Original Paper

Improving the Prognostic Evaluation Precision of Hospital Outcomes for Heart Failure Using Admission Notes and Clinical Tabular Data: Multimodal Deep Learning Model

Zhenyue Gao[1*], MS; Xiaoli Liu[2*], PhD; Yu Kang[3], MD; Pan Hu[2], MD; Xiu Zhang[3], MD; Wei Yan[2], MD; Muyang Yan[2], MD; Pengming Yu[3], MD; Qing Zhang[3], MD; Wendong Xiao[1], PhD; Zhengbo Zhang[2], PhD

[1]Beijing Engineering Research Center of Industrial Spectrum Imaging, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China

[2]Center for Artificial Intelligence in Medicine, The General Hospital of People's Liberation Army, Beijing, China

[3]Department of Cardiology, West China Hospital, Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Zhengbo Zhang, PhD
Center for Artificial Intelligence in Medicine
The General Hospital of People's Liberation Army
28 Fuxing Road
Beijing, 100853
China
Phone: 86 010 68295454
Email: zhangzhengbo@301hospital.com.cn

Abstract

Background: Clinical notes contain contextualized information beyond structured data related to patients’ past and current health status.

Objective: This study aimed to design a multimodal deep learning approach to improve the evaluation precision of hospital outcomes for heart failure (HF) using admission clinical notes and easily collected tabular data.

Methods: Data for the development and validation of the multimodal model were retrospectively derived from 3 open-access US databases, including the Medical Information Mart for Intensive Care III v1.4 (MIMIC-III) and MIMIC-IV v1.0, collected from a teaching hospital from 2001 to 2019, and the eICU Collaborative Research Database v1.2, collected from 208 hospitals from 2014 to 2015. The study cohorts consisted of all patients with critical HF. The clinical notes, including chief complaint, history of present illness, physical examination, medical history, and admission medication, as well as clinical variables recorded in electronic health records, were analyzed. We developed a deep learning mortality prediction model for in-hospital patients, which underwent complete internal, prospective, and external evaluation. The Integrated Gradients and SHapley Additive exPlanations (SHAP) methods were used to analyze the importance of risk factors.

Results: The study included 9989 (16.4%) patients in the development set, 2497 (14.1%) patients in the internal validation set, 1896 (18.3%) in the prospective validation set, and 7432 (15%) patients in the external validation set. The area under the receiver operating characteristic curve of the models was 0.838 (95% CI 0.827-0.851), 0.849 (95% CI 0.841-0.856), and 0.767 (95% CI 0.762-0.772), for the internal, prospective, and external validation sets, respectively. The area under the receiver operating characteristic curve of the multimodal model outperformed that of the unimodal models in all test sets, and tabular data contributed to higher discrimination. The medical history and physical examination were more useful than other factors in early assessments.

Conclusions: The multimodal deep learning model for combining admission notes and clinical tabular data showed promising efficacy as a potentially novel method in evaluating the risk of mortality in patients with HF, providing more accurate and timely decision support.

(J Med Internet Res 2024;26:e54363) doi: 10.2196/54363

KEYWORDS

heart failure; multimodal deep learning; mortality prediction; admission notes; clinical tabular data; tabular; notes; deep learning; machine learning; cardiology; heart; cardiac; documentation; prognostic; prognosis; prognoses; predict; prediction; predictions; predictive

Introduction

Heart failure (HF), a syndrome of impaired heart function, represents the advanced stage of various cardiac conditions [1-3]. With its substantial influence on both morbidity and mortality, HF poses a formidable challenge to human health and societal progress [4-6].

As a potentially life-threatening condition, particularly when accompanied by advanced organ dysfunction or severe complications, a considerable portion of patients with HF may require immediate access to advanced, high-technology, life-saving care, which is typically available only in intensive care units (ICUs) [7]. Studies have indicated that approximately 10% to 51% of patients with HF admitted to hospitals in the United States are subsequently admitted to ICUs [8,9]. It has also been found that ICU-admitted patients with HF experience significantly higher adjusted in-hospital mortality rates compared to those admitted solely to hospitals [10]. The in-hospital mortality rate for patients with HF receiving treatment in an ICU has been reported as 10.6%, in contrast to the overall in-hospital mortality rate of 4% for all patients with HF [11]. Given this substantially higher mortality rate, accurate prediction of in-hospital mortality could empower physicians to implement early interventions and tailor individualized treatments [12,13]. Consequently, there is an increasing need for the development of predictive models that can effectively identify individuals at a heightened risk of mortality in the ICU.

Most previous research works have applied statistical analysis or machine learning techniques using structured administrative data from electronic health records to identify significant risk predictors that trigger adverse outcomes [14-19]. However, HF disease often develops rapidly, and while some sensitive biomarkers, such as N-terminal pro–b-type natriuretic peptide (NT-proBNP), tend to increase in reactivity after the disease progresses, their efficiency is limited due to their high cost and inability to be measured in real time [20,21]. Recently, there has been a growing acknowledgment of the importance of clinical narratives in clinical decision-making [22,23]. The narrative notes at admission, such as chief complaint, history of present illness, physical examination, medical history, and admission medication, play a central role in health care communication. They represent a more comprehensive and personalized account of patient history and assessments [24]. Harnessing the potential of clinical narratives can largely enhance patient care and contribute to the improvement of predictive models for prognosis [25,26]. Exploiting the potential of clinical narratives and modeling them by multimodal deep learning (DL) approaches can enhance the precision of patient care and contribute to the improvement of predictive models for in-hospital mortality.

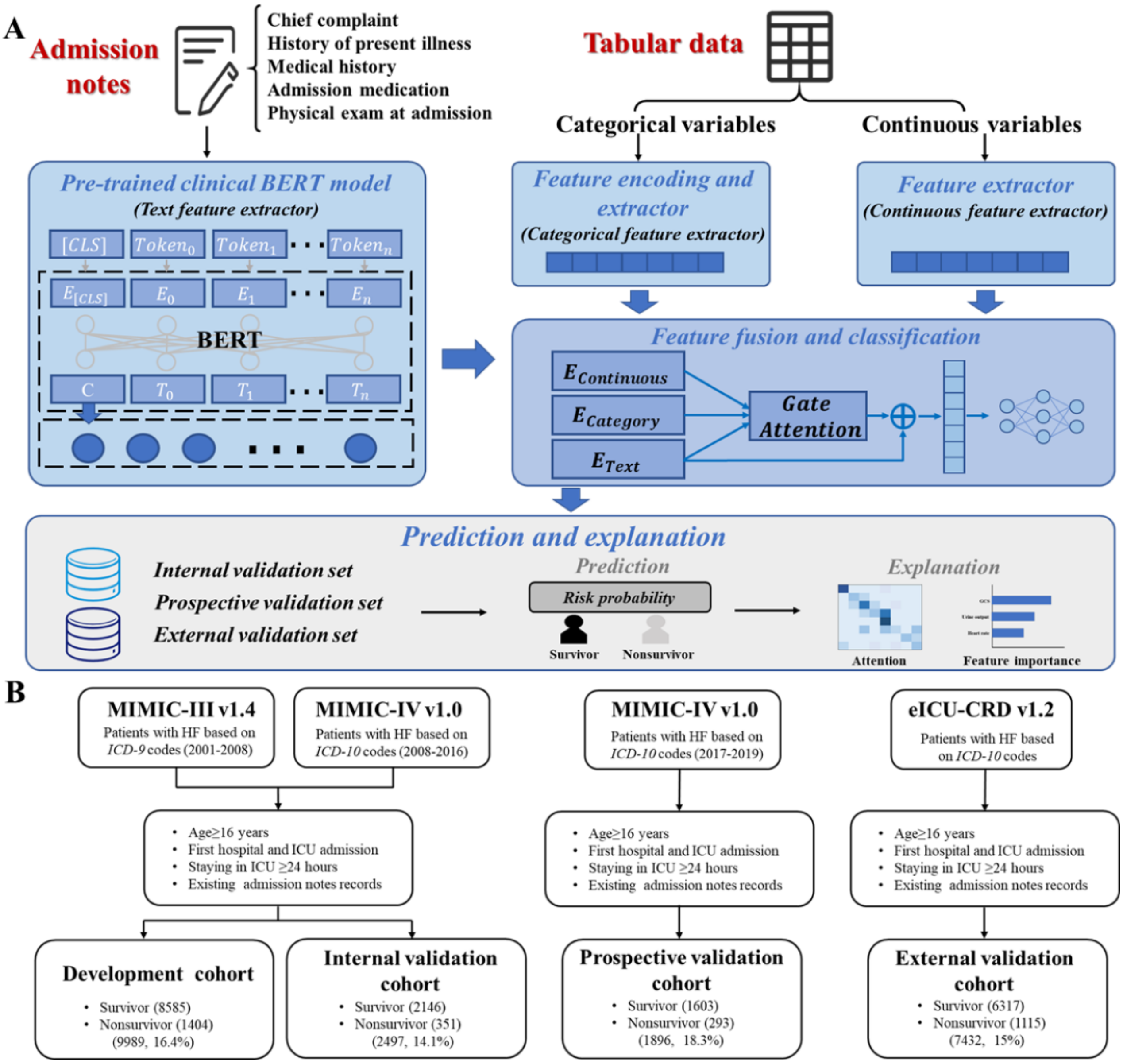
We aim to design a multimodal DL model and explore the infusion approaches to improve evaluation performance using tabular data and admission notes. The cross-modal model, which characterizes textual, categorical, and continuous variables separately, significantly outperforms the unimodal models on the multicenter and prospective validation sets. We believe our findings will motivate data-centric studies to more precisely characterize the illness severity of patients with HF.

Methods

Study Design

An overview of the study flow is shown in Figure 1A. First, we acquired patients' admission notes and tabular data, and these two single modalities were separately embedded to obtain the feature and status representation. The categorical and continuous variables in tabular data were characterized separately. Next, a feature-fusing DL network was applied to integrate the two modalities and achieved the model development. Then, a fully connected DL network was used to predict the in-hospital outcome, our primary outcome of interest. Two postexplanation approaches were adopted to increase the credibility of the model. Finally, the internal, prospective, and external validation with multiple evaluation metrics were accomplished. Our study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guidelines for prognosis studies.

Figure 1. (A) Workflow of the multimodal deep learning model with the fusion of textual and tabular data. (B) An overview of the inclusion criteria with all study cohorts. BERT: Bidirectional Encoder Representations from Transformers; eICU-CRD: eICU Collaborative Research Database; HF: heart failure; *ICD-10: International Classification of Diseases, Tenth Revision*; ICU: intensive care unit; MIMIC: Medical Information Mart for Intensive Care.



Data Sets and Cohorts

The cohorts for this multicenter retrospective cohort study were derived from 3 open-access clinical databases, including the Medical Information Mart for Intensive Care v1.4 (MIMIC-III; CareVue) and MIMIC-IV v1.0, collected from the Beth Israel Deaconess Medical Center in Boston from 2001 to 2008 and 2008 to 2019, respectively [27,28], and the eICU Collaborative Research Database v1.2 (eICU-CRD), collected from 208 hospitals in United States from 2014 to 2015 [29]. We included all first-time ICU admissions for patients with HF aged ≥16 years according to the International Classification of Diseases diagnostic codes. We excluded patients who stayed in the ICU for less than 24 hours and did not have admission notes. Patients were divided into 4 cohorts to support adequate model evaluation, including the development, internal validation, prospective validation, and external validation cohorts. These cohorts correspond to different stages of model development and evaluation. The development cohort consisted of a subset of data used to create or develop the predictive model. Internal validation, prospective validation, and external validation cohorts helped to check if the model had learned patterns that generalized well to new, unseen data. Each of these cohorts played a crucial role in different stages of model development and validation, ensuring that the predictive model was accurate, reliable, and applicable to new and diverse data sets or situations. The inclusion criteria of them are displayed in Figure 1B.

Data Extraction

Our target is to provide early clinical decision support during ICU admissions. The 5 types of commonly recorded notes were extracted, including chief complaint, history of present illness, medical history, admission medications, and physical exam. In

Figure S1 in Multimedia Appendix 1, an example of an admission note is shown with highlights. Meanwhile, 6 types of clinical variables were collected for model development, as follows: (1) basic information of age, gender, weight, BMI, and Charlson Comorbidity Index; (2) vital signs, such as Glasgow Coma Scale, heart rate, respiratory rate, and systolic blood pressure; (3) laboratory tests, including glucose, creatinine, white blood cell, and total bilirubin; (4) urine output; (5) treatments received, including mechanical ventilation; and (6) physical frailty assessments, including activity and fall risks. Representative statistical features were calculated based on the type of variable, such as the maximum, minimum, and mean values. The median value of each feature was used to impute missing values for continuous variables except for FiO_2 ([fraction of inspired oxygen] with the imputation of 21%), with a missing ratio limitation of less than 30%. Details about all types of candidate variables are provided in Table S1 in Multimedia Appendix 1. Their missing ratio is shown in Table S2 in Multimedia Appendix 1.

Model Development and Output

The model was constructed based on a supervised multimodal DL framework, which mainly included feature extractors and a feature fusion module. A pretrained Bidirectional Encoder Representations from Transformers (BERT) module was used for learning the presentation of clinical notes [30]. In the preliminary experiments, we used all the text chunks (the same subset from the training set for the model) to compare the performances of different pretrained clinical BERT models. We found that clinical BERT [31] demonstrated the best comprehensive performance (Table S3 in Multimedia Appendix 1). In the fusion module, a gate attention mechanism [32] was introduced to aggregate the embedded features of clinical notes and tabular data using attention scores; this module finally output the predicted risk probability of in-hospital deaths through a fully connected layer (Figure S2 in Multimedia Appendix 1). The maximum predicted value of all text chunks for a patient was adopted as the optimal risk prediction score. Further detailed information on model building and training is present in the Multimedia Appendix 1.

Model Explanation

In the pursuit of explicating the underlying mechanisms of the DL model and facilitating a comprehensive visualization of pivotal insights, we embarked on an intricate analysis of the pivotal terminologies instrumental in shaping predictions within the developed model. To achieve this, we used the Integrated Gradients (IG) technique [33] to enhance our comprehension of the BERT model’s inner workings and the rationale behind its predictions. This technique hinges on computing gradients with respect to input features, gauging each feature’s contribution to the model’s prediction. IG offers an intuitive understanding of model predictions by quantifying different features’ contributions, aiding clinicians and researchers in comprehending the model’s decisions [34,35]. At the same time, IG demonstrates stability across diverse samples and model architectures, yielding consistent explanatory outcomes, crucial in the face of clinical data diversity and complexity [36]. Consequently, it is considered a reliable analytical tool, helping to assess how each word in the input sequence influences the model's predictions for our research. Simultaneously, we harnessed the SHapley Additive exPlanations (SHAP) technique to unravel the importance of clinical variables in structured tabular data. We computed Shapley values to rank the important clinical variables. Shapley values involve a game theory–based approach to explain the prediction of DL models. They measure the contribution of a given feature value to the difference between the actual prediction and the mean prediction. It is important to note that higher SHAP values signify a heightened pertinence of specific terms in influencing the model’s predictions, whereas relatively diminished SHAP values connote a less pronounced impact. The IG technique exhibits a similar pattern.

Leveraging the IG and SHAP techniques offers valuable insights into the intricate relationship between input features and prediction outcomes, contributing to a more comprehensive understanding of the model’s decision-making process.

Model Validation

The discrimination performance of our prediction model was assessed on the internal (MIMIC, 2001-2016), prospective (MIMIC, 2017-2019) and external (eICU-CRD, 2014-2015) validation cohorts. This assessment compared the model against different single modalities covering notes, tabular data, and a combination of both. The importance of the 5 types of notes for outcome assessment was also examined separately. We trained 5 predictive models based on the tabular data and individual clinical notes. It should be mentioned that the chief complaint was absent in the external validation cohort. Three evaluation metrics were calculated along with their corresponding 95% CIs, the area under the receiver operating characteristic curve (AUROC), F_1-score, and the area under the precision-recall curve.

Statistics Analysis

The median (IQR) values for continuous variables are presented. The *t* test (2-tailed) or the Wilcoxon Rank Sum Test was used when appropriate to compare survivors and nonsurvivors of HF. Categorical variables were reported by total numbers and percentages. Two-sided *P* values of less than .05 were considered statistically significant.

Ethical Considerations

This study was exempt from institutional review board approval due to the retrospective design and lack of direct patient intervention. All data from patients were retrospectively collected from the electronic health care records systems (in the form of third-party public databases or hospital health care systems), which originated from daily clinical work.

All data were de-identified before the analysis. Third-party public databases (MIMIC-IV, MIMIC-III, and eICU-CRD) were used in this study. The institutional review boards of the Massachusetts Institute of Technology (number 0403000206) and Beth Israel Deaconess Medical Center (number 2001-P-001699/14) approved the use of the database for research.

The requirement for individual patient consent was waived because the study did not impact clinical care, all protected health information was deidentified, and all available data in the databases were anonymous.

Results

Patient Characteristics

A total of samples from 12,486 (14.1%) patients with HF were collected from MIMIC-III and MIMIC-IV joint data sets between 2001 and 2016; they were randomly divided into a development set and an independent internal validation set. Additionally, 1896 (18.3%) patients with HF were extracted from MIMIC-IV from 2017 to 2019 for a prospective validation set. For the external validation set, 7432 (15%) patients with HF were extracted from the eICU-CRD data set. Baseline characteristics are summarized in Table 1. The proportion of patients with in-hospital mortality in the 4 cohorts ranges from 14% to 19%. Detailed comparisons of survivors and nonsurvivors in all study cohorts are shown in Tables S4-S7 in Multimedia Appendix 1.

Table 1. The comparison of the total study cohorts for model development and validation.

Characteristics	Development set (n=9989)	Internal validation set (n=2497)	Prospective validation set (n=1896)	External validation set (n=7432)
Basic information				
Age (years), median (IQR)	75 (65-84)	75 (64-83)	74 (64-82)	73 (62-82)
Female sex, n (%)	4637 (46.4)	1151 (46.1)	786 (41.5)	3470 (46.7)
BMI (kg/m^2), median (IQR)	27.7 (24.1-32.6)	27.5 (23.8-32.6)	28.8 (24.4-33.8)	28.9 (24.4-35.0)
CCI[a] score, median (IQR)	7.0 (5.0-8.0)	7.0 (5.0-9.0)	7.0 (6.0-9.0)	5.0 (4.0-6.0)
Physical frailty (fall risk), n (%)	3334 (33.4)	836 (33.5)	1896 (100)	1374 (18.5)
Activity, n (%)				
Bed	7347 (73.8)	1801 (72.6)	1064 (56.2)	3955 (83.5)
Sit	1759 (17.7)	444 (17.9)	441 (23.3)	279 (5.9)
Stand	843 (8.5)	235 (9.5)	388 (20.5)	504 (10.6)
Notes recorded proportion, n (%)				
Chief complaint	8592 (86.0)	2146 (85.9)	1659 (87.5)	0 (0)
History of present illness	9866 (98.8)	2455 (98.3)	1680 (88.6)	7186 (96.7)
Medical history	9730 (97.4)	2427 (97.2)	1675 (88.3)	7188 (96.7)
Admission medication	9018 (90.3)	2257 (90.4)	1650 (87.0)	4387 (59.0)
Physical exam	9186 (92.0)	2303 (92.2)	1644 (86.7)	5104 (68.7)
Outcome				
Days before ICU[b] admission, median (IQR)	0.1 (0.0-1.1)	0.1 (0.0-1.3)	0.1 (0.0-0.9)	0.2 (0.1-0.8)
Days of ICU admission, median (IQR)	3.0 (1.8-5.5)	3.0 (1.8-5.5)	2.9 (1.7-5.3)	2.8 (1.8-4.9)
Days of hospital admission, median (IQR)	8.9 (5.7-14.6)	9.2 (5.8-14.7)	9.8 (6.0-15.7)	7.2 (4.3-11.9)
Death in hospital, n (%)	1404 (16.4)	351 (14.1)	293 (18.3)	1115 (15.0)

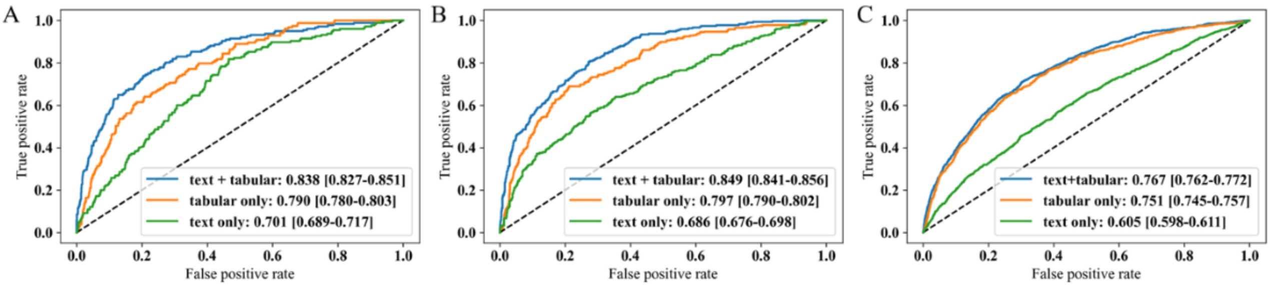
[a] CCI: Charlson Comorbidity Index.

[b] ICU: intensive care unit.

Model Performance Evaluation

We present the discrimination performance on internal, prospective, and external validation sets by receiver operating characteristic curves of the optimal models after tuning the hyperparameters (Figure 2). The AUROCs of the multimodal model were significantly higher than the two unimodal models in all 3 types of validation evaluations. They were 0.838 (95% CI 0.827-0.851), 0.849 (95% CI 0.841-0.856), and 0.767 [0.762-0.772] for the internal, prospective, and external validation sets, respectively. Specifically, the design details of the unimodal models were as follows: the text-based unimodal model used clinical BERT, leveraging its capabilities in contextualizing clinical text data; on the other hand, the tabular unimodal model used a fully connected model structure, tailored to effectively process structured tabular data from the tables. More comparisons on baseline models, such as random forest and logistic regression, and all evaluation metrics for these models in the 3 validation types are presented in Table S8 in Multimedia Appendix 1.

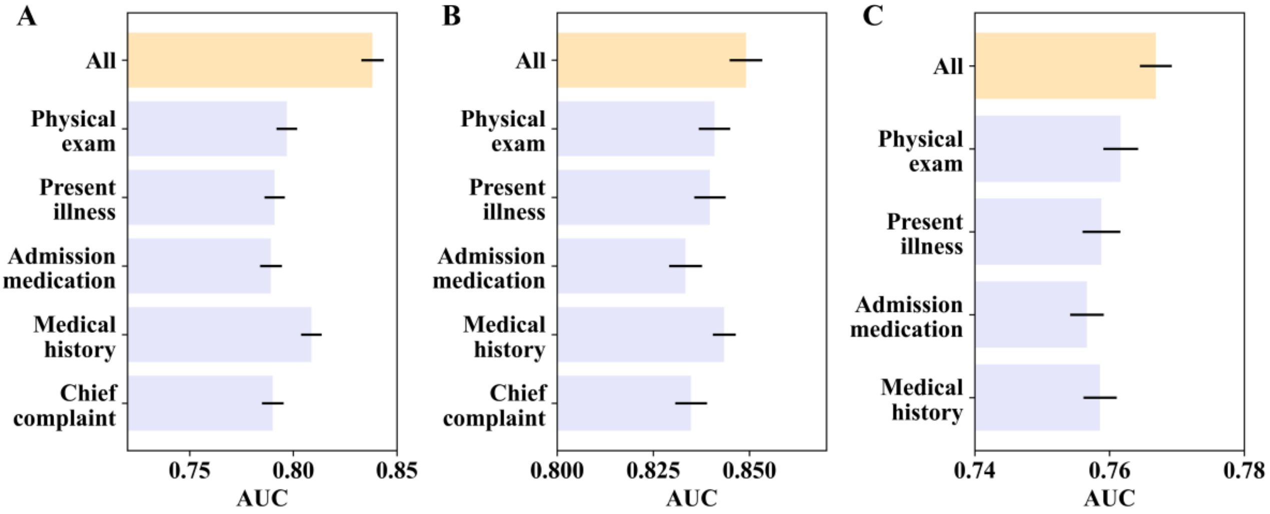
Figure 2. The receiver operating characteristic curve comparison of the optimal multimodal deep learning model against different single modalities in the (A) internal validation set, (B) prospective validation set, and (C) external validation set. Values inside brackets are 95% CIs.



Contribution of Individual Part in Clinical Notes

The performance contributions of the 5 types of clinical notes (including chief complaint, history of present illness, medical history, admission medication, and physical exam) were separately evaluated by combining them with clinical variables to retrain all prediction models. Figure 3 and Table S9 (Multimedia Appendix 1) display the AUROC comparisons with the full model. We found the individual contributions were much lower than the overall contribution in all validation cohorts. Specifically, medical history and physical exam contained more information that was useful in assessing the prognosis of patients with HF compared to other note types.

Figure 3. The reduction of discrimination when separately combining the individual clinical note with tabular data in the (A) internal validation set, (B) prospective validation set, and (C) external validation set. AUC: area under the receiver operating characteristic curve.

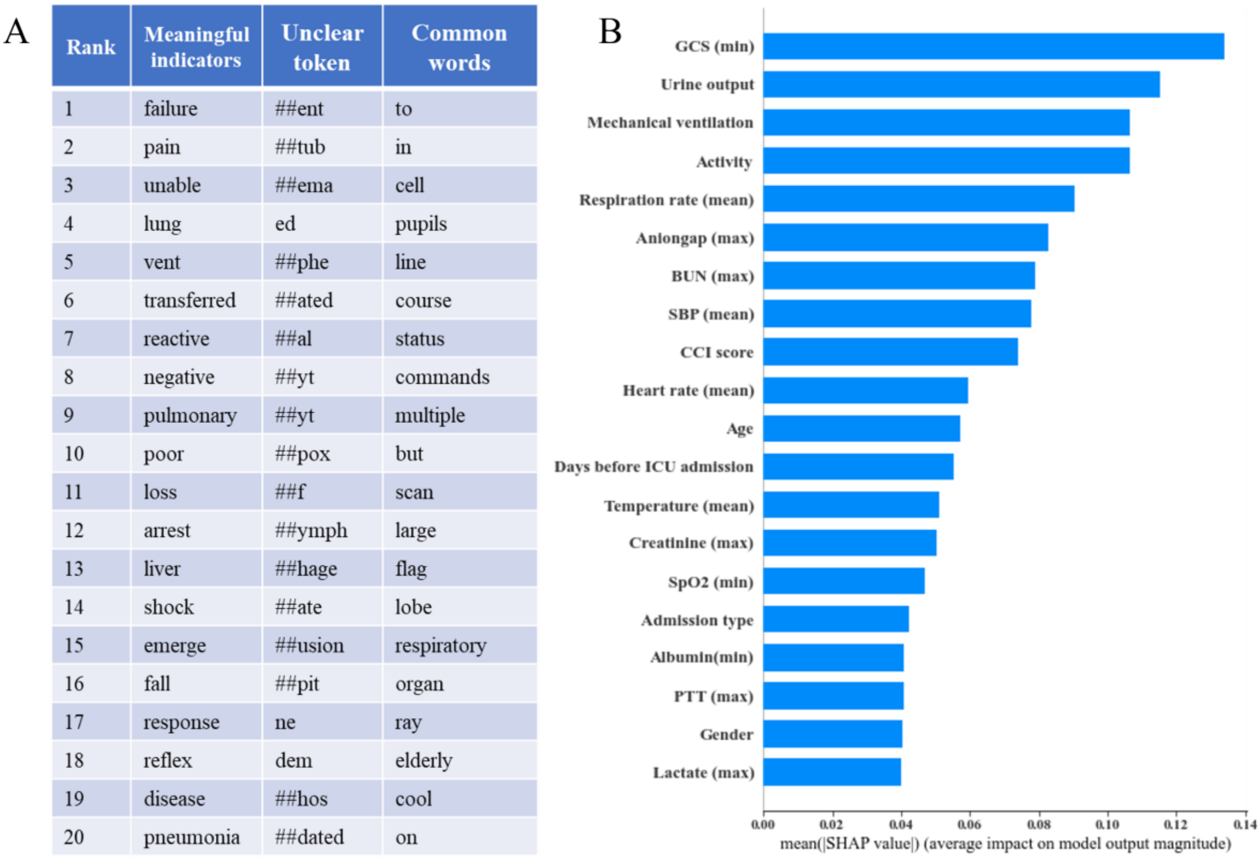


Clinical Notes Visualization and Interpretation

We applied the IG method to study the attribution of the prediction of a deep network to its input features, aiming to provide explanation for individual predictions. IG is computed based on the gradient of the prediction outputs with respect to the input words. Higher IG values denote the greater significance of a word to the model's prediction, whereas lower values indicate lesser importance. We derived IG values for all tokens present in the clinical notes of each patient within the test data set, extracting those tokens with higher IG values. It is important to note that, due to BERT's tokenization process, inputs are represented as tokens rather than individual words. For instance, the phrase "the patient has been extubated" is tokenized into "the patient has been ex ##tub ##ated" as the input sequence [35]. To enhance readability, we conducted postprocessing by excluding numbers, tokens with only 1 or 2 characters, and separators. A clinical expert assessed the clinical significance of tokens and their associated IG values in the context of mortality prediction. The sorted tokens are illustrated in Figure 4A.

The analysis identified commonly ubiquitous clinical terms like "in," "to," and "with," which were segregated due to their limited potential in distinguishing prognostic variations. Among the top 20 clinically meaningful indicators vital for mortality prediction, intriguing insights emerged upon clinical interpretation. For instance, "Failure" and "Pain," the leading predictors, denote prevalent symptoms within ICU care and can mirror disease severity and disability. Indicators 4 and 9 align with pulmonary pathology, their elevated importance reflecting the gravity of respiratory conditions and the necessity for ICU interventions, such as mechanical ventilation. Additional indicators such as "pneumonia" and "fall" manifest acute illness, carrying prognostic weight in mortality prediction. Clinical cues, such as "status," "reflex," and "shock," correspond to mental well-being, with their significance in prognosis attributed to the association of delirium with adverse outcomes.

Figure 4. Explanation of the developed deep learning model. (A) The top 20 tokens of clinical notes evaluated by the Integrated Gradient values in 3 types of language scenarios. (B) The top 20 clinical variables evaluated by the Shapley values. BUN: blood urea nitrogen; CCI: Charlson comorbidity index; GCS: Glasgow Coma Scale; ICU: intensive care unit; PTT: partial thromboplastin time; SBP: systolic blood pressure; SHAP: SHapley Additive exPlanations; SpO_2: oxygen saturation.



Rank	Meaningful indicators	Unclear token	Common words
1	failure	##ent	to
2	pain	##tub	in
3	unable	##ema	cell
4	lung	ed	pupils
5	vent	##phe	line
6	transferred	##ated	course
7	reactive	##al	status
8	negative	##yt	commands
9	pulmonary	##yt	multiple
10	poor	##pox	but
11	loss	##f	scan
12	arrest	##ymph	large
13	liver	##hage	flag
14	shock	##ate	lobe
15	emerge	##usion	respiratory
16	fall	##pit	organ
17	response	ne	ray
18	reflex	dem	elderly
19	disease	##hos	cool
20	pneumonia	##dated	on

Clinical Variables Feature Analysis

We ranked the important clinical variables using the SHAP technique. The top 20 out of 52 clinical variables (Figure 4) show that for structured tabular data, the highest ranked variables also correlate with disease severity and poorer prognosis. These variables represent clinically important information, such as mental status, using the Glasgow Coma Scale, urine output, mechanical ventilation, activity, and respiratory rate measurements.

Discussion

Principal Findings

This retrospective prognostic study aimed to develop, validate, and explain a multimodal DL prediction model for in-hospital outcomes in critically ill patients with HF. The model was constructed based on the admission notes and records from the first ICU admission day. Simultaneously, we compared the difference between multimodal and unimodal models and explored the individual importance of admission notes in the clinical practice of HF. We found that multimodality could further enhance the model's ability and credibility to evaluate outcomes compared to unimodality.

Emerging clinical data sets provide an opportunity for the DL techniques to study the problem of in-hospital mortality prediction. Compared to previous related work, which mostly considers single modality or simply concatenates embeddings from different modalities, our work demonstrates a novel approach. We separately embed texts as well as categorical and continuous variables to integrate multimodal knowledge and leverage clinical notes information for better predictions. Our comprehensive experiments demonstrate that our proposed model outperforms the models using single modality (text-only AUROC: 0.701; tabular-only AUROC: 0.790) by achieving high performance (AUROC: 0.838).

The fusion method we used integrated two different modalities—unstructured clinical notes and structured clinical variables—into a universal shareable space using a transformer block. It was efficient to leverage clinical notes and integrate tabular data. Meanwhile, the novel application of an attention mechanism on clinical data enhanced the model's ability to focus on evaluating the target in models when fusing multimodal information. Our ablation study, as shown in Figures S3-S5 (Multimedia Appendix 1), on domain adaptive pretraining and task adaptive fine-tuning with multiple BERTs verified the significance of pretraining and fine-tuning, when implementing BERT models on natural language text, especially on domain-specific clinical notes.

In the multimodal model, the proportion of clinical variables, especially continuous variables, was much higher compared to other parts (Figure S6 in Multimedia Appendix 1). The analysis and visualization of important words in clinical notes also yielded interesting findings. The ranking of words by IG values provided face validity, indicating that some of the important words used for prediction were clinically related to diseases trajectory, such as the severity of respiratory disease or mental status. Some of the unspecified words, such as "disease," used in diverse scenarios, were more difficult to interpret as isolated words. Lastly, some of the clinically meaningful words can change significantly with negation, such as "fall." In the future, we will use more techniques, such as the NegEx algorithm, to consider negation of keywords to better explain the clinical words' meanings.

There are some limitations in our study. First, our model leveraged the electronic health records data based on patients' ICU admission and the first day of admission to predict in-hospital death risk. It did not include recorded data during the treatment, which might reduce the evaluation performance of the model. Second, we simplify the feature extraction, using the maximum, minimum, or mean statistical values to characterize all data throughout the day. Such simplification ignored the changes in time series, and it might have caused the loss of useful information. Time series data will be considered in our future study. Finally, we recommend that the model need to be calibrated using local data to avoid assessment bias.

Conclusions

In this multicenter prognostic study, we developed and validated an attention-multimodal DL model for in-hospital outcome prediction of patients with HF and explored the approaches that can improve the evaluation precision by simultaneously characterizing both admission notes and tabular data. The AUROCs of our model were significantly higher than those of unimodal models in all validation sets. The clinical variables included in the study made a particularly significant contribution to the overall results, with the data from the clinical notes exhibiting a much lower contribution. The model shows good predictive and explainable performance to potentially support the precise decision-making and disease management of critically ill patients with HF.

Acknowledgments

We gratefully acknowledge the guidance and assistance of Dr Max Shen from Beth Israel Deaconess Medical Center (BIDMC).

This work was supported in part by the National Natural Science Foundations of China (NSFC) under grants 62173032 and 62171471, the Foshan Science and Technology Innovation Special Projects (grant BK22BF005), the Regional Joint Fund of the Guangdong Basic and Applied Basic Research Fund (grant 2022A1515140109).

Qing Zhang (qzhang2000cn@163.com), Wendong Xiao (wdxiao@ustb.edu.cn), and Zhengbo Zhang (zhangzhengbo@301hospital.com.cn) are co-corresponding authors of this manuscript.

Authors' Contributions

ZG and XL had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. They also contributed to the conceptualization and design of the study. YK, PH, XZ, and WY contributed to the acquisition, analysis, or interpretation of data. MY and PY conducted statistical analysis. ZG and XL drafted the manuscript. PY, QZ, ZZ, and WX obtained funding for the study and supervised the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional statistics.
[DOCX File , 466 KB-Multimedia Appendix 1]

References

1. Katz AM, Rolett EL. Heart failure: when form fails to follow function. Eur Heart J. Feb 01, 2016;37(5):449-454. [doi: 10.1093/eurheartj/ehv548] [Medline: 26497163]
2. Ormerod JO, Ashrafian H, Frenneaux MP. Impaired energetics in heart failure - a new therapeutic target. Pharmacol Ther. Sep 2008;119(3):264-274. [doi: 10.1016/j.pharmthera.2008.05.007] [Medline: 18602947]
3. Pagliaro BR, Cannata F, Stefanini GG, Bolognese L. Myocardial ischemia and coronary disease in heart failure. Heart Fail Rev. Jan 22, 2020;25(1):53-65. [doi: 10.1007/s10741-019-09831-z] [Medline: 31332663]
4. Murphy SP, Ibrahim NE, Januzzi JL. Heart failure with reduced ejection fraction: a review. JAMA. Aug 04, 2020;324(5):488-504. [doi: 10.1001/jama.2020.10262] [Medline: 32749493]
5. Savarese G, Becher P, Lund L, Seferovic P, Rosano G, Coats A. Global burden of heart failure: a comprehensive and updated review of epidemiology. Cardiovasc Res. Jan 18, 2023;118(17):3272-3287. [doi: 10.1093/cvr/cvac013] [Medline: 35150240]
6. Ziaeian B, Fonarow GC. Epidemiology and aetiology of heart failure. Nat Rev Cardiol. Dec 2016;13(6):368-378. [FREE Full text] [doi: 10.1038/nrcardio.2016.25] [Medline: 26935038]

7. Li F, Xin H, Zhang J, Fu M, Zhou J, Lian Z. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database. BMJ Open. Jul 23, 2021;11(7):e044779. [FREE Full text] [doi: 10.1136/bmjopen-2020-044779] [Medline: 34301649]
8. Safavi KC, Dharmarajan K, Kim N, Strait KM, Li S, Chen SI, et al. Variation exists in rates of admission to intensive care units for heart failure patients across hospitals in the United States. Circulation. Feb 26, 2013;127(8):923-929. [doi: 10.1161/circulationaha.112.001088]
9. van Diepen S, Bakal JA, Lin M, Kaul P, McAlister FA, Ezekowitz JA. Variation in critical care unit admission rates and outcomes for patients with acute coronary syndromes or heart failure among high - and low - volume cardiac hospitals. JAHA. Mar 10, 2015;4(3):e001708. [doi: 10.1161/jaha.114.001708]
10. Wunsch H, Angus DC, Harrison DA, Collange O, Fowler R, Hoste EAJ, et al. Variation in critical care services across North America and Western Europe*. Critical Care Medicine. 2008;36(10):2787-27e8. [doi: 10.1097/ccm.0b013e318186aec8]
11. Adams KF, Fonarow GC, Emerman CL, LeJemtel TH, Costanzo MR, Abraham WT, et al. ADHERE Scientific Advisory CommitteeInvestigators. Characteristics and outcomes of patients hospitalized for heart failure in the United States: rationale, design, and preliminary observations from the first 100,000 cases in the Acute Decompensated Heart Failure National Registry (ADHERE). Am Heart J. Feb 2005;149(2):209-216. [doi: 10.1016/j.ahj.2004.08.005] [Medline: 15846257]
12. Adler ED, Voors AA, Klein L, Macheret F, Braun OO, Urey MA, et al. Improving risk prediction in heart failure using machine learning. Eur J Heart Fail. Jan 2020;22(1):139-147. [FREE Full text] [doi: 10.1002/ejhf.1628] [Medline: 31721391]
13. Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, et al. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. JACC Heart Fail. Jan 2020;8(1):12-21. [FREE Full text] [doi: 10.1016/j.jchf.2019.06.013] [Medline: 31606361]
14. Soliman A, Agvall B, Etminani K, Hamed O, Lingman M. The price of explainability in machine learning models for 100-day readmission prediction in heart failure: retrospective, comparative, machine learning study. JMIR. Oct 27, 2023;25:e46934. [FREE Full text] [doi: 10.2196/46934] [Medline: 37889530]
15. White-Williams C, Rossi LP, Bittner VA, Driscoll A, Durant RW, Granger BB, et al. American Heart Association Council on CardiovascularStroke Nursing; Council on Clinical Cardiology;Council on EpidemiologyPrevention. Addressing social determinants of health in the care of patients with heart failure: a scientific statement from the American heart association. Circulation. Jun 02, 2020;141(22):e841-e863. [FREE Full text] [doi: 10.1161/CIR.0000000000000767] [Medline: 32349541]
16. Abraham WT, Fonarow GC, Albert NM, Stough WG, Gheorghiade M, Greenberg BH, et al. OPTIMIZE-HF InvestigatorsCoordinators. Predictors of in-hospital mortality in patients hospitalized for heart failure: insights from the Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF). J Am Coll Cardiol. Jul 29, 2008;52(5):347-356. [FREE Full text] [doi: 10.1016/j.jacc.2008.04.028] [Medline: 18652942]
17. Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, et al. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. JACC Heart Fail. Jan 2020;8(1):12-21. [FREE Full text] [doi: 10.1016/j.jchf.2019.06.013] [Medline: 31606361]
18. McGilvray MM, Heaton J, Guo A, Masood MF, Cupps BP, Damiano M, et al. Electronic health record-based deep learning prediction of death or severe decompensation in heart failure patients. JACC Heart Fail. Sep 2022;10(9):637-647. [FREE Full text] [doi: 10.1016/j.jchf.2022.05.010] [Medline: 36049815]
19. Li J, Liu S, Hu Y, Zhu L, Mao Y, Liu J. Predicting mortality in intensive care unit patients with heart failure using an interpretable machine learning model: retrospective cohort study. JMIR. Aug 09, 2022;24(8):e38082. [FREE Full text] [doi: 10.2196/38082] [Medline: 35943767]
20. Kaya SI, Cetinkaya A, Ozcelikay G, Samanci SN, Ozkan SA. Approaches and challenges for biosensors for acute and chronic heart failure. Biosensors (Basel). Feb 16, 2023;13(2):282. [FREE Full text] [doi: 10.3390/bios13020282] [Medline: 36832048]
21. Hudson SR, Chan D, Ng LL. Change in plasma volume and prognosis in acute decompensated heart failure: an observational cohort study. J R Soc Med. Sep 08, 2016;109(9):337-346. [FREE Full text] [doi: 10.1177/0141076816661316] [Medline: 27609799]
22. Goh KH, Wang L, Yeow AYK, Ding YY, Au LSY, Poh HMN, et al. Prediction of readmission in geriatric patients from clinical notes: retrospective text mining study. J Med Internet Res. Oct 19, 2021;23(10):e26486. [FREE Full text] [doi: 10.2196/26486] [Medline: 34665149]
23. Clapp MA, Kim E, James KE, Perlis RH, Kaimal AJ, McCoy TH, et al. Comparison of natural language processing of clinical notes with a validated risk-stratification tool to predict severe maternal morbidity. JAMA Netw Open. Oct 03, 2022;5(10):e2234924. [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.34924] [Medline: 36197662]
24. Özyılmaz E, Özkan Kuşçu Ö, Karakoç E, Boz A, Orhan Tıraşçı G, Güzel R, et al. Worse pre-admission quality of life is a strong predictor of mortality in critically ill patients. Turk J Phys Med Rehabil. Mar 01, 2022;68(1):19-29. [FREE Full text] [doi: 10.5606/tftrd.2022.5287] [Medline: 35949964]
25. Thapa NB, Nischay B, Sattar S, Sona T. Hospital readmission prediction using clinical admission notes. 2022. Presented at: ACSW '22: Proceedings of the 2022 Australasian Computer Science Week; February 14 - 18; Brisbane, Australia. [doi: 10.1145/3511616.3513115]

26. van Aken B, Papaioannou JM, Mayrdorfer M, Budde K, Gers F, Loeser A. Clinical outcome prediction from admission notes using self-supervised knowledge integration. arXiv. Preprint posted online on Feb 8, 2021. [doi: 10.48550/arXiv.2102.04110]
27. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. May 24, 2016;3:160035. [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
28. Johnson A, Stone D, Celi L, Pollard T. The MIMIC Code Repository: enabling reproducibility in critical care research. J Am Med Inform Assoc. Jan 01, 2018;25(1):32-39. [doi: 10.1093/jamia/ocx084] [Medline: 29036464]
29. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. Sep 11, 2018;5(1):180178. [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
30. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on Oct 11, 2018. [doi: 10.48550/arXiv.1810.04805]
31. Huang K, Altossar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv. Preprint posted online on Apr 10, 2019. [doi: 10.48550/arXiv.1904.05342]
32. Rahman W, Hasan MK, Lee S, Bagher Zadeh AA, Mao C, Morency LP, et al. Integrating multimodal information in large pretrained transformers. Proc Conf Assoc Comput Linguist Meet. Jul 2020;2020:2359-2369. [FREE Full text] [doi: 10.18653/v1/2020.acl-main.214] [Medline: 33782629]
33. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. 2017. Presented at: Proceedings of the 34th International Conference on Machine Learning; August 6 - 11;3319-3328; Sydney, Australia. [doi: 10.5555/3305890.3306024]
34. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat Rev Genet. Feb 03, 2023;24(2):125-137. [doi: 10.1038/s41576-022-00532-2] [Medline: 36192604]
35. Lyu W, Dong X, Wong R, Zheng S, Abell-Hart K, Wang F, et al. A multimodal transformer: fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction. AMIA Annu Symp Proc. 2022;2022:719-728. [FREE Full text] [Medline: 37128451]
36. Zhang Y, Tino P, Leonardis A, Tang K. A survey on neural network interpretability. IEEE Trans Emerg Top Comput Intell. Oct 2021;5(5):726-742. [doi: 10.1109/tetci.2021.3100641]

Abbreviations

AUROC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations from Transformers
DL: deep learning
eICU-CRD: eICU Collaborative Research Database
HF: heart failure
ICU: intensive care unit
IG: Integrated Gradients
MIMIC: Medical Information Mart for Intensive Care
NT-proBNP: N-terminal pro–b-type natriuretic peptide
SHAP: SHapley Additive exPlanations

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 07.11.23; peer-reviewed by E Kawamoto, MO Khursheed; comments to author 05.12.23; revised version received 01.01.24; accepted 19.03.24; published 02.05.24

Please cite as:
Gao Z, Liu X, Kang Y, Hu P, Zhang X, Yan W, Yan M, Yu P, Zhang Q, Xiao W, Zhang Z
Improving the Prognostic Evaluation Precision of Hospital Outcomes for Heart Failure Using Admission Notes and Clinical Tabular Data: Multimodal Deep Learning Model
J Med Internet Res 2024;26:e54363
URL: https://www.jmir.org/2024/1/e54363
doi: 10.2196/54363
PMID: