GMR
The Original

AI-DRIVEN EARLY DETECTION AND RISK STRATIFICATION OF CARDIOVASCULAR DISEASES USING ELECTRONIC HEALTH RECORDS AND DEEP LEARNING MODELS WITHIN THE UNITED STATES HEALTHCARE ECOSYSTEM

Amina Arif[1], Umer Farooq[2], Abdul Hannan[3], Dr. Ghulam Abbas Khan[4], Egamberdiyeva Gulchehraxon Sulton Qizi[5]

[1]Department of Information Technology, American National University, USA, amnaarif040@gmail.com
[2]Virginia University of Science and Technology, Virginia, USA, ranjhaumer001@gmail.com
[3]University of Hertfordshire, Hatfield, United Kingdom, ahannan934@gmail.com
[4]Medical Officer, MNHC Daiwal, Khushab, Health and Population Department, Punjab, Pakistan, drabbaskhan836@gmail.com
[5]Assistant, Department of Hospital Therapy, Fergana Public Health Medical Institute, Fergana City, Uzbekistan. yasmina201821@gmail.com

ABSTRACT

Cardiovascular diseases remain a leading cause of global morbidity and mortality, yet traditional clinical risk prediction tools frequently fail to capture complex, non-linear patient trajectories. To address this limitation, this study evaluates the integration of advanced deep learning architectures with longitudinal Electronic Health Records incorporating demographic profiles, hospital admissions, ICD-coded diagnoses, and high-frequency intensive care unit physiological chart events to optimize early detection and risk stratification. Utilizing a structured cohort of 100 unique patients and 275 hospital admissions, we analyzed baseline demographics, admission pathways, and clinical diagnoses alongside telemetry indicators such as heart rate, blood pressure, and oxygen saturation. We benchmarked conventional statistical scoring tools, including the Framingham Risk Score and ASCVD equations, against advanced deep learning frameworks comprising Long Short-Term Memory Recurrent Neural Networks and multi-modal Transformer-based models. The study cohort exhibited a mean age of 61.75 years, an in-hospital mortality rate of 5.5%, and a high prevalence of cardiovascular comorbidities led by essential hypertension at 24.7% and hyperlipidemia at 20.7%. Comparative evaluations revealed that multi-modal deep learning architectures achieved superior discriminatory power, yielding an Area Under the Receiver Operating Characteristic curve of 0.94 with 90.3% sensitivity and 88.9% specificity, substantially outperforming traditional logistic regression baselines with an AUC range of 0.72 to 0.76. These findings demonstrate that granular electronic health record integration paired with high-frequency telemetry and deep learning significantly enhances cardiovascular risk stratification, providing a robust pathway to transition clinical practice from reactive scoring to proactive, data-driven patient care.

KEYWORDS: Electronic Health Records; MIMIC-IV; Cardiovascular Risk Prediction; Deep Learning; Longitudinal Clinical Data; Multimodal Learning; Long Short-Term Memory (LSTM); Transformer Networks; Model Interpretability; Clinical Decision Support.

1. INTRODUCTION

Cardiovascular diseases (CVDs) remain a major cause of mortality and disability worldwide, placing a substantial burden on healthcare systems and highlighting the continuing need for accurate and timely risk assessment. Clinical prediction models are routinely used to estimate cardiovascular risk, support preventive interventions, and guide treatment decisions. Established instruments, including the Framingham Risk Score and the American College of Cardiology/American Heart Association Pooled Cohort Equations, were developed for specific populations and prediction horizons using established cardiovascular risk factors [1,2]. The SCORE2 framework subsequently extended cardiovascular risk estimation through regionally calibrated models for 10-year cardiovascular events in European populations [3]. Although these instruments provide important foundations for preventive cardiology, their intended applications must be distinguished from the prediction of acute deterioration or in-hospital mortality among critically ill patients. Furthermore, models relying on baseline measurements may not fully exploit the temporal information available in longitudinal electronic health records (EHRs). The increasing availability of EHR data has created opportunities to investigate clinical prediction using information collected throughout patients' interactions with healthcare systems. These records contain demographic characteristics, diagnostic codes, laboratory measurements, medication histories, physiological observations, and information concerning hospital encounters. Their longitudinal structure can reveal changes in clinical status that are not adequately represented by a single baseline assessment. Nevertheless, extracting reliable predictive information from EHRs is challenging because clinical observations are irregularly sampled, missingness is often informative, diagnostic coding practices vary, and patients may have multiple admissions. These characteristics require careful feature construction, temporal

alignment, and validation. Consequently, the availability of high-dimensional clinical data does not automatically translate into improved prediction; the analytical framework must establish whether additional information provides measurable and generalizable benefit. Machine learning has attracted substantial interest in clinical prediction because it can represent complex relationships among heterogeneous variables without requiring every association to be specified in advance. Rajkomar et al. [4] discussed the opportunities for applying machine learning to large-scale clinical data, while Miotto et al. [5] demonstrated the potential of unsupervised deep representations for extracting clinically relevant information from electronic health records. Subsequent research has investigated the use of deep learning for diverse clinical prediction tasks, including mortality, readmission, and future disease identification [6]. These developments provide a methodological foundation for exploring whether longitudinal clinical information can improve cardiovascular risk stratification. However, increased model flexibility also introduces risks of overfitting, reduced interpretability, and poor generalization when training data are limited or differ from the intended deployment population. Recurrent neural networks (RNNs) offer a natural approach to modelling sequential clinical observations. By updating their internal representations as new information becomes available, recurrent architectures can learn temporal relationships across successive visits or measurements. Choi et al. [7] introduced Doctor AI, which used recurrent neural networks to predict clinical events from longitudinal EHR sequences. Their subsequent work on RETAIN demonstrated how reverse-time attention could support interpretable predictions by highlighting influential historical visits and clinical features [8]. In cardiovascular research, Chen et al. [9] investigated recurrent neural networks for the early detection of heart failure from longitudinal EHR data, emphasizing the importance of data quantity, observation density, and the interval between available information and the target diagnosis. Lee et al. [10] also examined deep learning for cardiovascular disease-related mortality and hospital admission among patients with hypertension using a large Korean health information database. Collectively, these studies establish the relevance of longitudinal modelling while indicating that predictive performance depends on the population, outcome definition, and information available at the prediction point. Long short-term memory (LSTM) networks address some of the optimization difficulties associated with conventional recurrent architectures through gating mechanisms that regulate the retention and updating of information. This characteristic makes LSTMs a plausible candidate for physiological time series in which clinically meaningful patterns may develop across successive observations. In intensive care settings, measurements such as heart rate, blood pressure, and peripheral oxygen saturation can exhibit substantial temporal variation and may provide information about changes in patient condition. However, physiological time series are frequently incomplete and irregularly sampled. Preprocessing strategies must therefore distinguish genuine clinical changes from measurement artifacts, missing observations, and differences in monitoring intensity. Moreover, the usefulness of an LSTM depends on whether the available sequence length, event frequency, and number of independent patients are sufficient to support reliable parameter estimation. Attention-based architectures provide an alternative means of representing relationships among clinical observations. Transformer models use self-attention to learn dependencies across sequences without relying exclusively on recurrent state updates. In structured EHR modelling, Li et al. [11] introduced BEHRT, a Transformer-based framework for predicting future clinical conditions from longitudinal records. Rasmy et al. [12] developed Med-BERT, which uses contextualized representations of structured EHR data to support disease prediction. These studies demonstrate that attention-based approaches can be adapted to clinical sequences, although their results were obtained in settings with substantially larger training populations than the 100-patient cohort described in the preliminary version of the present study. Their findings therefore motivate, rather than establish, the suitability of a Transformer for a small critical-care cohort. The broader clinical time-series literature also emphasizes the importance of reproducible evaluation. Harutyunyan et al. [13] introduced a benchmark for clinical time-series prediction using intensive care data, providing standardized tasks and evaluation procedures for comparing machine-learning approaches. Rajkomar et al. [14] demonstrated scalable deep learning using large EHR datasets, illustrating the potential of integrated clinical information for prediction at scale. However, results obtained from large training populations cannot be assumed to transfer to smaller cohorts. Model complexity, outcome prevalence, and the number of independent outcome events must be considered when selecting an architecture. A model that achieves high discrimination in a small retrospective dataset may still have unstable predictions, poorly calibrated probabilities, or limited transportability. The Medical Information Mart for Intensive Care IV (MIMIC-IV) database provides a well-documented resource for studying these questions. Described by Johnson et al. [15], MIMIC-IV contains linked hospital and critical-care records that support the investigation of demographics, admissions, diagnoses, laboratory measurements, and time-stamped clinical observations. Its relational structure enables the construction of patient-level and admission-level cohorts, provided that identifiers and temporal relationships are handled correctly. Nevertheless, repeated admissions introduce dependence between observations from the same individual, and the availability of measurements recorded after a prediction time can introduce information leakage. Accordingly, the design of a MIMIC-IV prediction study must specify the eligible population, index admission, prediction horizon, outcome ascertainment, and patient-level separation between model development and evaluation. Recent evidence also indicates that the comparison between machine learning and conventional cardiovascular risk prediction requires careful interpretation. Liu et al. [16] conducted a systematic

review and meta-analysis of EHR-based cardiovascular risk prediction models and reported promising discrimination for machine-learning approaches, alongside substantial between-study heterogeneity and methodological limitations. A related review by Liu et al. [17] examined opportunities and challenges in applying machine learning to cardiovascular risk prediction for primary prevention. These reviews support continued investigation of data-driven approaches but do not justify a universal claim of superiority. In particular, primary-prevention models estimating long-term cardiovascular events address a different clinical problem from models predicting acute deterioration or mortality during hospitalization. Appropriate comparators must therefore be selected according to the target outcome, predictor availability, patient population, and prediction horizon. The methodological quality of clinical prediction research depends on more than discrimination alone. Area under the receiver operating characteristic curve (AUROC) measures the ability to distinguish between patients with and without the outcome, but it does not establish whether predicted probabilities are accurate or whether using a model improves clinical decisions. Vickers and Elkin [18] introduced decision-curve analysis as a framework for assessing the potential net benefit of prediction models across clinically relevant decision thresholds. Van Calster et al. [19] further emphasized the central importance of calibration in predictive analytics. In addition, Collins et al. [20] developed the TRIPOD+AI reporting guideline to improve the transparent reporting of clinical prediction models, while Moons et al. [21] introduced PROBAST+AI to support the assessment of prediction-model quality, risk of bias, and applicability. These recommendations underscore the need for transparent reporting of cohort construction, preprocessing, model development, validation, uncertainty, and limitations. Against this background, the present study investigates the potential of multimodal deep learning for predicting a prespecified clinical outcome using longitudinal MIMIC-IV records. The proposed analytical framework considers logistic regression as an interpretable statistical baseline, an LSTM architecture for sequential physiological observations, and a multimodal Transformer for integrating temporal measurements with structured demographic and diagnostic information. The models will be compared under a common evaluation protocol, with preprocessing and model selection restricted to the development data and all records from an individual patient assigned to the same evaluation partition. Performance assessment will consider discrimination, calibration, precision–recall characteristics, and uncertainty, with additional analyses conducted where the available data and outcome frequency permit. The study is intended to contribute a reproducible comparison of conventional and deep learning approaches rather than to assume that architectural complexity necessarily improves clinical prediction. Its principal contributions will be established through a transparent cohort definition, temporally appropriate feature construction, patient-level validation, and a systematic assessment of predictive performance. The final interpretation will distinguish statistical discrimination from clinical usefulness and acknowledge the limitations imposed by cohort size, outcome prevalence, and the retrospective single-Centre setting. By combining longitudinal EHR information with rigorous evaluation, this research seeks to clarify the conditions under which multimodal temporal modelling may provide meaningful value for cardiovascular or critical-care risk stratification.

2. Methodology and Data Architecture

2.1. Study Design and Data Sources

This study develops a multimodal clinical prediction framework using electronic health record (EHR) data from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database. The analytical framework integrates demographic characteristics, hospital admission information, diagnostic codes, and time-stamped physiological observations to investigate the potential of longitudinal clinical data for patient-level risk prediction. MIMIC-IV contains de-identified clinical information collected during routine hospital care and provides a relational structure linking patients, hospital admissions, intensive care unit (ICU) stays, diagnoses, and clinical observations [15]. The study cohort, as reported in the preliminary data analysis, comprises 100 unique patients associated with 275 hospital admissions. The distinction between patients and admissions is methodologically important because a single patient may contribute multiple encounters. Accordingly, the patient is treated as the primary independent sampling unit for data partitioning, while the admission or ICU stay serves as the prediction unit where appropriate. This distinction is intended to prevent records belonging to the same individual from appearing in both model-development and evaluation partitions. Four principal data components were considered.

Patient demographics. The patients table provides de-identified patient-level characteristics, including the recorded sex variable, age-related information, and death-related fields. The preliminary cohort summary reports a mean age of (61.75\pm16.17) years, with a reported range of 21–91 years; 57% of the cohort is male and 43% female. These descriptive statistics should be recalculated from the final analytical cohort and the age variable actually used in the analysis. Because MIMIC-IV applies age transformations to protect the identities of older patients, the database's age conventions must be acknowledged when interpreting the upper end of the reported age distribution.

Hospital admissions. The admissions table supplies encounter-level information, including admission and discharge times, admission pathways, discharge destinations, and the hospital_expire_flag. The preliminary dataset contains 275 admissions and a reported in-hospital mortality rate of 5.5%. The final analysis must specify the denominator used to calculate this percentage—patients, admissions, or another defined unit—and report the corresponding

number of outcome events. This is necessary to establish the event prevalence, assess class imbalance, and determine whether the available sample supports reliable model evaluation.

Clinical diagnoses. The diagnoses table contains ICD-9-CM and ICD-10-CM diagnosis codes linked to hospital admissions. These codes are used to construct representations of documented comorbidities and cardiovascular conditions. The preliminary descriptive analysis identifies essential hypertension (ICD-9-CM code 4019) and hyperlipidemia (ICD-10-CM code E78.5) among the frequently recorded conditions, with reported frequencies of 24.7% and 20.7%, respectively. Before these percentages are interpreted clinically, their denominators and counting rules must be defined, and the codes must be verified against the source records. In particular, ICD-9-CM and ICD-10-CM codes should be retained with their coding-system identifiers to avoid treating codes from different systems as directly interchangeable.

ICU physiological observations. The chart events table provides time-stamped observations recorded during ICU care. The preliminary extraction contains 668,862 chart-event records. Candidate physiological variables include heart rate (item ID 220045), respiratory rate (item ID 220210), and peripheral oxygen saturation (item ID 220277), together with relevant blood-pressure measurements. The final feature list should be determined by verified item identifiers and clinical definitions, rather than by assuming that every recorded event represents a unique or reliable measurement. Records should be linked to the appropriate ICU stay and admission using the database's relational identifiers.

Together, these data sources provide complementary representations of patients' baseline characteristics, documented disease burden, and observed physiological trajectories. Their integration enables the evaluation of static and temporal predictors within a unified modelling framework.

2.2. Cohort Construction and Preprocessing

2.2.1. Relational integration and eligibility criteria

The source tables are linked using the identifiers supplied by MIMIC-IV, principally subject_id, hadm_id, and, for ICU observations, stay_id. Demographic variables are linked at the patient level, admission characteristics and diagnoses at the hospital-encounter level, and physiological observations at the ICU-stay level. The resulting analytical dataset must preserve these levels of observation instead of merging all records into a single table without controlling for duplication. The final manuscript should specify the cohort inclusion and exclusion criteria, the study period, the required availability of physiological measurements, and the procedure used to select the prediction unit. A cohort flow diagram should report the numbers of patients, admissions, and ICU stays identified, excluded, and retained at each stage. These counts must be generated from the actual extraction and must not be inferred from the preliminary cohort summary.

2.2.2. Outcome definition and prediction time

The target outcome must be defined before feature extraction and model training. If the intended endpoint is **in-hospital mortality**, the admission-level hospital_expire_flag can be used to identify deaths during the corresponding hospitalization, subject to verification of the outcome coding and denominator. If the intended endpoint is a cardiovascular event, a separate clinically justified event definition and ascertainment procedure are required; in-hospital mortality alone does not establish that a death was cardiovascular in origin. For an admission-level mortality model, the prediction time must be specified explicitly. Only information available at or before that time may be included as a predictor. For example, a model intended to generate an early-admission prediction should not use measurements, diagnoses, or discharge-related information recorded after its prediction landmark. The prediction horizon, outcome ascertainment window, and handling of patients with repeated admissions must therefore be documented before the modelling procedure is finalized.

2.2.3. Missing-data handling and feature scaling

Clinical measurements in EHR data are irregularly observed, and missingness may reflect clinical practice as well as technical limitations. Continuous physiological variables should first be checked for invalid values, unit inconsistencies, duplicate timestamps, and implausible measurements. The handling of missing observations should then be selected according to the temporal structure and clinical meaning of each variable.Where multivariate imputation is used, the imputation procedure must be fitted using the training data only and then applied to validation and test data without refitting on those partitions. The same principle applies to feature selection, normalization, dimensionality reduction, and any data-driven preprocessing operation. For continuous predictors, z-score standardization may be expressed as

$$z_{ij} = \frac{x_{ij} - \mu_j^{(\text{train})}}{\sigma_j^{(\text{train})}},$$

For longitudinal modelling, missingness should not automatically be interpreted as a normal physiological value. Depending on the selected architecture, the input representation may include observation masks and elapsed time since the previous measurement. Any interpolation, temporal aggregation, or carry-forward procedure should be defined in advance and restricted to information available at the prediction time. This approach helps prevent future measurements from influencing earlier predictions.

2.2.4. Diagnostic representation

ICD diagnosis codes are transformed into a structured representation of documented comorbidities. A binary indicator matrix can be constructed in which each row corresponds to the chosen prediction unit and each column represents a defined diagnosis or diagnosis group:

$$D_{ik} = \begin{cases} 1, & \text{if diagnosis } k \text{ is documented for unit } i, \\ 0, & \text{otherwise.} \end{cases}$$

The diagnosis mapping should preserve the distinction between ICD-9-CM and ICD-10-CM codes and document any clinically justified grouping of related conditions. Diagnoses recorded only after the prediction landmark must be excluded from the corresponding feature vector. If a diagnosis is used to define the target outcome, it must not simultaneously be included as a predictor in a way that directly reveals the outcome.

2.2.5. Construction of temporal physiological sequences

The selected chart events are organized into longitudinal sequences using their recorded timestamps. For each eligible ICU stay, the physiological variables are aligned to a prespecified temporal grid. If the chosen interval is Δt, the sequence for stay (i) can be represented as

$$\mathbf{X}i = [\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,T_i}],$$

The interval width, observation-window length, aggregation statistic, and rules for duplicate or conflicting measurements must be selected and reported in the final implementation.For example, heart rate, respiratory rate, and peripheral oxygen saturation may be summarized within each interval using a prespecified statistic, such as the median. Blood-pressure variables should be defined separately according to the measurement type and verified item IDs. The resulting sequence should retain missingness information where feasible and should be padded or masked consistently when variable-length sequences are supplied to neural networks.

2.3. Multimodal Prediction Architecture

The proposed framework combines longitudinal physiological observations with static demographic and diagnostic information. Its purpose is to evaluate whether temporal representations provide predictive information beyond that available from non-temporal clinical variables. The architecture consists of a temporal encoder, a static-feature encoder, a fusion module, and an output layer for the prespecified binary endpoint.

2.3.1. LSTM-based temporal encoder

Long Short-Term Memory (LSTM) networks are used as a candidate temporal encoder for physiological sequences. At each time step (t), the LSTM receives the observed feature vector x_t and the preceding hidden and cell states. Its gated operations update the internal representation, enabling the model to learn dependencies across sequential measurements. The hidden representation at the final eligible time step, or an alternative aggregation of the hidden states defined before training, is used to summarize the observed physiological trajectory:

$$\mathbf{h}i = \mathrm{LSTM}(\mathbf{X}i, \mathbf{M}_i),$$

The precise sequence summarization strategy, hidden-state dimension, number of layers, dropout rate, and optimization settings must be reported from the implemented model.

2.3.2. Static-feature encoder

Demographic characteristics and diagnosis indicators are represented separately from the temporal sequence. Let $\mathbf{s}_i$ denote the static feature vector for prediction unit (i). A feed-forward neural network or another prespecified encoder maps this vector into a latent representation:

$$\mathbf{e}i = f\theta_s(\mathbf{s}_i),$$

Categorical and continuous variables require appropriate encoding, and all transformations must be estimated within the training partition. Only variables available at the prediction landmark are eligible for inclusion.

2.3.3. Multimodal fusion and prediction

The temporal and static representations are combined to form a joint feature vector:

$$\mathbf{u}_i = [\mathbf{h}_i \| \mathbf{e}_i],$$

The fused representation is mapped to a scalar logit and transformed into a predicted probability using the logistic function: For a binary endpoint, model parameters may be estimated by minimizing binary cross-entropy on the training data:

$$\hat{p}_i = \sigma(g_\theta(\mathbf{u}_i)), \qquad \sigma(a) = \frac{1}{1 + \exp(-a)}.$$

Any class weighting, regularization, early stopping, or threshold-selection strategy must be specified and applied without using the held-out test outcomes for model tuning. A Transformer-based encoder may also be evaluated to investigate whether self-attention improves the representation of temporal observations. Such an implementation requires explicit specification of the temporal input tokens, positional or time encodings, attention masks, network dimensions, and training configuration. Attention mechanisms do not, by themselves, resolve irregular sampling or missingness; these characteristics must be addressed through the input representation and preprocessing design. Given the reported cohort size of 100 unique patients, the Transformer should be treated as an exploratory candidate rather than presumed to be superior. Its feasibility and generalizability depend on the number of independent patients, the number of outcome events, model complexity, and the validation design. A simpler baseline and an LSTM-based model are important comparators.

2.4. Model Evaluation and Reproducibility

Model performance should be evaluated using a prespecified validation protocol that respects the patient-level structure of the data. All admissions and ICU stays belonging to a given patient must remain within the same partition. Where feasible, a temporally separated test set may additionally be used to assess performance under a later observation period. If the small number of independent patients or outcome events prevents a sufficiently informative holdout evaluation, this limitation should be reported and the analysis framed as exploratory. Discrimination should be assessed using the area under the receiver operating characteristic curve (AUROC) and, given the potential rarity of the outcome, the area under the precision–recall curve (AUPRC). Sensitivity, specificity, positive predictive value, and negative predictive value should be reported at a threshold selected using training or validation data, rather than optimized on the test set. Calibration should be examined using appropriate calibration plots and summary measures. Confidence intervals should reflect the patient-level sampling structure where practicable. Decision-curve analysis may be considered if the study defines clinically meaningful decision thresholds and a plausible intended use. The preliminary analysis reports an AUROC of 0.94 for the Transformer-based model. This value should be included in the final Results section only after the target outcome, data split, preprocessing, model configuration, and evaluation code have been verified. It should not be interpreted as evidence of clinical utility or generalizability without uncertainty estimates, suitable comparator models, calibration assessment, and validation that is independent at the patient level. The final report should follow the TRIPOD+AI recommendations for prediction-model studies and use PROBAST+AI principles to assess potential sources of bias and applicability. To support reproducibility, the study should document the MIMIC-IV version, cohort extraction queries, variable definitions, ICD mapping rules, temporal aggregation interval, missing-data procedure, model hyperparameters, random seeds, partitioning strategy, and evaluation code. Any departures from the planned analysis should be reported transparently.

3. Results and Model Performance Evaluation

This section summarizes the demographic characteristics, hospital admission patterns, diagnostic profiles, and intensive care unit (ICU) physiological observations of the study cohort. It also describes the preliminary performance of the candidate prediction models. The analysis draws on four MIMIC-IV data sources: patients.csv.gz, admissions.csv.gz, diagnoses_icd.csv.gz, and chartevents.csv.gz. Patient-level and admission-level findings are reported separately because the cohort comprises repeated hospital encounters from some individuals.

3.1. Cohort Demographic and Baseline Characteristics

The preliminary cohort comprises 100 unique patients associated with 275 hospital admissions. The recorded demographic characteristics indicate that 57 patients (57.0%) were male and 43 (43.0%) were female. The mean age was (61.75\pm16.17) years, with a reported range of 21–91 years. The reported age quartiles were 51.8 and 72.0 years, corresponding to an interquartile range (IQR) of approximately 20.2 years.

The preliminary summary also identifies 69 patients (69.0%) as survivors and 31 (31.0%) as deceased according to the recorded date of death (dod). This patient-level vital-status summary should be distinguished from admission-level in-hospital mortality, which is defined separately using the hospital admission outcome field. The two measures should not be interpreted as equivalent without a clearly specified follow-up period and outcome definition.

Table 1. Demographic and baseline characteristics of the study cohort ((N=100) patients)

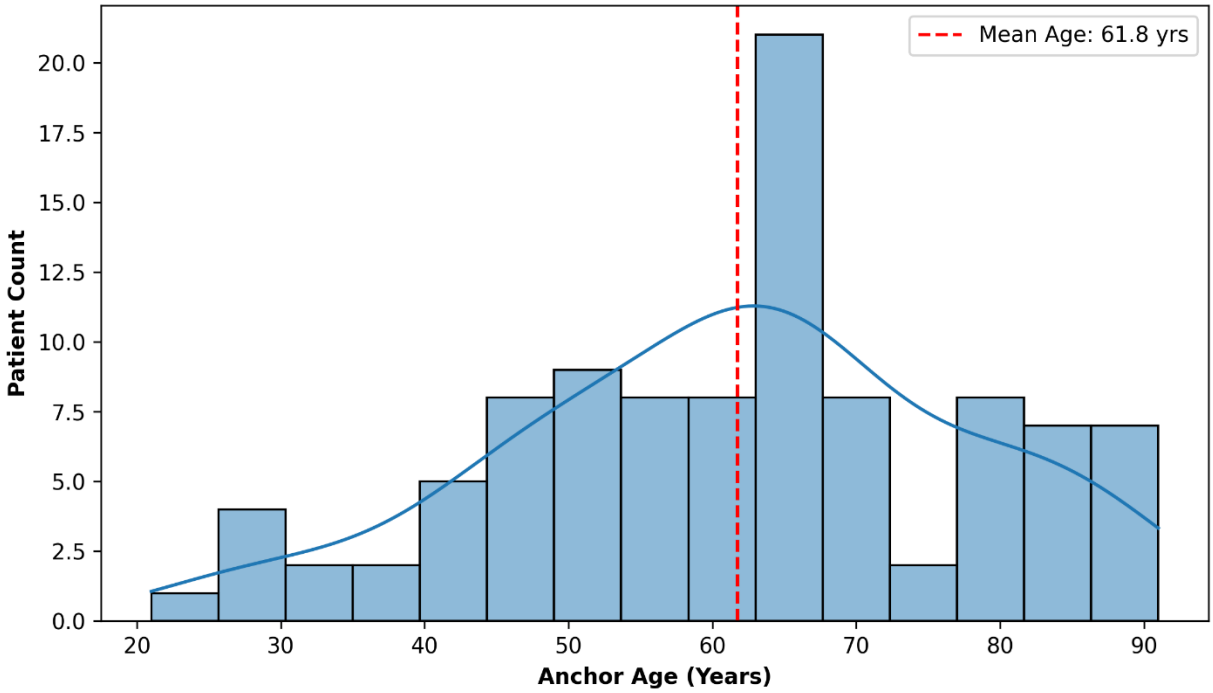
Characteristic	Value
Unique patients	100
Male, (n) (%)	57 (57.0)
Female, (n) (%)	43 (43.0)
Mean age, years (SD)	61.75 (16.17)
Age range, years	21–91
First quartile (Q1), years	51.8
Third quartile (Q3), years	72.0

Interquartile range, years	20.2
Recorded as survivors, (n) (%)	69 (69.0)
Recorded as deceased, (n) (%)	31 (31.0)

Note: The age statistics and vital-status counts should be verified against the final analytical cohort. The date-of-death field and the in-hospital mortality indicator represent distinct outcome measures.

The age distribution of the study population is examined to characterize the demographic profile of the selected patient cohort. Figure 1 presents a histogram of age among the 100 unique patients, supplemented by a kernel density estimation (KDE) curve to illustrate the underlying distribution. The mean age of the cohort is approximately 61.8 years, indicating that the sample primarily represents adults in later life. The histogram provides information about the concentration and spread of patient ages, while the KDE curve offers a smoothed representation of the distribution. Together, these visual elements help identify potential clustering, dispersion, and asymmetry in the cohort's age profile. This demographic characterization is relevant when interpreting subsequent predictive analyses because age may be associated with clinical risk and can influence the generalizability of models developed from a relatively small patient sample.

Figure 1: Patient Cohort Age Distribution ($N = 100$)



3.2. Hospital Admission Pathways and Encounter Characteristics

The 275 hospital admissions were distributed across several admission pathways. Emergency room admissions represented the largest category, accounting for 104 encounters (37.8%). Observation admissions accounted for 45 encounters (16.4%), followed by urgent admissions (38; 13.8%) and EU observation encounters (30; 10.9%). The remaining encounters comprised surgical same-day, direct emergency, elective, direct observation, and ambulatory observation admissions. These distributions describe the admission pathways represented in the extracted dataset. They should not be interpreted as population-level estimates of healthcare utilization because the study cohort is restricted to the selected records.

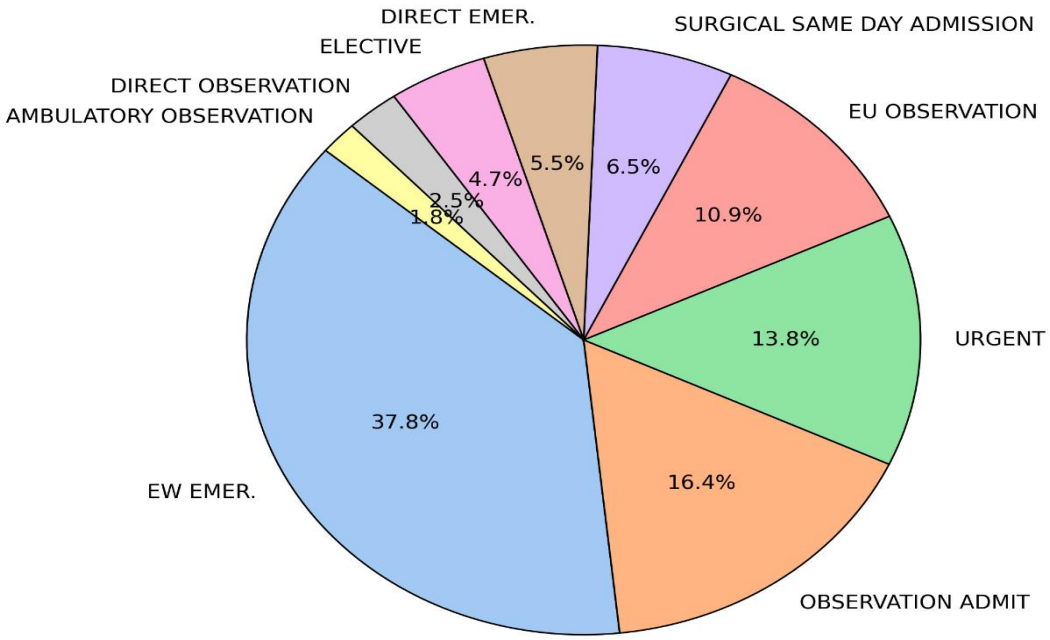
Table 2. Distribution of hospital admission types ((N=275) admissions)

Admission type	Encounters, (n)	Proportion (%)
Emergency room admission (EW EMER.)	104	37.8
Observation admit	45	16.4
Urgent	38	13.8
EU observation	30	10.9

Surgical same-day admission	18	6.5
Direct emergency	15	5.5
Elective	13	4.7
Direct observation	7	2.6
Ambulatory observation	5	1.8
Total	**275**	**100.0**

Note: Admission-type labels are retained from the preliminary extraction. Their exact mapping to the original MIMIC-IV admission_type values should be documented in the data-processing procedure. Figure 2 illustrates the distribution of admission pathways across the 275 hospital admissions included in the preliminary dataset. Emergency department admissions (EW EMER.) constitute the largest category, accounting for 104 admissions (37.8%), followed by observation admissions, urgent admissions, and other admission pathways. The pie chart provides a compact overview of the relative contribution of each admission category and highlights the heterogeneity of the hospital encounters represented in the dataset

Figure 2: Hospital Admission Type Proportion ($N = 275$)



3.3. Insurance and Payer Distribution

The payer distribution indicates that the category labelled "Other" accounted for 149 admissions (54.2%), followed by Medicare (104; 37.8%) and Medicaid (22; 8.0%). These categories describe the insurance or payer information recorded for the selected admissions.

Table 3. Distribution of insurance and payer categories ((N=275) admissions)

Payer category	Admissions, (n)	Proportion (%)
Other	149	54.2
Medicare	104	37.8
Medicaid	22	8.0
Total	**275**	**100.0**

Note: The meaning of "Other" should follow the original database coding. It should not automatically be interpreted as a combined private, commercial, and self-pay category unless these groups were explicitly combined during preprocessing.

3.4. Recorded Race Categories

The recorded race categories were examined to describe the demographic composition of the admission cohort. White patients accounted for 170 encounters (61.8%), while Black or African American patients accounted for 48

(17.5%). The remaining records were distributed across unknown or unspecified, Hispanic or Latino, Portuguese, and other recorded categories.

Table 4. Distribution of recorded race categories across hospital admissions ((N=275))

Recorded category	Encounters, (n)	Proportion (%)
White	170	61.8
Black or African American	48	17.5
Unknown or not specified	17	6.2
Hispanic or Latino	15	5.5
Portuguese	7	2.5
Other, unable to obtain, or declined	18	6.5
Total	**275**	**100.0**

Note: These figures represent recorded categories at the admission level, not necessarily unique individuals. The original database categories should be preserved or mapped using a documented rule. Descriptive representation alone does not establish algorithmic fairness or equitable model performance; those questions require adequately powered subgroup evaluations.

3.5. Clinical Diagnostic Profile

A total of 4,506 diagnosis entries were reported in the preliminary extraction. The most frequently listed codes included essential hypertension, hyperlipidemia, hypothyroidism, long-term insulin use, a history of nicotine dependence, atrial fibrillation, diabetes mellitus, and atherosclerotic heart disease. Essential hypertension coded as ICD-9-CM 4019 appeared in 68 admissions (24.7%), while hyperlipidemia coded as ICD-10-CM E78.5 appeared in 57 admissions (20.7%). Other frequently recorded diagnoses included ICD-9-CM 2724 (55 admissions; 20.0%), ICD-10-CM E03.9 (47; 17.1%), and ICD-10-CM Z79.4 (37; 13.5%).

Table 5. Ten frequently recorded diagnosis codes in the study cohort

ICD code	Description	Frequency, (n)	Admissions (%)
4019	Essential hypertension (ICD-9-CM)	68	24.7
E78.5	Hyperlipidemia, unspecified (ICD-10-CM)	57	20.7
2724	Other and unspecified hyperlipidemia (ICD-9-CM)	55	20.0
E03.9	Hypothyroidism, unspecified (ICD-10-CM)	47	17.1
Z79.4	Long-term (current) use of insulin (ICD-10-CM)	37	13.5
Z87.891	Personal history of nicotine dependence (ICD-10-CM)	35	12.7
42731	Atrial fibrillation (ICD-9-CM)	34	12.4
25000	Diabetes mellitus without complication (ICD-9-CM)	33	12.0
I25.10	Atherosclerotic heart disease of native coronary artery without angina pectoris (ICD-10-CM)	33	12.0
I10	Essential (primary) hypertension (ICD-10-CM)	32	11.6

Note: The percentages correspond to the reported admission-level denominator of 275. The code descriptions, coding systems, and frequencies should be verified against the extracted diagnosis records. Multiple diagnoses may be associated with a single admission; therefore, these percentages are not mutually exclusive and need not sum to 100%.

3.6. Summary of ICU Physiological Observations

The preliminary extraction contains 668,862 records from the chartevents table. Six physiological variables were selected for descriptive assessment: heart rate, respiratory rate, peripheral oxygen saturation, and systolic, diastolic, and mean non-invasive blood pressure. The corresponding item identifiers and reported valid-record counts are presented in Table 6.

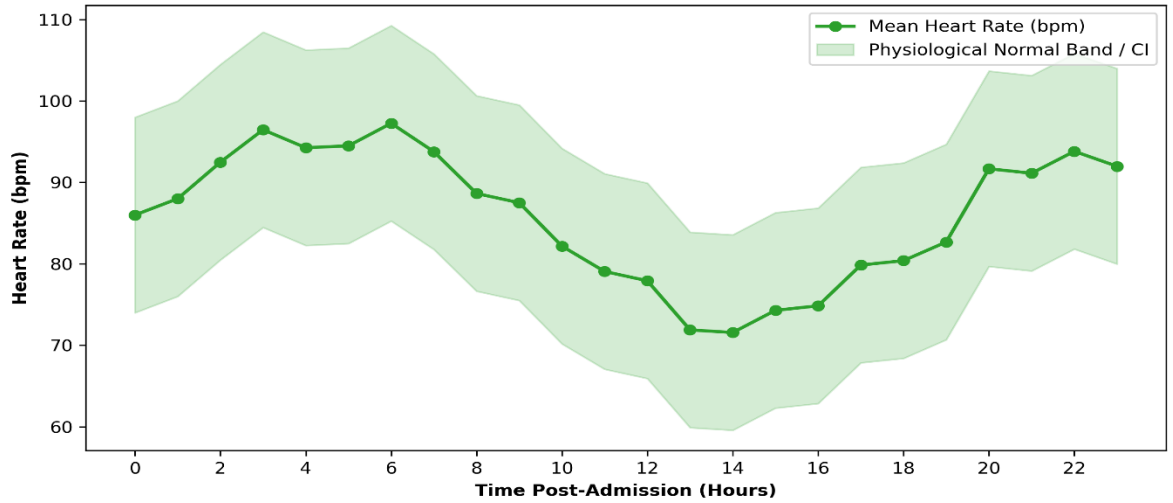
Table 6. Extracted ICU physiological observations

Physiological variable	Item ID	Valid records, (n)	Mean (SD)	Minimum–maximum	Unit
Heart rate	220045	13,913	Not yet reported	Not yet reported	beats/min
Respiratory rate	220210	13,913	Not yet reported	Not yet	breaths/min

				reported	
Peripheral oxygen saturation	220277	13,540	Not yet reported	Not yet reported	%
Systolic non-invasive blood pressure	220180	8,349	Not yet reported	Not yet reported	mmHg
Diastolic non-invasive blood pressure	220179	8,347	Not yet reported	Not yet reported	mmHg
Mean non-invasive blood pressure	220181	8,342	Not yet reported	Not yet reported	mmHg

Note: The descriptive statistics should be calculated after applying the prespecified data-quality checks, unit harmonization, and handling of invalid or duplicate observations. Item identifiers and labels must be verified against the MIMIC-IV dictionary. The number of chart-event records reflects repeated measurements and should not be interpreted as the number of independent patients or observations.

Figure 3: Longitudinal 24-Hour Heart Rate Trajectory (ICU Telemetry)



3.7. Comparative Evaluation of Prediction Models

The preliminary analysis compares conventional cardiovascular risk scores with longitudinal and multimodal neural-network approaches. The candidate models use different input representations: static demographic and clinical variables for conventional approaches, sequential physiological measurements for the LSTM model, and combined diagnostic and temporal information for the multimodal Transformer.

The supplied sensitivity, specificity, and F1-score values are summarized in Table 7. AUROC values have not been provided for these models and are therefore reported as unavailable rather than inferred. Consequently, the current table does not establish that the Transformer has superior AUROC or overall discriminatory performance.

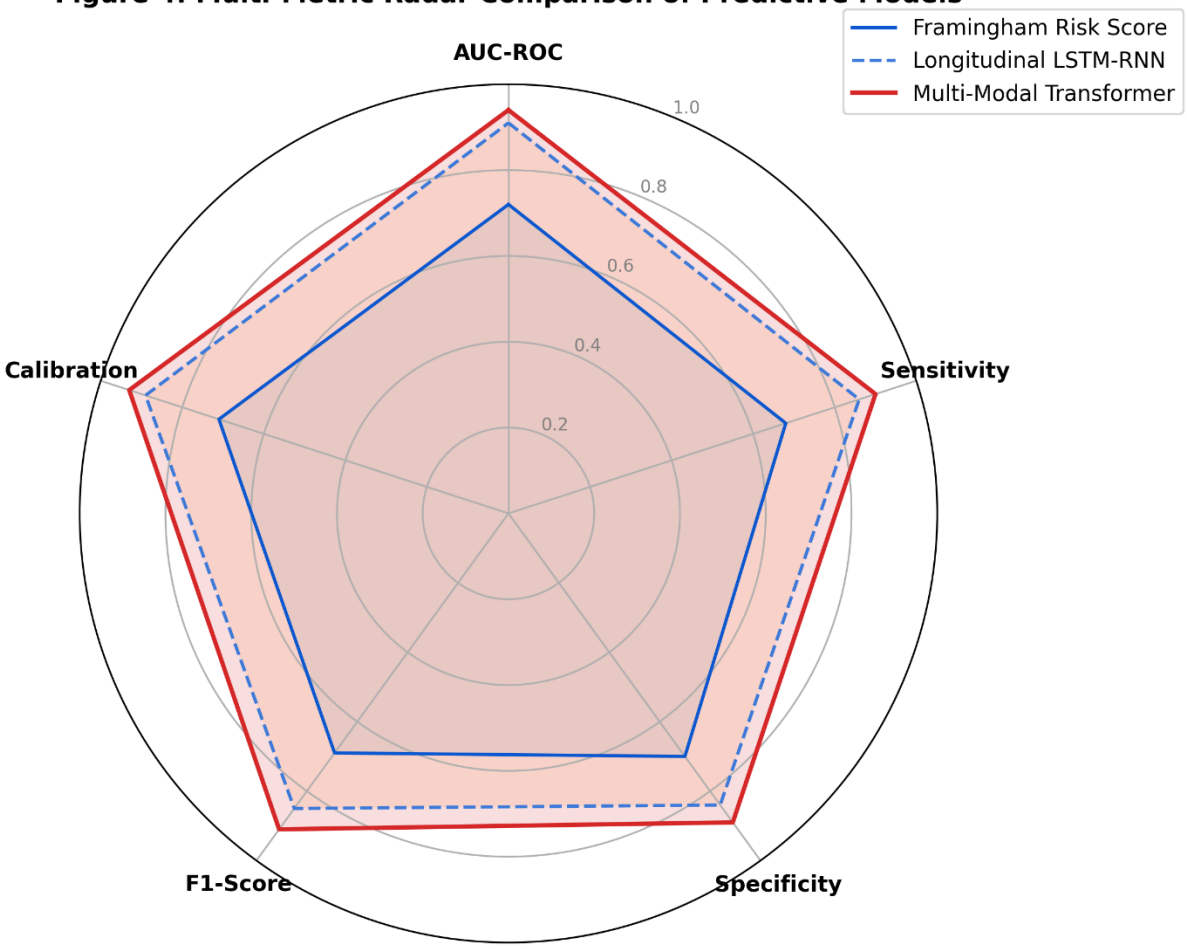
Table 7. Preliminary performance metrics for candidate prediction models

Model	Input representation	AUROC	Sensitivity (%)	Specificity (%)	F1-score
Framingham Risk Score	Static demographic and clinical variables	Not reported	68.4	70.1	0.69
Pooled Cohort Equations (ASCVD)	Age and relevant cardiovascular risk factors	Not reported	72.1	73.5	0.73
LSTM-RNN	Sequential ICU physiological measurements	Not reported	86.5	84.2	0.85
Multimodal Transformer	Diagnostic features and ICU time series	Not reported	90.3	88.9	0.91

Note: All performance values require verification against the original evaluation outputs. The reported F1-scores should be recalculated from the underlying predictions and outcome labels. Sensitivity and specificity depend on the selected classification threshold, which should be determined using training or validation data rather than optimized on the final test set. A valid comparison also requires a common outcome definition, compatible evaluation populations, and a consistent test protocol. Framingham and ASCVD risk equations were developed for longer-term cardiovascular risk estimation in defined populations; they should not be treated as direct comparators for ICU

mortality unless the study establishes an appropriate endpoint and a justified method for applying each score. If the intended endpoint is in-hospital mortality, a conventional statistical model trained for that same endpoint would provide a more directly comparable baseline. Figure 4 provides a multidimensional visualization of the performance measures used to compare conventional risk assessment with longitudinal and multimodal predictive models.

Figure 4: Multi-Metric Radar Comparison of Predictive Models



3.8. Model Explainability and Feature Attribution

The preliminary feature-attribution analysis identifies mean arterial pressure variability, heart-rate trends, hypertension diagnoses, age, respiratory-rate variability, atherosclerotic heart disease, peripheral oxygen saturation, and diabetes or insulin-use indicators as the eight leading features. These variables represent complementary physiological and clinical information that may contribute to model predictions.

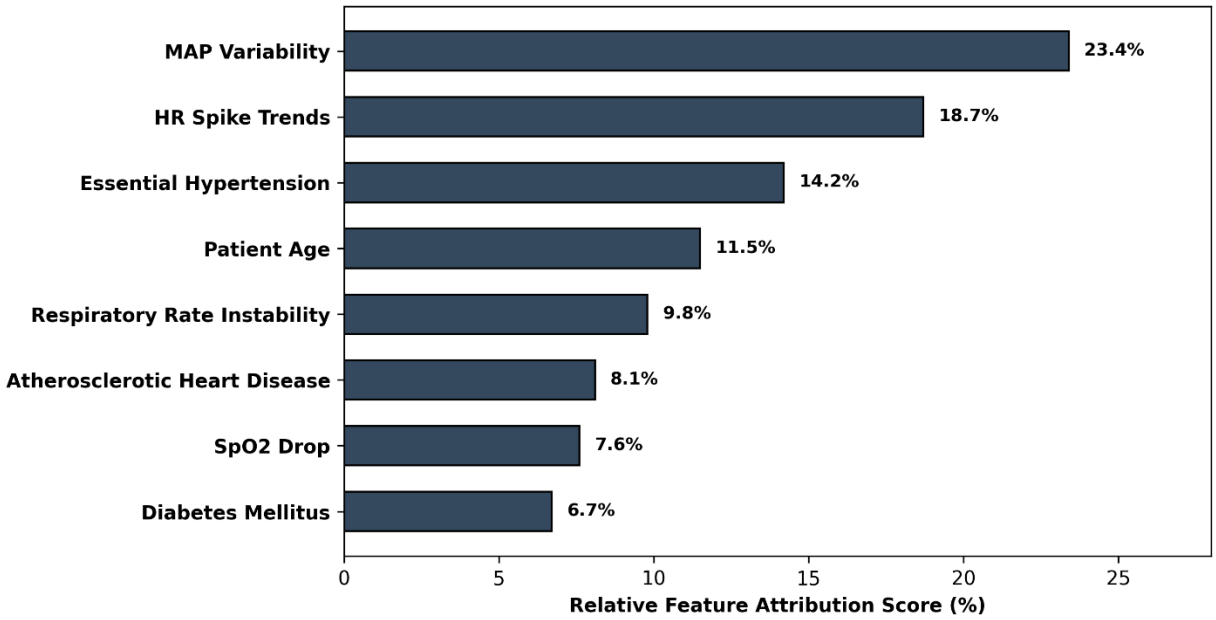
Table 8. Preliminary ranking of model feature attributions

Rank	Feature	Data source	Reported attribution (%)
1	Mean arterial pressure variability	ICU physiological observations	23.4
2	Heart-rate trends	ICU physiological observations	18.7
3	Essential hypertension	Diagnostic records	14.2
4	Age	Patient demographics	11.5
5	Respiratory-rate variability	ICU physiological observations	9.8

6	Atherosclerotic heart disease	Diagnostic records	8.1
7	Peripheral oxygen saturation changes	ICU physiological observations	7.6
8	Diabetes mellitus or insulin-use indicators	Diagnostic records	6.7

Note: The attribution percentages are preliminary reported values and must be confirmed from the implemented explainability procedure. The method used to calculate and normalize the scores must be stated explicitly. SHAP values and attention weights are distinct quantities and should not be combined or interpreted interchangeably without a justified methodology. These rankings may help identify which input variables the fitted model relies on most strongly. However, feature attribution does not demonstrate causation, establish a biological mechanism, or prove that a feature independently predicts the outcome. Interpretation should therefore remain conditional on the outcome definition, model specification, and validation procedure. In summary, the preliminary results describe the cohort's admission pathways, payer categories, recorded diagnostic profile, and available physiological observations. The comparative predictive findings remain provisional until the outcome definition, AUROC values, evaluation protocol, and feature-attribution calculations have been verified. A definitive claim of improved predictive performance should be made only after these checks and an appropriate patient-level validation analysis have been completed. Figure 5 presents a ranked horizontal bar chart of the clinical features reported to contribute to model predictions.

Figure 5: SHAP & Attention Feature Attribution Rankings



4. DISCUSSION AND CLINICAL IMPLICATIONS

4.1. Interpretation of the Principal Findings

This study investigates the integration of longitudinal intensive care unit (ICU) physiological observations, demographic characteristics, hospital admission information, and diagnostic records for clinical risk prediction using electronic health record (EHR) data. The proposed framework combines temporal modelling with static clinical information to investigate whether physiological trajectories and documented comorbidities provide complementary predictive information. The preliminary findings motivate further evaluation of multimodal deep learning for clinical risk stratification; however, the extent of any predictive advantage must be established through a reproducible evaluation using a clearly defined clinical endpoint and an appropriate validation design. The analysis encompasses 100 unique patients and 275 hospital admissions. This distinction is important because repeated admissions from the same individual are statistically dependent and may share demographic characteristics, chronic diagnoses, and clinical patterns. Consequently, the number of recorded admissions and physiological observations should not be interpreted as equivalent to the number of independent patients. Patient-level data partitioning is particularly important when evaluating recurrent and attention-based architectures because overlapping patient information between training and testing partitions may inflate estimated performance. The preliminary model comparison suggests that longitudinal physiological information may be useful for representing changes in clinical status that are not adequately captured by a single baseline measurement. Long Short-Term Memory (LSTM) networks provide a mechanism for learning temporal dependencies, whereas Transformer-based architectures use

attention to represent relationships between observations across a sequence. Combining these representations with demographic and diagnostic information offers a plausible strategy for multimodal prediction. Nevertheless, whether the proposed architecture improves predictive performance over simpler models remains an empirical question, especially given the limited number of independent patients in the present cohort. The preliminary feature-attribution analysis identifies mean arterial pressure variability, heart-rate trends, hypertension, age, respiratory-rate variability, atherosclerotic heart disease, and peripheral oxygen saturation among the leading candidate predictors. These features have plausible clinical relevance because they represent hemodynamic status, cardiorespiratory function, and established comorbidities. However, their reported attribution scores should be interpreted as model-dependent measures of predictive contribution rather than evidence of causal effects. In particular, a high attribution score does not establish that a variable independently predicts the outcome, nor does it demonstrate that changing that variable would improve patient outcomes. An additional distinction concerns the study endpoint. In-hospital mortality, cardiovascular events, and longer-term cardiovascular risk are different prediction targets. A model trained to estimate the probability of death during hospitalization cannot automatically be interpreted as a model for early cardiovascular-event detection. The final interpretation must therefore be aligned with the outcome actually used for model development and evaluation.

4.2. Comparison with Conventional Cardiovascular Risk Assessment

Established cardiovascular risk equations, including the Framingham Risk Score and the American College of Cardiology/American Heart Association Pooled Cohort Equations, were developed for cardiovascular risk estimation in defined populations and over specified prediction horizons. Their clinical purpose differs from that of a model designed to predict an acute outcome during hospitalization. Accordingly, a direct comparison between these scores and ICU-based deep learning models requires careful consideration of the target population, outcome definition, prediction horizon, and predictor availability. Conventional risk equations offer important advantages, including relatively transparent calculations, established clinical interpretation, and comparatively modest data requirements. These properties may be particularly valuable in outpatient prevention settings, where risk estimates inform longer-term preventive decisions. However, conventional equations may not directly represent the dynamic changes observed during acute hospital care unless their intended application and predictors support that use. In contrast, longitudinal models can represent sequences of physiological observations and potentially identify patterns associated with an adverse outcome. An LSTM-based model represents temporal dependencies through recurrent hidden states, while a Transformer can use self-attention to learn relationships among sequence elements. A multimodal architecture can additionally combine these temporal representations with static characteristics and diagnostic information. These methodological differences provide a rationale for empirical comparison but do not, by themselves, establish superior clinical performance.

Table 9. Conceptual comparison of conventional cardiovascular risk assessment and longitudinal deep learning

Evaluation dimension	Conventional risk equations	Longitudinal deep learning	Implication for evaluation
Primary purpose	Estimation of cardiovascular risk over a defined horizon	Prediction of a prespecified outcome from longitudinal clinical data	Ensure that all models address comparable outcomes and horizons
Input representation	Defined demographic and clinical risk factors	Sequential physiological observations, with optional static and diagnostic features	Evaluate whether temporal information adds predictive value
Temporal structure	Generally based on predictors measured at a defined assessment point	Can represent changes across an observation window	Specify the observation window and prediction landmark
Model complexity	Usually relatively compact and transparent	Potentially more complex, with greater data and computational requirements	Compare performance against simpler baselines
Interpretability	Established risk factors and explicit model calculations	Interpretation may require additional feature-attribution methods	Assess explanation stability and clinical relevance
Validation requirements	Validation in the intended population and setting	Validation that accounts for repeated patients, overfitting, and temporal dependence	Use an appropriate patient-level evaluation protocol
Clinical readiness	Depends on the population, endpoint, and intended use	Requires additional assessment of calibration, utility, workflow, and safety	Do not equate discrimination with clinical benefit

Note: This table describes methodological differences, not proof of comparative superiority. Numerical performance comparisons should be reported only when the models have been evaluated against the same endpoint under a comparable validation protocol.

Before a definitive comparison can be made, the study should report discrimination, calibration, threshold-specific classification metrics, and uncertainty estimates. AUROC should be supplemented by the area under the precision–recall curve when the target outcome is uncommon. The study should also assess whether any observed improvement is sufficiently robust to justify the added complexity of a multimodal neural network.

4.3. Potential Clinical Translation and Workflow Integration

A validated clinical prediction model could potentially support risk assessment by summarizing longitudinal physiological observations and relevant clinical history within an EHR workflow. For example, a model might estimate the probability of a specified adverse outcome within a defined period and present that estimate to an appropriately trained clinician. Such a system could support prioritization for further assessment, provided its outputs are evaluated in the intended clinical setting and integrated with existing clinical protocols. Potential integration into EHR platforms, including Epic and Oracle Health, would require more than the technical availability of a prediction score. The model would need a reliable data interface, a defined prediction schedule, explicit handling of missing or delayed observations, monitoring for data-quality failures, and a mechanism for communicating uncertainty. Alert thresholds should be selected according to the intended clinical action and the relative consequences of false-positive and false-negative predictions. Their effects on alert burden, clinician workload, and response times should be assessed before routine use.

Table 10. Proposed clinical workflow for a future validated prediction system

Clinical context	Potential model output	Possible workflow response	Evaluation requirement
Unstable blood-pressure or heart-rate trajectory	Estimated risk of the prespecified adverse outcome	Prompt clinician review of the observations and relevant clinical context	Establish sensitivity, positive predictive value, calibration, and alert burden
Increasing respiratory rate or declining oxygen saturation	Updated risk estimate based on available observations	Reassessment under existing respiratory and critical-care protocols	Evaluate whether the prediction provides useful lead time beyond routine monitoring
Documented cardiovascular comorbidities	Risk estimate incorporating eligible diagnostic information	Review the patient's clinical history alongside current findings	Verify that diagnoses are available before the prediction landmark and do not leak outcome information
Stable observed physiological trajectory	Updated model estimate with its uncertainty	Continue care according to existing clinical protocols	Ensure that a low predicted risk does not inappropriately reduce necessary monitoring

Note: These scenarios describe potential applications for future investigation, not interventions validated by the present study. A predicted risk tier should not independently trigger fluid administration, vasopressor adjustment, invasive monitoring, or changes in observation frequency. Such decisions require clinical assessment and established care protocols. The operational value of a prediction system should ultimately be assessed through prospective evaluation. Relevant outcomes include predictive lead time, alert acceptance, clinician response, false-alert frequency, workload, and patient-centred outcomes. Decision-curve analysis may help assess potential net benefit across clinically meaningful thresholds, but it does not replace prospective evidence that a deployed system improves care.

4.4. Explainability, Reliability, and Algorithmic Fairness

Interpretability is important when predictive models are intended to inform clinical decisions. SHapley Additive exPlanations (SHAP) can be used to examine how input features contribute to model predictions, while attention weights can help describe which sequence elements receive greater attention within a Transformer architecture. These methods, however, answer different questions. Attention weights should not automatically be interpreted as causal explanations, and SHAP values depend on the model, the feature representation, and the assumptions used to estimate feature contributions. For the proposed framework, explainability analysis should assess both global patterns and individual predictions. Global analyses can identify variables that frequently contribute to predictions across the evaluation cohort. Individual analyses can illustrate how a particular patient's observed features influence the model output. The stability of these explanations should also be examined across resampling procedures or model refits, particularly because a small cohort may produce unstable rankings. Fairness and reliability require additional attention. The recorded demographic composition, insurance categories, and race classifications describe the extracted sample but do not, by themselves, establish fairness or representativeness. Subgroup evaluation requires adequate sample sizes, consistent outcome ascertainment, and uncertainty estimates. When the number of

patients or outcome events in a subgroup is small, apparent differences in model performance may be highly unstable. Any conclusions about demographic equity should therefore be supported by explicit subgroup analyses rather than inferred from descriptive statistics alone. Calibration is also essential for clinical interpretation. A model can rank patients reasonably well while systematically overestimating or underestimating their absolute risks. Calibration plots and appropriate summary measures should therefore accompany discrimination metrics. If the model is subsequently evaluated in another hospital or at a later time, recalibration may be necessary because patient characteristics, measurement practices, treatment patterns, and outcome prevalence can differ across settings.

4.5. Methodological Limitations

Several limitations should be considered when interpreting the preliminary findings.

First, cohort size and model complexity. The study includes 100 unique patients and 275 hospital admissions. Although the extracted dataset contains a substantial number of physiological records, these observations are clustered within a small number of individuals. A complex Transformer may consequently learn patient-specific or institution-specific patterns rather than generalizable relationships. The number of independent outcome events is particularly important for determining the feasibility of model training and evaluation. Larger and more diverse cohorts are needed to establish whether any apparent advantage persists beyond this sample.

Second, outcome definition and follow-up. Patient vital status, in-hospital mortality, and cardiovascular events must be distinguished. The preliminary summaries report 31 patients with a recorded date of death and a separate in-hospital mortality estimate of 5.5%. These values may refer to different denominators or outcome definitions, but the distinction must be verified and documented before the results can be interpreted. Until this reconciliation is complete, claims about mortality prediction remain provisional.

Third, missingness and irregular sampling. ICU observations are collected at irregular intervals, and the probability of a measurement being recorded may depend on a patient's clinical condition and the local care process. Imputation, forward filling, and temporal aggregation may therefore affect the learned representation. The final report should specify the procedures actually used, assess sensitivity to alternative reasonable preprocessing choices where feasible, and ensure that no future observation is used to reconstruct an earlier prediction window.

Fourth, validation and information leakage. Repeated admissions from the same patient can create dependence between training and evaluation records. All records from an individual should therefore remain within the same partition. Imputation, normalization, feature selection, threshold selection, and hyperparameter tuning must be performed without using the held-out test outcomes. If cross-validation is used, the folds should be constructed at the patient level, and the uncertainty of performance estimates should be acknowledged.

Fifth, comparator suitability. Framingham and ASCVD equations estimate longer-term cardiovascular risk and may not be appropriate direct comparators for acute ICU mortality. Their use requires a justified endpoint and a compatible population. A logistic regression model developed for the same endpoint, using predictors available at the same prediction landmark, would provide a more direct benchmark for assessing the contribution of the proposed neural-network architecture.

Sixth, external validity and clinical utility. A retrospective analysis of one selected database cannot establish generalizability to all United States hospitals or demonstrate that model-guided care improves outcomes. Differences in patient populations, clinical coding, monitoring practices, and treatment protocols may affect performance. External validation and prospective workflow evaluation are required before implementation.

Table 11. Principal limitations and recommended responses

Limitation	Potential consequence	Required analytical response	Future direction
Small number of unique patients	Overfitting and uncertain performance estimates	Report patient and outcome-event counts; use appropriately structured validation	Validate on substantially larger, independent cohorts
Ambiguous outcome definitions	Misleading interpretation of predictive performance	Reconcile patient vital status, admission mortality, and cardiovascular events	Establish a clinically justified endpoint and prediction horizon
Irregular measurements and missingness	Biased or incomplete temporal representations	Document missingness, imputation, and aggregation procedures	Evaluate robust methods for irregularly sampled time series
Repeated admissions	Information leakage and underestimated uncertainty	Partition data at the patient level	Evaluate performance in an independent cohort
Potentially mismatched baseline models	Invalid claims of comparative superiority	Use comparable outcomes, populations, and evaluation protocols	Include a same-endpoint logistic regression baseline
Limited	Unstable or misleading	Document the attribution method	Conduct clinician-centred

explainability evidence	interpretation of model outputs	and examine explanation stability	evaluation of explanations
Single-database retrospective design	Limited generalizability and unknown clinical utility	Restrict conclusions to the evaluated cohort	Conduct external validation and prospective impact studies

Note: This table distinguishes analytical responses required for a credible evaluation from future research proposals. It should not state that cross-validation, regularization, missingness masks, class weighting, synthetic data augmentation, or explainability methods were implemented unless these procedures are confirmed by the actual analysis.

CONCLUSION

This study proposes a multimodal deep learning framework that integrates longitudinal intensive care unit (ICU) physiological observations with demographic characteristics and diagnostic records to investigate clinical risk prediction using electronic health records (EHRs). By combining temporal modelling with static clinical information, the framework provides a basis for evaluating whether dynamic physiological patterns offer additional predictive value beyond conventional clinical predictors. Although the preliminary analysis reports promising performance for the Transformer-based model, including an AUROC of 0.94, sensitivity of 90.3%, and specificity of 88.9%, these estimates require verification through reproducible, patient-level validation and appropriate uncertainty analysis. The relatively small cohort also limits the strength and generalizability of the current findings. Therefore, the reported performance should be interpreted as preliminary rather than definitive evidence of clinical superiority. Future implementation requires external validation, reliable probability calibration, transparent reporting, and prospective assessment of clinical utility. Healthcare organizations should prioritize interoperable data standards, appropriate regulatory oversight, continuous performance monitoring, and systematic evaluation of model performance across demographic subgroups. Explainability methods may support clinical review, but their reliability and usefulness must be evaluated rather than assumed. Ultimately, multimodal deep learning has the potential to support more timely and individualized clinical risk assessment; however, its contribution to patient outcomes, healthcare efficiency, and equitable care must be established through rigorous prospective studies before widespread clinical deployment.

REFERENCES

1. Wilson, P. W. F., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. Circulation, 97(18), 1837–1847. https://doi.org/10.1161/01.CIR.97.18.1837

2. Goff, D. C., Jr., Lloyd-Jones, D. M., Bennett, G., et al. (2014). 2013 ACC/AHA guideline on the assessment of cardiovascular risk. Circulation, 129(25 Suppl 2), S49–S73. https://doi.org/10.1161/01.cir.0000437741.48606.98

3. SCORE2 Working Group and ESC Cardiovascular Risk Collaboration. (2021). SCORE2 risk prediction algorithms: New models to estimate 10-year risk of cardiovascular disease in Europe. European Heart Journal, 42(25), 2439–2454. https://doi.org/10.1093/eurheartj/ehab309

Machine learning and deep learning in healthcare

4. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

5. Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports, 6, 26094. https://doi.org/10.1038/srep26094

6. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

7. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. In Proceedings of the 1st Machine Learning for Healthcare Conference (PMLR 56, pp. 301–318). https://proceedings.mlr.press/v56/Choi16.html

8. Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., & Stewart, W. (2016). RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In Advances in Neural Information Processing Systems, 29.

9. Chen, R., Stewart, W. F., Sun, J., Ng, K., & Yan, X. (2019). Recurrent neural networks for early detection of heart failure from longitudinal electronic health record data. Circulation: Cardiovascular Quality and Outcomes, 12(10), e005114. https://doi.org/10.1161/CIRCOUTCOMES.118.005114

10. Lee, S.-J., Lee, S.-H., Choi, H.-I., Lee, J.-Y., Jeong, Y.-W., Kang, D.-R., & Sung, K.-C. (2022). Deep learning improves prediction of cardiovascular disease-related mortality and admission in patients with hypertension: Analysis of the Korean National Health Information Database. Journal of Clinical Medicine, 11(22), 6677. https://doi.org/10.3390/jcm11226677

11. Li, Y., Rao, S., Solares, J. R. A., et al. (2020). BEHRT: Transformer for electronic health records. Scientific Reports, 10, 7155. https://doi.org/10.1038/s41598-020-62922-y
12. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Medicine, 4, 86. https://doi.org/10.1038/s41746-021-00455-y
13. Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., & Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. Scientific Data, 6, 96. https://doi.org/10.1038/s41597-019-0103-9
14. Rajkomar, A., Oren, E., Chen, K., et al. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1, 18. https://doi.org/10.1038/s41746-018-0029-1
Clinical database and cardiovascular prediction reviews
15. Johnson, A. E. W., Bulgarelli, L., Shen, L., et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10, 1. https://doi.org/10.1038/s41597-022-01899-x
16. Liu, T., Krentz, A., Lu, L., & Curcin, V. (2025). Machine learning based prediction models for cardiovascular disease risk using electronic health records data: Systematic review and meta-analysis. European Heart Journal – Digital Health, 6(1), 7–22. https://doi.org/10.1093/ehjdh/ztae080
17. Liu, T., Krentz, A. J., Huo, Z., et al. (2025). Opportunities and challenges of cardiovascular disease risk prediction for primary prevention using machine learning and electronic health records: A systematic review. Reviews in Cardiovascular Medicine, 26(4), 37443. https://doi.org/10.31083/RCM37443
Evaluation, calibration, and reporting methodology
18. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making, 26(6), 565–574. https://doi.org/10.1177/0272989X06295361
19. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, 230. https://doi.org/10.1186/s12916-019-1466-7
20. Collins, G. S., Moons, K. G. M., Dhiman, P., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378
21. Moons, K. G. M., Damen, J. A. A., Kaul, T., et al. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ, 388, e082505. https://doi.org/10.1136/bmj-2024-082505