Research Square

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

Bidirectional Prediction of Diabetes-Hypertension Comorbidity Transitions from Multimodal Outpatient EMRs: A Reference-Deviation-Aware Multimodal Stacking Framework (RD-MMS)

Shaofeng Tan
University of Science and Technology Beijing
Haoqiang Liu
University of Science and Technology Beijing
Luyao Tong
University of Science and Technology Beijing
Xiaoxuan Zhao
Beijing Pinggu Hospital
Jiamin Zhang
University of Science and Technology Beijing
LIsha Shuai
University of Science and Technology Beijing
Min Zhang
University of Science and Technology Beijing
Xiaolong Yang
yangxl@ustb.edu.cn
University of Science and Technology Beijing

Research Article

Keywords: multimodal learning, ensemble learning, reference interval encoding, data leakage control, chronic disease comorbidity prediction

Posted Date: July 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9517695/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Bidirectional Prediction of Diabetes–Hypertension Comorbidity Transitions from Multimodal Outpatient EMRs: A Reference-Deviation-Aware Multimodal Stacking Framework (RD-MMS)

Shaofeng Tan[1,2], Haoqiang Liu[1], Luyao Tong[1], Xiaoxuan Zhao[3], Jiamin Zhang[1], Lisha Shuai[1], Jialin Shi[1], Min Zhang[1], Xiaolong Yang[1*]

[1*]School of Computer and Communication Engineering, University of Science and Technology Beijing, 30 Xueyuan Road, Beijing, 100083, China.

[2]Department of Medical Engineering, Beijing Pinggu Hospital, No.59 Xinping North Road, Beijing, 101200, China.

[3]Information Management Center, Beijing Pinggu Hospital, No.59 Xinping North Road, Beijing, 101200, China.

*Corresponding author(s). E-mail(s): yangxl@ustb.edu.cn;
Contributing authors: winber@126.com; lhqbuaa@ustb.edu.cn; yaaooo1026@163.com; 373178067@qq.com; zhangjm_ustb@163.com; promethusls@outlook.com; shijl@ustb.edu.cn; zhangm@ustb.edu.cn;

Abstract

Background The highly specialized management of chronic diseases creates systematic blind spots in cross-disease risk identification. Existing predictive models predominantly rely on inpatient data and focus on unidirectional disease prediction, lacking systematic solutions tailored to the characteristics of real-world outpatient data.

Methods A large-scale real-world outpatient multimodal dataset spanning 2023 to 2024 is constructed, and the Reference-Deviation-Aware Multimodal Stacking (RD-MMS) framework is proposed to perform bidirectional prediction of transitions between diabetes mellitus (DM) and hypertension (HTN) as a representative comorbidity prediction task.

Results RD-MMS achieves an area under the receiver operating characteristic curve (AUROC) of 0.767 for DM-to-HTN and 0.735 for HTN-to-DM in 5-fold patient-level cross-validation, outperforming existing benchmarks in an outpatient setting where laboratory test coverage is only 13–18%. External validation on a 2025 holdout cohort confirms temporal stability, with AUROC values of 0.748 and 0.715 and temporal drift below 1.5%. Reference-Deviation (RD) semantic perturbation causes AUROC decreases of 0.028–0.030, confirming that the model depends on clinical semantics. The top 10% high-risk group achieves 2.9-fold enrichment in the event rate, and decision curve analysis (DCA) demonstrates consistently superior net benefit over extreme strategies across threshold probabilities of 2%–50%.
Conclusions Through reference-deviation-aware encoding, four-dimensional leakage control, and multimodal stacking, the RD-MMS framework achieves robust performance on low-quality outpatient data, with temporal stability confirmed by external validation. The core design is decoupled from specific disease types and supports task switching in a configuration-driven manner, providing a complete methodological paradigm for constructing reliable and deployable chronic comorbidity prediction models from real-world outpatient electronic medical records (EMR).

Keywords: multimodal learning; ensemble learning; reference interval encoding; data leakage control; chronic disease comorbidity prediction

1 Introduction

The growing burden of metabolic chronic diseases, exemplified by DM and HTN, continues to escalate. The 11th edition of the International Diabetes Federation (IDF) Diabetes Atlas projects that the global DM population will reach 589 million in 2025 and grow to 853 million by 2050 [1]. A pooled analysis by the NCD Risk Factor Collaboration of over 104 million participants demonstrates that HTN prevalence continues to rise, with substantial regional disparities in control rates [2]. The co-occurrence of DM and HTN is increasingly common [3, 4]. These two conditions engage in bidirectional pathophysiological reinforcement: DM promotes HTN through chronic hyperglycemia-induced endothelial injury, while HTN impairs insulin signaling and induces insulin resistance via renin-angiotensin-aldosterone system (RAAS) activation [5–8]. Comorbidity further multiplies the risk of cardiovascular events and end-stage renal disease [9, 10]. Early identification of patients at high risk of chronic disease transition, enabling a shift from passive treatment to proactive prevention, is a central goal of precision chronic disease management.

However, the progressive specialization of medicine means that DM patients are primarily managed by endocrinologists while HTN patients are seen by cardiologists. When patients with comorbidities seek care, each specialist attends to conditions within their own domain without systematically screening for cross-disease risks, leaving many patients at high transition risk unrecognized during the early stage of

comorbidity development. In primary care settings, this cross-disease diagnostic gap is even more pronounced due to limited technical expertise [11, 12].

In recent years, data-driven predictive models have advanced chronic disease risk identification; however, important limitations remain. Existing studies predominantly use inpatient records or population surveys, which differ substantially from real-world outpatient data in information density and quality [13–15]. Some studies using real-world data focus on complications rather than comorbidity transitions [16], and most prediction tasks adopt unidirectional or single-disease modeling that cannot reveal the task asymmetry between DM-to-HTN and HTN-to-DM transitions [17–19]. Systematic solutions remain lacking for the technical challenges specific to outpatient data, including cross-instrument heterogeneity in laboratory information system (LIS) data, hidden information leakage risks, and high sparsity [20, 21]. These issues limit model generalizability in real clinical environments, underscoring the need for predictive models built on real-world outpatient multimodal data that can simultaneously characterize bidirectional comorbidity transitions with adequate interpretability.

Hospital outpatient settings serve as the principal setting for chronic disease prevention and management. The longitudinal visit trajectory data recorded in outpatient EMR systems cross specialist boundaries and contain rich information for identifying disease transition risks. Nevertheless, constructing reliable predictive models from real-world outpatient data still faces four core technical challenges: the limitation of unidirectional modeling, which cannot reveal bidirectional task asymmetry [17–19]; reliance on single-modality data, with the predictive value of free text and visit behavior trajectories yet to be systematically quantified [15, 22]; technical pipeline challenges including systematic heterogeneity in LIS data and multiple hidden leakage pathways in longitudinal EMR [20, 21]; and insufficient clinical utility validation, lacking DCA, risk stratification, and other practical analyses [11, 12, 23–25].

2 Methods

To address these challenges, this study proposes the RD-MMS framework as an integrated solution that simultaneously addresses bidirectional comorbidity prediction, multimodal outpatient EMR integration, systematic leakage control, and robustness to sparse data within a unified architecture, with each design choice validated through perturbation experiments.

2.1 Overview of the RD-MMS Framework

Real-world outpatient EMRs pose three challenges that existing comorbidity prediction pipelines do not systematically address: cross-instrument heterogeneity in LIS data that conventional normalization cannot resolve, multiple subtle leakage pathways that inflate internal validation performance [20, 21], and high laboratory sparsity that undermines models relying solely on structured features [13, 14]. RD-MMS is designed to address all three within a unified architecture by encoding laboratory values relative to patient-specific reference intervals, enforcing a four-dimensional leakage control protocol, and fusing structured and text modalities so that clinical narratives actively compensate when laboratory coverage is low.

The framework consists of four core components (Fig.1). RD encoding invokes patient-specific reference intervals maintained by the LIS system to convert laboratory values into relative deviations: values within the normal range are encoded as zero, abnormal values as directional deviations, and missing values are preserved intact. This approach resolves cross-instrument comparability and enables tree-based models to learn from the absence of specific tests. Multimodal feature fusion concatenates RD numerical features, bidirectional encoder representations from transformers (BERT) text embeddings, term frequency-inverse document frequency (TF-IDF) representations, visit behavior statistics (density, intervals, and related quantities), and 20 guideline-based clinical threshold features, allowing structured and unstructured signals to complement each other. A four-dimensional leakage control boundary enforces patient-level fold splitting, temporal cutoff, sensitive field masking, and within-fold preprocessing fitting, ensuring that the model uses only information recorded prior to the index visit. A stacking ensemble then trains an L2-regularized logistic regression meta-learner on out-of-fold (OOF) predictions from multiple base models to reduce overfitting. The entire framework is driven by a configuration file that enables task switching and is not tied to specific disease types.

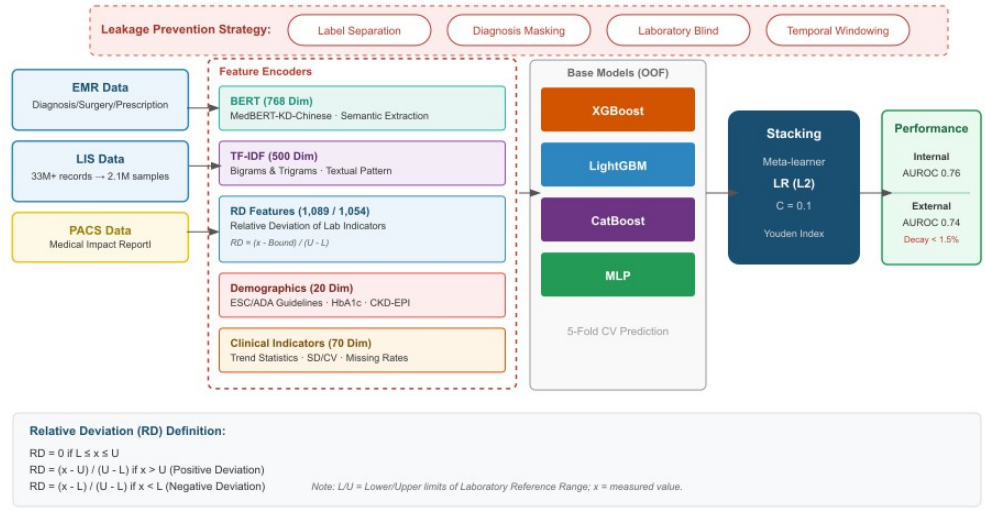


Fig. 1 RD-MMS: Reference-Deviation-Aware Multimodal Stacking Framework. Three input sources (EMR, LIS, and PACS) are encoded into five feature types: BERT semantic embeddings, character-level TF-IDF, Reference-Deviation (RD) encoded laboratory values, guideline-based demographic features, and clinical indicator features. A four-dimensional leakage prevention strategy (label separation, diagnosis masking, laboratory blind processing, and temporal windowing) is enforced throughout. Four base models are trained via 5-fold GroupKFold cross-validation, and a logistic regression meta-learner (L2, $C = 0.1$) integrates out-of-fold predictions into a final risk score. The RD encoding formula is shown in the lower panel: RD $= 0$ if $L \leq x \leq U$; $(x - U)/(U - L)$ if $x > U$; $(x - L)/(U - L)$ if $x < L$, where L and U are the lower and upper bounds of the patient-specific laboratory reference interval and x is the measured value.

2.2 Data Source and Cohort Construction

This study constructs a large-scale real-world outpatient multimodal dataset spanning 2023 to 2024, covering the primary data sources on which outpatient physicians base their diagnoses. The dataset is drawn from three sources: outpatient EMR (2,849,738 visit records containing structured diagnostic information and unstructured clinical narratives), LIS (33,211,443 raw test records, reduced to 2,188,770 records after quality filtering, covering 107 core laboratory indicators with patient-specific reference intervals fully retained), and a picture archiving and communication system (PACS) (633,489 imaging diagnostic reports). These three data types are aligned at the individual visit level, providing multidimensional coverage of outpatient clinical information. The dataset encompasses 276,392 patients, representing over 61% of the local permanent resident population, while the hospital accounts for approximately 30% of local outpatient volume (over 98% of patients are local residents), indicating strong population representativeness (Fig.2, Fig.**??**). Compared with existing similar studies that typically rely on inpatient records or targeted surveys [13–15], this dataset offers substantial advantages in scale, modality completeness, and real-world outpatient representativeness.

To focus on high-prevalence metabolic chronic diseases while remaining within practical computational constraints, patients with documented diagnoses of DM, HTN, or dyslipidemia are selected from the full patient pool to form the study cohort. The DM-to-HTN cohort includes patients with any confirmed DM diagnosis and no prior HTN record before the baseline date, with the first DM diagnosis set as the baseline date and the first incident HTN during follow-up as the positive outcome. The HTN-to-DM cohort is defined symmetrically. Disease names and targeted drug names in clinical text are masked with [UNK] tokens to prevent label leakage from clinical narratives. Negative samples are down-sampled at a 5:1 ratio [26]. The final cohort includes 13,290 patients for DM-to-HTN (positive rate 16.7%) and 13,086 patients for HTN-to-DM (positive rate 16.7%), comprising a total of 794,986 visit records.

2.3 Reference-Deviation Encoding

The core challenge in LIS data standardization is that even within a single hospital, the same laboratory indicator may correspond to multiple reference intervals due to differences in instrument platforms, reagent batches, and patient-specific characteristics [27, 28]. Conventional Z-score and min-max normalization share three limitations: they apply global summary statistics and cannot accommodate reference interval differences across instruments or patient populations; they map measurements within the normal range and abnormal values onto a common continuous scale, erasing the clinically meaningful zero-deviation signal encoded in normal-range values; and they require mean imputation for missing values, which destroys the ability of tree-based models to exploit "not tested" as a weak negative signal [29–31]. Generative adversarial imputation networks (GAIN) and similar methods can produce more realistic imputed values [32], but the imputation process itself introduces additional distributional assumptions that may obscure the clinical behavioral signal conveyed by "not

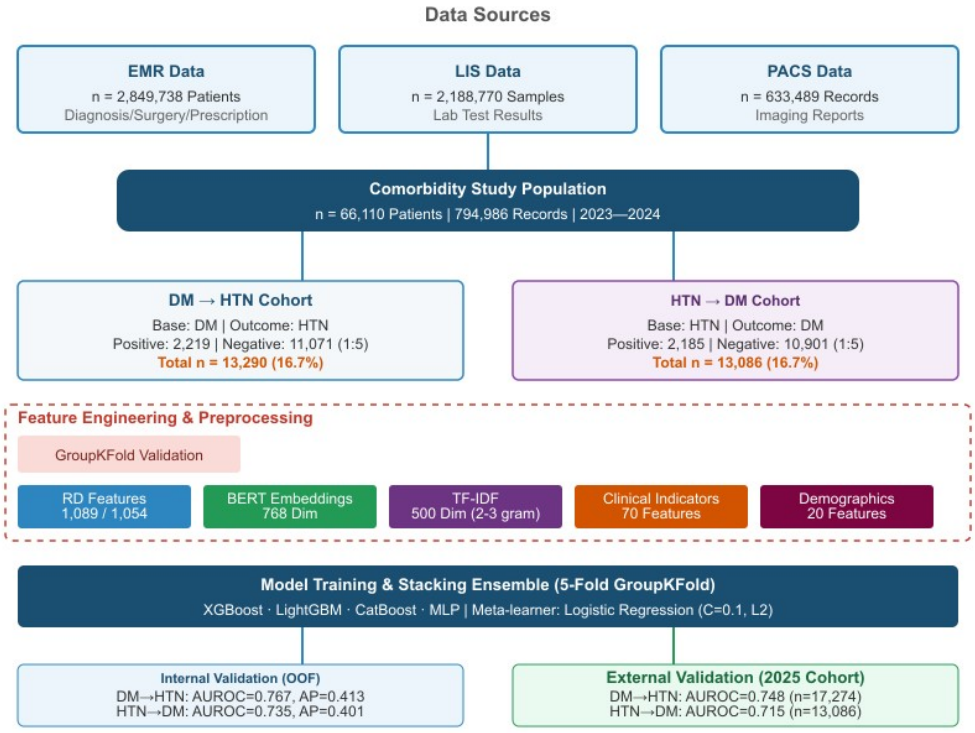


Fig. 2 Study Design and Patient Flow Diagram. Multi-source data (EMR, LIS, and PACS) from 2023–2024 are integrated into a comorbidity study population of 66,110 patients. Two bidirectional cohorts are derived: DM-to-HTN (n = 13,290, positive rate 16.7%) and HTN-to-DM (n = 13,086, positive rate 16.7%), each with negative samples down-sampled at a 1:5 ratio. Five feature modalities are assembled within a GroupKFold validation scheme, and the stacking ensemble is evaluated by internal OOF validation and an independent 2025 holdout cohort.

tested." The present framework retains NaN values intact so that tree-based models can natively exploit missingness patterns.

RD encoding is motivated by the routine clinical practice of interpreting laboratory results relative to reference intervals. This study formalizes that interpretive convention into a machine learning feature engineering scheme and systematically validates its predictive utility through perturbation experiments [27, 28]. Each measurement x_{ij}

is encoded as a clinically meaningful relative deviation:

$$\mathrm{RD}_{ij} = \begin{cases} \dfrac{x_{ij} - U_{ij}}{U_{ij} - L_{ij}}, & \text{if } x_{ij} > U_{ij} \\ 0, & \text{if } L_{ij} \leq x_{ij} \leq U_{ij} \\ \dfrac{x_{ij} - L_{ij}}{U_{ij} - L_{ij}}, & \text{if } x_{ij} < L_{ij} \end{cases} \tag{1}$$

where U_{ij} and L_{ij} denote the upper and lower bounds of the patient-specific reference interval for laboratory test j of patient i, respectively.

RD encoding addresses each of the aforementioned limitations. By invoking the patient-specific reference interval corresponding to each individual test, it fundamentally resolves cross-instrument heterogeneity. Encoding normal-range values as zero and abnormal values as directional continuous deviations aligns with the granularity of clinical judgment. Preserving NaN values for untested items enables tree-based models to simultaneously exploit both "what was tested" and "what was not tested" as informative signals [29–31]. After filtering columns with missingness exceeding 95% (sensitivity analyses at 70%, 80%, and 90% missingness thresholds show AUROC changes smaller than 0.005), 1,089 numerical features are retained for DM-to-HTN and 1,054 for HTN-to-DM.

2.4 Multimodal Feature Engineering

Throughout this study, the term "multimodal" refers to the integration of heterogeneous data types originating from distinct clinical source systems, specifically structured numerical records, free-text narratives, and behavioral logs, following the convention adopted in the medical informatics literature [17, 18].

This study integrates five categories of heterogeneous features. BERT semantic embeddings (768 dimensions) are obtained by concatenating six types of text fields and feeding them into a pretrained Chinese BERT model (bert-base-chinese), followed by mean pooling to capture deep clinical semantics. This approach leverages contextual representations shown to be highly effective in clinical text modeling and prediction tasks [33, 34]. Character-level TF-IDF (500 dimensions, 2–3-gram) is fitted independently within each training fold to prevent data leakage. Visit behavior features (7 dimensions) include visit density, mean and standard deviation of visit intervals, coefficient of variation, and visit count in the past 90 days; these features carry no missing values and have zero acquisition cost, making them particularly useful in primary care settings [35]. For deeper modeling of longitudinal visit sequences, Transformer-based architectures such as Bidirectional Encoder Representations from Transformers for Electronic Health Records (BEHRT) offer an alternative [36], but this framework instead adopts computationally less expensive and more interpretable visit behavior statistics. Guideline-based clinical threshold features (20 dimensions) are constructed primarily according to the European Society of Cardiology (ESC)/European Society of Hypertension (ESH), American Diabetes Association (ADA), and Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) guidelines, directly encoding key

clinical abnormality states such as blood pressure grading, glycemic threshold crossings, and estimated glomerular filtration rate (eGFR) staging [37–39]. The above features are concatenated into a final feature set of approximately 2,357 dimensions for DM-to-HTN and approximately 2,322 dimensions for HTN-to-DM.

2.5 Four-Dimensional Leakage Control System

Data leakage is the primary systematic threat to clinical predictive models, as it leads to inflated internal validation performance, unreliable external generalization, and reproducibility failures. To strictly simulate real clinical decision-making constraints and ensure that the model uses only information recorded prior to the index visit [20, 21], four complementary controls are enforced throughout the pipeline. Patient-level splitting through group-stratified k-fold cross-validation (GroupKFold) ensures no patient overlap between folds, preventing the model from implicitly memorizing individual patient patterns across longitudinal visit records, a form of leakage that random record-level splitting would introduce and that is frequently overlooked in studies using longitudinal EMR data [20, 21]. Temporal cutoff restricts features to records dated before the baseline date. Semantic masking adds LIS year fields and other fields containing implicit temporal information to a restricted list, and replaces disease diagnosis names and targeted drug names in clinical text with [UNK] tokens. Within-fold leakage prevention ensures that TF-IDF, Synthetic Minority Oversampling Technique (SMOTE), and StandardScaler are fitted exclusively within the training fold, and the stacking meta-learner is trained on OOF predictions [40, 41].

2.6 Model Architecture and Evaluation

This study adopts differentiated preprocessing strategies to address class imbalance and data heterogeneity. Conventional models (multilayer perceptron [MLP], random forest [RF], and logistic regression [LR]) apply SMOTE oversampling combined with StandardScaler normalization within the training fold. Gradient boosting trees (XGBoost, LightGBM, and CatBoost) use built-in class weights and retain missing values. Six heterogeneous base models are trained in parallel using 5-fold patient-level GroupKFold cross-validation [29–31, 42–44]. The stacking meta-learner is an L2-regularized logistic regression ($C = 0.1$) [40, 41], taking as input the OOF prediction probability matrix from the four base models with the highest OOF AUROC and outputting the final integrated risk score.

The evaluation framework uses AUROC as the primary discriminative metric, with average precision (AP), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score as secondary metrics; the classification threshold is determined by the Youden index. Robustness is assessed through three systematic perturbation experiments: sparsity robustness, which randomly masks laboratory values to simulate missingness rates from 0% to 90% in increments; RD semantic perturbation, which randomly pairs reference intervals from one patient with laboratory values from a different patient; and temporal order perturbation, which randomly shuffles visit sequences. Clinical utility is comprehensively evaluated through DCA [23, 25], Hosmer-Lemeshow calibration testing, top-10% risk stratification net

benefit, and subgroup analyses by age, sex, and baseline disease severity. External temporal validation uses an independent 2025 holdout cohort (DM-to-HTN: $n = 17{,}274$; HTN-to-DM: $n = 13{,}086$), strictly following the future-validation principle to avoid temporal leakage. Interpretability analysis is based on SHAP TreeExplainer [45, 46], deriving feature contributions and interaction effects from 2,000 randomly sampled cases.

To evaluate the incremental contribution of each design layer in the RD-MMS framework, three baseline methods are constructed and evaluated under the same patient-level GroupKFold cross-validation scheme ($k = 5$) and leakage-prevention rules as the main pipeline. B1 is a logistic regression model using only demographic and visit behavior features. B2 is a logistic regression model with raw laboratory values added to B1. B3 is an XGBoost model using all structured features but excluding clinical text. All baselines use the same data partitioning as the main pipeline to ensure fair comparison. Statistical significance of AUROC differences between methods is assessed using the DeLong test (two-sided).

3 Results

Results are organized in five parts: overall model performance and external validation (Section 3.1), comparison against baseline methods (Section 3.2), ablation of design decisions and modality contributions (Sections 3.3–3.4), perturbation validation (Section 3.5), SHAP feature importance analysis (Section 3.6), and clinical utility analysis (Section 3.7). Baseline characteristics of both development cohorts are summarized in Table 1.

3.1 Model Prediction Performance

Table 2 presents the baseline characteristics of both task development cohorts. Patients who experience comorbidity transition (positive cases) are significantly older at first visit than negative cases (DM-to-HTN: 60.8 ± 12.7 years vs. 53.6 ± 15.3 years; HTN-to-DM: 61.4 ± 13.1 years vs. 60.0 ± 13.7 years; both $p < 0.001$). Positive cases have a lower median visit count than negative cases (4 vs. 6), but higher visit density and smaller standard deviation of visit intervals, reflecting more regular and concentrated visit patterns among positive cases.

The full-modality stacking model achieves the best performance: DM-to-HTN AUROC = 0.767 (AP = 0.413, sensitivity = 0.708, specificity = 0.689, NPV = 0.923); HTN-to-DM AUROC = 0.735 (AP = 0.401, sensitivity = 0.679, specificity = 0.658, NPV = 0.912). Gradient boosting tree models consistently outperform MLP and LR [29–31], and stacking improves upon the best single model (LightGBM) by +0.010 to +0.018 [40, 41]. AP values for both tasks are approximately 2.4–2.5 times the random baseline (0.167) [26] (Table 2, Fig.**??**, Fig.3). Compared with similar existing studies [17, 47, 48], the present study achieves superior performance using real-world outpatient laboratory data with coverage of only 13–18%, a quality substantially lower than that of the inpatient or survey data used in those studies. Temporal drift on the holdout external cohort remains below 1.5%, demonstrating the robustness of the framework under low-quality real-world data conditions.

External validation on the 2025 holdout cohort yields AUROC values of 0.748 and 0.715, with temporal drift of 1.3%–1.5% (Fig.**??**), confirming the effectiveness of the leakage control system [20, 21]. The external AUROC of RF drops from 0.737 to 0.615 (a decrease of 0.122), while the stacking ensemble declines by only 0.013, demonstrating the systematic suppression of overfitting by the OOF mechanism [40, 41].

Table 1 Baseline characteristics of the study cohorts (development cohort, 2023–2024)

Characteristic	Positive	Negative	p
DM→HTN cohort ($n_{\text{pos}} = 1{,}817$; $n_{\text{neg}} = 9{,}085$)			
Age at first visit (years, mean ± SD)	60.8 ± 12.7	53.6 ± 15.3	<0.001
Age at last visit (years, mean ± SD)	61.2 ± 12.7	54.2 ± 15.3	<0.001
Visit count (median [IQR])	4 [2–10]	6 [2–13]	<0.001
Visit density (visits/month, median [IQR])	1.52 [0.54–4.29]	1.20 [0.51–3.20]	<0.001
Visit interval SD (days, median [IQR])	8.0 [0.0–24.5]	13.5 [0.0–36.2]	<0.001
LIS test coverage (patient-level aggregate)	51.1%	51.5%	0.310
HTN→DM cohort ($n_{\text{pos}} = 1{,}846$; $n_{\text{neg}} = 9{,}230$)			
Age at first visit (years, mean ± SD)	61.4 ± 13.1	60.0 ± 13.7	<0.001
Age at last visit (years, mean ± SD)	61.8 ± 13.0	60.7 ± 13.7	<0.001
Visit count (median [IQR])	4 [2–9]	6 [2–13]	<0.001
Visit density (visits/month, median [IQR])	1.39 [0.55–4.83]	1.04 [0.49–2.43]	<0.001
Visit interval SD (days, median [IQR])	7.5 [0.0–26.8]	17.2 [0.0–39.7]	<0.001
LIS test coverage (patient-level aggregate)	51.3%	51.6%	0.283

Notes: Positive: comorbidity transition occurred during follow-up; Negative: no transition. LIS test coverage denotes the proportion of LIS items with recorded values among all LIS items across a patient's visit history (single-visit coverage 13%–18%, see main text). Between-group comparisons for continuous variables use the Mann-Whitney U test. SD: standard deviation; IQR: interquartile range; LIS: laboratory information system.

3.2 Performance Comparison with Baseline Methods

Table 3 presents the performance comparison between RD-MMS and the three baseline methods under the same data partitioning.

Baseline B1, a logistic regression model using only demographic and visit behavior features, achieves AUROC values of 0.690 and 0.633 for the DM-to-HTN and HTN-to-DM tasks, respectively, demonstrating that longitudinal visit behavior patterns carry an independent predictive signal.

Baseline B2, which augments B1 with raw laboratory values, does not improve performance on the DM-to-HTN task (AUROC 0.673 vs. 0.690 for B1; $\Delta\text{AUROC} = -0.094$ relative to RD-MMS). This indicates that raw laboratory measurements, without appropriate feature transformation, can introduce noise when processed by a linear classifier.

Table 2 Predictive performance of all models (5-fold patient-level OOF evaluation)

Task	Model	AUROC	AP	Sens	Spec	PPV	NPV	F1
DM→HTN	Stacking	0.767	0.413	0.708	0.689	0.316	0.923	0.436
	LightGBM	0.757	0.401	0.697	0.678	0.308	0.920	0.425
	XGBoost	0.754	0.396	0.648	0.726	0.320	0.912	0.430
	CatBoost	0.750	0.387	0.685	0.691	0.308	0.918	0.428
	Random Forest	0.737	0.352	0.751	0.601	0.275	0.925	0.401
	MLP	0.725	0.349	0.693	0.648	0.282	0.915	0.402
	LR	0.671	0.275	0.643	0.614	0.251	0.898	0.361
HTN→DM	Stacking	0.735	0.401	0.679	0.658	0.282	0.912	0.398
	LightGBM	0.717	0.377	0.641	0.677	0.284	0.905	0.394
	CatBoost	0.716	0.385	0.638	0.674	0.281	0.904	0.390
	XGBoost	0.713	0.374	0.578	0.734	0.304	0.898	0.399
	Random Forest	0.687	0.321	0.621	0.647	0.261	0.896	0.368
	MLP	0.684	0.316	0.561	0.719	0.285	0.892	0.378
	LR	0.646	0.266	0.498	0.718	0.262	0.879	0.345

*Notes:*All models are trained on the full feature set (RD-encoded laboratory tests, clinical text embeddings, and structured clinical features) using 5-fold patient-level GroupKFold OOF evaluation. The classification threshold is determined by the Youden index. AUROC: area under the ROC curve; AP: average precision; Sens: sensitivity; Spec: specificity; PPV: positive predictive value; NPV: negative predictive value; LR: logistic regression; MLP: multilayer perceptron; OOF: out-of-fold.

Replacing logistic regression with XGBoost and incorporating all structured features in B3 yields substantial performance gains (DM-to-HTN AUROC = 0.737; HTN-to-DM AUROC = 0.697), confirming the necessity of nonlinear modeling for high-dimensional laboratory data.

The complete RD-MMS framework achieves AUROC values of 0.767 and 0.735 for the two tasks, representing improvements of 0.030 and 0.037 over B3. These incremental gains are attributed to the synergistic effect of RD semantic encoding, clinical text embedding features, and the stacking ensemble. All pairwise comparisons between baseline methods and RD-MMS are statistically significant (DeLong test, all $p < 0.001$).

The degradation from B1 to B2 indicates that raw laboratory values without semantic normalization introduce net noise into a linear classifier, directly motivating the RD encoding and nonlinear modeling choices in RD-MMS.

3.3 Design Decision Ablation and Encoding Comparison

Ablation experiments systematically validate the effect of key engineering decisions on model performance (Table 4), using the configuration after removing the LIS year field as the reference point (DM-to-HTN: 0.760; HTN-to-DM: 0.721). The effects of individual decisions vary considerably across configurations.

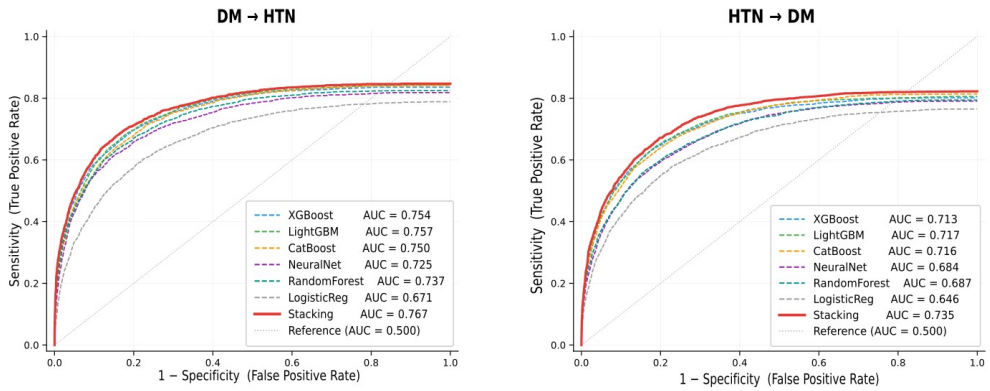


Fig. 3 ROC Curves: Internal Validation (5-fold Patient-Level OOF). Left panel: DM-to-HTN task; right panel: HTN-to-DM task. The solid red curve represents the Stacking ensemble (AUROC = 0.767 and 0.735, respectively). Dashed lines show the six base learners: XGBoost, LightGBM, CatBoost, NeuralNet (MLP), Random Forest, and Logistic Regression. The dotted diagonal line represents the no-discrimination reference. All curves are derived from complete 5-fold patient-level OOF prediction vectors.

Table 3 Performance comparison of RD-MMS against baseline methods (5-fold patient-level GroupKFold OOF evaluation)

Task	Method	AUROC	AP	Sens	Spec	PPV	NPV	F1	ΔAUROC	p
DM→HTN	B1	0.690	0.301	0.690	–	0.584	–	0.366	−0.077	<0.001
	B2	0.673	0.270	0.643	–	0.638	–	0.373	−0.094	<0.001
	B3	0.737	0.356	0.763	–	0.587	–	0.399	−0.030	<0.001
	RD-MMS	0.767	0.413	0.708	0.689	0.316	0.923	0.436	Ref	Ref
HTN→DM	B1	0.633	0.240	0.652	–	0.544	–	0.331	−0.102	<0.001
	B2	0.652	0.263	0.611	–	0.628	–	0.352	−0.083	<0.001
	B3	0.697	0.316	0.656	–	0.622	–	0.370	−0.037	<0.001
	RD-MMS	0.735	0.401	0.679	0.658	0.282	0.912	0.398	Ref	Ref

Notes: B1: LR using demographics and visit behavior features only. B2: LR with B1 features plus raw laboratory values. B3: XGBoost using all structured features, excluding clinical text. RD-MMS: proposed full-modality stacking framework. Specificity (Spec) and NPV are not reported (−) for B1, B2, and B3 as these baselines do not produce a matched specificity under the same threshold scheme. ΔAUROC: the AUROC of each method minus the AUROC of RD-MMS; p: DeLong test (two-sided). Sens: sensitivity;LR: logistic regression; Ref: reference.

The most prominent finding involves SMOTE misuse: when SMOTE oversampling is incorrectly applied to all models including gradient boosting trees, AUROC for the HTN-to-DM task drops by 0.041 (from 0.721 to 0.680). This decline arises because the interpolation mechanism of SMOTE disrupts the "not tested" weak negative signals captured by the native missing-value handling of tree-based models [29–31, 49], a clinically discriminative behavioral pattern that existing studies have largely overlooked in chronic disease screening [20, 21]. Restricting SMOTE to MLP, LR, and RF

restores performance to the reference level, confirming the necessity of differentiated preprocessing.

The leakage control effect ranks second in magnitude: systematic masking of the LIS year field yields approximately 0.030 AUROC improvement across both tasks, revealing the concealed nature of implicit leakage pathways in EMR temporal data. Such time-related metadata typically fall outside the scope of intuitive auditing, and existing studies lack standardized field-level leakage screening procedures.

Feature engineering gains are stable but show diminishing returns at the margin. Upgrading text features from word-level TF-IDF to 500-dimensional character-level 2–3-gram representations yields incremental improvements of +0.002 for DM-to-HTN and +0.006 for HTN-to-DM. Adding 20 guideline-based clinical threshold features to the final configuration yields improvements of 0.007–0.014 AUROC over the reference, confirming the value of embedding domain knowledge. These features, constructed according to ESC/ESH, ADA, and CKD-EPI guidelines, directly encode key clinical abnormality states such as blood pressure grading, glycemic threshold crossings, and estimated glomerular filtration rate (eGFR) staging as explicit features.

Table 4 Design decision ablation study (sequential steps, 5-fold OOF evaluation)

Configuration	DM→HTN AUROC	HTN→DM AUROC	Primary effect
Reference (LIS year field removed)	0.760	0.721	Engineering baseline
SMOTE applied to all models (incl. trees)	0.738	0.680	NaN signal disrupted (max ↓0.041)
SMOTE applied to MLP/LR/RF only (correct)	0.760	0.722	NaN signal restored
TF-IDF upgraded to 500-dim char 2–3-gram	0.762	0.728	+0.002/+0.006
20 clinical threshold features added	0.767	0.735	+0.005/+0.007

Notes: Each row represents an incremental step relative to the reference configuration. OOF: out-of-fold; MLP: multilayer perceptron; LR: logistic regression; RF: random forest; TF-IDF: term frequency-inverse document frequency.

3.4 Modality Ablation and Sparsity Robustness

Modality ablation experiments (Fig. 4 and Table 5) reveal the asymmetric contributions of different information sources to the two prediction tasks. The gain from text features is substantially larger for HTN-to-DM (ΔAUROC = +0.038) than for DM-to-HTN (+0.020), quantifying the fundamental difference in information structure between the two transition directions [13, 50]. This asymmetry cannot be identified in unidirectional modeling, highlighting the methodological necessity of the bidirectional framework.

The independent value of visit behavior features is particularly noteworthy: using only 7-dimensional visit behavior metrics (with zero additional examination cost),

AUROC values of 0.606–0.611 are achieved, substantially above chance level, demonstrating that patient visit behavior itself carries an independent signal for disease transition risk. This finding suggests that behavioral trajectory analysis can serve as a low-cost screening tool in resource-limited settings, with important practical implications for primary care institutions.

Table 5 Modality ablation study (LightGBM, 5-fold OOF evaluation)

Feature configuration	DM→HTN AUROC	HTN→DM AUROC
Visit behavior features only (7 dim)	0.606	0.611
BERT embeddings only (768 dim)	0.682	0.631
TF-IDF only (500 dim, char 2–3-gram)	0.693	0.653
Numerical only (RD LIS + BP + behavior)	0.747	0.697
Numerical + TF-IDF (BERT removed)	0.759	0.726
Numerical + BERT (TF-IDF removed)	0.750	0.709
Full-modality Stacking (final)	0.767	0.735

*Notes:*Full-modality Stacking uses LightGBM, XGBoost, CatBoost, and Random Forest as the four base models. TF-IDF is fitted independently within each training fold (within-fold leakage prevention). BERT: bidirectional encoder representations from transformers; TF-IDF: term frequency-inverse document frequency; RD: reference deviation; LIS: laboratory information system; BP: blood pressure; dim: dimensions.

The marginal contribution of text features is modality-dependent. Using BERT semantic embeddings alone achieves AUROC values of only 0.631–0.682, with limited improvement over visit behavior features alone (+0.025–+0.071), while carrying the risk of high-dimensional sparsity. Their predictive contribution is contingent on the anchoring effect of structured laboratory data; the final multimodal fusion configuration achieves AUROC gains of 0.085–0.103 over the single-text modality.

Sparsity robustness experiments (Fig.5, Table 6) show that as laboratory missingness increases from 0% to 90%, the numerical-only configuration degrades substantially: DM-to-HTN AUROC falls from 0.747 to 0.598 ($\Delta = -0.149$), while the full-modality configuration declines only to 0.708 ($\Delta = -0.059$), representing a 60.4% reduction in degradation.

The text compensation effect expands monotonically, with the compensation gain increasing from +0.020 under complete data to +0.110 under 90% missingness, a 5.5-fold increase. This indicates that text and structured data form a strong complementary relationship rather than redundant overlap [48, 51], with text features actively capturing substitute signals from clinical narratives when laboratory data are sparse. This property has critical practical value for primary care settings: even for patients who have never undergone laboratory testing, the narrative information recorded by physicians in outpatient records, including chief complaints, physical examination findings, and medication adjustments, maintains discriminative ability at approximately AUROC 0.61 (Fig.**??**). The framework does not degrade when patients have few test records. This robustness derives from the reference-interval prior embedded in RD encoding: even when laboratory values are absent, "not tested" can itself participate

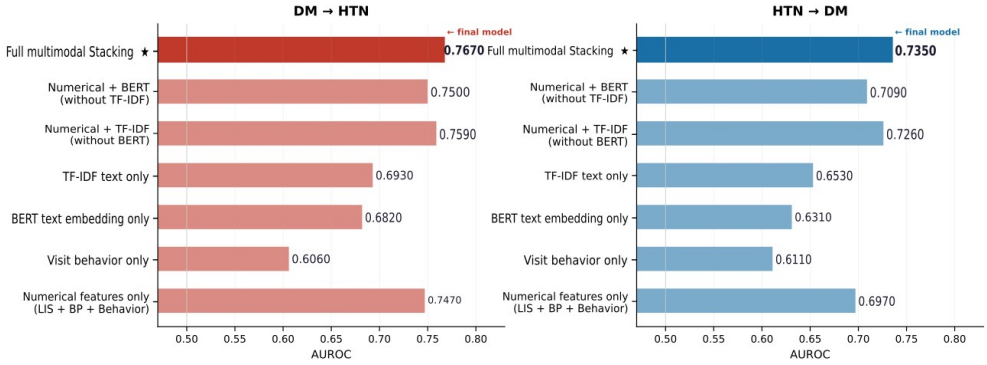


Fig. 4 Modality Ablation Study: AUROC by Feature Configuration (LightGBM, 5-fold OOF). Horizontal bar charts compare AUROC across seven feature configurations for the DM-to-HTN (left, red) and HTN-to-DM (right, blue) tasks. The gain from adding text features to the numerical-only baseline is +0.020 for DM-to-HTN and +0.038 for HTN-to-DM, revealing an asymmetry in the predictive value of text features that is undetectable in unidirectional models.

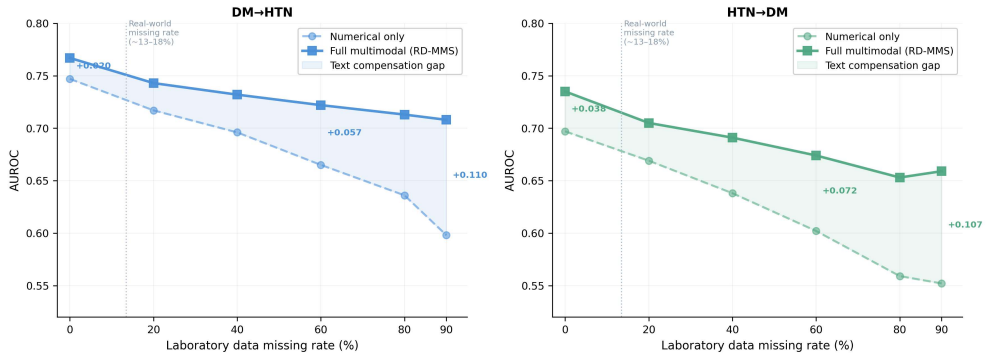


Fig. 5 Robustness to Missing Laboratory Data. AUROC as a function of laboratory data missing rate (0%–90%) for the numerical-only (dashed, circles) and full multimodal RD-MMS (solid, squares) configurations in the DM-to-HTN (left, blue) and HTN-to-DM (right, green) tasks. The shaded region represents the text compensation gap, which expands monotonically as missingness increases. The vertical dotted line indicates the approximate real-world single-visit missing rate (13%–18%).

in prediction through the encoding mechanism, forming complementary redundancy alongside semantic compensation from the text modality.

The asymmetric text gain across tasks reflects a structural difference in how each transition manifests in outpatient records, a pattern that unidirectional modeling would average away and therefore miss entirely.

Table 6 Sparsity robustness experiment: AUROC at varying laboratory data missingness rates

Configuration	0%	20%	40%	60%	80%	90%
DM→HTN, numerical only	0.747	0.717	0.696	0.665	0.636	0.598
DM→HTN, full modality	0.767	0.743	0.732	0.722	0.713	0.708
Text compensation gain (Δ)	+0.020	+0.026	+0.036	+0.057	+0.077	+0.110
HTN→DM, numerical only	0.697	0.669	0.638	0.602	0.559	0.552
HTN→DM, full modality	0.735	0.705	0.691	0.674	0.653	0.659
Text compensation gain (Δ)	+0.038	+0.036	+0.053	+0.072	+0.094	+0.107

*Notes:*Missingness is introduced by randomly masking laboratory values at the specified rate. Full-modality configuration uses the complete RD-MMS feature set. AUROC: area under the receiver operating characteristic curve.

3.5 Perturbation Validation

For RD semantic perturbation, laboratory values and reference intervals from different patients are randomly mismatched, preserving the numerical value distribution while destroying patient-specific clinical semantic alignment. DM-to-HTN AUROC decreases from 0.767 to 0.737 ($\Delta = -0.030$) and HTN-to-DM from 0.735 to 0.707 ($\Delta = -0.028$), demonstrating that performance gains do not derive from simple numerical scale transformation but rather from the clinical prior knowledge embedded in patient-specific reference intervals (Table 7).

For temporal order perturbation, visit sequences are randomly shuffled to eliminate temporal dependencies, reducing DM-to-HTN AUROC to 0.721 ($\Delta = -0.046$) and HTN-to-DM to 0.698 ($\Delta = -0.037$). This indicates that longitudinal visit trajectories represent a more fundamental predictive signal than laboratory semantics, and that models based on cross-sectional or single-visit data [14, 25] cannot capture this dynamic evolutionary pattern. This finding explains why the framework achieves performance comparable to inpatient cohort studies in the lower-quality outpatient setting: the critical factor is thorough modeling of longitudinal care-seeking behavior rather than reliance on data completeness alone.

3.6 SHAP Feature Importance Analysis

SHapley Additive exPlanations (SHAP) interpretability analysis reveals fundamental differences in feature dependency patterns between the two transition tasks [45, 46] (Table 8), providing mechanistic justification for modeling the two directions independently.

The DM-to-HTN task is dominated by vascular aging-related features, with age at first visit ($|\text{SHAP}| = 0.196$) and blood pressure trajectory constituting the core predictors, consistent with the progressive vascular stiffening and endothelial dysfunction observed in DM patients with advancing age [5, 52, 53]. The HTN-to-DM task presents a markedly different information structure: visit density ($|\text{SHAP}| = 0.246$) ranks first, and text-extracted glucose anomaly mentions (exam_glucose, $|\text{SHAP}| = 0.142$) rank third, both exceeding traditional laboratory biomarkers. This finding aligns

Table 7 Perturbation experiments: RD semantic perturbation and temporal order perturbation

Perturbation type	DM→HTN AUROC	HTN→DM AUROC	Clinical interpretation
Original RD encoding (reference)	0.767	0.735	–
RD semantic perturbation (reference intervals randomly mismatched)	0.737 (↓0.030)	0.707 (↓0.028)	Performance derives from clinical semantics, not numerical scale
Temporal order perturbation (visit sequence randomly shuffled)	0.721 (↓0.046)	0.698 (↓0.037)	Longitudinal visit trajectory is the core predictive signal

with clinical practice, in which increased visit frequency among hypertensive patients often reflects poor blood pressure control, while sustained blood pressure fluctuation induces insulin resistance through RAAS activation and oxidative stress [6, 7]. Glucose abnormalities documented by physicians in clinical notes frequently precede formal diagnosis, constituting a cross-disease early warning signal that specialist-centered management is most likely to miss [48, 51].

Feature importance is markedly asymmetric across tasks. In a unidirectional DM-to-HTN model, the contribution of text features would be substantially underestimated (gain of only +0.020), whereas text features are a critical driver in the HTN-to-DM direction (gain of +0.038). This directional dependency can only be identified within a bidirectional unified framework. Visit behavior features surpass traditional laboratory indicators in both tasks, independently validating the predictive value of healthcare utilization behavior as a digital phenotype [45, 46] (Figs.**??**,**??**,**??**). These findings have been confirmed through clinical expert review as consistent with the observed disease progression patterns in chronic comorbidity development.

That temporal shuffling causes greater performance loss than semantic mismatching suggests that longitudinal visit trajectories carry more fundamental predictive information than any individual laboratory deviation.

3.7 Clinical Utility and Risk Stratification

Clinical utility is assessed across four dimensions: decision support, probability calibration, risk stratification, and population heterogeneity.

Decision curve analysis (Fig.6) shows that across the full range of clinical decision threshold probabilities from 2% to 50%, the net benefit of the framework consistently exceeds both the treat-all and treat-none strategies [23, 25], with the advantage most pronounced in the 10%–30% threshold range. This indicates that the model provides the greatest decision support value when physicians must choose between close monitoring and initiating intervention.

Regarding probability calibration, the model is well calibrated in the low-to-moderate predicted risk range (0–0.60) but shows mild overestimation above 0.60.

Table 8 Top 10 SHAP feature importances (XGBoost, $n = 2{,}000$ random samples)

Rank	DM→HTN feature	|SHAP|	Rank	HTN→DM feature	|SHAP|
1	Age at first visit	0.196	1	Visit density (visits/-month)	0.246
2	Visit density	0.157	2	Visit interval SD	0.157
3	Visit interval SD	0.138	3	Text-extracted glucose (exam_glucose)	0.142
4	Chronic disease prescription flag	0.123	4	Total visit count	0.095
5	Age at last visit	0.079	5	Diastolic BP at last visit	0.088
6	Maximum SBP (sbp_max)	0.076	6	Visit interval CV	0.075
7	Mean SBP (sbp_mean)	0.070	7	Proportion of elevated BP readings	0.071
8	Comorbidity count	0.063	8	Mean heart rate	0.059
9	Visit count in past 90 days	0.058	9	Age at first visit	0.057
10	Mean visit interval	0.054	10	Heart rate at last visit	0.056

Notes: exam_glucose is a numerical glucose value (mmol/L) extracted from outpatient clinical text via regular expression and is classified as a text-derived feature. Complete top-20 feature importance bar charts and beeswarm plots are shown in Supplementary Fig. S5. SHAP: SHapley Additive exPlanations; SD: standard deviation; BP: blood pressure; SBP: systolic blood pressure; CV: coefficient of variation.

This bias arises from the sparsity of positive cases in the training set; Platt scaling or isotonic regression is recommended for post-hoc calibration before clinical deployment [26].

Risk stratification validation demonstrates that the top 10% high-risk group achieves actual positive rates of 48.0% (DM-to-HTN) and 45.9% (HTN-to-DM), representing 2.9-fold and 2.8-fold enrichment over the baseline incidence, respectively. The low-risk group (bottom 70%) has positive rates of 9.3% and 10.2%, yielding an approximately fivefold difference in incidence between high- and low-risk groups (RR = 4.8–5.2). Concentrating limited specialist follow-up resources on the top 10% high-risk patients alone would cover approximately 50% of true transition cases while achieving nearly a threefold improvement in follow-up efficiency, offering significant cost-effectiveness advantages for primary care institutions with scarce specialist resources [12, 25].

Subgroup analysis shows that AUROC is maintained above 0.713 across all age groups, demonstrating robust model generalizability. Patients aged 65 years and older show the highest incidence rates (DM-to-HTN: 25.1%; HTN-to-DM: 18.7%), and text feature contribution is elevated in this subgroup (+0.008 AUROC), possibly reflecting more detailed clinical notes for elderly patients. This population should be considered a priority intervention target [53, 54].

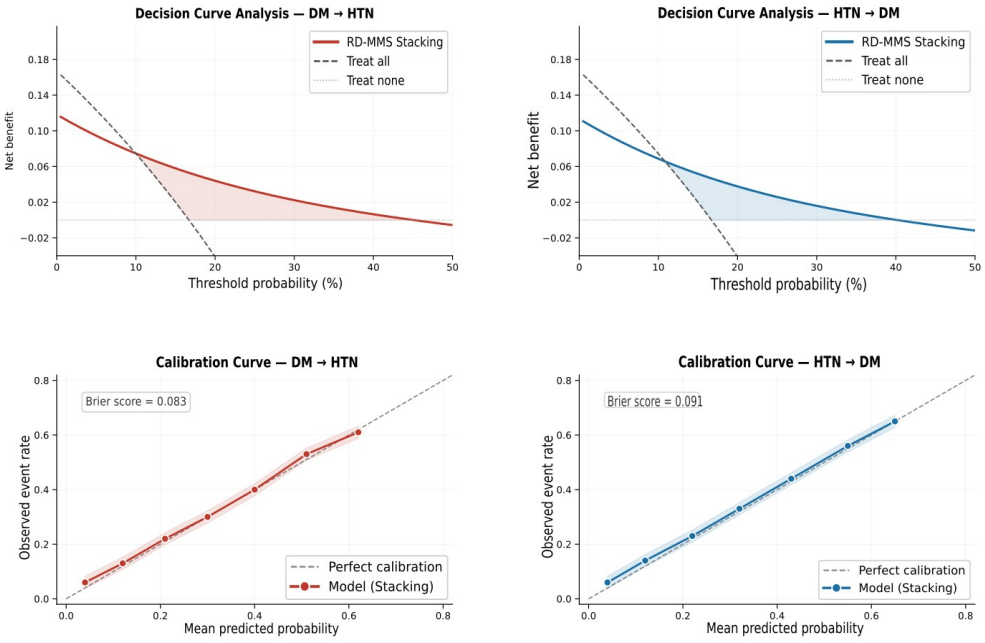


Fig. 6 Clinical Utility Analysis: Decision Curve Analysis and Calibration. Upper panels show DCA for the DM-to-HTN (left) and HTN-to-DM (right) tasks. The solid colored curve is the Stacking model; the dashed line is treat-all; the dotted line is treat-none. The shaded region indicates net benefit gain over both extreme strategies. Lower panels show calibration curves; the dashed diagonal represents perfect calibration. Both tasks show good agreement between predicted probability and observed event rate in the low-to-moderate risk range, with mild overestimation at higher predicted probabilities.

4 Discussion

4.1 Fundamental Distinction of RD Encoding from Existing Normalization Approaches

Compared with Z-score normalization, which relies on global distributional statistics, and min-max normalization, which relies on global extremes, the fundamental distinction of RD encoding from these methods lies in its grounding in clinical cognition. RD encoding replicates the reasoning process by which clinicians interpret laboratory reports using the reference interval as the coordinate origin, encoding normal values as zero and degrees of deviation as directional relative quantities, thereby achieving cross-instrument and cross-reagent clinical semantic consistency [27, 28].

RD encoding differs from existing methods in three respects. Unlike Z-score and min-max normalization, which cannot accommodate cross-instrument reference interval differences, RD encoding directly invokes patient-specific reference intervals, resolving cross-instrument heterogeneity at its source. Conventional normalization methods also require mean imputation for missing values, which destroys the ability

of tree-based models to exploit "not tested" as a weak negative signal [29–31]; RD encoding retains NaN values so that the model can exploit both "what was tested" and "what was not tested" as informative signals. Binary abnormality flags avoid the imputation problem but discard the degree-of-deviation information that captures pathological severity, whereas RD encoding retains continuous deviation quantities consistent with the granularity of clinical interpretation.

The RD semantic perturbation experiment, causing AUROC decreases of 0.028–0.030, demonstrates that performance gains originate from clinical semantic alignment rather than numerical scale transformation. Temporal order perturbation causes greater loss (0.037–0.046), revealing that longitudinal visit trajectories are a more fundamental predictive signal than single-visit laboratory tests, consistent with the clinical characteristics of chronic diseases as conditions with insidious onset and protracted course [5, 6]. Models based on cross-sectional or single-visit data [14, 25] cannot capture this signal, explaining why the framework achieves comparable performance in the lower-quality outpatient setting. Furthermore, RD encoding is decoupled from specific disease types and can be directly extended to other chronic disease prediction tasks such as chronic kidney disease and coronary artery disease [37, 55]. Binary abnormality flags are a common clinically interpretable baseline in rule-based decision systems, and the superior performance of RD encoding over this baseline indicates that continuously modeling the direction and magnitude of deviations provides more predictive value than a binary normal-versus-abnormal classification.

4.2 NaN Signal Preservation and Leakage Control

Ablation experiments reveal two engineering pitfalls that are prevalent in existing research. One is the inappropriate application of SMOTE to tree-based models, with AUROC impact as large as -0.041. Studies addressing class imbalance typically apply SMOTE uniformly to all models [49] without distinguishing between the missing-value handling mechanisms of tree-based and non-tree-based models, systematically destroying the "not tested" weak negative signals that tree-based models can natively exploit. A more insidious issue is implicit leakage through EMR temporal fields, with LIS year field masking yielding an impact of approximately -0.030. Such leakage is linked to temporal associations with disease progression and introduces hidden information; existing studies acknowledge its prevalence but rarely implement systematic field-level auditing [20, 21]. The four-dimensional leakage control system integrates these scattered engineering lessons into a reproducible systematic methodology, providing a replicable standard for similar studies.

The practical consequences of leakage are directly illustrated by the temporal cross-validation results for Random Forest: its internal OOF AUROC is 0.737, but it degrades to 0.615 on the 2025 holdout external validation (a decrease of 0.122). Random Forest does not employ the OOF mechanism, so its internal validation is systematically inflated due to patient overlap across folds. In contrast, the stacking ensemble shows drift of only 0.013 on the same external cohort, a tenfold difference. This comparison demonstrates that without rigorous leakage control, internal validation metrics are poor proxies for clinical deployment performance; even when a single model performs well internally, serious generalization risks may still exist [40, 41].

4.3 Methodological Implications of Baseline Analysis

The layered baseline comparisons reveal an important pattern: in the DM-to-HTN task, adding raw laboratory values to logistic regression B2 produces lower performance than the demographic-plus-visit-behavior baseline B1 alone (AUROC 0.673 vs. 0.690). This is consistent with existing findings that high-dimensional clinical measurements require nonlinear models or explicit feature engineering to generate predictive value [29, 30]. This finding supports two core design choices in RD-MMS: the semantically normalized encoding of reference intervals through RD encoding, and the nonlinear modeling architecture based on tree ensembles. Together, these transform the heterogeneity of raw measurements into discriminative feature representations.

The incremental improvement of RD-MMS over B3, an XGBoost model using the same structured features, with ΔAUROC of +0.030 and +0.037 respectively, further demonstrates that clinical text records carry independent information not captured by structured data. This corroborates the findings from feature ablation experiments and supports the inclusion of free text in chronic comorbidity prediction models built on Chinese community outpatient EMR data.

The baseline comparisons in this study focus on methodological-level internal ablations rather than performance competition with externally published models. Since no existing study conducts the same bidirectional prediction task on a fully homogeneous Chinese community EMR dataset, cross-heterogeneous-dataset comparisons do not constitute fair benchmarks and are therefore not used as the primary line of argument.

4.4 Bidirectional Task Asymmetry

The HTN-to-DM task achieves an AUROC of 0.735, slightly below that of the DM-to-HTN task (0.767). The difference in information structure between the two directions explains this asymmetry and simultaneously reveals systematic limitations of existing unidirectional modeling studies.

In terms of information density, glucose test coverage for HTN patients in the outpatient LIS (approximately 12.8%) is substantially lower than blood pressure monitoring coverage for DM patients (approximately 18.3%), leading to fewer available structured signals in the HTN-to-DM direction. Mechanistically, the transition from HTN to DM involves multiple parallel pathways including RAAS activation and diuretic effects [6, 7, 56], resulting in more dispersed signals.

This asymmetry has important methodological implications. In a unidirectional DM-to-HTN model, the contribution of text features would be underestimated (+0.020), in turn underestimating the overall value of multimodal fusion. In the HTN-to-DM direction, text features are a critical compensatory source (+0.038), as physician-documented glucose abnormalities in clinical notes (exam_glucose, ranked third by SHAP analysis) frequently precede formal diagnosis [48, 51]: these cross-disease early warning signals are precisely what specialist-centered management is most likely to miss. The BERT interaction dependence plot in Supplementary Fig.**??** further shows that the SHAP effect of exam_glucose is significantly modulated by

BERT semantic dimensions, providing direct evidence that text semantics and numerical features form genuine interactions rather than merely additive contributions.

4.5 Clinical Deployment Prospects

When a confirmed DM patient presents for an endocrinology follow-up, the system can automatically output an HTN transition risk score in the hospital information system (HIS) interface. The physician simultaneously receives a cross-disease risk alert while formulating the diabetes management plan, enabling a decision about whether to incorporate blood pressure monitoring or initiate a referral during the current visit. This design embeds cross-specialty screening into the existing visit workflow without additional burden, directly targeting the cross-disease blind spot inherent in specialist-centered management.

A three-tier risk management strategy is proposed. The top 10% high-risk patients (actual positive rates 48%–46%) are recommended for shortened follow-up intervals and enhanced monitoring of relevant indicators. Intermediate-risk patients (10%–30%) are the priority target for proactive health education interventions. Low-risk patients maintain routine follow-up frequency, freeing limited medical resources [12, 25]. Concentrating intensive follow-up on high-risk patients alone would cover approximately half of all true transition cases while achieving a nearly threefold improvement in the concentration of follow-up resources.

The framework is also well-suited for primary care deployment. It relies solely on data sources universally available in primary care institutions, namely outpatient EMR, LIS records, and visit logs. Single-inference latency is below 50 ms (CPU environment, excluding offline BERT inference), and no specialist interpretation is required. The 7-dimensional visit behavior features (zero additional collection cost, AUROC 0.606–0.611) provide a lightweight alternative for resource-constrained settings. This makes the framework practically feasible for embedded deployment within primary care HIS systems, precisely where the cross-disease diagnostic gap caused by specialized fragmented management is most acute.

A natural extension of this work targets community-level high-risk populations. Given the well-established familial aggregation of chronic diseases, first-degree relatives constitute an important group for prospective screening. A lightweight derivative of this framework, combining wearable device data (blood pressure, heart rate, step count), family history questionnaires, and a small number of self-measurable indicators (body weight, waist circumference, fasting blood glucose), holds potential for the development of simplified screening tools for community and household high-risk populations, extending the intervention window further upstream to the pre-onset period.

4.6 Framework Generalizability and Scalability

RD-MMS uses DM-HTN bidirectional prediction as a proof of concept. Unlike existing task-specific models, the modular design of the framework, in which task switching is driven by a configuration file, supports extension to other chronic comorbidity pairs such as CKD-DM, HTN-coronary artery disease, and dyslipidemia-DM [57]: the RD

encoding mechanism and leakage control system are both decoupled from specific disease types, and only the prediction labels and semantic masking list need to be redefined to configure a new task. Complete cross-disease generalizability validation awaits prospective multi-disease studies, which represent the core research direction for the next phase of this framework.

4.7 Limitations

Several limitations of this study warrant acknowledgment. Data are derived from a single center, and cross-institutional generalizability requires validation in prospective multi-center studies [58]. Laboratory test coverage in the outpatient setting is low (approximately 13%–18%), reflecting an inherent sparsity constraint in real-world data. Mild probability overestimation in the high-risk range necessitates Platt scaling calibration before clinical deployment [26]. The prediction time window definition and independent predictive performance for specific DM subtypes require dedicated follow-up studies [13]. Visit behavior features do not distinguish between voluntary visits and scheduled follow-ups, potentially losing information about clinical urgency [58]. AUROC comparisons with existing studies are influenced by dataset differences in scale, positive rate, and data quality, and should be interpreted with caution; multi-center independent validation constitutes a more rigorous external reference.

5 Conclusion

The highly specialized management of chronic diseases creates systematic blind spots in cross-disease risk identification, while longitudinal outpatient data provide a natural foundation for addressing this gap. Through three core design elements, namely reference-deviation-aware encoding, four-dimensional leakage control, and multimodal stacking, the RD-MMS framework systematically addresses the three principal challenges of real-world outpatient EMR: LIS standardization, information leakage, and laboratory sparsity. It achieves predictive performance comparable to or better than existing similar studies under conditions of substantially lower data quality than those typical of inpatient cohort data.

Methodological comparisons demonstrate that a bidirectional unified framework is a necessary condition for identifying task-level information asymmetry, a difference that unidirectional modeling cannot reveal. The OOF-stacking mechanism exhibits markedly superior temporal generalizability compared with single models (external validation drift of 0.013 vs. 0.122 for RF). RD semantic perturbation (AUROC decrease of 0.028–0.030) provides key methodological evidence confirming that performance gains originate from clinical semantic alignment. Temporal order perturbation (decrease of 0.037–0.046) demonstrates that longitudinal visit trajectories are a more fundamental predictive signal than single-visit laboratory tests.

Regarding clinical utility, the top 10% high-risk group achieves 2.9-fold and 2.8-fold enrichment in the event rate, and DCA demonstrates net benefit superior to extreme strategies across the full threshold range, constituting a complete evidence chain from technical performance to clinical practical value. The framework relies solely on data

sources universally available in primary care institutions and is practically feasible for embedded deployment in outpatient and primary care HIS systems. It offers a practical technical pathway for addressing the cross-disease diagnostic blind spot arising from specialized management, and a complete methodological paradigm for constructing reliable and deployable chronic comorbidity prediction models from real-world outpatient data.

Abbreviations

ADA	American Diabetes Association
AP	Average precision
AUROC	Area under the receiver operating characteristic curve
BEHRT	Bidirectional Encoder Representations from Transformers for Electronic Health Records
BERT	Bidirectional encoder representations from transformers
BP	Blood pressure
CatBoost	Categorical boosting
CKD	Chronic kidney disease
CKD-EPI	Chronic Kidney Disease Epidemiology Collaboration
DCA	Decision curve analysis
DM	Diabetes mellitus
EMR	Electronic medical record
ESC	European Society of Cardiology
ESH	European Society of Hypertension
eGFR	Estimated glomerular filtration rate
F1	F1 score
GAIN	Generative adversarial imputation networks
GroupKFold	Group-stratified k-fold cross-validation
HIS	Hospital information system
HTN	Hypertension
IDF	International Diabetes Federation
IQR	Interquartile range
LIS	Laboratory information system
LightGBM	Light gradient boosting machine
LR	Logistic regression
MLP	Multilayer perceptron
NaN	Not a number (missing value indicator)
NCD-RisC	NCD Risk Factor Collaboration
NPV	Negative predictive value
OOF	Out-of-fold
PACS	Picture archiving and communication system
PPV	Positive predictive value
RAAS	Renin-angiotensin-aldosterone system
RD	Reference deviation
RD-MMS	Reference-Deviation-Aware Multimodal Stacking framework
RF	Random forest
ROC	Receiver operating characteristic
SD	Standard deviation
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
TF-IDF	Term frequency-inverse document frequency
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
XGBoost	Extreme gradient boosting

Acknowledgements. Not applicable.

Declarations

Ethics Approval and Consent to Participate

This study was approved by the Medical Ethics Committee of Beijing Pinggu Hospital (Approval No.: 2025-ke-054-01). The study uses fully de-identified retrospective data; informed consent was waived in accordance with the Personal Information Protection Law (2021).

Consent for Publication

Not applicable. All data are fully de-identified and contain no personally identifiable information.

Availability of Data and Materials

The raw data are subject to the Data Security Law (2021) and the Personal Information Protection Law (2021) and cannot be shared publicly. The complete training code will be made available on GitHub upon acceptance, subject to a supervised-access arrangement. Trained model weights (containing no patient data) and aggregate summary statistics are available from the corresponding author upon reasonable request. This study adheres to the TRIPOD reporting guidelines [47]; the complete TRIPOD checklist is submitted as supplementary material.

Competing Interests

All authors declare no competing interests.

Funding

Not applicable.

Authors' Contributions

ST: Writing – original draft, Writing – review & editing;HL: Supervision, Writing – review & editing;LT: Data curation;XZ: Investigation;JZ: Data curation;LS: Visualization;JS: Supervision, Writing – review & editing;MZ: Writing – review & editing;XY: Conceptualization, Methodology, Supervision,Writing – review & editing.All authors read and approved the final manuscript. W

References

1. International Diabetes Federation.: IDF Diabetes Atlas, 11th ed. International Diabetes Federation. Brussels: International Diabetes Federation. Available from: https://diabetesatlas.org.

2. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. Lancet. 2021;398(10304):957–980. https://doi.org/10.1016/S0140-6736(21)01330-1.

3. Wang L, Peng W, Zhao Z, Zhang M, Shi Z, Song Z, et al. Prevalence and treatment of diabetes in China, 2013–2018. JAMA. 2021;326(24):2498–2506. https://doi.org/10.1001/jama.2021.22208.

4. Xu Y, Lu J, Li M, Wang T, Wang K, Cao Q, et al. Diabetes in China part 1: epidemiology and risk factors. Lancet Public Health. 2024;9(12):e1089–e1097. https://doi.org/10.1016/S2468-2667(24)00250-0.

5. Petrie JR, Guzik TJ, Touyz RM. Diabetes, hypertension, and cardiovascular disease: clinical insights and vascular mechanisms. Can J Cardiol. 2018;34(5):575–584. https://doi.org/10.1016/j.cjca.2018.01.024.

6. Sinha S, Haque M. Insulin resistance is cheerfully hitched with hypertension. Life (Basel). 2022;12(4):564. https://doi.org/10.3390/life12040564.

7. Ferrannini E, Cushman WC. Diabetes and hypertension: the bad companions. Lancet. 2012;380(9841):601–610. https://doi.org/10.1016/S0140-6736(12)60987-8.

8. Cheung BMY, Li C. Diabetes and hypertension: is there a common metabolic pathway? Curr Atheroscler Rep. 2012;14(2):160–166. https://doi.org/10.1007/s11883-012-0227-2.

9. Yuan Y, Isasi CR, Al-Rousan T, et al. Associations of concurrent hypertension and type 2 diabetes with mortality outcomes: a prospective study of US adults. Diabetes Care. 2025;48(7):1241–1250. https://doi.org/10.2337/dca24-0118.

10. Martin SS, Aday AW, Almarzooq ZI, et al. 2024 heart disease and stroke statistics: a report of US and global data from the American Heart Association. Circulation. 2024;149(8):e347–e913. https://doi.org/10.1161/CIR.0000000000001209.

11. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31–38. https://doi.org/10.1038/s41591-021-01614-0.

12. Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–1219. https://doi.org/10.1056/NEJMp1606181.

13. Yao L, Li Q, Zhou Z, Yin J, Wang T, Liu Y, et al. Machine learning models for predicting multimorbidity trajectories in middle-aged and elderly adults. Sci Rep. 2025;15:24711. https://doi.org/10.1038/s41598-025-07060-z.

14. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4):e0174944. https://doi.org/10.1371/journal.pone.0174944.

15. Yang H, Fan C, Liu C. Development and validation of a machine learning model to predict comorbid hypertension in patients with type 2 diabetes. Front Med. 2026;13:1754916. https://doi.org/10.3389/fmed.2026.1754916.

16. Lin F, Chen J, Xiao L. Predicting the risk of activities of daily living dysfunction in middle-aged and older adults with comorbid hypertension and diabetes: a national population-based survey analysis. BMC Med Inform Decis Mak. 2026;26:37. https://doi.org/10.1186/s12911-025-03336-6.

17. Alsaleh MM, Allery F, Choi JW, et al. Prediction of disease comorbidity using explainable artificial intelligence and machine learning techniques: a systematic review. Int J Med Inform. 2023;175:105088. https://doi.org/10.1016/j.ijmedinf.2023.105088.

18. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record analysis. IEEE J Biomed Health Inform. 2018;22(5):1589–1604. https://doi.org/10.1109/JBHI.2017.2767063.

19. Bernett J, Blumenthal DB, Grimm DG, et al. Guiding questions to avoid data leakage in biological machine learning applications. Nat Methods. 2024;21(8):1444–1453. https://doi.org/10.1038/s41592-024-02362-y.

20. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. https://doi.org/10.1016/j.patter.2023.100804.

21. Battineni G, Sagaro GG, Chinatalapudi N, Amenta F. Applications of machine learning predictive models in the chronic disease diagnosis. J Pers Med. 2020;10(2):21. https://doi.org/10.3390/jpm10020021.

22. Ding JE, Thao PNM, Peng WC, Wang JZ, Chug CC, Hsieh MC, et al. Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records. Sci Rep. 2024;14:20774. https://doi.org/10.1038/s41598-024-71020-2.

23. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352:i6. https://doi.org/10.1136/bmj.i6.

24. Grout R, Gupta R, Bryant R, et al. Predicting disease onset from electronic health records for population health management: a scalable and explainable deep learning approach. Front Artif Intell. 2024;6:1287541. https://doi.org/10.3389/

frai.2023.1287541.

25. Muse VP, Placido D, Haue AD, Brunak S. Seasonally adjusted laboratory reference intervals to improve the performance of machine learning models for classification of cardiovascular diseases. BMC Med Inform Decis Mak. 2024;24:62. https://doi.org/10.1186/s12911-024-02467-6.

26. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2.

27. Katayev A, Balciza C, Seccombe DW. Establishing reference intervals for clinical laboratory test results: is there a better way? Am J Clin Pathol. 2010;133(2):180–186. https://doi.org/10.1309/AJCPN5BMTSF1CDYP.

28. Jones GRD, Haeckel R, Loh TP, Bullock DG, Chaudhary N, Florkowski CM, et al. Indirect methods for reference interval determination: review and recommendations. Clin Chem Lab Med. 2019;57(1):20–29. https://doi.org/10.1515/cclm-2018-0073.

29. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco; 2016. p. 785–794.

30. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30:3146–3154.

31. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst. 2018;31:6638–6648.

32. Yoon J, Jordon J, van der Schaar M. GAIN: missing data imputation via generative adversarial nets. In: Proceedings of the 35th International Conference on Machine Learning (ICML); 2018. p. 5689–5698.

33. Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:190405342. 2019;.

34. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT; 2019. p. 4171–4186.

35. Verhoeff M, de Groot J, Burgers JS, van Munster BC, Bajpai RC. Development and internal validation of prediction models for future hospital care utilization by patients with multimorbidity using electronic health record data. PLoS One. 2022;17(3):e0260829. https://doi.org/10.1371/journal.pone.0260829.

36. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: transformer for electronic health records. Sci Rep. 2020;10:7155. https://doi.org/10.1038/s41598-020-62922-y.

37. Mancia G, Kreutz R, Brunström M, Burnier M, Grassi G, Januszewicz A, et al. 2023 ESH guidelines for the management of arterial hypertension. J Hypertens. 2023;41(12):1874–2071. https://doi.org/10.1097/HJH.0000000000003480.

38. American Diabetes Association Professional Practice Committee. Standards of care in diabetes—2024. Diabetes Care. 2024;47(Suppl 1):S1–S321. https://doi.org/10.2337/dc24-S001.

39. Inker LA, Eneanya ND, Coresh J, Tighiouart H, Wang D, Sang Y, et al. New equations to estimate GFR. N Engl J Med. 2021;385(19):1737–1749. https://doi.org/10.1056/NEJMoa2102953.

40. Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241–259. https://doi.org/10.1016/S0893-6080(05)80023-1.

41. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol. 2007;6(1):Article 25. https://doi.org/10.2202/1544-6115.1309.

42. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.

43. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov RR. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–1958.

44. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–357. https://doi.org/10.1613/jair.953.

45. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. Explainable AI for trees: from local explanations to global understanding. Nat Mach Intell. 2020;2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9.

46. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NeurIPS); 2017. p. 4765–4774.

47. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.

48. Paneni F, Diaz Cañestro C, Libby P, Lüscher TF, Camici GG. The aging cardiovascular system: understanding it at the cellular and clinical levels. J Am Coll Cardiol. 2017;69(15):1952–1967. https://doi.org/10.1016/j.jacc.2017.01.064.

49. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013;14:106. https://doi.org/10.1186/1471-2105-14-106.

50. Zhang D, Bashar MA, Nayak R, Cuttle L. A novel modality contribution confidence-enhanced multimodal deep learning framework for multiomics data. BMC Bioinformatics. 2025;26:271. https://doi.org/10.1186/s12859-025-06219-9.

51. Uddin S, Khan A, Lu H, Wang H, Moni MA. Multimorbidity prediction using deep learning: a systematic review and meta-analysis. Expert Syst Appl. 2022;189:116034. https://doi.org/10.1016/j.eswa.2021.116034.

52. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19.1 million participants. Lancet. 2017;389:37–55. https://doi.org/10.1016/S0140-6736(16)31919-5.

53. Sinclair AJ, Dunning T, Rodríguez-Mañas L. Diabetes in older people: new insights and remaining challenges. Lancet Diabetes Endocrinol. 2015;3(4):275–285. https://doi.org/10.1016/S2213-8587(14)70235-0.

54. Nghiem N, Wilson N, Krebs J, Tran T. Predicting the risk of diabetes complications using machine learning and social administrative data in a country with ethnic inequities in health: Aotearoa New Zealand. BMC Med Inform Decis Mak. 2024;24:274. https://doi.org/10.1186/s12911-024-02678-x.

55. Tangri N, Grams ME, Levey AS, et al. Multinational assessment of accuracy of equations for predicting risk of kidney failure: a meta-analysis. JAMA. 2016;315(2):164–174. https://doi.org/10.1001/jama.2015.18202.

56. Kautzky-Willer A, Harreiter J, Pacini G. Sex and gender differences in risk, pathophysiology and complications of type 2 diabetes mellitus. Endocr Rev. 2016;37(3):278–316. https://doi.org/10.1210/er.2015-1137.

57. Salisbury C, Man MS, Bower P, Guthrie B, Chaplin K, Gaunt DM, et al. Management of multimorbidity using the 3D approach: a mixed methods feasibility study in primary care. Lancet. 2018;392(10141):41–50. https://doi.org/10.1016/S0140-6736(18)31994-3.

58. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140. https://doi.org/10.1136/bmj.i3140.

Supplementary Files

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

- Supplementarymaterial.docx