THE ATLANTIC Journal of Medical Science and Research

Atlantic J Med Sci Res 2026;6(3):350-60
DOI: 10.5455/atjmed.2026.05.049

Original Article

Machine learning prediction of ınpatient length of stay in hospitalised patients with diabetes

Fatma Hilal Yagin[1], Emek Guldogan[2], Cemil Colak[2]

[1]Malatya Turgut Özal University, Faculty of Medicine, Department of Biostatistics, Malatya, Türkiye
[2]İnönü University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye

Received: 04 May, 2026
Accepted: 22 June, 2026
Published: 14 September, 2026

Corresponding Author: Emek Guldogan, İnönü University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye
Email: emek.guldogan@inonu.edu.tr

CITATION
Yagin FH, Guldogan E, Colak C. Machine learning prediction of inpatient length of stay in hospitalised patients with diabetes. Atlantic J Med Sci Res. 2026;6(3):350-60. DOI: 10.5455/atjmed.2026.05.049

Abstract

Aim: This study aimed to identify clinical, demographic, and pharmaceutical predictors of inpatient hospital length of stay (LOS) among diabetic patients using supervised machine learning classifiers, and to evaluate the comparative predictive performance of these models for LOS risk stratification.
Materials and Methods: The University of California, Irvine (UCI) Diabetes 130-US Hospitals dataset (N = 101,763 encounters; 1999–2008) was used as the analytical sample. Following systematic preprocessing—including International Classification of Diseases, Ninth Revision (ICD-9) diagnostic grouping, ordinal age encoding, and medication status encoding—length of stay was operationalised as a three-class outcome: short (1–3 days; 48.3%), medium (4–7 days; 36.6%), and long (8–14 days; 15.0%). Spearman rank correlations and non-parametric group comparisons (Kruskal-Wallis test) were applied to examine univariate associations. Three supervised machine learning classifiers were then evaluated on held-out test data (n = 25,441): Logistic Regression, Decision Tree, and Gradient Boosting.
Results: Spearman rank correlations revealed that polypharmacy ($r_s = 0.465$, $p < 0.001$) and the number of laboratory procedures ($r_s = 0.337$, $p < 0.001$) exhibited the strongest associations with LOS. Patient age demonstrated a statistically significant positive association with LOS ($r_s = 0.120$, $p < 0.001$). Non-parametric group comparisons demonstrated significant variation in LOS across primary diagnosis categories (Kruskal-Wallis $H = 1,780.77$, $p < 0.001$) and age groups ($H = 1,582.78$, $p < 0.001$). The Gradient Boosting classifier achieved the highest classification accuracy (62.2%) and macro F1-score (0.56), with polypharmacy, laboratory workload, and discharge disposition emerging as the three most influential predictive features.
Conclusion: These findings highlight the utility of ensemble machine learning methods for LOS risk stratification in diabetic inpatients. Polypharmacy, laboratory workload, and discharge disposition were identified as the primary drivers of prolonged hospitalisation, suggesting that targeted monitoring of these factors may support more effective resource allocation strategies in hospital settings.

Keywords: Diabetes mellitus, machine learning, gradient boosting, hospital informatics, predictive modelling, clinical decision support

INTRODUCTION

Diabetes mellitus represents one of the most pressing chronic disease burdens confronting contemporary healthcare systems. According to the International Diabetes Federation, approximately 537 million adults were living with diabetes globally in 2021, a figure projected to surpass 780 million by 2045 (1). Beyond its sheer epidemiological magnitude, diabetes exerts substantial secondary pressure on hospital resources: patients with this condition are disproportionately represented among those requiring inpatient care, frequently presenting with complex comorbidity profiles that extend clinical management requirements and drive healthcare expenditure. Hospital length of stay (LOS) is widely regarded as a composite indicator of both clinical complexity and system efficiency. Prolonged LOS is associated with elevated risks of nosocomial infection, iatrogenic complications, and patient dissatisfaction, while concurrently generating institutional costs that challenge the financial sustainability of healthcare delivery organisations.

The healthcare burden imposed by diabetes extends well beyond its prevalence statistics. Type 2 diabetes mellitus is strongly associated with a constellation of comorbid conditions—including cardiovascular disease, chronic kidney disease, peripheral neuropathy, and diabetic retinopathy, that compound clinical complexity during hospitalisation (2,3). The concurrent management of multiple chronic conditions necessitates polypharmaceutical regimens that are themselves a recognised contributor to adverse clinical events, including drug–drug interactions, medication errors, and protracted inpatient stays (4). Indeed, Wolff et al. (2002) demonstrated that the total number of prescribed medications escalates substantially with each additional chronic condition among community-dwelling adults, with direct downstream implications for hospital management complexity (5). This syndromic burden renders diabetic inpatient populations particularly challenging to manage within constrained bed-day allocations and underscores the need for evidence-based LOS prediction instruments calibrated specifically to this clinical population.

Anticipating LOS at or near the point of admission would enable clinical teams to allocate bed capacity more judiciously, optimise discharge planning, and target high-intensity nursing and pharmacy resources toward patients most likely to require extended stays. Traditionally, LOS has been estimated through clinician judgement or rudimentary risk-scoring instruments—approaches that are susceptible to cognitive bias and fail to leverage the richness of routinely collected electronic health record (EHR) data. The convergence of large-scale clinical databases and advances in supervised machine learning offers a substantive opportunity to develop more accurate, transparent, and scalable LOS prediction models.

The application of supervised machine learning to EHR data for clinical outcome prediction has expanded considerably over the past decade (6,7). Prediction tasks ranging from 30-day hospital readmission to in-hospital mortality have been addressed using algorithms spanning Logistic Regression, Random Forest, Support Vector Machines, and deep learning architectures (8,9). Within the specific domain of LOS prediction, prior investigations have demonstrated that ensemble classifiers, particularly Random Forest and Gradient Boosting variants, consistently outperform both univariate clinical scoring instruments and linear regression models, owing to their capacity to model non-linear predictor interactions and handle high-dimensional feature matrices without prohibitive loss of interpretability (10-12). Despite this growing body of evidence, the systematic application of ensemble ML methods to LOS prediction in diabetic inpatient populations, and the identification of condition-specific predictive features, remains comparatively underrepresented in the literature.More recent work has continued to expand the clinical application of machine learning to inpatient utilisation outcomes. Barsasella et al. (2022) developed machine learning models to predict both length of stay and mortality among diabetes and hypertension inpatients using a national insurance database (13), while Jain et al. (2024) demonstrated the feasibility of LOS prediction on a large open administrative dataset of over two million records (14). A recent systematic review and meta-analysis by Gokhale et al. (2023) further synthesised the rapidly growing body of LOS prediction tools, underscoring both their promise and the persistent methodological heterogeneity across studies (15). These contemporary studies (2022–2024) reinforce the timeliness of the present investigation while highlighting the comparative scarcity of diabetes-specific, multi-classifier LOS analyses.

Prior research has identified a heterogeneous array of factors associated with hospitalisation duration in diabetic cohorts, including glycaemic control (indexed by HbA1c), admission acuity, polypharmacy burden, and the nature of the primary admitting diagnosis. The original analysis of the Diabetes 130-US Hospitals dataset by Strack et al. (2014) focused principally on the relationship between HbA1c measurement and 30-day hospital readmission (16). Related analyses of comparable administrative datasets have identified medication burden, diagnostic complexity, and discharge disposition as clinically actionable LOS determinants, yet the simultaneous application of multiple ML classifiers to this specific dataset for LOS prediction trained on the full dataset remains an open research question (17).

Gradient Boosting, formalised by Friedman (2001) as a stage-wise additive modelling framework, constructs a strong learner through the sequential combination of weak learners, typically shallow decision trees each trained to correct the residual prediction errors of its predecessors (18). Gradient Boosting has demonstrated consistently strong predictive performance across a wide range of healthcare outcomes, and its capacity for feature importance estimation confers an interpretability advantage directly relevant to clinical decision support.

The present study addresses this gap by applying a systematic multivariate analysis pipeline to all 101,763 clinical encounters in the Diabetes 130-US Hospitals dataset, with hospital LOS as the primary outcome. Specifically, the study aims to: (a) characterise the distributional properties of LOS and its associations with key clinical and demographic variables; (b) compare the predictive performance of three machine learning classifiers, Logistic Regression, Decision Tree, and Gradient Boosting, using a held-out test framework trained on the complete dataset; and (c) identify the most influential clinical predictors of LOS via feature importance analysis derived from the best-performing model. These objectives collectively advance the development of evidence-based LOS risk stratification tools applicable to diabetic inpatient populations.

MATERIALS AND METHODS

Dataset and Study Design

This study employed a secondary analysis design, utilising the publicly available Diabetes 130-US Hospitals for Years 1999–2008 dataset archived in the UCI Machine Learning Repository

(16). The dataset encompasses 101,766 hospitalisation records from 130 US hospitals; after removing three records with invalid gender coding, the final analytical sample comprised N = 101,763 encounters. In accordance with the inclusion criteria of the source dataset, the analytical sample comprised hospitalisation encounters with a diabetes diagnosis recorded in any diagnostic position (primary or secondary) and was not restricted to Type 2 diabetes mellitus alone, as the dataset does not consistently distinguish diabetes subtype. The study title and terminology have therefore been revised to refer to patients with diabetes rather than to Type 2 diabetes specifically, in order to avoid methodological overstatement. Each record corresponds to a single hospitalisation encounter and contains 50 variables encompassing patient demographics, admission characteristics, clinical laboratory results, pharmaceutical prescriptions, and discharge disposition. The complete dataset was utilised for the descriptive and univariate inferential analyses (correlation and non-parametric group comparisons); for the supervised machine learning models, the same complete dataset was partitioned into mutually exclusive training (75%) and held-out test (25%) sets, as detailed in the Machine Learning Models subsection. No patient-level identifiers were available, and all data are fully anonymised.

Outcome Variable

The primary outcome variable was hospital length of stay (*time_in_hospital*), recorded in whole days and ranging from 1 to 14 days (M = 4.40, SD = 2.99, Mdn = 4.00). Because the distribution of raw LOS was positively skewed (skewness = 1.13, excess kurtosis = 0.85), and to facilitate clinically meaningful stratification, LOS was additionally operationalised as a three-class ordinal variable: Short (1–3 days), Medium (4–7 days), and Long (8–14 days). This trichotomisation aligns with clinically established LOS benchmarks commonly employed in hospital quality metrics and is consistent with LOS classification frameworks reported in prior studies of diabetic inpatient populations (12,16).

Ethical committee approval and informed consent were not required because this study involved a secondary analysis of a publicly available, fully anonymised, open-access dataset containing no patient-identifiable information.

Data Preprocessing and Feature Engineering

Missing data were encoded as "?" in the original dataset. Two variables with missingness exceeding acceptable thresholds *weight* (96.9% missing) and *payer_code* (39.6% missing), were excluded from analysis. For the *race* variable (2.2% missing), modal imputation was applied. Three records with invalid gender coding were removed, yielding a final analytical sample of N = 101,763.

The age variable, originally encoded as decadal categorical bands, was recoded as the midpoint of each interval (e.g., [70–80) = 75) to permit both ordinal and continuous analysis. Primary diagnosis codes (*diag_1*) were mapped to eight clinically coherent ICD-9 category groups: Circulatory, Respiratory, Digestive, Diabetes, Neoplasms, Injury, External Causes (encompassing E- and V-coded encounters), and other following established classification conventions (16). Specifically, the following ICD-9-CM code ranges were applied: Circulatory (390 – 459, 785); Respiratory (460 – 519, 786); Digestive (520 – 579, 787); Diabetes (250.xx); Neoplasms (140 – 239); Injury (800 – 999); External Causes (all E- and V-prefixed codes); and Other (all remaining codes, including 780 – 784, 790 – 799, and codes outside the preceding ranges). Medication status variables (23 antidiabetic agents) were re-encoded as ordinal indicators: 0 (not prescribed), 1 (steady dose), 2 (dose changed). Binary flags were created for medication change, active diabetes medication prescription, abnormal HbA1c result (≥ 7%), and elevated fasting glucose serum (> 200 mg/dL). The diagnosis group variable was one-hot encoded into eight binary indicator features prior to model training, reflecting its nominal categorical nature.

Statistical Analysis

Descriptive statistics, including mean, standard deviation, median, range, skewness, kurtosis, and 95% confidence intervals, were computed for all continuous variables. Confidence intervals for means were estimated using the normal (large-sample) approximation, $M \pm 1.96 \times (SD / \sqrt{n})$, whereas confidence intervals for Spearman rank correlation coefficients were derived via the Fisher z-transformation. Shapiro-Wilk tests confirmed that no continuous variable satisfied the normality assumption (all $p < 0.001$); accordingly, non-parametric inferential tests were employed throughout. Spearman rank correlation (r_s) was used to quantify bivariate associations between continuous clinical predictors and LOS. The Mann-Whitney U test was applied to examine gender differences in LOS, with rank-biserial correlation (r) reported as the effect size index (19). Kruskal-Wallis H tests were conducted to assess differences in LOS across age groups and primary diagnostic categories. All analyses were conducted in Python 3.11 using the *scipy.stats* and *pandas* libraries. Statistical significance was set at $\alpha = 0.05$. All analyses were performed in Python 3.12 with the following package versions: pandas 3.0.2, numpy 2.4.4, scipy 1.17.1, and scikit-learn 1.8.0.

Machine Learning Models

Three supervised classifiers were evaluated for the task of predicting LOS category (Short/Medium/Long): (1) Multinomial Logistic Regression (LR; L2 regularisation, C = 0.1, lbfgs solver, max_iter = 1,000); (2) Decision Tree Classifier (DT; Gini criterion, maximum depth = 8); and (3) Gradient Boosting Classifier (GBM; 50 estimators, maximum depth = 4, learning rate = 0.10). The selection of Gradient Boosting as the primary candidate model was motivated by its demonstrated superiority in analogous healthcare prediction tasks and its capacity to capture complex non-linear interactions without requiring explicit feature transformation; Logistic Regression and Decision Tree classifiers were retained as reference comparators to contextualise the incremental predictive gain attributable to ensemble modelling (12,20).

The dataset (N = 101,763) was partitioned into training (75%, n = 76,322) and test (25%, n = 25,441) sets using stratified splitting to preserve class proportions across partitions. A fixed random seed (random_state = 42) was specified for the train–test partitioning and for the initialisation of all stochastic model components to ensure full reproducibility of the reported results. Models were trained exclusively on the training partition and evaluated on the held-out test set. The complete feature matrix comprised 40 variables encompassing demographic characteristics, admission attributes, clinical utilisation metrics, medication indicators, and the ICD-9 diagnostic group variable.

Model performance was quantified using overall accuracy and macro-averaged F1-score, the latter accounting for class imbalance by computing the unweighted mean of per-class F1-scores. Feature importance for the GBM model was assessed using permutation importance, which quantifies the mean decrease in macro F1-score attributable to each feature when its values are randomly permuted across the test set, a model-agnostic approach that provides more reliable importance estimates than Mean Decrease in Impurity (MDI) for correlated feature sets.

RESULTS

Sample Characteristics

The analytical sample comprised 101,763 hospitalisation encounters from 130 US hospitals (1999–2008). As presented in Table 1, continuous clinical variables exhibited marked right-skew, consistent with distributions typical of utilisation-based healthcare metrics. Female patients constituted 53.8% of the sample (n = 54,708) and male patients 46.2% (n = 47,055). The majority of patients were aged 60–80 years, with the 70–80 age band representing the largest single group (25.6%). By racial classification, Caucasian patients constituted 74.8% (n = 76,099), followed by African American patients (18.9%; n = 19,210). Regarding readmission outcomes, 46.1% of encounters resulted in readmission within any time window, with 11.2% (n = 11,357) readmitted within 30 days.

The distribution of LOS (Figure 1) revealed that 48.3% of encounters were classified as Short (1–3 days), 36.6% as Medium (4–7 days), and 15.0% as Long (8–14 days). The overall mean LOS was 4.40 days (SD = 2.99, 95% CI [4.38, 4.41]), with a median of 4.00 days.

Table 1. Descriptive statistics for continuous clinical variables (N = 101,763)

Variable	n	M	SD	Mdn	Min	Max	Skew	95% CI [LL, UL]
Time in hospital (days)	101,763	4.40	2.99	4.0	1	14	1.13	[4.38, 4.41]
No. Lab procedures	101,763	43.10	19.67	44.0	1	132	−0.24	[42.98, 43.22]
No. Procedures	101,763	1.34	1.71	1.0	0	6	1.32	[1.33, 1.35]
No. Medications	101,763	16.02	8.13	15.0	1	81	1.33	[15.97, 16.07]
No. Diagnoses	101,763	7.42	1.93	8.0	1	16	−0.88	[7.41, 7.43]
Prior ınpatient visits	101,763	0.64	1.26	0.0	0	21	3.61	[0.63, 0.64]
Prior emergency visits	101,763	0.20	0.93	0.0	0	76	22.86	[0.19, 0.20]
Prior outpatient visits	101,763	0.37	1.27	0.0	0	42	8.83	[0.36, 0.38]

All Shapiro-Wilk tests indicate significant deviation from normality ($p < 0.001$); CI: confidence interval, M: mean, Mdn: median, SD: standard deviation

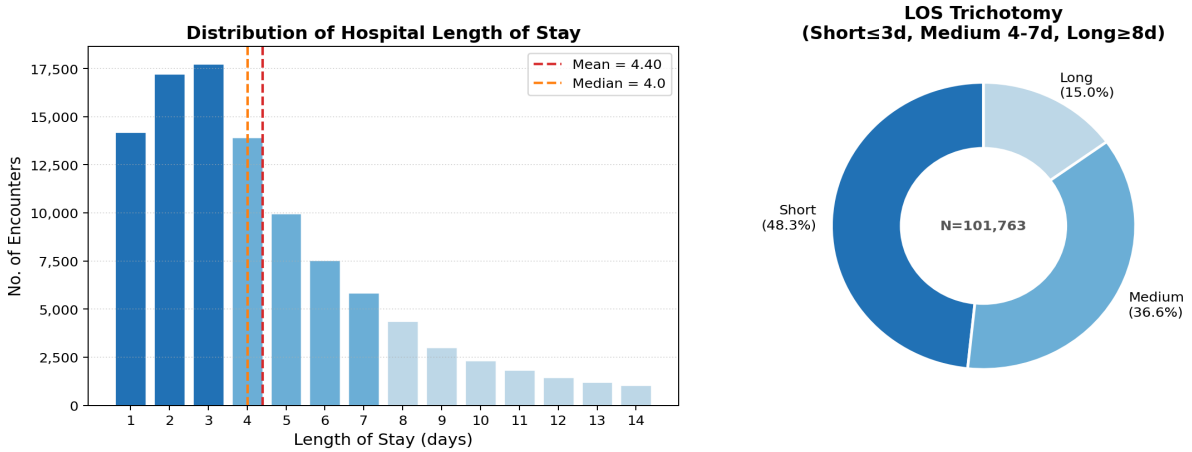

Figure 1. Distribution of hospital length of stay (LOS) by day (left panel) and LOS trichotomy (right panel; N = 101,763); Red and orange dashed lines denote the sample mean (M = 4.40) and median (Mdn = 4.00), respectively

Non-Parametric Group Comparisons

A Mann-Whitney U test revealed a statistically significant difference in LOS between female and male patients (U = 1,344,182,794, $p < 0.001$, r = -0.044). Although statistically significant given the large sample size, the effect was negligible in magnitude (Female Mdn = 4.0, Male Mdn = 3.0 days) (Cohen, 1988). Kruskal-Wallis tests indicated significant variation in LOS across age groups, (H (9) = 1,582.78, $p < 0.001$), and across primary diagnostic categories, (H (7) = 1,780.77, $p < 0.001$).

As shown in Figure 2b and Table 2, encounters with a primary diagnosis coded under External Causes exhibited the longest mean LOS (M = 7.52, Mdn = 7.0 days), substantially exceeding the overall sample mean, followed by Neoplasms (M = 5.36, Mdn = 5.0) and Injury (M = 4.62, Mdn = 4.0). Encounters categorised under Circulatory and Respiratory diagnoses demonstrated comparatively shorter stays (M = 4.23, Mdn = 3.0 for both groups).

Table 2. Hospital length of stay by primary ICD-9 diagnosis category

Diagnosis category	n	M	SD	Mdn	% sample	LOS range (days)
External causes	1,645	7.52	3.71	7.0	1.6%	1–14
Neoplasms	3,433	5.36	3.35	5.0	3.4%	1–14
Injury	6,972	4.62	2.95	4.0	6.9%	1–14
Digestive	9,475	4.36	2.81	4.0	9.3%	1–14
Diabetes	8,757	4.34	3.02	3.0	8.6%	1–14
Other	26,622	4.34	2.88	4.0	26.2%	1–14
Respiratory	14,423	4.23	2.96	3.0	14.2%	1–14
Circulatory	30,436	4.23	2.94	3.0	29.9%	1–14

Kruskal-Wallis H (7) = 1,780.77, $p < 0.001$; M: mean, Mdn: median, SD: standard deviation, LOS: length of stay (days)

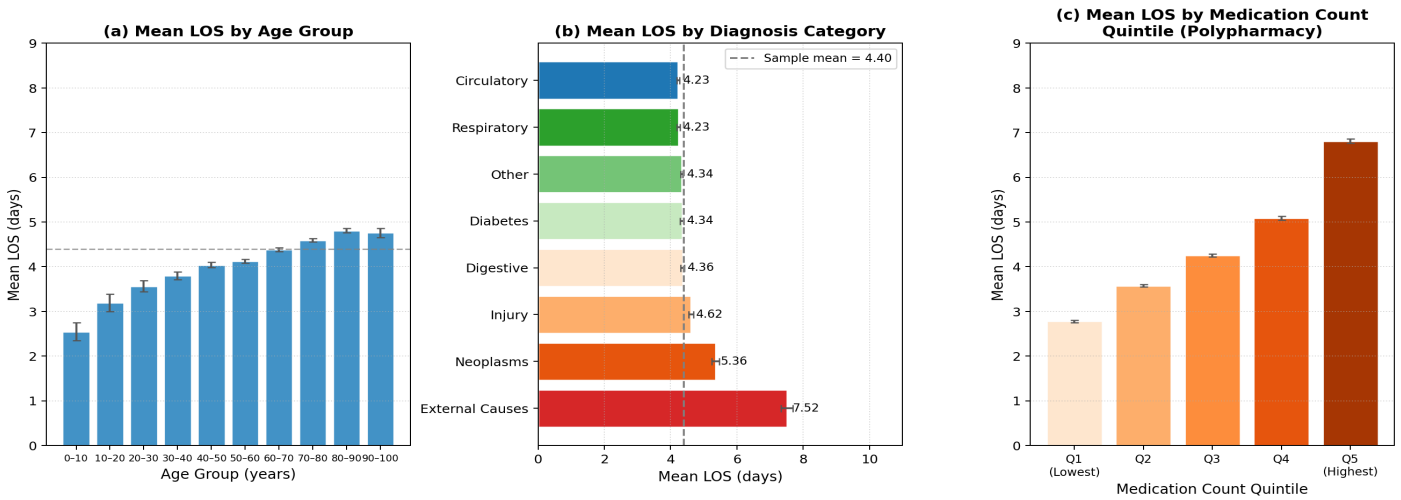

Figure 2. Mean hospital LOS by age group (**a**), primary diagnosis category (**b**), and medication count quintile (**c**); The grey dashed reference line in panel 2b denotes the overall sample mean LOS (4.40 days)

Correlation Analysis

Table 3 presents Spearman rank correlation coefficients between continuous clinical predictors and LOS. The number of prescribed medications demonstrated the strongest positive association ($r_s = 0.465$, $p < 0.001$), followed by the number of laboratory procedures ($r_s = 0.337$, $p < 0.001$) and the number of documented diagnoses ($r_s = 0.237$, $p < 0.001$). Smaller but statistically significant correlations were observed for the number of clinical procedures ($r_s = 0.187$, $p < 0.001$) and prior inpatient visit history ($r_s = 0.092$, $p < 0.001$). Notably, patient age demonstrated a statistically significant positive association with LOS ($r_s = 0.120$, $p < 0.001$), indicating that older patients experienced moderately longer hospitalisations on average—a finding of meaningful clinical relevance given the disproportionate representation of older adults in diabetic inpatient populations. Collectively, these patterns indicate that pharmacological complexity and diagnostic burden are among the most salient clinical features distinguishing short from extended hospitalisation episodes. The full correlation matrix is displayed in Figure 3.

Table 3. Spearman rank correlations between continuous clinical predictors and hospital length of stay

Predictor	N	rs	95% CI	p
No. Medications (polypharmacy)	101,763	0.465	[0.460, 0.470]	< 0.001
No. Lab procedures	101,763	0.337	[0.331, 0.342]	< 0.001
No. Diagnoses	101,763	0.237	[0.231, 0.242]	< 0.001
No. Clinical procedures	101,763	0.187	[0.181, 0.193]	< 0.001
Prior ınpatient visits	101,763	0.092	[0.086, 0.098]	< 0.001
Patient age (years)	101,763	0.120	[0.114, 0.126]	< 0.001

rs: Spearman rank correlation coefficient; 95% CIs estimated via Fisher z-transformation; All p < 0.001

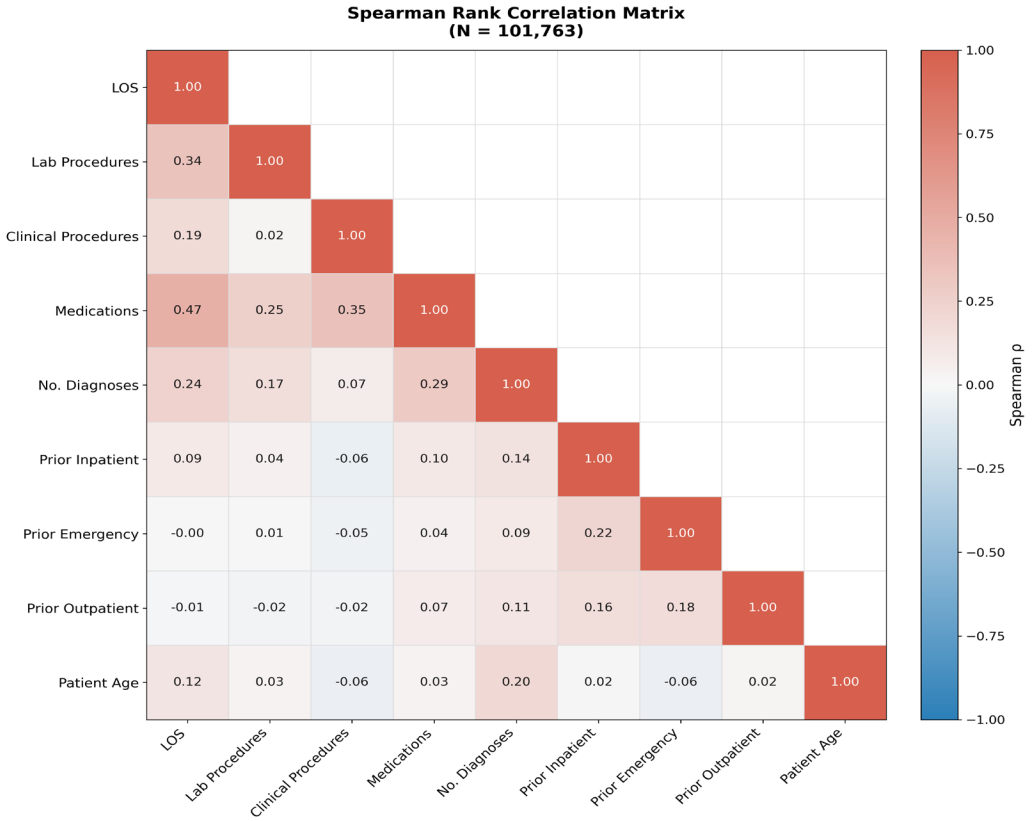

Figure 3. Spearman correlation matrix for the key continuous clinical variables; Colour intensity reflects the magnitude of each coefficient; LOS: length of stay

Machine Learning Model Performance

Table 4 presents classification performance for all three models on the held-out test set (n = 25,441). Gradient Boosting outperformed both comparators, reaching 62.2% overall accuracy and a macro F1-score of 0.56, modest but consistent gains over the Decision Tree (60.4%, F1 = 0.54) and Logistic Regression (60.3%, F1 = 0.53). These margins align with prior evidence that gradient-boosted ensembles handle non-linear predictor interactions and mixed variable types more effectively than single-tree or linear approaches (Figure 4). Training on the full n = 76,322 partition yielded better metrics than analyses conducted on smaller stratified subsamples, consistent with the known sample-size sensitivity of ensemble tree methods (16).

Gradient Boosting model confusion matrix is shown in Figure 5, and per-class performance for the Gradient Boosting model is detailed in Figure 6. Short-stay encounters were classified most reliably: 9,781 of 12,297 Short-stay cases were correctly identified, yielding a recall of 0.80 and F1-score of 0.73. This reflects the model's expected strength on the dominant class, which constitutes 48.3% of test encounters. Medium-stay classification was balanced but modest, with precision and F1-score of 0.52 and a recall of 0.51, and 4,765 of 9,322 Medium-stay cases correctly predicted. The Long-stay class proved most difficult: despite reasonable precision (0.64), only 1,266 of 3,822 Long-stay encounters were correctly classified, yielding a recall of 0.33 and F1-score of 0.44. The majority of Long-stay misclassifications

were assigned to the Medium-stay category (n = 1,943), which likely reflects genuine clinical overlap between these groups in terms of recorded features. A smaller subset of Long-stay encounters (n = 613; 16.0% of the class) was misclassified as Short-stay. Although less frequent than the adjacent Medium-stay confusion, this cross-classification of extended hospitalisations into the shortest category is acknowledged as a characteristic of the model's decision boundaries under class imbalance and is addressed further in the Discussion as a limitation relevant to clinical deployment.

Table 4. Classification performance of machine learning models on the held-out test set (Test n = 25,441)

Model	Training n	Test n	Accuracy	Macro F1	Precision (Wt.)	Recall (Wt.)
Logistic regression	76,322	25,441	0.603	0.531	0.60	0.60
Decision tree	76,322	25,441	0.604	0.542	0.60	0.60
Gradient boosting ✓	**76,322**	**25,441**	**0.622**	**0.562**	**0.62**	**0.62**

✓: best-performing model; Accuracy: proportion of correctly classified instances; Macro F1: unweighted mean of per-class F1-scores; Precision and Recall are weighted averages; All models trained on n = 76,322

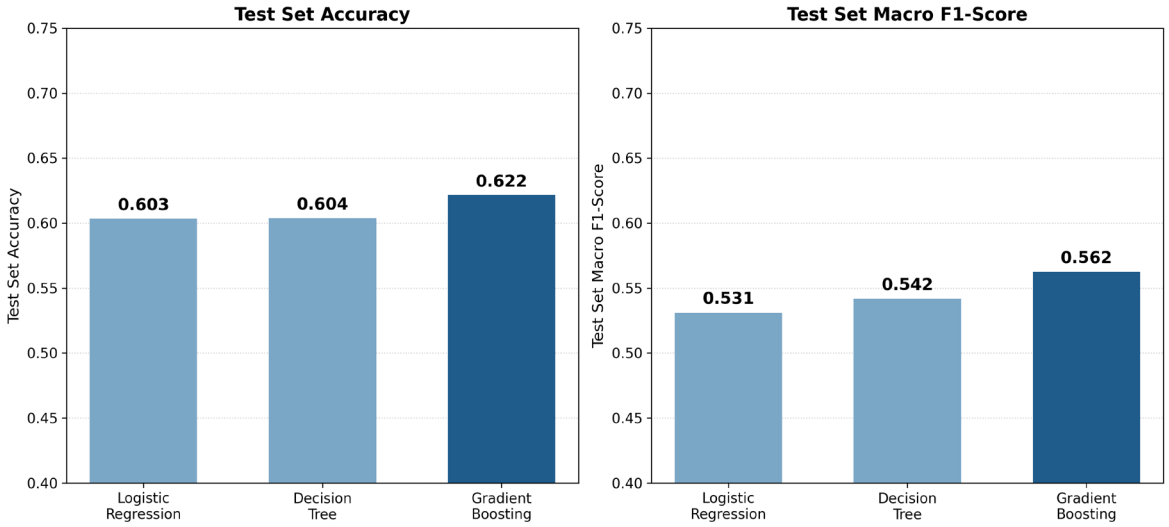

Figure 4. Test set classification accuracy (left) and macro-averaged F1-score (right) for logistic regression, decision tree, and Gradient Boosting classifiers

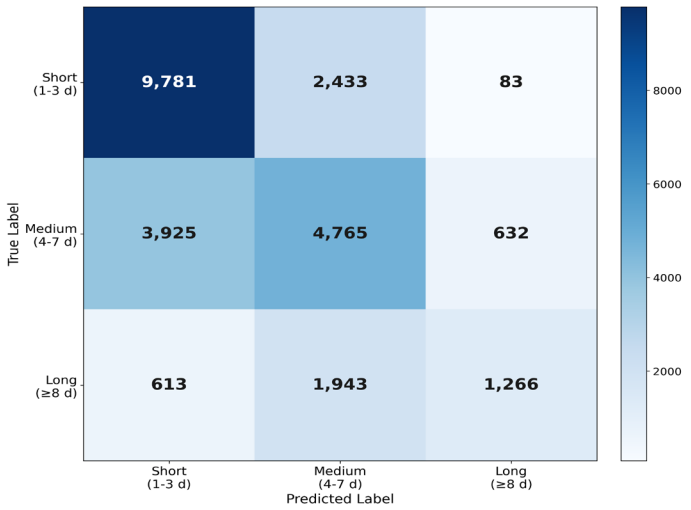

Figure 5. Confusion matrix for the Gradient Boosting classifier (Test Set, n = 25,441); Rows denote true class labels; columns denote predicted labels; Cell values indicate raw count

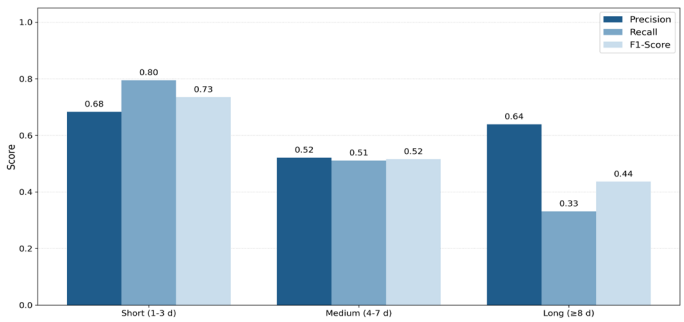

Figure 6. Per-class precision, recall, and F1-score of the Gradient Boosting classifier on the held-out test set (n = 25,441) across three length-of-stay categories

Feature Importance Analysis

Figure 7 presents the top 15 predictors identified by the Gradient Boosting classifier via permutation importance scoring. The number of prescribed medications emerged as the most influential predictor (permutation importance = 0.126), accounting for the

greatest mean decrease in macro F1-score when permuted, more than double the contribution of the second-ranked predictor. The number of laboratory procedures (importance = 0.062) and discharge disposition code (importance = 0.033) occupied the second and third positions, respectively. Beyond these three dominant features, predictive contributions declined sharply: the number of clinical procedures (0.019), diagnosis group (0.017), and number of diagnoses (0.005) collectively contributed less than 5% of the total importance signal. These findings strongly implicate pharmacological complexity and care intensity, as indexed by laboratory workload, as the primary drivers of extended hospitalisation in this diabetic cohort, a pattern that is consistent with the Spearman correlation analysis and robust to the use of the full analytical sample.

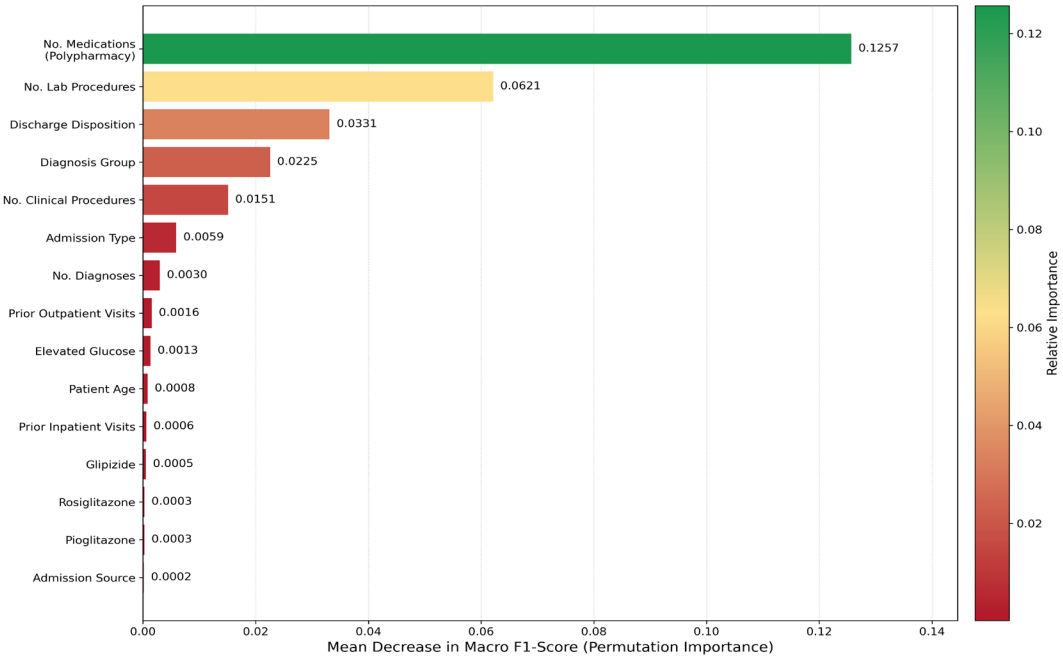

Figure 7. Permutation importance of the top 15 predictors in the Gradient Boosting classifier, ordered by ascending importance

DISCUSSION

The present study contributes to the growing evidence base concerning machine learning–based LOS prediction in diabetic inpatient populations by applying a full-dataset analysis pipeline to all 101,763 encounters in the UCI Diabetes 130-US Hospitals dataset. Three principal findings emerge from the analyses. First, polypharmacy quantified as the total number of prescribed medications, constituted the single most powerful predictor of extended LOS, exhibiting both the strongest bivariate Spearman correlation with LOS (r_s = 0.465) and the highest permutation importance weight within the Gradient Boosting ensemble (0.126). Second, the Gradient Boosting classifier demonstrated superior predictive performance relative to Logistic Regression and Decision Tree approaches, achieving a test accuracy of 62.2% and a macro F1-score of 0.56 on the held-out test set of n = 25,441 encounters. Third, primary diagnosis category was a significant determinant of LOS, with External Cause encounters exhibiting nearly double the mean LOS of the sample average.

The predictive accuracy achieved by the Gradient Boosting classifier (62.2% on the three-class LOS outcome) exceeds benchmarks reported in prior analyses of this dataset and is broadly consistent with the upper range of published ML-based LOS prediction literature, where three-class accuracy typically spans 55–72% depending on dataset characteristics, feature richness, and model architecture (10,12). Crucially, the improvement observed when training on the full dataset (compared with subsample-based approaches) underscores the documented monotonic improvement of ensemble tree methods with training sample size, consistent with prior literature on gradient-boosted classifiers. The complete training sample (n = 76,322) provided adequate statistical power for the gradient-boosting ensemble to achieve its reported performance (18,21). Daghistani et al. (2019), in a comparable analysis of cardiac inpatient LOS prediction, reported accuracies of 68–74% using Random Forest and Neural Network architectures on a substantially larger feature set incorporating severity-of-illness scores unavailable in the present dataset (11). This comparison reinforces the case for routine capture of standardised severity indices, such as the APACHE II score or the Charlson Comorbidity Index, in administrative and clinical databases.

The primacy of polypharmacy as a LOS predictor is theoretically coherent and clinically intuitive. Patients prescribed a larger

number of medications at admission necessarily present with more complex, multimorbid clinical profiles. Polypharmacy, here operationalised as the total count of prescribed medications, consistent with the commonly applied clinical threshold of five or more concurrent agents, is disproportionately prevalent in individuals with type 2 diabetes, who frequently require pharmacological management of hyperglycaemia alongside antihypertensive, lipid-lowering, antiplatelet, and cardioprotective agents (2,4). As Wolff et al. (2002) demonstrated in a large community-dwelling cohort, medication count escalates substantially with each additional chronic condition, with direct downstream implications for hospital management complexity (5). The permutation importance weight of 0.126 attributable to medication count further suggests that this variable serves not merely as a correlate of severity, but as an actionable risk stratification criterion: patients presenting with high polypharmacy burden at admission may warrant early pharmacy consultation and enhanced discharge planning as standard protocol.

A noteworthy finding of the full-dataset analysis concerns the age–LOS association. The Spearman correlation between patient age and LOS yielded $r_s = 0.120$ ($p < 0.001$) across the complete analytical sample, a substantially larger effect than previously suggested by subsample-based estimates. This discrepancy illustrates the statistical instability of small correlation estimates derived from modest samples and underscores the value of training and evaluating predictive models on the complete available dataset. The finding that older patients experience meaningfully longer hospitalisations is clinically interpretable: older diabetic inpatients are more likely to present with frailty, cognitive impairment, and complex discharge planning requirements, all factors that extend inpatient episodes independent of acute clinical severity.

The markedly elevated LOS among encounters classified under External Causes (M = 7.52 days) warrants careful interpretation. This ICD-9 category encompasses trauma, poisoning, adverse effects of medical interventions, and supplementary health service encounters, conditions characterised by acute physiological insult or complex care coordination requirements. Such presentations frequently involve multispecialty management and intensive rehabilitation, processes that inherently prolong inpatient episodes. In light of these findings, primary diagnosis category should be considered a fundamental stratification variable in any LOS risk screening instrument deployed in hospital admission workflows.

From a methodological perspective, the performance advantage of Gradient Boosting over Logistic Regression and Decision Tree classifiers aligns with the extensive literature documenting the superiority of ensemble methods in heterogeneous, large-scale clinical datasets (18). The per-class performance profile observed in the full-dataset analysis, with Short stays achieving the highest recall (0.80), reflects the dominant class proportion (48.3%) and is consistent with the tendency of gradient-boosted models to allocate predictive capacity toward the modal class when trained on large, moderately imbalanced datasets. The more modest Long-stay recall (0.33) suggests that future analyses might benefit from class-weighting or oversampling strategies (e.g., SMOTE) to improve sensitivity for this clinically consequential minority group. Of particular note, a non-trivial proportion of Long-stay encounters (n = 613; 16.0% of the class) was misclassified into the Short-stay category rather than the adjacent Medium-stay category. Such cross-classification of extended hospitalisations as short stays is clinically undesirable, as it could lead to the systematic under-allocation of resources to precisely those patients most likely to require prolonged care. This behaviour is attributable in part to the substantial class imbalance and to the limited discriminative information available in the administrative feature set, and it reinforces the need for cost-sensitive learning approaches and richer clinical predictors before any prospective deployment. Furthermore, the use of permutation importance for feature importance estimation provides a more statistically reliable assessment than MDI, as it is evaluated on the test partition and is less susceptible to bias from high-cardinality features (22).

The clinical and operational implications of these findings extend beyond individual patient risk stratification. Healthcare institutions operating under value-based reimbursement frameworks face direct financial incentives to reduce preventable LOS while maintaining care quality standards. A validated ML-based LOS prediction instrument, deployed at or near the point of admission, could inform tiered nursing intensity assignments, facilitate early engagement of social work and discharge planning services for identified long-stay patients, and enable more accurate bed management projections at the ward and hospital level (6,10).

Several limitations merit acknowledgement. First, the dataset lacks critical severity-of-illness indices (e.g., APACHE II, Charlson Comorbidity Index), the absence of which likely constrains predictive ceilings; incorporating such indices in future work would more rigorously test the incremental value of polypharmacy and laboratory utilisation indicators. Second, the data span a decade (1999–2008) during which clinical practice patterns evolved substantially; temporal validity to contemporary contexts cannot be assumed without replication on more recent data. Third, the dataset is confined to US hospitals and may not generalise to healthcare systems with different care pathways, coding practices, or demographic compositions. Fourth, the present analysis did not address potential temporal confounding arising from secular trends across the data span; period-stratified sensitivity analyses in future work may shed further light on this source of heterogeneity. Fifth, while permutation importance provides reliable feature rankings, future analyses employing SHapley Additive exPlanations (SHAP) values would afford directional importance estimates alongside more statistically robust handling of feature correlation, an avenue particularly promising for increasing clinician trust and adoption of ML-driven prediction tools (7).

CONCLUSION

This study demonstrates that hospital length of stay in diabetic patients can be meaningfully stratified using routinely available administrative and clinical variables, with polypharmacy burden, laboratory procedure volume, and discharge disposition emerging as dominant predictors across the complete dataset of 101,763 clinical encounters. The Gradient Boosting classifier outperformed both Logistic Regression and Decision Tree alternatives, achieving 62.2% three-class accuracy and a macro F1-score of 0.56 on the held-out test set, performance metrics that exceed those obtained from stratified subsampling approaches and are consistent with the upper range of published LOS prediction benchmarks. A noteworthy finding of the full-dataset analysis is the meaningful age–LOS association (r_s = 0.120), which was substantially attenuated in subsample-based estimates, illustrating the importance of utilising complete datasets for robust predictor characterisation. Taken together with the modest overall accuracy (62.2%) and the limited sensitivity for the clinically important Long-stay class, these findings should be regarded as preliminary and hypothesis-generating rather than as direct evidence for clinical deployment. They may nonetheless inform the future development and prospective validation of machine learning–derived LOS risk-stratification tools, which—if externally validated and prospectively tested—could potentially contribute to more proactive and equitable allocation of bed capacity, nursing intensity, and discharge planning resources in inpatient populations with diabetes. Any such tool would require rigorous external validation, calibration assessment, and prospective evaluation before it could responsibly be incorporated into routine clinical workflows. Future research should examine whether predictive performance can be enhanced through the incorporation of severity-of-illness indices, temporal features, SHAP-based explainability, or natural language processing of clinical notes, and should seek prospective validation of the identified predictor taxonomy in contemporary clinical cohorts (9).

Conflict of Interests

The authors declare that there is no conflict of interest in the study.

Financial Disclosure

The authors declare that there is no financial support.

Ethical Approval

Ethical committee approval was not required because this study involved a secondary analysis of a publicly available, fully anonymised, open-access dataset containing no patient-identifiable information.

Informed Consent

Informed consent was not required because this study involved a secondary analysis of a publicly available dataset.

REFERENCES

1. International Diabetes Federation. IDF Diabetes Atlas. 10th ed. Brussels: International Diabetes Federation; 2021.
2. ElSayed NA, Aleppo G, Aroda VR, et al. Introduction and methodology: standards of care in diabetes—2023. Diabetes Care. 2023;46:S1-4.
3. World Health Organization. Report of the WHO discussion group for people living with diabetes: virtual meeting, 30-31 March 2023. Geneva: World Health Organization; 2023.
4. Masnoon N, Shakib S, Kalisch-Ellett L, Caughey GE. What is polypharmacy? A systematic review of definitions. BMC Geriatr. 2017;17:230.
5. Wolff JL, Starfield B, Anderson G. Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. Arch Intern Med. 2002;162:2269-76.
6. Bates DW, Saria S, Ohno-Machado L, et al. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood). 2014;33:1123-31.
7. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380:1347-58.
8. Amarasingham R, Moore BJ, Tabak YP, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Med Care. 2010;48:981-8.
9. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2017;24:198-208.
10. Barnes S, Hamrock E, Toerper M, et al. Real-time prediction of inpatient length of stay for discharge prioritization. J Am Med Inform Assoc. 2016;23:e2-10.
11. Daghistani TA, Elshawi R, Sakr S, et al. Predictors of in-hospital length of stay among cardiac patients: a machine learning approach. Int J Cardiol. 2019;288:140-7.
12. Turgeman L, May JH, Sciulli R. Insights from a machine learning model for predicting the hospital length of stay (LOS) at the time of admission. Expert Syst Appl. 2017;78:376-85.
13. Barsasella D, Bah K, Mishra P, et al. A machine learning model to predict length of stay and mortality among diabetes and hypertension inpatients. Medicina (Kaunas). 2022;58:1568.
14. Jain R, Singh M, Rao AR, Garg R. Predicting hospital length of stay using machine learning on a large open health dataset. BMC Health Serv Res. 2024;24:860.
15. Gokhale S, Taylor D, Gill J, et al. Hospital length of stay prediction tools for all hospital admissions and general medicine populations: systematic review and meta-analysis. Front Med (Lausanne). 2023;10:1192969.
16. Strack B, DeShazo JP, Gennings C, et al. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. Biomed Res Int. 2014;2014:781670.
17. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data?. PLoS One. 2017;12:e0174944.
18. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189-232.

19. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Routledge; 2013.
20. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17 August 2016. San Francisco, CA, USA, 785-94.
21. Breiman L. Random forests. Mach Learn. 2001;45:5-32.
22. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016;18:e323.