

Article

A Comparative Study of Structured and Narrative EHR Data for 30-Day Readmission Risk Assessment

Zhengxiao He ¹, Huayu Li ¹, Siyuan Tian ², Jinghao Wen ³, Geng Yuan ⁴ and Ao Li ^{1,5,*}

¹ Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ 85721, USA; zhengxiaohe@arizona.edu (Z.H.); hl459@arizona.edu (H.L.)

² Microsoft Research Asia, Beijing 100080, China; fringstian@gmail.com

³ Department of Electrical and Computer Engineering, Villanova University, Villanova, PA 19085, USA; jwen01@villanova.edu

⁴ School of Computing, University of Georgia, Athens, GA 30602, USA; geng.yuan@uga.edu

⁵ BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA

* Correspondence: aoli1@arizona.edu

Abstract

Hospital readmissions within 30 days of discharge are a key metric of healthcare quality and a major driver of cost. Accurate risk stratification enables targeted home care interventions, such as remote monitoring, timely nurse follow-up, and medication adherence programs, designed to mitigate preventable readmissions. Thus, we conducted a systematic comparison of structured electronic health record (EHR) data and unstructured discharge summaries for predicting 30-day unplanned readmissions. Using the MIMIC-IV database, we integrated admissions, emergency department records, laboratory values, and discharge notes, and we restricted the cohort to ED-based readmissions. Following rigorous preprocessing, we balanced the dataset and split it into training, validation, and test sets at the patient level. Structured features, including the LACE score (length of stay, acuity of admission, comorbidities, and ED utilization) and triage vitals, were modeled with logistic regression, random forest, XGBoost, and LightGBM. Discharge notes were analyzed with ClinicalLongformer to generate contextual embeddings and with Qwen2.5-7B-Instruct for few-shot classification. ClinicalLongformer achieved the best discrimination (AUROC = 0.72, F1 = 0.68), outperforming classical ML baselines (AUROC $\approx$ 0.65–0.67, F1 $\approx$ 0.64–0.65). The LLM (Qwen2.5-7B-Instruct) yielded moderate discrimination (AUROC = 0.66, F1 = 0.65) while providing interpretable chain-of-thought rationales. These findings highlight the value of narrative data for risk stratification and suggest that transformer-based language models have the potential to enable scalable, explainable early-warning systems to guide home care and prevent avoidable readmissions.

Keywords: patient readmission; large language model; explainable artificial intelligence



Academic Editor: Enzo Pasquale Scilingo

Received: 19 September 2025

Revised: 5 October 2025

Accepted: 11 October 2025

Published: 14 October 2025

Citation: He, Z.; Li, H.; Tian, S.; Wen, J.; Yuan, G.; Li, A. A Comparative Study of Structured and Narrative EHR Data for 30-Day Readmission Risk Assessment. *Electronics* **2025**, *14*, 4033. <https://doi.org/10.3390/electronics14204033>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Hospital readmissions within 30 days after discharge are widely recognized as a critical indicator of healthcare quality. High readmission rates not only disrupt patient recovery, but they also impose significant financial strain on healthcare systems and add stress to caregivers. In the U.S., approximately 20% of patients are readmitted within 30 days after discharge [1]. In the United Kingdom, emergency readmissions accounted for 15.5% of all admissions in 2020/21, costing the National Health Service an estimated GBP 1.6 billion annually [2]. Similar patterns have been observed in Canada, where hospital

readmissions represent a persistent burden despite system-wide initiatives to improve care transitions [3]. An accurate readmission risk alert is crucial for targeted home care and the corresponding monitoring to prevent further complications. Identifying high-risk patients in a timely manner enables care teams to allocate resources efficiently, initiate remote follow-up, and implement evidence-based interventions during the vulnerable post-discharge period. This integration of predictive analytics has the potential to reduce avoidable readmissions, improve patient outcomes, and reduce healthcare costs.

To address this challenge, policies such as the Hospital Readmissions Reduction Program (HRRP), established under the Affordable Care Act in the U.S., were designed to penalize hospitals with higher-than-expected readmission rates, thus incentivizing improvements in care coordination and post-discharge follow-up [4]. Despite these efforts, predicting and preventing readmissions remains a multifactorial problem shaped by clinical complexity at the patient level, social determinants of health (SDOH), and systemic limitations in healthcare delivery [5]. Therefore, there is a need to develop predictive models for early alerts of readmission risk to prevent further economic and mental damage, as shown in Figure 1.



Figure 1. Workflow of early alerts For readmission prevention.

A wide range of predictive models have been developed to identify patients at risk of readmission. Traditional tools, such as the LACE index [6], MEWS [7], and the HOSPITAL score [8], utilize structured electronic health record (EHR) features, including demographics, comorbidities, and admission characteristics, to generate interpretable risk estimates. Although simple and widely used, their granularity is limited, as they condense complex patient data into a few coarse features. This simplification can omit nuanced clinical patterns in real practice. Moreover, heterogeneity between institutions can degrade their precision; a model derived in one hospital population may not generalize well to another due to differences in patient demographics, care practices, or coding standards [9]. Subsequent efforts have applied machine learning (ML) techniques to structured EHR data, demonstrating moderate improvements in predictive performance. Instead of relying on a fixed score, these approaches can utilize dozens of variables from structured EHR data. Common techniques include logistic regression, decision trees, ensemble methods (e.g., random forests and gradient boosting), and support vector machines [10]. Studies frequently report that tree-based models and regularized logistic regression are among the best-performing methods for this task. For example, Caruana et al. [11] developed intelligible models for 30-day hospital readmission by combining logistic regression with decision tree ensembles, using a wide range of clinical characteristics. Such ML models tend to outperform simplistic scores like LACE in internal validation, albeit often by a moderate margin. A recent scoping review found that most ML-based readmission models achieved Area Under the ROC Curve (AUROC) values with an average of 0.67 [10], indicating fair predictive power, although not dramatically higher than traditional scores. These efforts illustrate the potential gains of combining readmission risk scores and clinical characteristics of the EHR.

However, even sophisticated ML models face challenges. Generalizability remains difficult due to differences between hospitals in patient populations and EHR data quality. Models trained on data from one system can see a performance drop when applied elsewhere, reflecting the same heterogeneity problem in a new form [9]. Additionally, purely data-driven models still rarely include social and behavioral variables unless they are explic-

itly collected, meaning that important context can be missing. In summary, traditional ML models that use structured EHR data have somewhat improved readmission risk prediction compared to simple scores, but their predictive performance is still only moderate.

One promising alternative is to leverage unstructured clinical narratives, particularly discharge summaries, to predict readmission risks. Unlike structured EHR fields, discharge notes contain rich contextual information written by clinicians at the time of care transitions. These narratives often document the severity of the disease, the response to treatment, social circumstances, and follow-up plans, factors that are difficult to capture in structured fields but highly relevant to the risk of readmission. By analyzing discharge notes, models can access nuanced signals, such as concerns about patient adherence, psychosocial stressors, or detailed explanations of comorbidities, thus complementing structured data and potentially improving predictive performance.

Recent advances in natural language processing (NLP) have opened promising avenues to make use of unstructured records, such as discharge notes. The latest technology is widely used in fields related to healthcare care [12–16]. Early work with ClinicalBERT demonstrated that transformer-based models trained on clinical text could outperform traditional structured data models in predicting readmissions [17]. More recently, hybrid architectures that combine textual embeddings with structured EHR features, such as DeepNote-GNN [18] and ClinicalT5 [19], have achieved state-of-the-art performance on benchmark datasets like MIMIC-III, with AUROC values exceeding 0.85. These advances underscore the importance of leveraging both structured and unstructured data to fully capture patient risk. With the rise of decoder-only large language models (LLMs) such as ChatGPT-4o [20,21], the processing of long clinical narratives has become more efficient and flexible, allowing models to integrate contextual information across entire patient histories. However, applications of these techniques to hospital readmission prediction remain limited.

In this study, we focus on comparing the predictive value of discharge notes using the latest available NLP technologies (e.g., Longformer [22], Large Language Models) versus structured EHR data for 30-day readmission prediction. By examining the two data modalities side by side, our objective is to quantify their relative strengths and limitations in capturing patient risk. A head-to-head comparison provides critical insight into (i) whether narrative data alone can match or exceed the performance of established structured-data models, and (ii) which data modality might be most scalable and impactful for deployment in real-world settings, particularly where one type of data may be sparse or inconsistent. Thus, our work not only advances the methodological understanding in developed healthcare systems but also contributes to ongoing discussions on how narrative and structured data should be prioritized in future health informatics policies and machine learning implementations. In summary, our study makes the following contributions.

- **Systematic comparison of data modalities.** Using the MIMIC-IV database, we performed a head-to-head evaluation of structured EHR characteristics (LACE score, vital signs, and laboratory values) and unstructured discharge summaries for the prediction of unplanned readmission within 30 days. Our results show that transformer-based text models (ClinicalLongformer) achieve higher discrimination (AUROC = 0.72) than classical ML models trained on structured features alone (AUROC $\approx$ 0.65–0.67), quantifying the added value of narrative data.
- **Advancing NLP for clinical prediction.** We applied state-of-the-art transformer architectures capable of processing long clinical narratives, including ClinicalLongformer for contextual embeddings and Qwen2.5-7B-Instruct for few-shot direct classification. This dual approach demonstrates that modern NLP techniques can capture nuanced

clinical signals such as follow-up instructions, disease severity, and psychosocial risk factors that are often missing from structured fields.

- **Unstructured text as an alternative data source.** While structured EHR variables remain the conventional foundation for predictive modeling, our findings show that discharge summaries can serve as a viable alternative when structured data are incomplete, inconsistent, or unavailable. By demonstrating that narrative text alone supports competitive prediction performance, this work highlights the complementary role of unstructured documentation in broadening the data modalities available for clinical risk stratification.
- **Clinical implications.** The latest decoder-only large language models (e.g., Qwen2.5-7B-Instruct) not only provide comparable predictive performance but also generate explicit reasoning paths (chain-of-thought explanations) for each prediction. This interpretability enables clinicians to understand the basis of the model’s decision, increasing trust and facilitating its integration into post-discharge risk management workflows.

Together, these contributions position our work at the intersection of methodological innovation and clinical utility, emphasizing both the scientific advancement and the real-world relevance of large language model-based readmission prediction.

2. Materials and Methods

2.1. Data Source

This study uses the MIMIC-IV deidentified dataset, which contains deidentified electronic health records (EHRs) from Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA, USA [23]. Access was granted through PhysioNet [24] after completing the required CITI training. MIMIC-IV consists of the following two primary modules: hosp (the focus of this study) and icu, capturing hospital and intensive care unit stays. The hosp module predominantly records inpatient admissions, with partial coverage of emergency department visits. In addition, we incorporated clinical notes from the MIMIC-IV-Note extension and MIMIC-IV-ED [23]. For this study, we extracted admissions, patient tables from the hospital module, discharge note tables from MIMIC-IV-Note, ed stays, triage and diagnosis tables from MIMIC-IV-ED. Together, these sources provided comprehensive demographic, admission, diagnostic, and physiological data in all modalities. To facilitate analysis, the extracted tables were merged into a single integrated database that represents unique hospital stays, as illustrated in Figure 2.

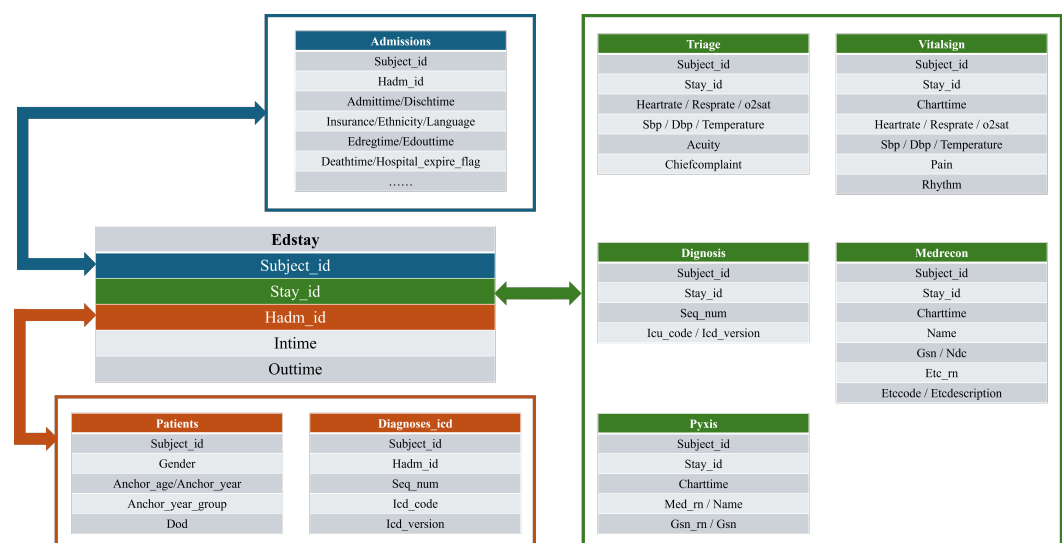


Figure 2. Database linkage.

Before downstream modeling, rigorous preprocessing and data cleaning procedures were performed to ensure data quality and consistency. The data preprocessing pipeline is illustrated in Figure 3. The demographic and clinical information of the patients was first aligned with records from the emergency department, since our analysis specifically targeted readmissions from emergency care. Subsequently, these records were merged with discharge summaries and only encounters containing complete clinical narratives were retained.

To address class imbalance, we adopted the following resampling strategy: (1) the majority (non-readmission) class was first undersampled to reduce extreme imbalance; (2) we then validated the resampled cohort by confirming that the distributions of key numerical features (e.g., age, vital signs, and laboratory results) and categorical variables (e.g., gender) remained consistent with those observed in the original, imbalanced cohort. This ensured that resampling did not introduce spurious biases and that the resampled dataset preserved clinically meaningful characteristics of the population. The corresponding characteristics before and after subsampling are shown in Appendix D.

For missing values, numerical variables were imputed with the median, while categorical variables were imputed with their most frequent category. In addition, we performed rigorous data quality control by removing physiologically implausible values (e.g., systolic blood pressure outside 60–250 mmHg, diastolic blood pressure outside 30–150 mmHg, oxygen saturation outside 0–100%, respiratory rate outside 0–200 breaths/min, pain score outside 0–18, and temperature outside 95–104°F).

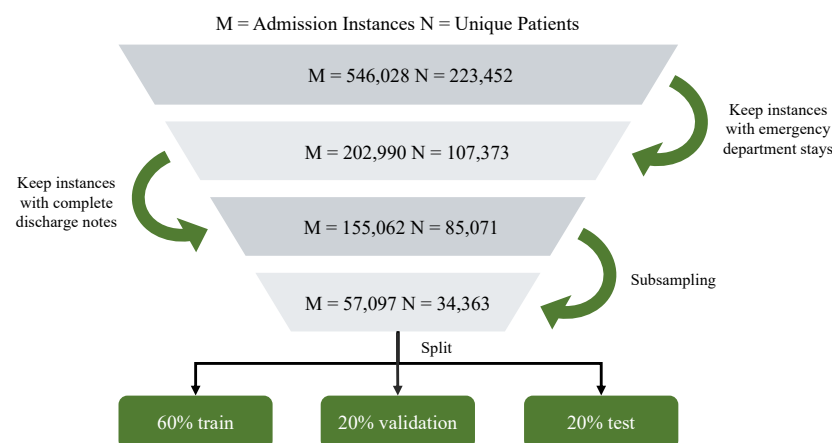


Figure 3. Flowmap of data selection.

2.2. Label Assignment

For label assignment, we defined the outcome variable as 30-day hospital readmission, consistent with the previous literature on healthcare quality assessment. Each patient encounter was assigned a binary label, where a value of 1 indicated that the patient was readmitted within 30 days of discharge, and 0 indicated the patient was not readmitted or the readmitted was outside 30 days of discharge. The assignment was determined based on the temporal relationship between consecutive admissions recorded in the hospital database, with the discharge and subsequent admission timestamps serving as reference points. Importantly, in line with the focus of the study, we restricted the definition of the outcome to readmissions through the emergency department, thus excluding planned or elective hospitalizations and ensuring that the labels reflected unplanned utilization related to acute care.

2.3. Dataset Splitting

To enable robust model training and evaluation, the data set was randomly divided into training, validation, and testing subsets following a 6:2:2 ratio at the patient level. Specifically, 60% of the patient encounters were assigned to the training set for model development, 20% were assigned to the validation set for hyperparameter tuning and model selection, and the remaining 20% were reserved for the held-out test set to assess generalization performance. The splitting was performed in a stratified manner to preserve the proportion of positive and negative readmission cases across all subsets, thus mitigating potential sampling bias. Importantly, patient identifiers were used to ensure that the records of the same individual did not appear in multiple subsets, thus preventing information leakage between the training and evaluation phases.

2.4. Methods

As illustrated in Figure 4, we developed a comparative framework for 30-day hospital readmission prediction by integrating structured EHR with unstructured discharge summaries. The structured EHR module encompassed diagnosis codes, previous ED encounters, length of stay, admission department, and demographic characteristics (for example, age). From these variables, we derived validated indices such as the LACE score, in addition to individual demographic predictors. The resulting feature set was modeled using established machine learning algorithms, including logistic regression, to provide a robust baseline.

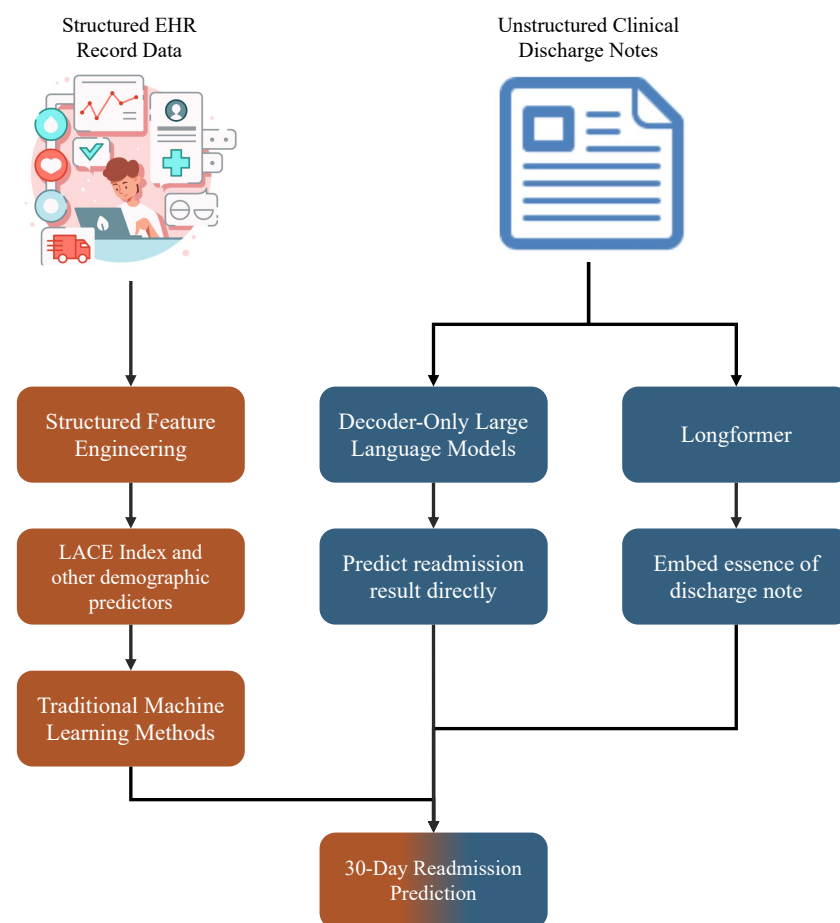


Figure 4. Comparative modeling framework using structured and narrative EHR data.

In parallel, we used the semantic richness of unstructured clinical narratives. Discharge summaries contain detailed accounts of patient conditions, therapeutic interventions,

and follow-up instructions, which extend beyond the information available in structured fields. To harness these signals, we explored two complementary strategies. First, we used domain-adapted transformer architectures designed for long sequences and pre-trained in biomedical corpora (e.g., ClinicalLongformer [25]), generating contextual embeddings that were subsequently integrated into predictive models. Second, we evaluated decoder-only LLMs, which directly inferred readmission outcomes by conditioning on the raw text, bypassing explicit feature engineering. This dual strategy allowed us to systematically compare traditional feature-based modeling with modern representation learning approaches tailored to unstructured narratives.

As for the decoder-only LLMs, we selected Qwen2.5-7B-Instruct [26], a well-established and competitive baseline within the 7B parameter class. This model offers a favorable trade-off between performance and efficiency, demonstrating strong results on reasoning and instruction-following tasks while maintaining computational requirements that are feasible in medical research settings with moderate resources. Its robust instruction tuning and multilingual capabilities further make it suitable for evaluating readmission prediction tasks based on heterogeneous clinical narratives.

For long-sequence modeling, we used the ClinicalLongformer, which was pre-trained on large-scale biomedical corpora and supports extended input windows. This allows the model to capture contextual dependencies in lengthy discharge summaries and other clinical documents, which are crucial for accurate readmission prediction.

In addition, we included several traditional machine learning models as baselines, namely, logistic regression, random forest, XGBoost, and LightGBM, to benchmark the performance of modern transformer-based architectures against the established approaches widely used in clinical predictive modeling.

3. Results

3.1. Feature Analysis

We first examine the distribution of demographic, clinical, and encounter-level variables across the study cohort, as shown in Table 1. Table 2 summarizes the baseline characteristics in the training, validation, and test cohorts at the level of the subject index, stratified by readmission status at 30 days. Categorical variables (for example, sex) were compared using Pearson's χ^2 test, while continuous variables (for example, age, vital signs, and LACE score) were evaluated with Shapiro–Wilk and Levene's tests to determine the appropriate use of t tests or Mann–Whitney U tests. All tests were two-sided, with $p < 0.05$ considered significant.

In all splits, the mean patient age was approximately 62 years, with no significant differences between the readmitted and non-readmitted groups. The proportion of males was modestly higher among readmitted patients (approximately 51–53%) compared to non-readmitted patients (approximately 49%), a difference reaching statistical significance in the train and test cohorts.

The distributions of vital signs in the ED triage were broadly comparable between the groups, but several consistent and clinically significant trends emerged. Readmitted patients had significantly elevated heart rates (≈ 2 to 3 bpm higher) and significantly lower systolic and diastolic blood pressures in all splits, suggesting early hemodynamic instability at presentation. This pattern, tachycardia accompanied by hypotension, is clinically concerning, as it can reflect underlying cardiovascular stress or imminent deterioration, aligning with MEWS [7]. SpO₂ values showed negligible group differences. Pain scores were modestly higher in the readmission group (mean 3.6 vs. 3.5), although this difference was not uniformly significant in splits.

The most pronounced separation was observed in the LACE score, which was consistently elevated in readmitted patients (mean 8.6 to 8.7) compared to non-readmitted patients (mean ≈ 8.1), with highly significant p -values ($<10^{-28}$ across all splits). This aligns with the established role of the LACE index as a validated marker of readmission risk.

Laboratory markers provided additional discriminative signals. Creatinine levels were slightly but significantly lower among readmitted patients, which may reflect a complex relationship between renal function and readmission risk. Hemoglobin values were markedly lower in readmitted patients (10.7 vs. 11.2 g/dL, $p < 10^{-12}$), consistent with anemia as a known risk factor for poor outcomes and readmission. WBC counts did not show significant group differences, indicating that acute infection or systemic inflammation may not have been a dominant driver of readmission in this cohort.

Taken together, these results suggest that although most demographic and physiological characteristics were broadly comparable between groups, readmitted patients consistently exhibited tachycardia, reduced blood pressure, elevated LACE scores, and laboratory evidence of anemia. In particular, the modest discriminative power of single-timepoint triage measures underscores the critical need for the longitudinal modeling of vital sign and laboratory trajectories to capture dynamic risk patterns more effectively. These observations provide a clear rationale for prioritizing blood pressure, heart rate, LACE score, creatinine, and hemoglobin as key predictors in subsequent machine learning models. The corresponding odds ratio and the values p estimated by logistic regression are reported in the Appendix C.

Table 1. Summary of features extracted from structured EHR and discharge summaries.

Category	Feature	Description
Demographics	Age at ED	Age at the time of ED admission (0–120 years).
	Gender	Sex of the patient (male/female).
Admission	Admission type	Admission category (e.g., emergency, urgent).
	Length of stay	Derived from admission and discharge timestamps (days).
Triage features	Temperature	Body temperature at ED triage ($^{\circ}\text{C}$).
	Heart rate	Beats per minute (bpm).
	Respiratory rate	Breaths per minute.
	Oxygen saturation	SpO ₂ , percentage.
	Systolic BP	Systolic blood pressure (mmHg).
	Diastolic BP	Diastolic blood pressure (mmHg).
Lab events (close to discharge)	Pain score	Self-reported pain level.
	Creatinine	Serum creatinine level measured close to discharge (mg/dL).
	Hemoglobin	Hemoglobin concentration measured close to discharge (g/dL).
Comorbidity and risk	White blood cell count	WBC count measured close to discharge (k/ μL).
	LACE score	Composite index: Length of stay, Acuity, Comorbidity, ED visits.
Clinical narratives	Discharge summary	Free-text note of hospital course, treatment, and follow-up.

Table 2. Cohort characteristics at subject-index level by 30-day readmission across splits. Continuous variables are reported as the mean $\pm$ SD; Sex is n (%). Index admission is defined as each subject's earliest admission within the split. p compares No vs. Yes within each split.

Characteristic	Train + Validation			Test		
	No (0)	Yes (1)	p	No (0)	Yes (1)	p
N (subjects)	19,488	7864	—	5001	1988	—
Male, n (%)	9525 (48.87)	4026 (51.20)	5.16×10^{-4} ***	2459 (49.17)	1051 (52.87)	6.52×10^{-3} **
Age (years)	62.57 ± 18.87	62.38 ± 18.37	2.72×10^{-1}	62.40 ± 19.06	62.45 ± 18.36	9.06×10^{-1}
Temperature ($^{\circ}$ F) ¹	98.05 ± 3.08	98.13 ± 2.40	8.30×10^{-2}	98.08 ± 2.70	98.15 ± 2.35	1.62×10^{-1}
Heart rate (bpm)	86.43 ± 18.32	87.96 ± 18.84	1.94×10^{-10} ***	85.96 ± 18.38	88.40 ± 18.46	4.92×10^{-7} ***
Respiratory rate (breaths/min)	17.96 ± 2.95	18.00 ± 2.89	1.53×10^{-1}	18.00 ± 2.84	18.00 ± 2.80	7.69×10^{-1}
SpO ₂ (%)	97.74 ± 2.67	97.79 ± 2.59	9.37×10^{-2}	97.74 ± 2.65	97.66 ± 2.48	3.52×10^{-1}
SBP (mmHg) ²	134.34 ± 23.58	132.78 ± 23.25	2.91×10^{-6} ***	134.54 ± 23.46	132.96 ± 23.02	6.15×10^{-3} **
DBP (mmHg) ²	75.42 ± 15.38	74.90 ± 15.27	4.75×10^{-3} **	75.45 ± 15.54	74.49 ± 14.92	5.75×10^{-2}
Pain	3.47 ± 4.05	3.62 ± 4.12	2.17×10^{-2} *	3.52 ± 4.07	3.53 ± 4.11	9.55×10^{-1}
LACE score (points)	8.16 ± 1.96	8.67 ± 1.97	1.09×10^{-87} ***	8.08 ± 1.94	8.61 ± 1.92	1.15×10^{-28} ***
Creatinine (mg/dL)	1.15 ± 1.24	1.12 ± 1.18	1.45×10^{-3} **	1.16 ± 1.21	1.11 ± 1.24	4.38×10^{-3} **
Hemoglobin (g/dL)	11.16 ± 2.08	10.73 ± 2.07	4.99×10^{-55} ***	11.11 ± 2.08	10.73 ± 2.05	6.1×10^{-12} ***
WBC (k/ μ L)	8.26 ± 6.31	8.22 ± 6.29	5.48×10^{-1}	8.33 ± 8.82	8.57 ± 11.35	5.23×10^{-1}

¹ Temperature from ED triage records (Fahrenheit). Convert to Celsius if preferred. ² SBP and DBP occasionally contain extreme values due to data recording artifacts. Notes: * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

3.2. Evaluation Metrics

We evaluated model performance using the following standard classification metrics: recall, precision, accuracy, F_1 score, and the area under the AUROC.

Recall (sensitivity or true positive rate) measures the proportion of actual positive cases correctly identified:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}.$$

Precision quantifies the accuracy of positive predictions, focusing on minimizing false positives, as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}.$$

Accuracy reflects the overall proportion of correctly classified samples, as follows:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}}.$$

F_1 Score is the harmonic mean of precision and recall, balancing both metrics, as follows:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

The AUROC provides a threshold-independent measure of a model's ability to distinguish positive and negative classes, with higher values indicating better discrimination.

For ClinicalLongformer and classical machine learning models, we report the precision, recall, specificity, accuracy, F_1 score, and AUROC to capture both threshold-dependent and threshold-independent performance.

For LLMs, we report the same set of metrics. The AUROC is calculated using token-level probability scores as surrogate confidence estimates, consistent with previous work. This allows a numerical comparison across modeling paradigms.

3.3. Implementation Details

All classical machine learning models were implemented with the scikit-learn 1.7.1. To ensure fairness and rigor, we performed systematic hyperparameter optimization using Optuna (TPE sampler, 80 trials per model) [27]. The search objective was the area under

the ROC curve (AUROC), with the decision threshold jointly optimized under a recall constraint of 0.7. This setting reflects the clinical use case of automated alerts, where missed readmissions are more costly than false alarms. The selected operating point therefore prioritizes sensitivity while maintaining acceptable precision, consistent with previous work on early warning systems in emergency care. The best hyperparameters for each model are reported in detail in Table 3.

Transformer-based models were implemented with the transformers 4.56.1. Training was carried out using AdamW optimization with a linear learning rate schedule, with a learning rate of 1×10^{-4} , batch sizes of 32, and early stopping based on validation F_1 .

Decoder-only LLMs such as Qwen2.5-7B-Instruct were used in few-shot inference modes without gradient updates. For each discharge summary, the raw text was rewritten with a standardized prompt that requested a binary classification of readmission risk. A temperature of 0.7 was applied and the final label was parsed from the model output. The full prompt template used for large language model inference is provided in Appendix A.

All experiments were conducted on a single NVIDIA 3090 GPU with 24 GB of memory, utilizing Python 3.10 and mixed precision to accelerate training.

Table 3. Hyperparameter configurations for baseline machine learning models. All models were implemented in scikit-learn (v1.7.1), except where noted. Preprocessing (median imputation, standardization, and one-hot encoding) was identical across models.

Model	Key Hyperparameters	Notes
Logistic Regression	solver = saga	Linear baseline; interpretable coefficients
	penalty = L2	
	max_iter = 5000	
	n_jobs = -1	
Random Forest	C = 89.14	Bagged ensemble of decision trees
	n_estimators = 300	
	max_depth = 9	
	min_samples_split = 17	
	min_samples_leaf = 9	
	n_jobs = -1	
	random_state = 42	
XGBoost	max_features = sqrt	Gradient boosting
	n_estimators = 500	
	max_depth = 4	
	learning_rate = 0.015	
	subsample = 0.90	
	colsample_bytree = 0.65	
	reg_lambda = 0.03	
	tree_method = hist	
	n_jobs = -1	
	random_state = 42	

Table 3. *Cont.*

Model	Key Hyperparameters		Notes
LightGBM	n_estimators	= 2000	Leaf-wise boosting (LightGBM)
	learning_rate	= 0.02	
	max_depth	= -1	
	bagging_fraction	= 0.75	
	feature_fraction	= 0.65	
	reg_lambda	= 0.05	
	n_jobs	= -1	
	random_state	= 42	

3.4. Experimental Results

3.4.1. Overall Performance

Table 4 summarizes the predictive performance of all models in the holdout test set. For traditional ML models and ClinicalLongformer, we report the AUROC, accuracy, precision, recall, and F_1 -score. For LLMs, we report the same metrics; the AUROC is computed from token-level probability scores as surrogate confidence estimates, following prior work. The bold values indicate the best performance within each column. Figure 5 presents the confusion matrices for all evaluated methods, providing a direct comparison of predictive performance between models. To highlight the added value of domain-specific pre-training, we attach the head-to-head result for ClinicalLongformer and its general version Longformer in Appendix E.

Table 4. Performance comparison of baseline models, transformer-based models, and LLMs on the test set. The best results are shown in bold.

Model	AUROC	Accuracy	Precision	Recall	F1
Logistic Regression	0.65	0.61	0.59	0.70	0.64
Random Forest	0.66	0.62	0.60	0.70	0.64
XGBoost	0.67	0.61	0.59	0.70	0.64
LightGBM	0.67	0.62	0.60	0.71	0.65
ClinicalLongformer	0.72	0.66	0.65	0.70	0.68
Qwen2.5-7B-Instruct (LLM)	0.66	0.66	0.57	0.74	0.65

In general, transformer-based models substantially outperformed classical machine learning baselines trained on structured EHR features. ClinicalLongformer achieved the highest discrimination (AUROC = 0.72) and F1 score (0.68), demonstrating its ability to leverage the semantic richness of long discharge narratives. In contrast, logistic regression, random forest, XGBoost, and LightGBM yielded only moderate AUROC values (≈ 0.65 – 0.67), reflecting the limited granularity of structured variables, such as the LACE score and single-timepoint vital signs.

Interestingly, Qwen2.5-7B-Instruct exhibited competitive recall (0.74), surpassing all other methods and aligning with the clinical priority of minimizing false negatives in early warning systems. However, its precision was lower than that of ClinicalLongformer, suggesting that while LLM is highly sensitive to high-risk cases, it also produces more false alarms. In particular, the performance of decoder-only LLMs is sensitive to prompt design and decoding strategies, which can introduce variability in predictions.

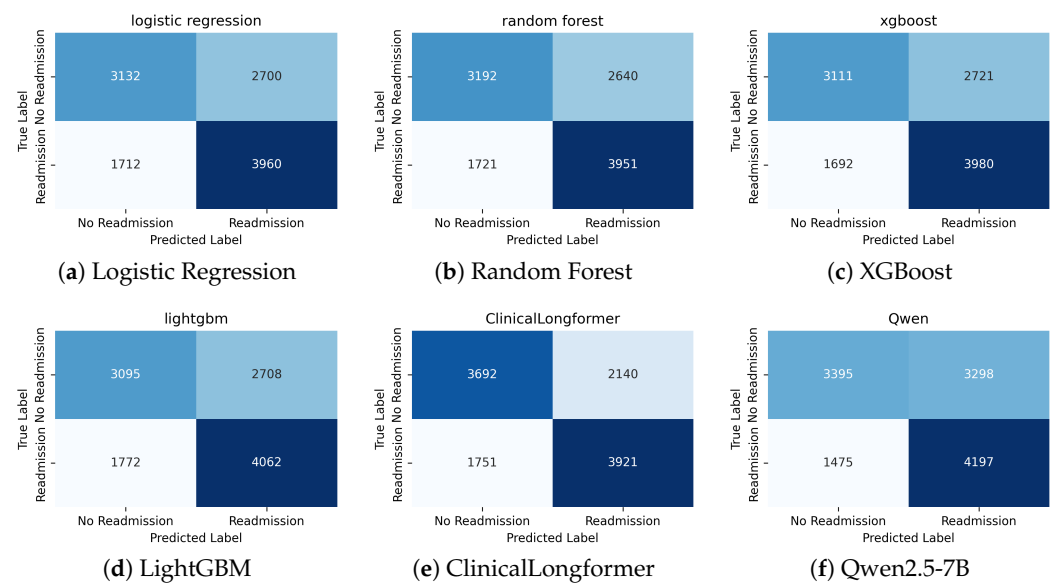


Figure 5. Confusion matrices of different models on the held-out test set.

Together, these results highlight the modalities of the narrative data that hold hidden value. Although structured features provide stable, interpretable baselines, transformer-based models unlock context-specific risk factors embedded in free text, improving predictive power. In addition, the chain-of-thought rationales produced by the LLM offer a novel opportunity for transparent model auditing and decision support, paving the way for explainable readmission risk alerts in practice. As shown in Appendix B, LLM can retrieve relevant and critical information from discharge notes and explicitly present the rationale behind its prediction, which facilitates clinician validation and trust.

3.4.2. Error Analysis Results

We would like to make use of the Brier score to conduct an error analysis [28]. The Brier score quantifies the difference between the estimated risk levels and the empirical risk levels. The calibration results highlighted clear differences among models, as shown in Figure 6. ClinicalLongformer achieved the lowest Brier score (0.2138) and showed near-perfect alignment with the diagonal in all probability bins, indicating a well-calibrated risk estimation. Logistic regression (Brier = 0.2312) also performed consistently, with only mild deviations in the lowest probability range. In contrast, tree-based methods showed systematic miscalibration. Random Forest (Brier = 0.2368) substantially underestimated readmission risk in the low-probability bins (<0.3) and became overconfident at the high end (>0.7), where predicted probabilities were higher than the observed outcomes. XGBoost (Brier = 0.2341) and LightGBM (Brier = 0.2343) followed a similar pattern, producing S-shaped curves that alternated between underestimation and overestimation. Qwen2.5-7B (Brier = 0.2367) demonstrated systematic overestimation of risk in all ranges, with predicted confidence values consistently exceeding the observed event rates, although it is not acceptable for tasks that require high recall.

These differences suggest that domain-specific language models like ClinicalLongformer provide more reliable probability outputs, whereas ensemble trees and large language models (decoder-based) suffer from structural biases in probability estimation. Such miscalibration is particularly problematic in clinical applications where accurate risk stratification is critical. Therefore, further optimization methods should be considered to address these issues.

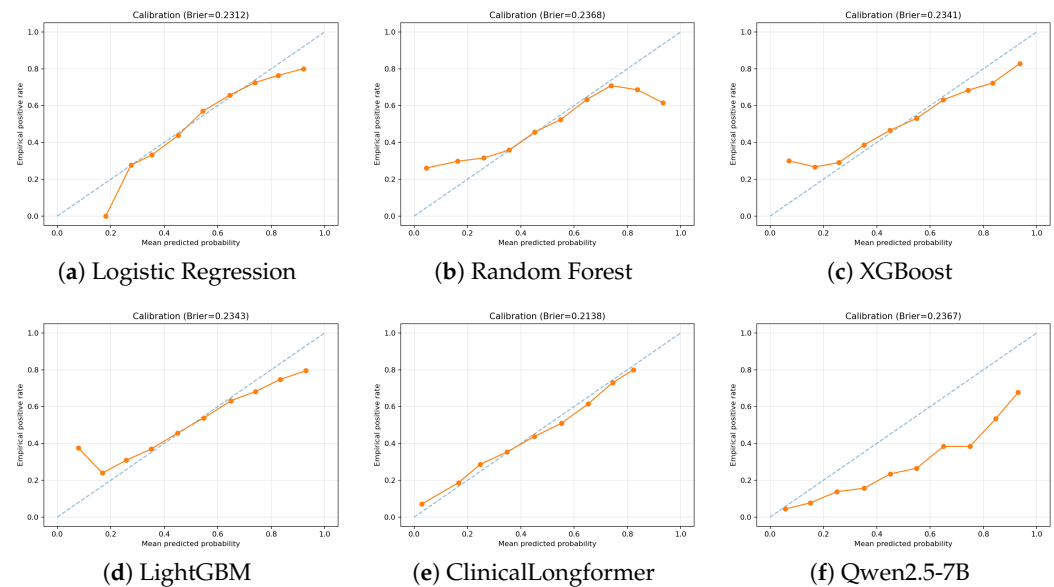


Figure 6. Calibration curves of different models on the held-out test set. The dashed line represents perfect calibration (ideal diagonal), while the orange line shows the model's empirical calibration performance.

4. Discussion

In this study, we systematically compared structured EHR features and unstructured discharge summaries for predicting 30-day unplanned readmissions. Our results show that transformer-based models that use narrative data substantially outperform classical machine learning baselines trained solely on structured features, which supports our claim that unstructured text can work as an alternative data source for structured EHR variables. ClinicalLongformer achieved the highest discrimination (AUROC = 0.72, F1 = 0.68), confirming that long-form discharge narratives contain clinically relevant risk signals beyond those captured by conventional indices such as the LACE score. These findings directly address our main research question and highlight the added value of text-based modeling for early readmission risk stratification. Our results suggest that the use of NLP models integrated with unstructured text for readmission alerts is now a feasible and promising approach.

Our results are consistent with previous studies reporting incremental gains from incorporating free-text notes into predictive models [17,19]. However, by employing long-sequence transformers capable of processing entire discharge summaries, our approach overcomes the truncation limitations of earlier ClinicalBERT-style models and captures richer context such as disease severity, follow-up plans, and psychosocial risk factors. In addition, our focus on unplanned ED-based readmissions provides a clinically actionable target population, in contrast to previous work that often aggregates elective and planned readmissions.

From a clinical perspective, the superior recall achieved by Qwen2.5-7B-Instruct (0.74) is particularly valuable, as minimizing false negatives is critical for early warning systems where missed high-risk patients may lead to preventable adverse outcomes. More importantly, Qwen's chain-of-thought rationales provide interpretable reasoning for each prediction, enabling clinicians to review the basis of the model's decision and increasing trust in automated alerts. This explainability has the potential to support clinician-in-the-loop risk management and to improve patient outcomes by facilitating targeted interventions at discharge. However, similar to previous studies researching the precision of the data, zero-shot large language models for readmission risk [29], in order to achieve a higher

accuracy for the prediction of a specific group of patients, lightweight fine-tuning may be needed, and a reliable method must be developed to calibrate the prediction threshold.

Despite these strengths, several limitations should be acknowledged. First, the analysis was performed on a single data set. MIMIC-IV is one of the largest and most widely used critical care databases worldwide, offering rich, high-quality data, making a wide range of impactful studies possible. However, the generalizability to other institutions remains to be validated. Second, the performance of decoder-only LLMs is sensitive to prompt formulation and decoding strategies, which can introduce variability and limited reproducibility. The lower precision observed relative to ClinicalLongformer suggests that additional calibration is needed to reduce false alarms and enhance clinical utility. Finally, we did not explore the longitudinal modeling of vital signs and laboratory trajectories, which could further improve risk prediction. Future research should also systematically investigate prompt engineering and parameter-efficient adaptation strategies (e.g., LoRA, adapters, and instruction tuning) to improve the precision and stability of LLM-based predictions. External validation on multi-institutional datasets and prospective studies in real-world care transitions are needed to confirm clinical impact and cost-effectiveness. Together, these suggestions could help translate transformer-based predictive models into practical and trustworthy tools to reduce preventable hospital readmissions.

In conclusion, our study demonstrates that long-sequence transformers applied to unstructured discharge summaries deliver superior performance compared to classical machine learning models trained on structured EHR variables, establishing free text notes as a viable alternative data source for readmission prediction. Beyond improved discrimination, these models capture rich contextual information, including disease severity, follow-up instructions, and psychosocial risks, often overlooked by conventional indices. The superior recall achieved by LLMs is particularly valuable in minimizing missed high-risk patients, and the interpretability of their chain-of-thought rationales opens opportunities for clinician-in-the-loop decision support. Timely and targeted home care becomes possible with the integration of NLP models into discharge workflows, enabling the earlier initiation of community nursing visits, medication reconciliation, and remote monitoring programs that can reduce preventable readmissions and improve patient quality of life.

Author Contributions: Conceptualization, Z.H., H.L., and A.L.; methodology, Z.H., H.L., and A.L.; software, Z.H., H.L., and A.L.; validation, Z.H., H.L., A.L., G.Y., S.T., and J.W.; formal analysis, Z.H.; investigation, Z.H. and H.L.; resources, A.L.; data curation, Z.H.; writing original draft preparation, Z.H.; writing—review and editing, H.L., A.L., G.Y., J.W., and S.T.; visualization, Z.H., J.W., and S.T.; supervision, A.L.; project administration, A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The MIMIC-IV and related datasets are available through <https://physionet.org/content/mimic-iv-ed/2.2/> (accessed on 15 September 2025), <https://physionet.org/content/mimic-iv-note/2.2/> (accessed on 15 September 2025), and <https://physionet.org/content/mimiciv/3.1/> (accessed on 15 September 2025), following the submission and approval of the required credentialing process.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Full Prompt Template

The following template was used for LLM inference. The placeholder {DISCHARGE_TEXT} is replaced with the actual discharge summary.

LLM Full Prompt (Few-shot Version)

You are a clinical readmission assessment expert.

You will be provided with the discharge summary of a patient. You need to predict the readmission status of the patient within 30 days after discharge. And you will also be provided with some cases for your reference.

Prediction categories:

- **yes:** The patient will be readmitted within 30 days after discharge. - **no:** The patient will not be readmitted within 30 days after discharge.

Output Format Strictly follow the XML format below and do not add any extra text:

<Explanation>One or two clinically grounded sentences explaining the readmission status and your reasoning.</Explanation>

<Answer>yes/no</Answer>

Examples:

Example 1 Discharge Summary:{DISCHARGE_TEXT}

Answer: Corresponding Answer

Example 2 Discharge Summary:{DISCHARGE_TEXT}

Answer: Corresponding Answer

Test Case: Discharge Summary: {DISCHARGE_TEXT}

Appendix B. Trustworthy Explanation of LLM for Specific Patient

Case Summary: The patient was discharged with a diagnosis of transverse myelitis, which is a rare condition involving spinal cord inflammation. Although the spinal tap showed mild inflammation without active infection, key diagnostic tests (e.g., VZV and MS oligoclonal bands) remain pending. The patient is currently stable, ambulatory, and has no new symptoms or weakness.

LLM-Generated Readmission Assessment:

Explanation: The patient was discharged with a diagnosis of transverse myelitis. The pending diagnostic tests introduce some uncertainty, but the patient is clinically stable and ambulatory with no acute complications.

Prediction: no (not likely to be readmitted within 30 days).

Trustworthiness Considerations: This explanation balances clinical findings (ambulatory status, absence of weakness, and stable discharge condition) against potential risks (pending CSF results and possible recurrence). The LLM output aligns with clinical reasoning by recognizing that follow-up is warranted, but there is no immediate instability.

Appendix C. Odds Ratio of Logistic Regression

We report the estimated coefficients and the corresponding odds ratios (ORs) of the logistic regression model trained on structured EHR features. The positive class here is the readmission case.

Positive coefficients indicate that higher predictor values are associated with a higher likelihood of readmission at 30 days, whereas negative coefficients indicate protective effects. Odds ratios are obtained by exponentiating the coefficients, i.e., $OR = e^{\beta}$.

As shown in Table A1, the LACE score exhibited the strongest effect, with each increase in units associated with an increase of 58% in the odds of 30-day readmission ($OR = 1.58$,

95% CI [1.54–1.62], $p < 0.001$). Lower hemoglobin and WBC were protective (OR = 0.81 and 0.94, respectively; both $p < 0.001$). Male sex was modestly associated with increased odds (OR = 1.06), whereas female sex had a small protective effect (OR = 0.94). Higher heart rate was associated with increased risk (OR = 1.06 per bpm). In contrast, higher systolic blood pressure was inversely associated with readmission (OR = 0.95 per mmHg), consistent with previous observations that a lower systolic BP may reflect hemodynamic compromise. Creatinine was positively associated with risk (OR = 1.04), reflecting the contribution of renal dysfunction to readmission. Diastolic blood pressure did not reach statistical significance ($p = 0.135$).

Table A1. Logistic regression coefficients, odds ratios, and 95% confidence intervals for selected predictors.

Feature	Coefficient (β)	OR	95% CI	p -Value
LACE score	0.46	1.58	[1.54, 1.62]	<0.001
Hemoglobin	−0.21	0.81	[0.79, 0.83]	<0.001
WBC	−0.06	0.94	[0.92, 0.96]	<0.001
Gender (male)	0.06	1.06	[1.04, 1.09]	<0.001
Gender (female)	−0.06	0.94	[0.92, 0.96]	<0.001
HeartRate	0.05	1.06	[1.03, 1.08]	<0.001
Systolic BP	−0.05	0.95	[0.92, 0.97]	0.001
Creatinine	0.04	1.04	[1.01, 1.06]	0.0021
Diastolic BP	−0.02	0.98	[0.95, 1.01]	0.135

Appendix D. Comparison Between Original Dataset and Downsampled Dataset

Table A2. Comparison of cohort characteristics before and after subsampling. Continuous variables are expressed as the mean $\pm$ SD, and categorical variables are expressed as counts (percentage). Balance was evaluated using standardized mean differences (SMD) and Kolmogorov–Smirnov (KS) statistics. Values of SMD < 0.1 and KS < 0.1 indicate negligible imbalance.

Characteristic	Before Subsampling (No)	After Subsampling (No)	SMD	KS Stat
N (subjects)	82,659	24,459	–	–
Male, n (%)	40,503 (49%)	11,985 (49%)	–	–
LACE score	8.12 $\pm$ 1.90	8.12 $\pm$ 1.91	0.008	0.034
Age at ED (years)	61.27 $\pm$ 19.27	62.25 $\pm$ 18.99	0.051	0.021
DBP (mmHg)	75.83 $\pm$ 15.08	75.53 $\pm$ 15.49	0.020	0.102
Respiratory rate (breaths/min)	17.86 $\pm$ 2.74	18.01 $\pm$ 2.71	0.014	0.094
SpO ₂ (%)	97.78 $\pm$ 2.75	97.60 $\pm$ 2.90	0.012	0.093
Heart rate (bpm)	86.16 $\pm$ 18.28	86.55 $\pm$ 18.55	0.005	0.094
Pain score	3.96 $\pm$ 3.79	3.94 $\pm$ 3.83	0.003	0.098
Temperature (°F)	98.08 $\pm$ 2.77	98.09 $\pm$ 2.72	0.003	0.091
SBP (mmHg)	134.27 $\pm$ 22.92	134.22 $\pm$ 23.17	0.002	0.095
Creatinine (mg/dL)	1.15 $\pm$ 1.27	1.15 $\pm$ 1.13	0.000	0.029
Hemoglobin (g/dL)	11.13 $\pm$ 2.08	11.10 $\pm$ 2.10	0.014	0.008
WBC (k/ μ L)	8.31 $\pm$ 9.12	8.30 $\pm$ 7.13	0.001	0.064

Appendix E. Comparison Between Language Models Trained on Medical Context and Its General Version

Table A3. Performance comparison of language models trained on medical context and a general language model. Bold values indicate the best performance.

Model	AUROC	Accuracy	Precision	Recall	F1
Longformer	0.70	0.64	0.68	0.70	0.66
ClinicalLongformer	0.72	0.66	0.65	0.70	0.68

References

1. Wadhera, R.K.; Maddox, K.E.J.; Kazi, D.S.; Shen, C.; Yeh, R.W. Hospital revisits within 30 days after discharge for medical conditions targeted by the Hospital Readmissions Reduction Program in the United States: National retrospective analysis. *BMJ* **2019**, *366*, l4563. [CrossRef]
2. Friebel, R.; Hauck, K.; Aylin, P.; Steventon, A. National trends in emergency readmission rates: A longitudinal analysis of administrative data for England between 2006 and 2016. *BMJ Open* **2018**, *8*, e020325. [CrossRef]
3. Jha, A.K. What readmission rates in Canada tell us about the Hospital Readmissions Reduction Program. *JAMA Cardiol.* **2019**, *4*, 453–454. [CrossRef]
4. Wadhera, R.K.; Yeh, R.W.; Maddox, K.E.J. The hospital readmissions reduction program—Time for a reboot. *N. Engl. J. Med.* **2019**, *380*, 2289. [CrossRef]
5. Lax, Y.; Martinez, M.; Brown, N.M. Social determinants of health and hospital readmission. *Pediatrics* **2017**, *140*, e20171427. [CrossRef]
6. Van Walraven, C.; Dhalla, I.A.; Bell, C.; Etchells, E.; Stiell, I.G.; Zarnke, K.; Austin, P.C.; Forster, A.J. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Can. Med. Assoc. J.* **2010**, *182*, 551–557. [CrossRef]
7. Subbe, C.P.; Kruger, M.; Rutherford, P.; Gemmel, L. Validation of a modified Early Warning Score in medical admissions. *QJM* **2001**, *94*, 521–526. [CrossRef]
8. Donze, J.; Aujesky, D.; Williams, D.; Schnipper, J.L. Potentially avoidable 30-day hospital readmissions in medical patients: Derivation and validation of a prediction model. *JAMA Intern. Med.* **2013**, *173*, 632–638. [CrossRef] [PubMed]
9. Kansagara, D.; Englander, H.; Salanitro, A.; Kagen, D.; Theobald, C.; Freeman, M.; Kripalani, S. Risk prediction models for hospital readmission: A systematic review. *JAMA* **2011**, *306*, 1688–1698. [CrossRef] [PubMed]
10. Huang, Y.; Talwar, A.; Chatterjee, S.; Aparasu, R.R. Application of machine learning in predicting hospital readmissions: A scoping review of the literature. *BMC Med. Res. Methodol.* **2021**, *21*, 96. [CrossRef] [PubMed]
11. Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 1721–1730.
12. He, Z.; Cai, M.; Li, L.; Tian, S.; Dai, R.J. EEG-EMG FAConformer: Frequency Aware Conv-Transformer for the fusion of EEG and EMG. In Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 3–6 December 2024; pp. 3258–3261.
13. Li, H.; He, Z.; Chen, X.; Zhang, C.; Quan, S.F.; Killgore, W.D.S.; Wung, S.F.; Chen, C.X.; Yuan, G.; Lu, J.; Li, A. Smarter Together: Combining Large Language Models and Small Models for Physiological Signals Visual Inspection. *J. Healthc. Inform. Res.* **2025**. [CrossRef]
14. He, Z.; Wen, J.; Li, H.; Tian, S.; Li, A. NeuroHD-RA: Neural-Distilled Hyperdimensional Model with Rhythm Alignment. *arXiv* **2025**, arXiv:2507.14184.
15. Dai, R.; Hu, K.; Yin, H.; Lu, B.; Zheng, W. Self-Supervised EEG Representation Learning Based on Temporal Prediction and Spatial Reconstruction for Emotion Recognition. In Proceedings of the 47th Annual Meeting of the Cognitive Science Society (CogSci 2025); Cognitive Science Society: San Francisco, CA, USA, 2025; in press.
16. He, Z.; Li, H.; Yuan, G.; Killgore, W.D.S.; Quan, S.F.; Chen, C.X.; Li, A. Multimodal Cardiovascular Risk Profiling Using Self-Supervised Learning of Polysomnography. *arXiv* **2025**, arXiv:2507.09009. [CrossRef]
17. Huang, K.; Altsaier, J.; Ranganath, R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv* **2019**, arXiv:1904.05342.
18. Golmaei, S.N.; Luo, X. DeepNote-GNN: Predicting hospital readmission using clinical notes and patient network. In Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Virtual, 1–4 August 2021; pp. 1–9.
19. Pandey, S.R.; Tile, J.D.; Oghaz, M.M.D. Predicting 30-day hospital readmissions using ClinicalT5 with structured and unstructured electronic health records. *PLoS ONE* **2025**, *20*, e0328848. [CrossRef] [PubMed]
20. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 technical report. *arXiv* **2025**, arXiv:2505.09388. [CrossRef]
21. OpenAI. ChatGPT (Sept 6 Version). Large Language Model. 2025. Available online: <https://chat.openai.com/> (accessed on 15 September 2025).
22. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The long-document transformer. *arXiv* **2020**, arXiv:2004.05150. [CrossRef]
23. Johnson, A.E.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T.J.; Hao, S.; Moody, B.; Gow, B.; et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **2023**, *10*, 1. [CrossRef]

24. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220. [[CrossRef](#)]
25. Li, Y.; Wehbe, R.M.; Ahmad, F.S.; Wang, H.; Luo, Y. A comparative study of pretrained language models for long clinical text. *J. Am. Med. Inform. Assoc.* **2023**, *30*, 340–347.
26. Team, Q. Qwen2 technical report. *arXiv* **2024**, arXiv:2407.10671.
27. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631.
28. Rufibach, K. Use of Brier score to assess binary predictions. *J. Clin. Epidemiol.* **2010**, *63*, 938–939. [[CrossRef](#)] [[PubMed](#)]
29. Naliyatthaliyazchayil, P.; Muthyala, R.; Gichoya, J.W.; Purkayastha, S. Evaluating the Reasoning Capabilities of Large Language Models for Medical Coding and Hospital Readmission Risk Stratification: Zero-Shot Prompting Approach. *J. Med. Internet Res.* **2025**, *27*, e74142. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.