom descriptions, key-population records, insurance linkage data, hospital or medical-record diagnoses, and geographic accessibility measures.

Quality control included checking date order, patient-delay distribution, registration classification, implausible symptom years, missingness patterns, rare category frequencies, variable types and outcome completeness. Derived predictors were constructed before modeling using prespecified public-health logic, while leakage variables directly derived from patient_delay were removed before model fitting.

Item	Description
Original dataset	analysis_df
Initial records	67,371
Raw variables	217
Primary analytical dataset	ml_df_final_imputed after cohort restriction, feature engineering and missing-value handling
Primary outcome	Patient delay >=14 days
Sensitivity outcome	Patient delay >7 days
Primary modeling design	Stratified random 70:30 split; five-fold cross-validation in the training set; independent held-out test-set evaluation

Note. This overview documents the source-data structure and the main modeling design. It is intended to make the supplementary appendix independently understandable without requiring access to the analysis scripts.

Interpretation. The dataset is large enough for machine-learning development, but the outcome and predictor definitions required careful cleaning because patient delay, insurance linkage, symptoms and comorbidity fields contained administrative or clinical missingness.

2. Study population and cohort construction

The target population was restricted to newly registered tuberculosis patients. This restriction was applied because patient delay in new patients has a clear temporal interpretation: the interval between symptom onset and first medical visit. Relapse, return and initial-treatment-failure records may include old symptom dates or treatment-process dates, making patient delay conceptually non-comparable.

Records with negative patient_delay values were excluded because these represented logical date-order contradictions. Values greater than 365 days were winsorized to 365 days to reduce the influence of potential recall bias while preserving the information that the patient experienced very long delay.

Step	Criterion	Rationale	Result
Initial dataset	All records in analysis_df	Starting population	67,371 records
Registration restriction	registration_classification == newly registered patients	Ensures comparable definition of patient delay	61,040 records
Negative delay exclusion	patient_delay <0 excluded	Invalid date order	564 records excluded after new-patient restriction
Extreme delay handling	patient_delay >365 winsorized to 365	Reduces recall-bias influence	2,089 records capped
Final analytical cohort	Outcome and engineered predictors available	Used for model development	60,778 records

Note. The cohort construction steps identify the analytic population used for model development. New patients were used because the patient-delay interval was conceptually valid only in first-episode disease records.

Interpretation. After applying the eligibility and delay-cleaning rules, 60,778 patients remained for the primary modeling analysis. This cohort size supports stable model training and held-out evaluation.

3. Outcome definition and abnormal value handling

patient_delay was defined as the number of days between symptom onset and the first healthcare visit. The original distribution was highly skewed, with negative values and extreme positive values. Negative values were considered logically impossible because the first visit cannot occur before symptom onset under the intended definition.

Among newly diagnosed patients, 564 records had negative patient_delay values. These were treated as invalid and excluded from the analytic cohort. Extreme values greater than 365 days were considered likely to reflect recall or recording error and were winsorized at 365 days in the cleaned delay variable.

The primary binary outcome was patient_delay_14d_clean. Patients were coded as delayed if the cleaned patient delay was at least 14 days; all others were coded as non-delayed. The event level was Yes. The 7-day sensitivity outcome was defined analogously using a threshold of patient delay greater than 7 days.

Raw patient_delay statistic	Value
Minimum	-1,092 days
25th percentile	7 days
Median	23 days
Mean	73.78 days
75th percentile	57 days
Maximum	17,517 days
Negative values in initial dataset	638 records (0.95%)
Extreme values >365 days in initial dataset	2,723 records (4.04%)

The primary binary outcome was patient_delay_14d_clean. Patients were coded as delayed if the cleaned delay was at least 14 days and non-delayed otherwise. The event level was Yes. A seven-day delay definition was used as a threshold sensitivity analysis.

Outcome	Definition	Use
patient_delay_14d_clean	Yes: patient_delay >=14 days; No: patient_delay <14 days	Primary analysis
patient_delay_7d_clean	Yes: patient_delay >7 days; No: patient_delay <=7 days	Sensitivity analysis

Primary outcome category	N	%
No delay or shorter delay (<14 days)	22,106	36.5
Patient delay (>=14 days)	38,370	63.5

Note. The outcome definition follows a binary classification framework. The 14-day threshold was used for the primary model, while the 7-day threshold was used to evaluate threshold robustness.

Interpretation. The primary event rate was 63.5%, indicating moderate class imbalance. Therefore, both ROC-AUC and PR-AUC were reported, and threshold-dependent metrics were interpreted at a prespecified Youden-index threshold.

4. Candidate predictors and feature engineering

Candidate predictors were organized into domains including demographics, household registration and migration, insurance type, comorbidities, temporal features, socioeconomic status, patient source, symptoms, key populations and geographic accessibility. Feature engineering aimed to preserve public-health interpretability while reducing sparse categories and retaining clinically meaningful missingness.

Domain	Predictors	Feature engineering
Demographics	gender_clean, age, age_squared, age_group, occupation_group, ethnicity_group, is_minority	Cleaned sex; age retained as continuous; age squared and age group created; occupation and ethnicity grouped; minority indicator derived.
Household registration and migration	Residence status	Original seven-category classification retained to reflect local and migrant status.
Insurance type	insurance types	Original insurance types simplified into local resident, local employee, non-local resident and non-local employee; missing values

Domain	Predictors	Feature engineering
		inferred using household registration and occupation rules.
Comorbidities	hiv_aids_final, diabetes_final, cci_score, cci_cat	HIV/AIDS and diabetes derived from structured fields plus hospital/medical-record diagnoses and diagnosis-code matching; Unknown retained for untested or undocumented status; Charlson comorbidity index derived from hospitalization diagnoses.
Temporal features	symptom_year, symptom_month, symptom_season, is_winter	Symptom onset year/month and seasonality retained; winter indicator derived.
Socioeconomic status	is_poverty_final	Poverty status derived from assistance records; missing poverty records coded as non-poor according to administrative logic.
Patient source	patient_source_group, is_referral, is_direct_visit	Patient source grouped into referral, direct visit, passive detection, active screening and other; binary indicators created.
Symptoms	symptom_severity_group, is_asymptomatic_final, has_hemoptysis_final, has_fever_final, has_chest_pain_final, has_typical_symptoms_final	Missing symptom text interpreted using symptom-severity logic; symptom flags derived by text matching.
Key population	is_key_population_final and 10 subtype indicators	Overall key-population indicator missing coded as No; subtype missing coded as Unknown to retain systematic

Domain	Predictors	Feature engineering
		missingness.
Geographic accessibility	nearest_road_time, avg_road_time, nearest_road_dist, avg_road_dist, nearest_straight_dist, avg_straight_dist	Road time, road distance and straight-line distance used to represent spatial access; log(x+1) transformation and/or geographic dimension reduction applied during modeling.

Note. The table summarizes the candidate predictor domains and the feature-engineering rules used before model development.

Interpretation. The predictor set covers individual, clinical, temporal, care-pathway and spatial-access dimensions. HIV/AIDS and diabetes were not based solely on structured fields; diagnosis records and relevant diagnosis codes were also used to improve ascertainment.

4.1 Demographic predictors

Age was modeled both as a raw continuous feature and through nonlinear transformations. In the modeling recipe, natural cubic splines were applied to age to capture nonlinearity. Occupation and ethnicity were grouped to reduce sparse categories. Minority status was encoded as a binary variable based on ethnicity.

4.2 Insurance and migration features

Insurance-type missingness was common and likely reflected incomplete insurance linkage, non-enrollment or incomplete records. Missing insurance information was not treated as random. A rule-based imputation strategy was used based on household registration type and occupation: local household registration implied local insurance; migrant status implied non-local insurance; employee-type occupations implied employee insurance; other occupations implied resident insurance.

4.3 Comorbidity features

HIV/AIDS and diabetes were defined using both structured comorbidity fields and diagnosis-code matching. HIV-related codes included B20.x and relevant HIV diagnostic codes. Diabetes-related codes included E10.x–E14.x. Unknown categories were retained because absence of testing or documentation could not be interpreted as disease absence.

4.4 Symptom features

The free-text symptom field had high missingness. Cross-checking showed that missing symptom text was aligned with the no-symptom category in symptom_severity_group. Therefore, missing symptom text was coded according to symptom-specific logic rather than handled as random missingness. Typical symptoms were defined as concurrent cough and sputum.

4.5 Geographic accessibility features

Geographic accessibility variables represented nearest and average travel burden using road-network time, road-network distance and straight-line distance. These variables were expected to be correlated but conceptually non-identical. The raw-variable Spearman matrix was therefore used to document the correlation structure among these accessibility indicators.

5. Missing data handling

Missing data were handled according to source-specific meaning. Missingness was not automatically treated as random. Variables with clinically or administratively meaningful missingness were recoded to preserve information, while variables with excessive non-informative missingness were removed. Model-based imputation steps were learned within training folds only.

Variable/group	Missingness or issue	Handling	Rationale
Negative patient_delay	0.92% after new-patient restriction	Deleted	Logical date error
patient_delay >365 days	3.42% after new-patient restriction	Winsorized at 365 days	Potential recall bias
is_poverty	80.9% missing	NA coded as non-poor	No poverty-assistance record usually indicates not in poverty-assistance file
hiv_aids	42.0% missing	Diagnosis-code enrichment; remaining NA coded as Unknown	Untested or undocumented status cannot be

Variable/group	Missingness or issue	Handling	Rationale
			interpreted as negative
diabetes	39.1% missing	Diagnosis-code enrichment; remaining NA coded as Unknown	Untested or undocumented status cannot be interpreted as negative
Symptom text	84.0% missing	Interpreted using symptom_severity_group and symptom flags	NA aligned with no-symptom coding logic
Key-population subtypes	~50% missing	NA coded as Unknown	Preserves systematic missingness signal
marital_group	~90% missing	Removed	Excessive missingness and limited interpretability
Residual numeric missingness in model recipe	Variable-specific	Median imputation within training folds	Prevents information leakage
Residual nominal missingness in model recipe	Variable-specific	Mode imputation, Unknown category or rare-category pooling	Ensures complete model matrix

Note. This table documents both pre-modeling data cleaning and missing-value handling inside the modeling workflow.

Interpretation. The missingness strategy was primarily rule-based, not purely statistical. Unknown categories were retained when absence of documentation could not safely be interpreted as absence of disease or risk status.

6. Modeling dataset and preprocessing pipeline

After cohort restriction, delay cleaning and feature engineering, the primary modeling dataset contained 60,778 patients. Direct leakage variables derived from patient_delay were removed before model fitting. The primary split was a stratified random 7:3 training-test split, preserving the outcome distribution.

The preprocessing pipeline included median imputation for numeric predictors, mode imputation for nominal predictors, log(x+1) transformation for distance/time variables, natural splines for age, low-frequency category pooling, GLM target encoding for selected high-cardinality predictors, one-hot encoding for remaining nominal variables, zero-variance filtering and numeric normalization. All steps were implemented within a tidymodels workflow.

Preprocessing step	Variables affected	Purpose
Median imputation	Numeric predictors	Handle residual numeric missingness without using test-set information
Mode imputation	Nominal predictors	Handle residual categorical missingness
log(x+1) transformation	Distance and travel-time variables	Reduce right skew and stabilize model fitting
Natural splines	Age	Capture nonlinear age effects
Low-frequency category pooling	Nominal predictors	Reduce sparse dummy variables
GLM target encoding	Selected high-cardinality categorical variables	Improve representation while retaining predictive signal
One-hot encoding	Remaining nominal variables	Generate model matrix
Zero-variance filtering	All predictors	Remove uninformative features
Normalization	Numeric predictors	Ensure comparability for scale-sensitive models

Note. All preprocessing steps were embedded in the modeling workflow and learned using training data only.

Interpretation. Embedding preprocessing inside resampling reduces leakage risk and makes

cross-validated performance closer to the expected performance in unseen patients.

7. Model development and validation

7.1 Predictor selection and retention strategy

Candidate predictors were defined a priori based on clinical relevance, public-health interpretability, routine availability, data quality and feature-engineering rules. We did not use univariable screening or LASSO preselection before fitting the machine-learning models, because linear preselection may discard predictors with nonlinear, threshold-dependent or interaction effects. Instead, predictor reduction and stabilization were achieved through clinically guided feature engineering, rare-category pooling, zero-variance filtering, model-specific regularization and model-based importance assessment.

The final prediction model was selected through comparative model development and validation rather than through a separate pre-model feature-selection step. LightGBM was selected as the primary model because it achieved the best discrimination in the held-out test set. Elastic Net logistic regression was included as a regularized linear comparator to assess whether a simpler sparse linear model could achieve comparable discrimination and to provide a supplementary view of predictors retained under linear regularization.

In the final Elastic Net comparator, the selected hyperparameters were penalty = 1e-05 and mixture = 0.2. A total of 70 non-zero coefficients were retained. The largest coefficients were mainly related to geographic accessibility, symptom year and month, occupation group, typical symptoms, HIV/AIDS and diabetes status, age spline terms, and patient-source/referral variables. These results were broadly consistent with the variable-importance patterns observed in the tree-based models, supporting the robustness of the main predictor domains.

7.2 Model development and validation

Candidate models included LightGBM, XGBoost, Random Forest, Elastic Net and a mean-probability ensemble. Hyperparameters were optimized in the training data using stratified five-fold cross-validation. The final model was refitted on the entire training set and evaluated on the held-out test set. Threshold-dependent metrics were calculated at the threshold selected by maximizing Youden’s index in cross-validated predictions.

LightGBM was selected as the primary model for interpretation because it achieved the highest test-set ROC-AUC and PR-AUC among candidate models. XGBoost showed comparable discrimination, Random Forest provided an additional tree-based benchmark, Elastic Net served as a linear regularized benchmark, and mean ensemble was used to evaluate simple probability averaging.

Model	Role in analysis	Reason for inclusion
LightGBM	Primary model	Highest held-out test-set performance and efficient tree boosting
XGBoost	Comparator tree boosting model	Widely used gradient-boosted tree baseline
Random Forest	Non-boosted tree ensemble comparator	Robust nonlinear baseline
Elastic Net	Regularized linear comparator	Assesses performance of simpler parametric model
Mean Ensemble	Simple probability ensemble	Tests whether probability averaging improves discrimination

Note. The model set intentionally included nonlinear boosted trees, a non-boosted ensemble and a regularized linear comparator.

Interpretation. LightGBM and XGBoost performed similarly, suggesting that the predictive signal was not specific to one implementation of gradient boosting.

8. Performance metrics, calibration and decision curve analysis

Discrimination was assessed using ROC-AUC and PR-AUC. Because the delayed group accounted for more than half of the cohort, PR-AUC was interpreted relative to the event prevalence. Classification metrics included accuracy, sensitivity, specificity, PPV and NPV at the Youden-index threshold. Calibration was assessed graphically using calibration plots.

Decision curve analysis was conducted to quantify public-health decision utility. Net benefit was calculated as $TP/N - FP/N \times pt/(1 - pt)$, where pt denotes the threshold probability. The model strategy was compared against treat-all and treat-none strategies across clinically plausible thresholds.

9. Model interpretability and correlation analysis

Model interpretability was assessed using variable importance and SHAP output. The SHAP beeswarm plot is planned for the main manuscript and is not duplicated in this supplementary appendix. The raw-variable Spearman correlation matrix was calculated using continuous, numeric and binary raw predictors only. Multicategory nominal variables were not arbitrarily encoded.

10. Sensitivity and supplementary analyses

The completed sensitivity analysis redefined patient delay using a 7-day threshold and retrained the LightGBM model under the same modeling framework. Additional optional analyses include 21-day threshold sensitivity, temporal validation, complete-case analysis, subgroup performance analysis and removal of temporal predictors.

II. Supplementary Tables

Supplementary Table S1. Baseline characteristics of the final analytic cohort

domain	variable	level	Overall	No delay / shorter delay (<14 days)	Patient delay (>=14 days)
Demographics	age	NA	54 [36, 66]	49 [31, 64]	56 [40, 67]
Temporal features	symptom_year	NA	2021 [2019, 2023]	2021 [2019, 2023]	2021 [2019, 2023]
Temporal features	symptom_month	NA	6 [3, 9]	7 [4, 9]	6 [3, 9]
Geographic accessibility	nearest_road_time	NA	2126 [1239, 3389]	2035.5 [1205.75, 3199]	2126 [1262, 3469]
Geographic accessibility	avg_road_time	NA	3673.67 [3003.33, 4611.58]	3564.67 [2879, 4484.33]	3690 [3054.33, 4670.58]
Geographic accessibility	nearest_road_dist	NA	14358 [6711, 32907.25]	12448 [6491, 30531.75]	15024 [6946, 34016.75]
Geographic accessibility	avg_road_dist	NA	33652.17 [21812, 48944.58]	28194.17 [21220.58, 46856.92]	35952.67 [22370, 49804]
Geographic accessibility	nearest_straight_dist	NA	10656.77 [4476, 23416.93]	8910.31 [4158.58, 21722.04]	11244.76 [4690.63, 24167.52]
Geographic accessibility	avg_straight_dist	NA	25148.09 [18153.12, 34509.67]	22799.22 [17493.63, 33257.35]	26431.09 [18657.01, 35128.93]
Demographics	gender_clean	Male	39803 (65.5%)	14152 (63.8%)	25651 (66.5%)
Demographics	gender_clean	Female	20975 (34.5%)	8036 (36.2%)	12939 (33.5%)
Demographics	age_group	35-64	29433 (48.4%)	10169 (45.8%)	19264 (49.9%)
Demographi	age_group	>=65	17358 (28.6%)	5329 (24%)	12029 (31.2%)

domain	variable	level	Overall	No delay / shorter delay (<14 days)	Patient delay (>=14 days)
cs					
Demographics	age_group	<35	13987 (23%)	6690 (30.2%)	7297 (18.9%)
Demographics	occupation_group	Farmers	32920 (54.2%)	9474 (42.7%)	23446 (60.8%)
Demographics	occupation_group	Homemakers or unemployed	11504 (18.9%)	5004 (22.6%)	6500 (16.8%)
Demographics	occupation_group	Officials, employees or retirees	4896 (8.1%)	1891 (8.5%)	3005 (7.8%)
Demographics	occupation_group	Other	4552 (7.5%)	2137 (9.6%)	2415 (6.3%)
Demographics	occupation_group	Students	3337 (5.5%)	1780 (8%)	1557 (4%)
Demographics	occupation_group	Service workers	1632 (2.7%)	943 (4.3%)	689 (1.8%)
Demographics	occupation_group	Industrial workers	1026 (1.7%)	488 (2.2%)	538 (1.4%)
Demographics	occupation_group	Professional or technical workers	911 (1.5%)	471 (2.1%)	440 (1.1%)
Demographics	ethnicity_group	Han	35110 (57.8%)	14254 (64.2%)	20856 (54%)
Demographics	ethnicity_group	Zhuang	23366 (38.4%)	7126 (32.1%)	16240 (42.1%)
Demographics	ethnicity_group	Other ethnic minorities	1335 (2.2%)	488 (2.2%)	847 (2.2%)
Demographics	ethnicity_group	Miao	451 (0.7%)	150 (0.7%)	301 (0.8%)
Demographics	ethnicity_group	Dong	306 (0.5%)	91 (0.4%)	215 (0.6%)
Demographics	ethnicity_group	Bouyei	114 (0.2%)	44 (0.2%)	70 (0.2%)
Demographics	ethnicity_group	Yi	96 (0.2%)	35 (0.2%)	61 (0.2%)
Demographics	is_minority	0	35110 (57.8%)	14254 (64.2%)	20856 (54%)

domain	variable	level	Overall	No delay / shorter delay (<14 days)	Patient delay (>=14 days)
Demographics	is_minority	1	25668 (42.2%)	7934 (35.8%)	17734 (46%)
Residence and insurance	Residence status	Local	39410 (64.8%)	15612 (70.4%)	23798 (61.7%)
Residence and insurance	Residence status	Intra-city migrant	10605 (17.4%)	3570 (16.1%)	7035 (18.2%)
Residence and insurance	Residence status	Inter-city migrant within province	10266 (16.9%)	2768 (12.5%)	7498 (19.4%)
Residence and insurance	Residence status	Inter-provincial migrant	492 (0.8%)	237 (1.1%)	255 (0.7%)
Residence and insurance	Residence status	Foreign national	3 (0%)	0 (0%)	3 (0%)
Residence and insurance	Residence status	Hong Kong/Macau /Taiwan	2 (0%)	1 (0%)	1 (0%)
Residence and insurance	Insurance type	Local resident insurance	31563 (51.9%)	11967 (53.9%)	19596 (50.8%)
Residence and insurance	Insurance type	Non-local resident insurance	22060 (36.3%)	7229 (32.6%)	14831 (38.4%)
Residence and insurance	Insurance type	Local employee insurance	4968 (8.2%)	2192 (9.9%)	2776 (7.2%)
Residence and insurance	Insurance type	Non-local employee insurance	2187 (3.6%)	800 (3.6%)	1387 (3.6%)
Clinical comorbidity	hiv_aids_final	No	34944 (57.5%)	12438 (56.1%)	22506 (58.3%)
Clinical comorbidity	hiv_aids_final	Unknown	24232 (39.9%)	9143 (41.2%)	15089 (39.1%)
Clinical comorbidity	hiv_aids_final	Yes	1602 (2.6%)	607 (2.7%)	995 (2.6%)
Clinical	diabetes_final	No	31952 (52.6%)	11583 (52.2%)	20369 (52.8%)

domain	variable	level	Overall	No delay / shorter delay (<14 days)	Patient delay (>=14 days)
comorbidity					
Clinical comorbidity	diabetes_final	Unknown	21978 (36.2%)	8417 (37.9%)	13561 (35.1%)
Clinical comorbidity	diabetes_final	Yes	6848 (11.3%)	2188 (9.9%)	4660 (12.1%)
Temporal features	symptom_season	Spring	16066 (26.4%)	5764 (26%)	10302 (26.7%)
Temporal features	symptom_season	Summer	15984 (26.3%)	6318 (28.5%)	9666 (25%)
Temporal features	symptom_season	Winter	14825 (24.4%)	4755 (21.4%)	10070 (26.1%)
Temporal features	symptom_season	Autumn	13903 (22.9%)	5351 (24.1%)	8552 (22.2%)
Temporal features	is_winter	0	45953 (75.6%)	17433 (78.6%)	28520 (73.9%)
Temporal features	is_winter	1	14825 (24.4%)	4755 (21.4%)	10070 (26.1%)
Patient source and care pathway	patient_source_group	Referral	40962 (67.4%)	13142 (59.2%)	27820 (72.1%)
Patient source and care pathway	patient_source_group	Direct care-seeking	10976 (18.1%)	4872 (22%)	6104 (15.8%)
Patient source and care pathway	patient_source_group	Passive detection	8426 (13.9%)	3919 (17.7%)	4507 (11.7%)
Patient source and care pathway	patient_source_group	Active screening	396 (0.7%)	250 (1.1%)	146 (0.4%)
Patient source and care pathway	patient_source_group	Other	18 (0%)	5 (0%)	13 (0%)
Patient source and care pathway	is_referral	1	40962 (67.4%)	13142 (59.2%)	27820 (72.1%)
Patient source and care pathway	is_referral	0	19816 (32.6%)	9046 (40.8%)	10770 (27.9%)
Patient source and	is_direct_visit	0	49802 (81.9%)	17316 (78%)	32486 (84.2%)

domain	variable	level	Overall	No delay / shorter delay (<14 days)	Patient delay (>=14 days)
care pathway					
Patient source and care pathway	is_direct_visit	1	10976 (18.1%)	4872 (22%)	6104 (15.8%)
Symptoms	symptom_severity_group	Asymptomatic	52122 (85.8%)	19329 (87.1%)	32793 (85%)
Symptoms	symptom_severity_group	Mild	7233 (11.9%)	2222 (10%)	5011 (13%)
Symptoms	symptom_severity_group	Moderate	1221 (2%)	533 (2.4%)	688 (1.8%)
Symptoms	symptom_severity_group	Severe	202 (0.3%)	104 (0.5%)	98 (0.3%)
Symptoms	is_asymptomatic_final	1	52122 (85.8%)	19329 (87.1%)	32793 (85%)
Symptoms	is_asymptomatic_final	0	8656 (14.2%)	2859 (12.9%)	5797 (15%)
Symptoms	has_typical_symptoms_final	0	54032 (88.9%)	20254 (91.3%)	33778 (87.5%)
Symptoms	has_typical_symptoms_final	1	6746 (11.1%)	1934 (8.7%)	4812 (12.5%)
Key population	is_key_population_final	No	57493 (94.6%)	20542 (92.6%)	36951 (95.8%)
Key population	is_key_population_final	Yes	3285 (5.4%)	1646 (7.4%)	1639 (4.2%)
Key population	Close contact	No	35425 (58.3%)	12599 (56.8%)	22826 (59.2%)
Key population	Close contact	Unknown	25086 (41.3%)	9452 (42.6%)	15634 (40.5%)
Key population	Close contact	Yes	267 (0.4%)	137 (0.6%)	130 (0.3%)
Key population	Health-care worker	No	35369 (58.2%)	12568 (56.6%)	22801 (59.1%)
Key population	Health-care worker	Unknown	25194 (41.5%)	9519 (42.9%)	15675 (40.6%)
Key population	Health-care worker	Yes	215 (0.4%)	101 (0.5%)	114 (0.3%)
Key	School/kindergar	No	34795 (57.2%)	12269 (55.3%)	22526 (58.4%)

domain	variable	level	Overall	No delay / shorter delay (<14 days)	Patient delay (>=14 days)
population	ten personnel				
Key population	School/kindergarten personnel	Unknown	23538 (38.7%)	8649 (39%)	14889 (38.6%)
Key population	School/kindergarten personnel	Yes	2445 (4%)	1270 (5.7%)	1175 (3%)
Key population	Custodial personnel	No	31981 (52.6%)	10529 (47.5%)	21452 (55.6%)
Key population	Custodial personnel	Unknown	28650 (47.1%)	11601 (52.3%)	17049 (44.2%)
Key population	Custodial personnel	Yes	147 (0.2%)	58 (0.3%)	89 (0.2%)
Key population	Livestock worker	No	32089 (52.8%)	10573 (47.7%)	21516 (55.8%)
Key population	Livestock worker	Unknown	28689 (47.2%)	11615 (52.3%)	17074 (44.2%)
Key population	Dust-exposed worker or pneumoconiosis patient	No	35408 (58.3%)	12598 (56.8%)	22810 (59.1%)
Key population	Dust-exposed worker or pneumoconiosis patient	Unknown	25259 (41.6%)	9551 (43%)	15708 (40.7%)
Key population	Dust-exposed worker or pneumoconiosis patient	Yes	111 (0.2%)	39 (0.2%)	72 (0.2%)
Key population	Psychiatric-hospital patient	No	35319 (58.1%)	12548 (56.6%)	22771 (59%)
Key population	Psychiatric-hospital patient	Unknown	25320 (41.7%)	9568 (43.1%)	15752 (40.8%)
Key population	Psychiatric-hospital patient	Yes	139 (0.2%)	72 (0.3%)	67 (0.2%)
Key population	Nursing-home resident	No	32089 (52.8%)	10573 (47.7%)	21516 (55.8%)
Key	Nursing-home	Unknown	28689 (47.2%)	11615 (52.3%)	17074 (44.2%)

domain	variable	level	Overall	No delay / shorter delay (<14 days)	Patient delay (>=14 days)
population	resident				
Key population	Welfare-institution resident	No	32086 (52.8%)	10573 (47.7%)	21513 (55.7%)
Key population	Welfare-institution resident	Unknown	28689 (47.2%)	11615 (52.3%)	17074 (44.2%)
Key population	Welfare-institution resident	Yes	3 (0%)	0 (0%)	3 (0%)
Key population	Other congregate-setting population	No	32088 (52.8%)	10572 (47.6%)	21516 (55.8%)
Key population	Other congregate-setting population	Unknown	28688 (47.2%)	11615 (52.3%)	17073 (44.2%)
Key population	Other congregate-setting population	Yes	2 (0%)	1 (0%)	1 (0%)

Note. Continuous variables are summarized as median [interquartile range], and categorical variables are summarized as n (%). Patient delay was defined as an interval of ≥14 days between symptom onset and first medical visit. This table is descriptive and was not used for univariable predictor screening. Candidate predictors were defined a priori based on clinical relevance, public-health interpretability, data availability and feature-engineering rules.

Domain	Predictors	Type	Handling
Demographics	gender_clean; age; age_squared; age_group; occupation_group; ethnicity_group; is_minority	Mixed: numeric, binary and categorical	Sex cleaned; age retained as continuous; age squared and age group derived; occupation and ethnicity grouped; minority indicator derived.
Household registration	Residence status	Categorical	Original seven-category classification

Domain	Predictors	Type	Handling
and migration			retained to represent local and migrant status.
Insurance type	Insurance type	Categorical	Raw insurance categories simplified into local/non-local and resident/employee groups; missing values inferred using household registration and occupation rules.
Comorbidities	hiv_aids_final; diabetes_final; cci_score; cci_cat	Categorical and numeric	HIV/AIDS and diabetes derived from structured fields plus hospital/medical-record diagnoses and diagnosis-code matching; unresolved status coded Unknown; CCI derived from hospitalization diagnoses.
Temporal features	symptom_year; symptom_month; symptom_season; is_winter	Numeric and categorical	Symptom year/month retained; season and winter indicators derived.
Socioeconomic status	is_poverty_final	Binary/categorical	Poverty-assistance record coded poor; missing/no record coded non-poor based on

Domain	Predictors	Type	Handling
			administrative logic.
Patient source	patient_source_group; is_referral; is_direct_visit	Categorical and binary	Patient source grouped into referral, direct visit, passive detection, active screening and other; pathway indicators derived.
Symptoms	symptom_severity_group; is_asymptomatic_final; has_hemoptysis_final; has_fever_final; has_chest_pain_final; has_typical_symptoms_final	Categorical and binary	Missing symptom text interpreted using severity logic; symptom flags derived by text matching.
Key population	is_key_population_final and 10 subtype variables	Categorical/binary	Overall flag missing coded No; subtype missing coded Unknown to preserve systematic missingness.
Geographic accessibility	nearest_road_time; avg_road_time; nearest_road_dist; avg_road_dist; nearest_straight_dist; avg_straight_dist	Continuous	Road time, road distance and straight-line distance used to represent spatial access; log(x+1) transformation and/or geographic dimension reduction applied during modeling.

Supplementary Table S2. Data dictionary of final candidate predictors

Note. This data dictionary summarizes the final candidate predictors used in the modeling

dataset and the main feature-engineering rules.

Interpretation. The predictor set is intentionally multidimensional. Comorbidity variables were enhanced using diagnosis records rather than relying only on incomplete structured fields, while geographic variables were retained as interpretable raw constructs despite their expected correlation.

Supplementary Table S3. Missing data and abnormal value handling

Variable/group	Issue	Handling	Rationale
Negative patient_delay	0.92% after new-patient restriction	Deleted	Logical date error
patient_delay >365 days	3.42% after new-patient restriction	Winsorized at 365 days	Potential recall bias
is_poverty	80.9% missing	NA coded as non-poor	No poverty-assistance record usually indicates not in poverty-assistance file
hiv_aids	42.0% missing	Diagnosis-code enrichment; remaining NA coded as Unknown	Untested or undocumented status cannot be interpreted as negative
diabetes	39.1% missing	Diagnosis-code enrichment; remaining NA coded as Unknown	Untested or undocumented status cannot be interpreted as negative
Symptom text	84.0% missing	Interpreted using symptom_severity_group and symptom flags	NA aligned with no-symptom coding logic
Key-population	~50% missing	NA coded as Unknown	Preserves

Variable/group	Issue	Handling	Rationale
subtypes			systematic missingness signal
marital_group	~90% missing	Removed	Excessive missingness and limited interpretability

Note. This table separates implausible outcome values from predictor missingness and documents the rule used for each major variable group.

Interpretation. The handling strategy aimed to avoid misclassifying unknown clinical status as disease absence. HIV/AIDS and diabetes retained Unknown categories after diagnosis-code enrichment because incomplete testing or documentation is informative in routine data.

Predictor domain / variable group	Representative retained predictors	Interpretation
Geographic accessibility	avg_road_dist, avg_straight_dist, nearest_straight_dist, nearest_road_dist, avg_road_time, nearest_road_time	Spatial-access variables were retained by the regularized linear model, consistent with the tree-based model importance ranking.
Temporal features	symptom_year, symptom_month	Calendar-year and monthly variation contributed to prediction under the linear benchmark.
Demographic/social factors	occupation_group, age spline terms	Occupation and nonlinear age patterns were retained as relevant predictors.
Symptom features	has_typical_symptoms_final	Typical symptom status contributed to the linear benchmark.
Comorbidities	hiv_aids_final, diabetes_final	Comorbidity-related variables were retained after diagnosis-code enrichment and

Predictor domain / variable group	Representative retained predictors	Interpretation
		missingness coding.
Patient pathway	patient_source_group, is_referral	Patient source and referral pathway variables were retained, consistent with care-seeking pathway relevance.
Key population indicators	selected key-population subtype indicators	Several key-population subtype variables were retained, although individual dummy-level coefficients should be interpreted cautiously.

Supplementary Table S4. Non-zero coefficients from Elastic Net logistic regression.

Note. Elastic Net logistic regression was included as a regularized linear comparator rather than as a pre-screening feature-selection method for the final LightGBM model. The final Elastic Net model used penalty = 1e-05 and mixture = 0.2, retaining 70 non-zero coefficients. Because several categorical predictors were expanded into dummy variables during preprocessing, individual dummy-level coefficients should be interpreted cautiously. The table summarizes retained predictor domains rather than serving as the variable-selection basis for the final LightGBM model.

Interpretation. The Elastic Net comparator retained predictors from the same broad domains identified by the primary LightGBM model, especially geographic accessibility, temporal factors, occupation, symptom characteristics, comorbidities and patient pathway variables. This supports the robustness of the main predictor domains while confirming that the final model was not dependent on a single nonlinear tree-based algorithm.

Supplementary Table S5. Top 30 non-zero Elastic Net coefficients.

term	estimate	abs_estimate	direction
Average road-network distance	-0.751	0.751	Negative
Average straight-line distance	0.590	0.590	Positive
Symptom-onset year	0.332	0.332	Positive
Nearest straight-line distance	-0.287	0.287	Negative
Occupation	-0.215	0.215	Negative

Nearest road-network distance	0.182	0.182	Positive
Average road-network travel time	0.163	0.163	Positive
Symptom-onset month	0.141	0.141	Positive
Typical TB symptoms	-0.131	0.131	Negative
HIV/AIDS status: no	0.125	0.125	Positive
HIV/AIDS status: unknown	-0.121	0.121	Negative
Residence status	-0.117	0.117	Negative
Age, nonlinear component 2	-0.103	0.103	Negative
Travel time to nearest TB-designated facility	0.101	0.101	Positive
Diabetes status: no	0.100	0.100	Positive
Age, nonlinear component 1	-0.078	0.078	Negative
Health-care worker: unknown	0.071	0.071	Positive
School/kindergarten personnel: no	0.067	0.067	Positive
School/kindergarten personnel: unknown	-0.062	0.062	Negative
Patient source: other	0.059	0.059	Positive
Nursing-home resident: no	-0.051	0.051	Negative
Welfare-institution resident: no	-0.048	0.048	Negative
Psychiatric-hospital patient: unknown	-0.048	0.048	Negative
Psychiatric-hospital patient: other	0.044	0.044	Positive
Livestock worker: unknown	0.043	0.043	Positive
Key population: no	0.042	0.042	Positive
Nursing-home resident: unknown	0.042	0.042	Positive
Patient source: referral	-0.039	0.039	Negative
Referral source	-0.038	0.038	Negative
Welfare-institution resident: unknown	0.038	0.038	Positive

Coefficients from Elastic Net represent conditional effects in the transformed model matrix. Because several geographic accessibility variables were strongly correlated, coefficient signs should not be interpreted as isolated marginal associations.

Supplementary Table S6. Final model performance in the independent test set

Model	Threshold	PR-AUC	ROC-AUC	Accuracy	Sensitivity	Specificity
LightGBM	0.620	0.796	0.703	0.651	0.678	0.603
XGBoost	0.620	0.794	0.702	0.649	0.678	0.599
MeanEnsemble	0.500	0.792	0.697	0.686	0.894	0.326
RandomForest	0.610	0.781	0.687	0.641	0.678	0.577
ElasticNet	0.650	0.764	0.662	0.612	0.601	0.631

Note. The primary outcome was patient delay >=14 days. The classification threshold was selected using Youden's index from cross-validated predictions in the training data.

Interpretation. LightGBM achieved the best overall discrimination in the held-out test set, with ROC-AUC 0.703 and PR-AUC 0.796. XGBoost had nearly identical performance, while Random Forest and Elastic Net were lower, supporting nonlinear modeling but indicating moderate rather than high separability.

Supplementary Table S7. Comparison of single candidate models

Model	Threshold	PR-AUC	ROC-AUC	Accuracy	Sensitivity	Specificity
LightGBM	0.620	0.796	0.703	0.651	0.678	0.603
XGBoost	0.620	0.794	0.702	0.649	0.678	0.599
RandomForest	0.610	0.781	0.687	0.641	0.678	0.577
ElasticNet	0.650	0.764	0.662	0.612	0.601	0.631

Note. This table reports individual candidate models evaluated on the same held-out test set and excludes the mean ensemble.

Interpretation. The ranking of single models was consistent with the final model table: LightGBM and XGBoost showed the highest discrimination, Random Forest was

intermediate and Elastic Net was lowest. This pattern suggests a meaningful nonlinear signal in the predictor set.

Supplementary Table S8. Decision curve analysis net benefit

Threshold probability	Model	Treat all	Treat none
0.10	0.594	0.594	0.000
0.20	0.545	0.544	0.000
0.30	0.484	0.478	0.000
0.40	0.412	0.392	0.000
0.50	0.322	0.270	0.000

Note. Decision curve analysis was performed to evaluate the net benefit of using the final LightGBM model to identify patients at risk of delayed care-seeking, compared with default strategies of treating all patients as delayed or treating no patients as delayed. Net benefit was calculated across clinically relevant threshold probabilities.

Interpretation. At thresholds from 0.10 to 0.50, the model showed consistently higher net benefit than the treat-none strategy and slightly to moderately higher net benefit than the treat-all strategy. The incremental advantage over treat-all became more evident at higher threshold probabilities, suggesting that model-guided risk stratification may provide operational utility over non-selective intervention strategies.

Supplementary Table S9. Seven-day delay threshold sensitivity analysis

Outcome	Model	Threshold	Event rate	ROC-AUC	PR-AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
Patient delay >7 days	LightGBM	0.740	0.739	0.727	0.874	0.655	0.648	0.676	0.850	0.405

Note. The outcome was redefined as patient delay >7 days, and the LightGBM model was retrained under the same general modeling framework.

Interpretation. The seven-day model achieved ROC-AUC 0.727 and PR-AUC 0.874, indicating that model discrimination was not weakened when a shorter delay threshold was

used. Because the event rate was 73.9%, the PR-AUC should be interpreted relative to this higher prevalence baseline.

Supplementary Table S10. Threshold selection in the seven-day sensitivity analysis

.threshold	.metric	.estimator	.estimate
0.74	j_index	binary	0.327
0.75	j_index	binary	0.326
0.72	j_index	binary	0.326
0.73	j_index	binary	0.325
0.71	j_index	binary	0.324
0.76	j_index	binary	0.324
0.70	j_index	binary	0.322
0.77	j_index	binary	0.321
0.78	j_index	binary	0.316
0.69	j_index	binary	0.316
0.68	j_index	binary	0.312
0.79	j_index	binary	0.309
0.67	j_index	binary	0.303
0.80	j_index	binary	0.300
0.66	j_index	binary	0.294
0.65	j_index	binary	0.289
0.64	j_index	binary	0.283
0.63	j_index	binary	0.276
0.62	j_index	binary	0.267
0.61	j_index	binary	0.260
0.60	j_index	binary	0.251

.threshold	.metric	.estimator	.estimate
0.59	j_index	binary	0.240
0.58	j_index	binary	0.229
0.57	j_index	binary	0.220
0.56	j_index	binary	0.209
0.55	j_index	binary	0.201
0.54	j_index	binary	0.190
0.53	j_index	binary	0.182
0.52	j_index	binary	0.174
0.51	j_index	binary	0.163
0.50	j_index	binary	0.155
0.49	j_index	binary	0.146
0.48	j_index	binary	0.136
0.47	j_index	binary	0.128
0.46	j_index	binary	0.119
0.45	j_index	binary	0.111
0.44	j_index	binary	0.105
0.43	j_index	binary	0.099
0.42	j_index	binary	0.092
0.41	j_index	binary	0.085
0.40	j_index	binary	0.077
0.39	j_index	binary	0.070
0.38	j_index	binary	0.064
0.37	j_index	binary	0.060
0.36	j_index	binary	0.052
0.35	j_index	binary	0.046

.threshold	.metric	.estimator	.estimate
0.34	j_index	binary	0.042
0.33	j_index	binary	0.037
0.32	j_index	binary	0.031
0.31	j_index	binary	0.027
0.30	j_index	binary	0.022
0.29	j_index	binary	0.018
0.28	j_index	binary	0.014
0.27	j_index	binary	0.011
0.26	j_index	binary	0.007
0.25	j_index	binary	0.005
0.24	j_index	binary	0.004
0.23	j_index	binary	0.002
0.22	j_index	binary	0.001
0.21	j_index	binary	0.000
0.20	j_index	binary	0.000

Note. The table shows the top 10 thresholds ranked by Youden's index in the seven-day sensitivity analysis. The full threshold grid can be retained in the analysis output file if required.

Interpretation. The optimal threshold was 0.74, with Youden's index 0.327. Neighboring thresholds from 0.71 to 0.76 produced similar values, suggesting that classification performance was not highly sensitive to a single exact threshold.

Supplementary Table S11. Raw variables included in Spearman correlation analysis

Variable	Reason for inclusion
age	Age-related patient-level factor

Variable	Reason for inclusion
age_squared	Nonlinear age term
is_referral	Patient pathway marker
is_minority	Ethnicity-related social factor
gender_clean	Demographic factor
nearest_road_time	Geographic accessibility
nearest_straight_dist	Geographic accessibility
nearest_road_dist	Geographic accessibility
avg_road_time	Geographic accessibility
avg_straight_dist	Geographic accessibility
avg_road_dist	Geographic accessibility
symptom_month	Temporal factor
symptom_year	Temporal factor
has_typical_symptoms_final	Symptom pattern
is_asymptomatic_final	Symptom pattern

Note. Only continuous, numeric and binary raw predictors were included in the Spearman correlation matrix. Multicategory nominal predictors were not forced into arbitrary ordinal codes.

Interpretation. The included variables were selected to represent interpretable raw constructs, especially geographic accessibility, age, care pathway, symptoms and time. This avoids the interpretability problems caused by correlating dummy-expanded model features.

Supplementary Table S12. SHAP global importance ranking

Rank	Feature	Mean absolute SHAP
1	Geographic accessibility PC1	0.3561
2	Geographic accessibility PC2	0.1675
3	Occupation	0.1282

Rank	Feature	Mean absolute SHAP
4	Symptom-onset month	0.0995
5	Symptom-onset year	0.0789
6	Age, nonlinear component 1	0.0681
7	Referral source	0.0648
8	Typical TB symptoms	0.0552
9	Residence status	0.0518
10	Geographic accessibility PC3	0.0517
11	Age squared	0.0511
12	Patient source: referral	0.0468
13	School/kindergarten personnel: no	0.0406
14	Diabetes status: no	0.0339
15	HIV/AIDS status: unknown	0.0297
16	Close contact: unknown	0.0279
17	Close contact: no	0.0258
18	Symptom-onset season: summer	0.0234
19	Custodial personnel: unknown	0.0208
20	Age, nonlinear component 3	0.0200

Note. This table lists the top 20 predictors ranked by mean absolute SHAP values from the available SHAP output.

Interpretation. The ranking provides a numeric supplement to the SHAP beeswarm plot. The beeswarm plot itself is planned for the main manuscript and is therefore not duplicated in the supplementary appendix.

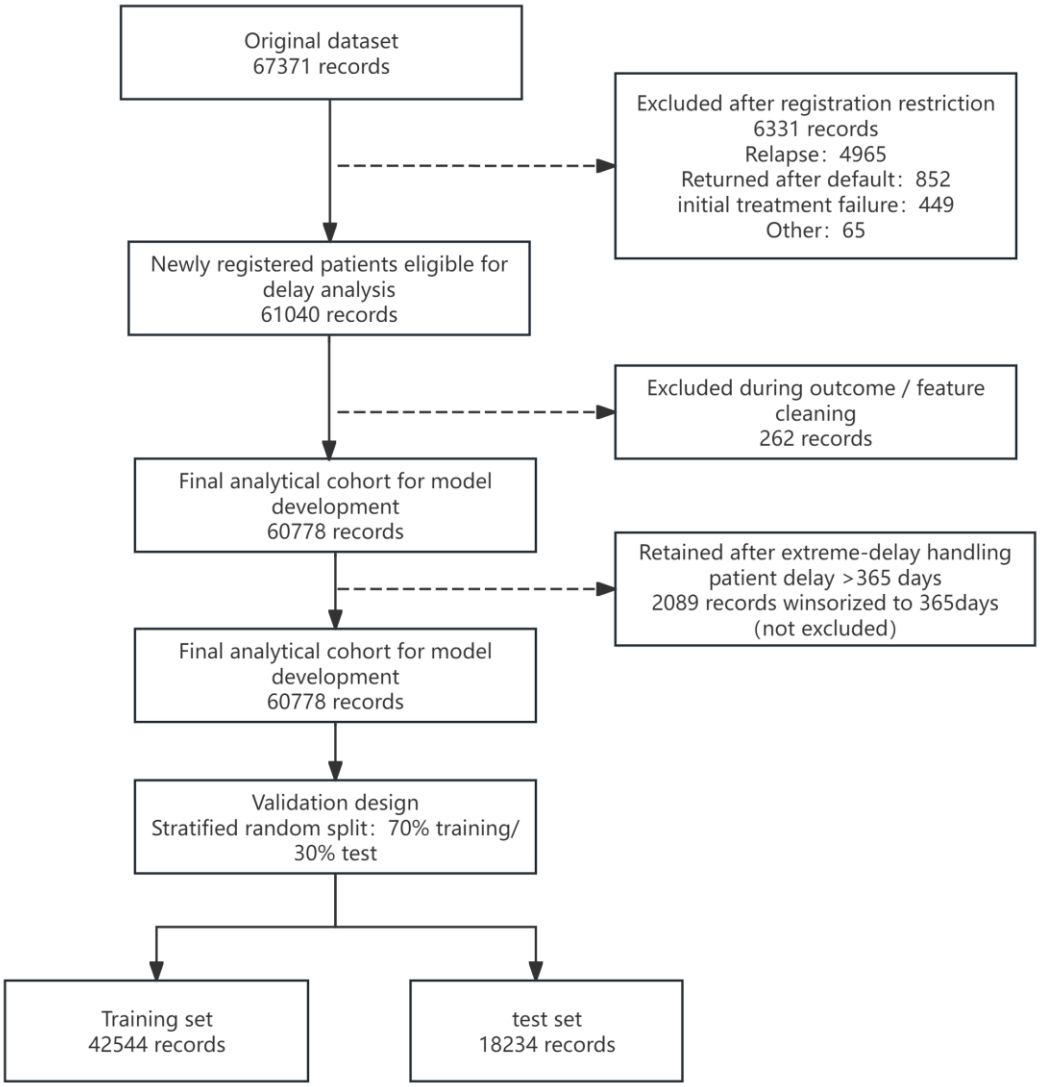
Supplementary Table S13. Subgroup performance of the final model

subgroup_type	level	n	events	event_rate	auc	auc_low	auc_high

Age group	35-64	8745	5745	0.657	0.692	0.680	0.703
	<35	4209	2202	0.523	0.690	0.674	0.706
	>=65	5280	3630	0.688	0.688	0.673	0.703
Sex	Female	6234	3805	0.610	0.706	0.693	0.720
	Male	12000	7772	0.648	0.699	0.690	0.709
Residence status	Migrant	6410	4489	0.700	0.697	0.683	0.711
	Local	11824	7088	0.599	0.696	0.687	0.706

III. Supplementary Figures / 补充图

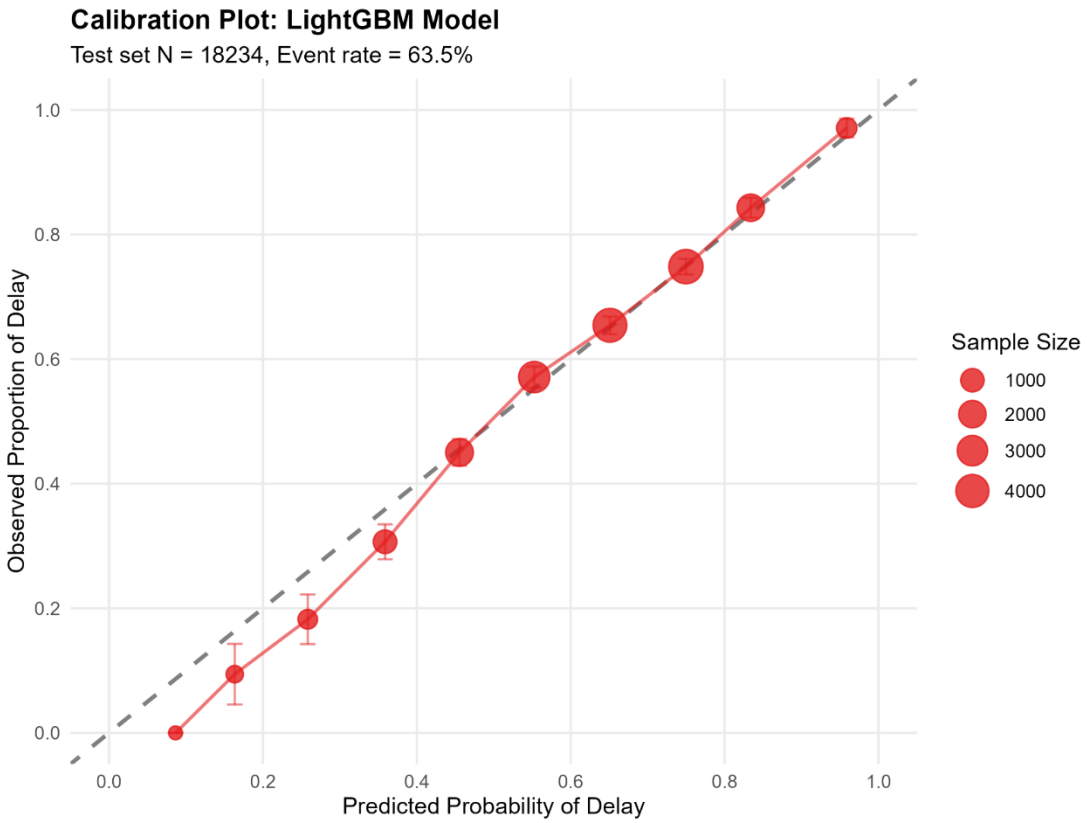
Supplementary Figure S1. Participant flow diagram



Note. Reserved position for a patient-flow diagram showing raw records, restriction to newly diagnosed patients, exclusion of negative delay records, winsorization of extreme delay values and final analytical cohort.

Interpretation. A flow diagram is useful for the main manuscript or supplementary appendix. If the main text includes the flow diagram, this supplementary placeholder should be removed to avoid duplication.

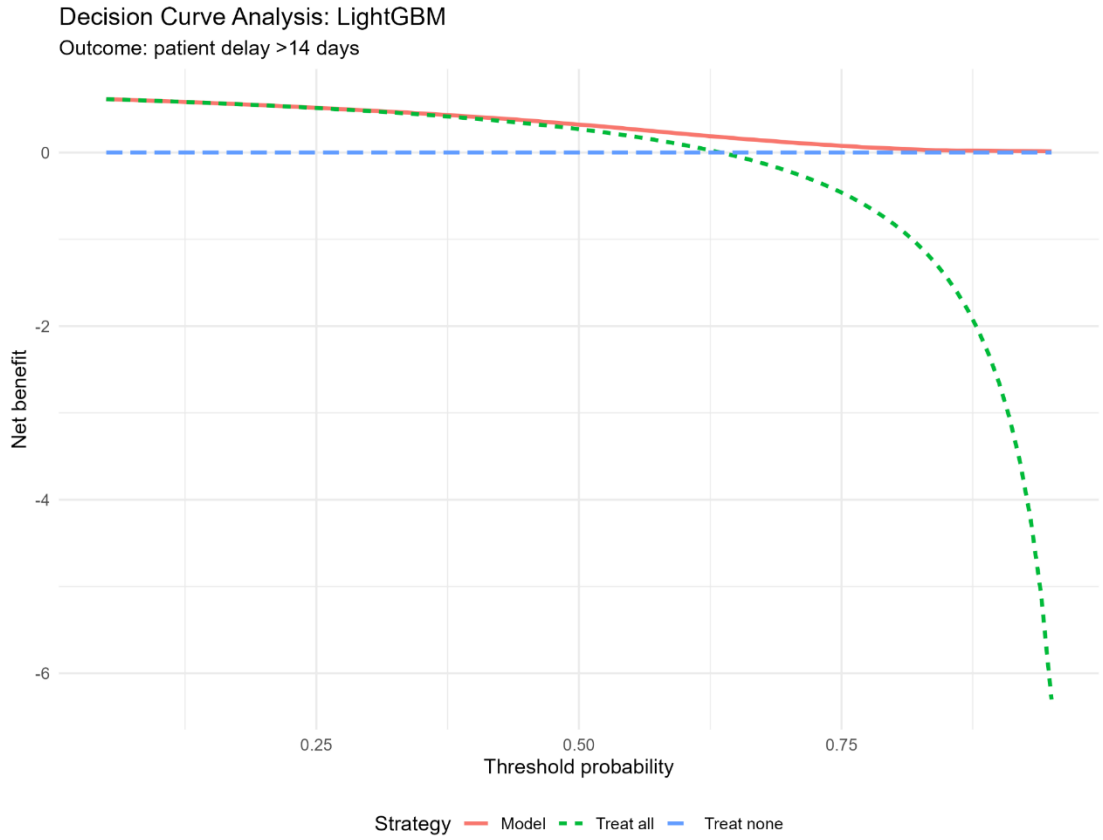
Supplementary Figure S2. Calibration plot for the primary LightGBM model



Note. The calibration plot compares grouped predicted probabilities with observed proportions of patient delay in the held-out test set.

Interpretation. The curve closely follows the diagonal over most of the risk range, indicating acceptable graphical calibration. Mild deviations in lower-risk strata should be interpreted cautiously because sample sizes are smaller in the extreme probability bins.

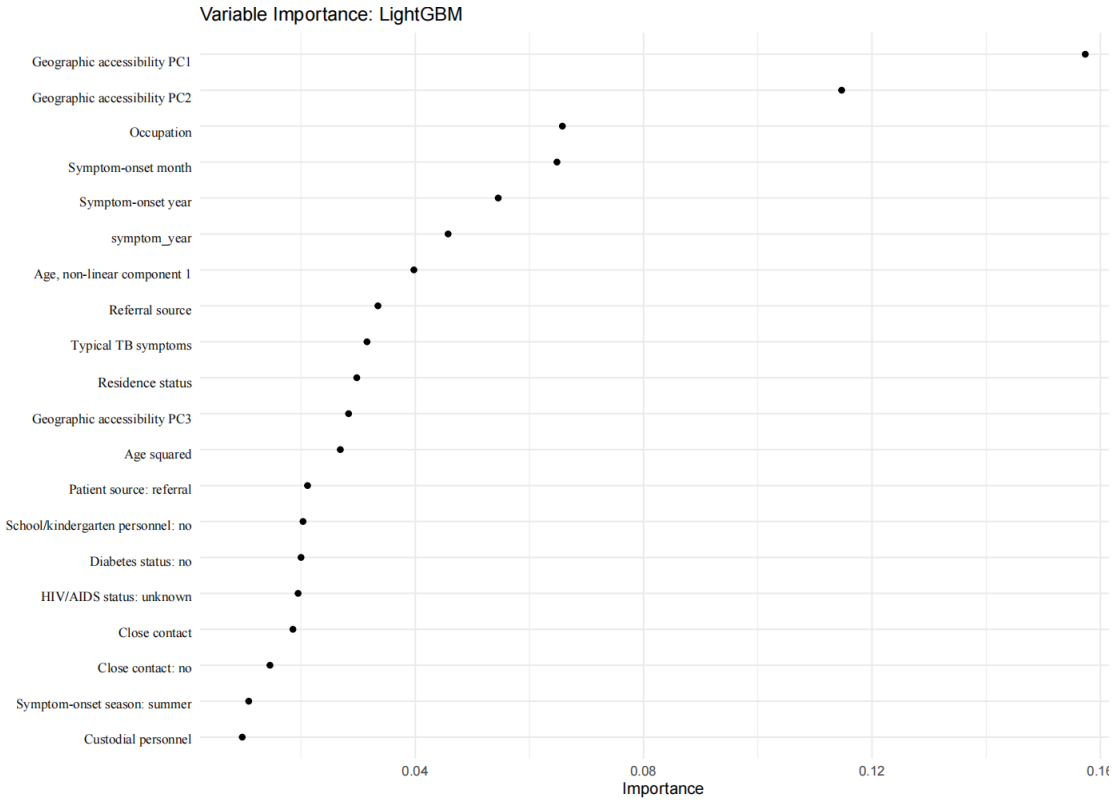
Supplementary Figure S3. Decision curve analysis for the primary LightGBM model



Note. Decision curve analysis compares the model strategy with treat-all and treat-none strategies across threshold probabilities.

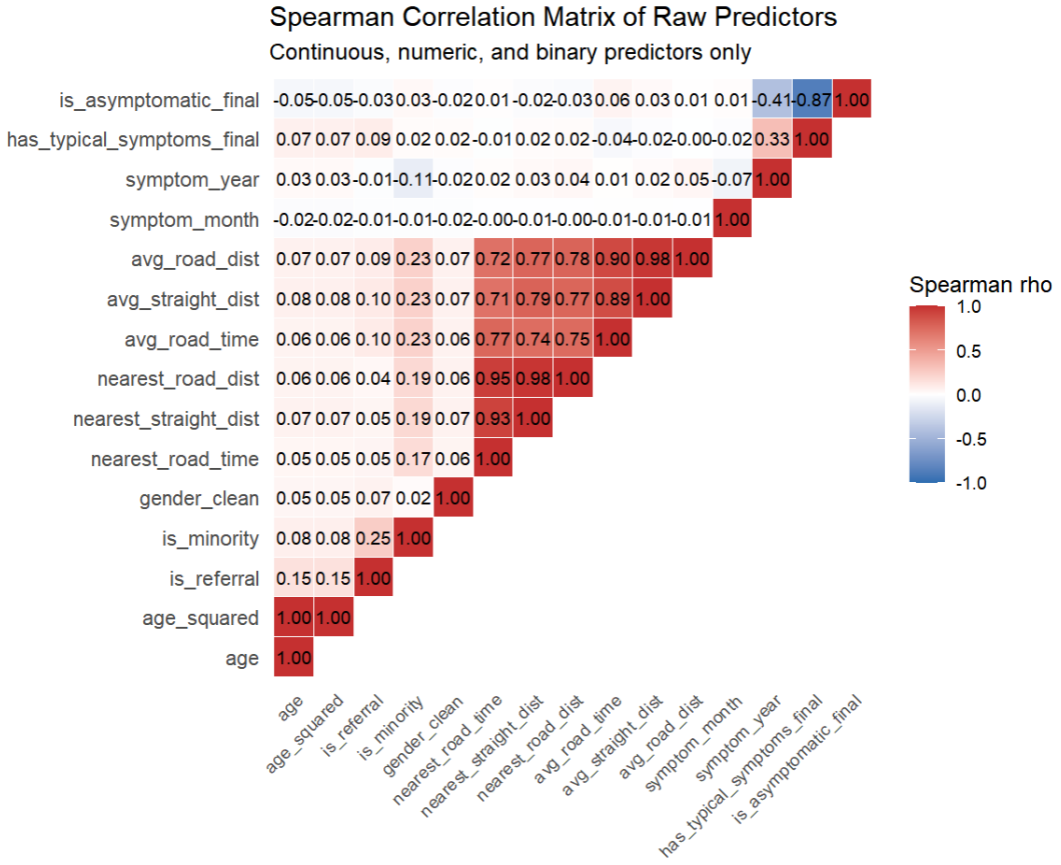
Interpretation. The model curve remained above treat-none and generally above treat-all across clinically plausible thresholds. This supports the potential incremental net benefit of risk-stratified intervention, especially when the operational threshold for intervention is not extremely low.

Supplementary Figure S4. Variable importance of the primary model



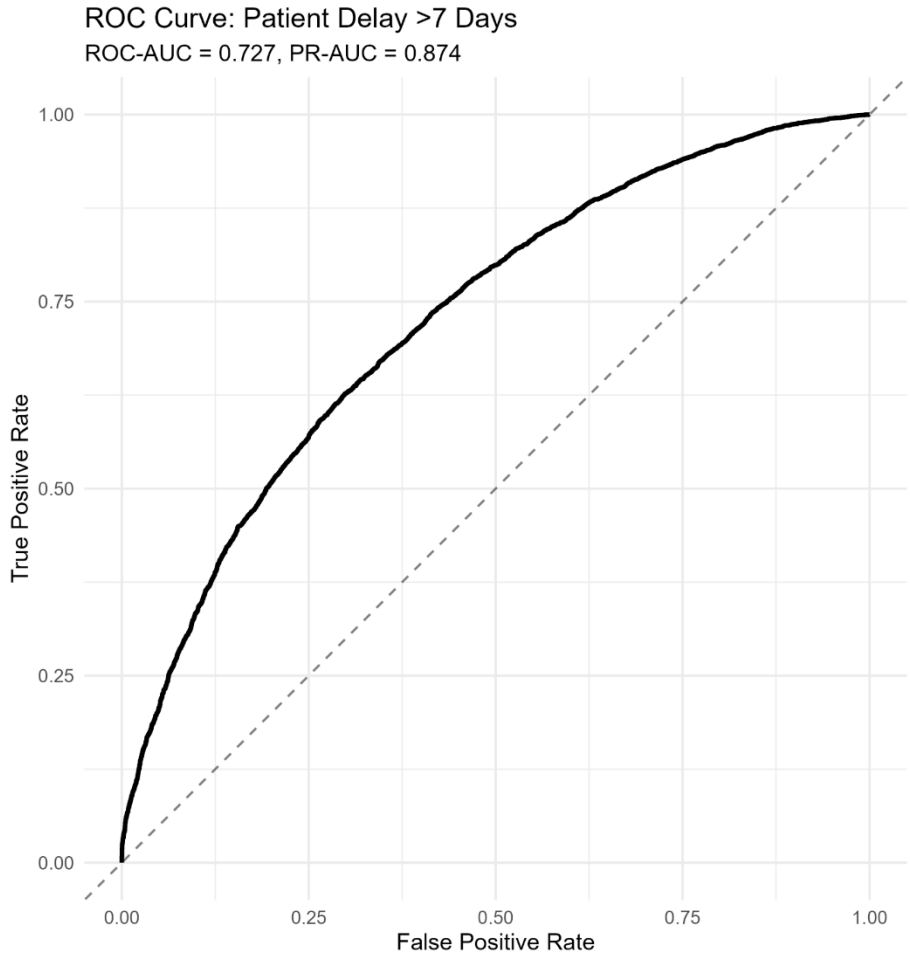
Interpretation. This figure should be used as a supplementary explanation of the primary model. It should not replace the SHAP beeswarm plot in the main text because variable importance does not show effect direction or patient-level heterogeneity.

Supplementary Figure S5. Spearman correlation matrix of key raw predictors



Interpretation. The raw-variable matrix should emphasize that geographic accessibility variables are highly correlated but conceptually represent related components of spatial access. Age and age-squared, as well as typical symptoms and asymptomatic status, are expected to show strong correlations due to deterministic or clinical coding logic.

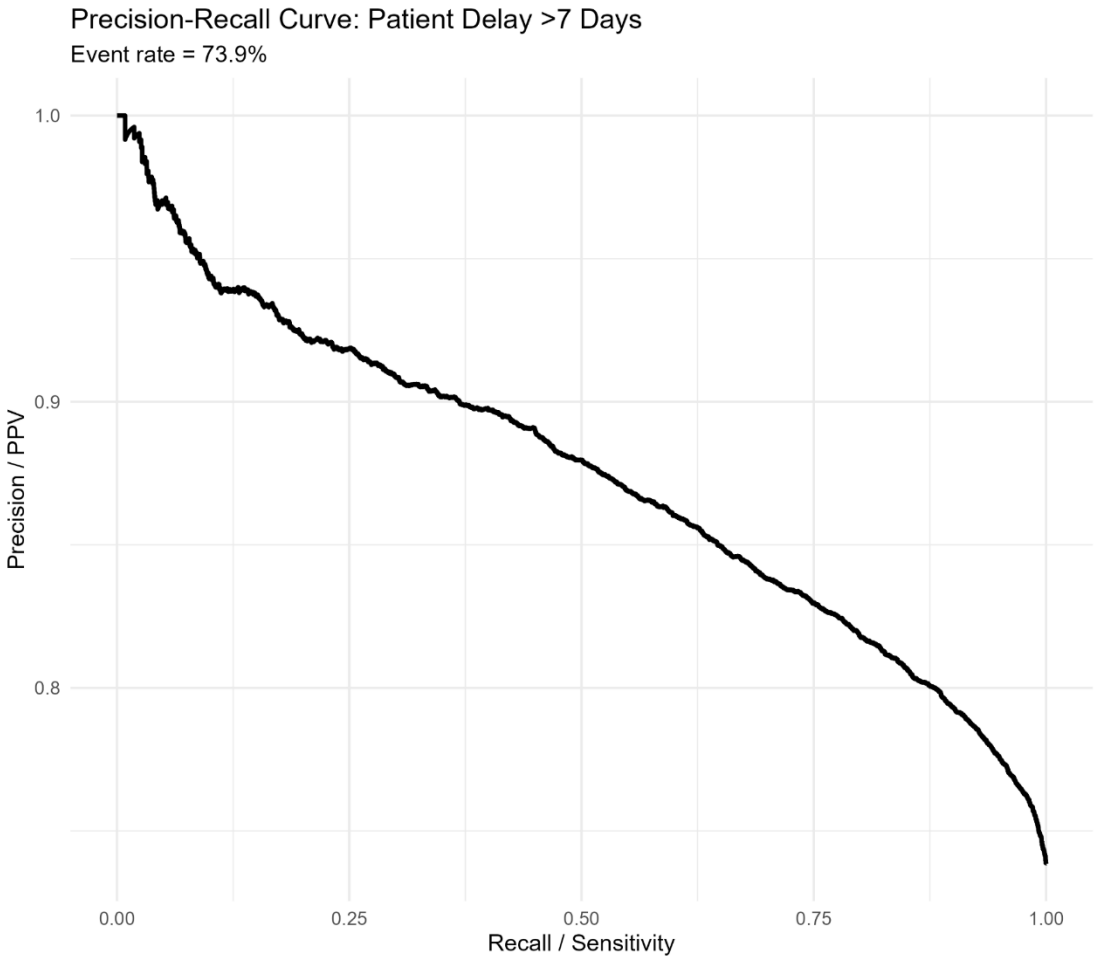
Supplementary Figure S6. ROC curve in the seven-day delay sensitivity analysis



Note. The ROC curve evaluates the LightGBM model after redefining the outcome as patient delay >7 days.

Interpretation. The ROC-AUC of 0.727 suggests that the model retained moderate discrimination under the shorter delay definition. This supports the robustness of the modeling framework to an alternative threshold.

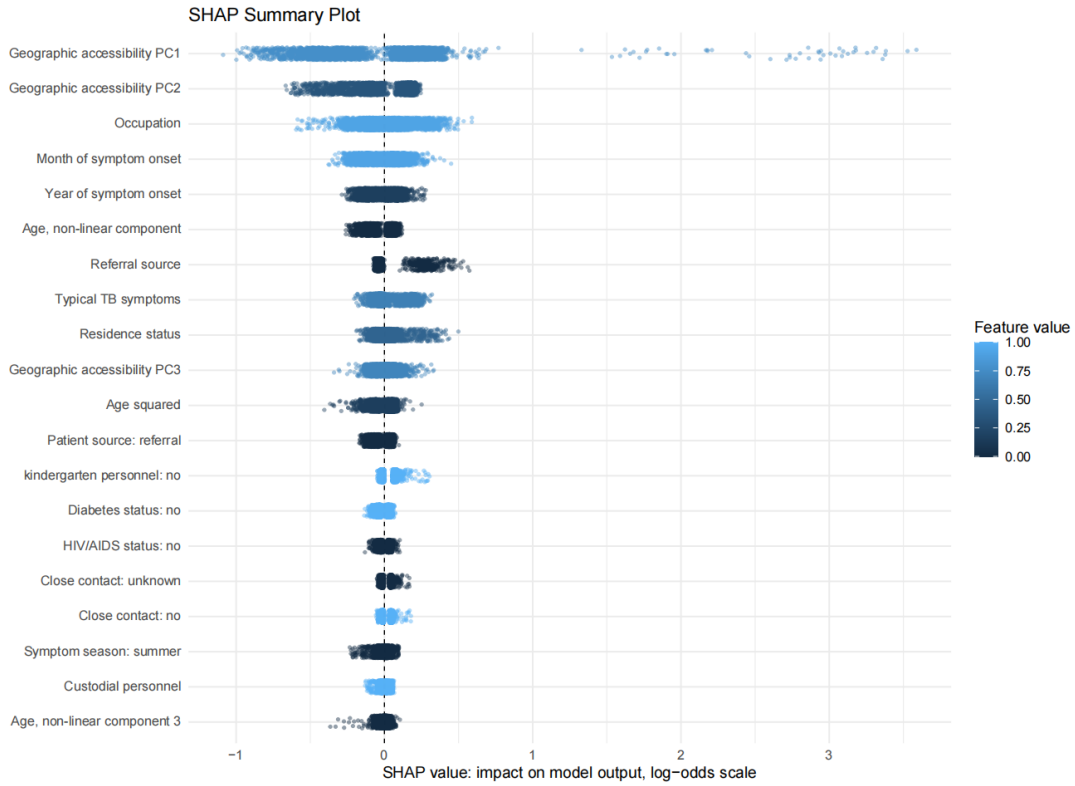
Supplementary Figure S7. Precision-recall curve in the seven-day delay sensitivity analysis



Note. The PR curve summarizes precision and recall for the seven-day delay model. It should be interpreted relative to the 73.9% event rate.

Interpretation. The PR-AUC of 0.874 reflects strong precision-recall performance, but because the event was common under the seven-day definition, the high PR-AUC should be interpreted relative to the high prevalence baseline rather than as evidence of near-perfect prediction.

Supplementary Figure S8. SHAP summary plot / beeswarm



Interpretation. The SHAP beeswarm is the preferred interpretability figure because it simultaneously presents feature importance, direction of association with predicted delay and heterogeneity across patients.

Supplementary Figure S9. Subgroup performance plot

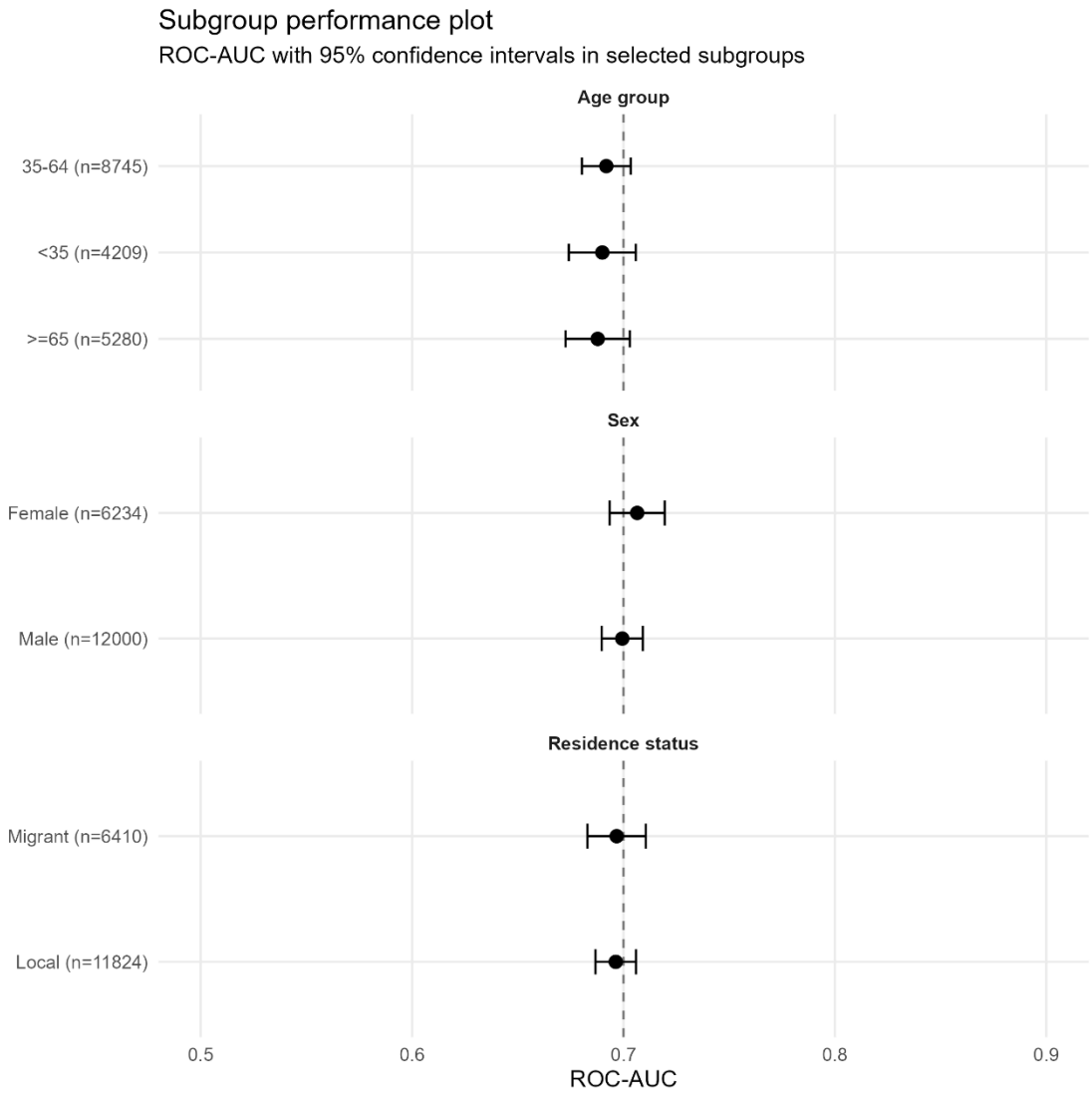


Figure note.

The plot shows ROC-AUC estimates and 95% confidence intervals for the final model across selected subgroups. The dashed vertical line indicates the overall test-set ROC-AUC of the primary model.

Interpretation.

Model performance was generally stable across the examined subgroups. Female patients showed a slightly higher point estimate than male patients, while age and residence-status subgroups showed highly comparable discrimination. The overlapping confidence intervals indicate that these differences should be interpreted cautiously.