scientific reports

Check for updates

OPEN

Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records

Jun-En Ding[1,12], Phan Nguyen Minh Thao[2,12], Wen-Chih Peng[2], Jian-Zhe Wang[2], Chun-Cheng Chug[2], Min-Chen Hsieh[2], Yun-Chien Tseng[2], Ling Chen[3], Dongsheng Luo[4], Chenwei Wu[8], Chi-Te Wang[11], Chih-Ho Hsu[5], Yi-Tui Chen[7], Pei-Fu Chen[9,10], Feng Liu[1] & Fang-Ming Hung[6,7]✉

Type 2 diabetes mellitus (T2DM) is a prevalent health challenge faced by countries worldwide. In this study, we propose a novel large language multimodal models (LLMMs) framework incorporating multimodal data from clinical notes and laboratory results for diabetes risk prediction. We collected five years of electronic health records (EHRs) dating from 2017 to 2021 from a Taiwan hospital database. This dataset included 1,420,596 clinical notes, 387,392 laboratory results, and more than 1505 laboratory test items. Our method combined a text embedding encoder and multi-head attention layer to learn laboratory values, and utilized a deep neural network (DNN) module to merge blood features with chronic disease semantics into a latent space. In our experiments, we observed that integrating clinical notes with predictions based on textual laboratory values significantly enhanced the predictive capability of the unimodal model in the early detection of T2DM. Moreover, we achieved an area greater than 0.70 under the receiver operating characteristic curve (AUC) for new-onset T2DM prediction, demonstrating the effectiveness of leveraging textual laboratory data for training and inference in LLMs and improving the accuracy of new-onset diabetes prediction.

Type 2 diabetes mellitus (T2DM) and chronic metabolic diseases are health challenges faced by countries worldwide. In recent years, the prevalence of chronic complications associated with T2DM has increased, including obesity, hypertension, hyperlipidemia, and heart disease[1,2]. As the risk of T2DM gradually increases worldwide, the World Health Organization (WHO) has proposed a common T2DM covenant and indicators to prevent long-term complications associated with the disease[3]. According to statistics from the Taiwan Health Promotion Administration, among the three most common chronic diseases in Taiwan, the prevalence of hyperglycemia/T2DM among people aged 65 and over was 27.8% from 2017 to 2020, and the prevalence of hyperlipidemia during this time was 37.9%[4].

In recent years, electronic health records (EHRs) have become the primary tool for recording patients' medical conditions. This information is necessary to make medical decisions and includes the patient's medical history, laboratory results, and imaging reports. As part of standard medical practice, this information is incorporated into a doctor's notes to document and summarize patient care. Some studies on EHRs employ Support Vector Machines (SVM)[5] for feature classification of T2DM or establish sequential deep learning methods to predict

[1]School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, USA. [2]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu City, Taiwan. [3]Institute of Hospital and Health Care Administration, National Yang Ming Chiao Tung University, Taipei City, Taiwan. [4]School of Computing and Information Science, Florida International University, Miami, USA. [5]Department of Surgery, Far Eastern Memorial Hospital, New Taipei City, Taiwan. [6]Surgical Trauma Intensive Care Unit, Far Eastern Memorial Hospital, New Taipei City, Taiwan. [7]Smart Healthcare Interdisciplinary College, National Taipei University of Nursing and Health Sciences, Taipei City, Taiwan. [8]Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA. [9]Department of Anesthesiology, Far Eastern Memorial Hospital, New Taipei City, Taiwan. [10]Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan. [11]Center of Artificial Intelligence, Far Eastern Memorial Hospital, New Taipei City, Taiwan. [12]These authors contributed equally: Jun-En Ding and Phan Nguyen Minh Thao. ✉email: philip@mail.femh.org.tw

T2DM based on the sequence of patients' treatment records[5]. Despite the potential of machine learning for diagnostic prediction, extracting data from EHRs presents a significant challenge. EHRs often contain a mix of data in numeric, categorical, and other formats, making it challenging for fundamental machine learning models to handle medical terminology, complex sentence structures, textual ambiguity/uncertainty, and contextual understanding[6].

In particular, more research has focused on using EHR and natural language processing (NLP) to predict chronic disease[7,8]. With the increasing volume of EHRs, structured and unstructured data types (such as text, CT scans, MRI images, etc.) are more frequently applied in deep learning research[9]. For example, researchers use the NLP model to identify cardiovascular disease (CVD) from EHRs[10] or research in allergy, asthma, and immunology clinics[11]. Additionally, the existing NLP models developed over the past few years, rely solely on text data or ICD code, making it difficult to diagnose T2DM or chronic metabolic diseases[12,13].

In recent years, large language models (LLMs) have been trained successfully using large corpuses and have shown significant effectiveness in natural language processing tasks[14,15]. The most popular research used open datasets, such as the MIMIC series collected by the Medical Information Mart for Intensive Care (MGH)[16]. Such datasets include numerical values, categories, and other formats. However, the MIMIC data are limited by the small sample size and do not adequately represent the diversity of data formats needed for LLMs clinics and training tasks. Numerous medical studies on LLMs face constraints arising from the restricted availability of corpus samples of clinical notes[17], such as the MIMIC or UK Biobank dataset[18], or from inherent imbalances possibly related to specific diseases[19]. These constraints may lead to biases in the predictive capabilities and usability of LLMs in clinical settings. These models undergo training on large amounts of textual data, enabling them to discern intricate statistical relationships embedded within words and phrases. Furthermore, researchers have begun to combine modality data with LLMs[20]. This method addresses the complexities of data extraction and the challenges associated with textual modeling utilizing tabular data. Such NLP applications include text classification[21,22] and even extend into the realm of clinical prediction within the intricate landscape of the medical field[23,24].

In this study, we propose a novel large language multimodal models (LLMMs) framework that integrates clinical notes and textual laboratory values for new-onset T2DM prediction. The main contributions of our work are as follows:

- We have collected five years of EHRs and laboratory results to research the use of LLMs and multimodal data for predicting new-onset T2DM.
- We propose a method for converting laboratory values to text and evaluating its effectiveness in training LLMs. This approach addresses missing patient data and improves LLM contextual learning.
- We propose a method for post hoc explanation and disease risk assessment using LLMs combined with Shapley Additive exPlanations (SHAP)[25] values to visualize textual laboratory values.

The sections of this paper are organized as follows. In "Related work" section, we summarize the limitations of machine learning (ML) techniques and briefly present the existing works applying LLMs in the healthcare domain. "Data collection and study design" section provides an overview of data collection. Our proposed approach is given in "Method" section. To demonstrate the effectiveness of our model, we conducted extensive experiments and early prediction of new-onset T2DM in "Results" section. Finally, in "Interpretable attention in textual laboratory results" section, we conduct a textual interpretable risk assessment of LLMMs in "Interpretable attention in textual laboratory results" section.

Related work

The limitations of machine learning methods

This section examines the shortcomings of classical ML techniques, like SVM and XGBoost[26], when handling large-scale EHRs. These methods are challenged by inherent complexities within EHR data such as missing entries, skewed sample sizes, and the computational burden of processing massive datasets[26]. While XGBoost's tree-based structure alleviates some of these challenges, a significant limitation persists with traditional ML methods; they are incapable of effectively modeling and predicting diseases using a variety of data modalities, including text, images, and tabular data.

Predictive assessments in clinical settings are crucial for estimating a patient's risk of developing diseases, their potential response to treatment, and the likely course of their condition[27]. Traditionally, ML methods such as logistic regression[28] and random forest[29] have been used for these disease prediction tasks. However, a key limitation of these approaches is their inability to effectively model the time-dependent nature of medical events, such as the order in which diagnoses, procedures, and medications occur. Instead, they often focus primarily on whether these events are present or absent as features, without considering the importance of their sequence.

Large language models

Most LLMs require prior knowledge of specific domains and are trained for specific tasks and data[30]. These models undergo extensive training using large datasets and have shown impressive capabilities in various NLP applications including language generation, machine translation, and answering questions[31]. LLMs have the potential to help healthcare professionals identify medical conditions[32]. By examining patient information, including medical history and test results, these models can produce diagnoses and propose additional tests[33–35]. This contributes to reducing diagnostic errors, streamlining diagnostic procedures, and improving the overall standard of healthcare[36].

Moreover, LLMs can revolutionize various aspects of medical practice, including improving diagnostic precision, forecasting disease progression, and aiding clinical decision-making[37,38]. By analyzing extensive medical datasets, LLMs can quickly acquire specialized expertise in various medical fields such as radiology, pathology, and oncology[39,40]. These models can be refined using domain-specific medical literature, maintaining their currency and relevance. In addition, its adaptability to various languages and contexts promotes enhanced global access to medical knowledge and expertise.

Data collection and study design

In this study, we collected five-year EHRs from the Far Eastern Memorial Hospital (FEMH) Taiwan hospital database from 2017 to 2021, including 1,420,596 clinical notes, 387,392 laboratory results, and more than 1505 laboratory test items. The database included clinical notes and laboratory results, as described in Table 2. The study was approved by the FEMH Research Ethics Review Committee (https://www.femh-irb.org/) and data has been de-identified. All ethics review work and data collection were carried out in accordance with the ethics committee's standard guidelines and regulations (https://www.femh-irb.org/index.php/regulations).

In this research, we employed a multi-stage filtering process to focus on clinically relevant information, specifically for patients with new-onset T2DM. Our data collection and preprocessing workflow was as described in Fig. 2A,C, and followed these steps: First, we filtered the patient's visit history to include only outpatient visits. Next, we identified the outpatient visit with the smallest time difference between the first onset testing records, as shown in Fig. 2B. This visit likely represents the closest encounter to the initial detection of T2DM. Finally, we included the records closest to the new onset of T2DM in our training samples. We identified individuals as positive samples if they had two successive abnormal laboratory values recorded prior to T2DM diagnosis. Specifically, these values are hemoglobin (HbA1c) $\geqslant$ 6.5% and fasting plasma glucose (FPG) $\geqslant$ 126 mg/dL. Table 1 presents the demographic information of new-onset diabetes patients with comprehensive biochemical testing. This filtered data served as the foundation for pre-training LLMs in our research, and we also extended 31 standard T2DM-related indicators as input features in our LLMMs for prediction; these detailed indicators are listed in "Appendix". To enhance the ability of LLMs to process unstructured data, we will incorporate value-to-text encoding for patient laboratory values. Further details on this approach will be provided in "4.4" section. Table 2 illustrates a brief overview of our input data format, detailing clinical notes as well as numeric and textual laboratory values. The remaining group consists of numerical data that explicitly includes items related to laboratory results.

Method

Large language multimodal models

The majority of LLMs can be trained on large-scale text data before being applied as downstream models. However, most EHRs contain numerical information (e.g., age, length of hospital stay, and laboratory values) and categorical information, limiting LLMs in prediction tasks on modality data. In our study, we investigated two methods of pre-training: (1) we used a multimodal technique that combines text embedding encoders with multi-head attention mechanisms fused on laboratory data; (2) we transformed the laboratory results of patients with chronic conditions into textual data and tokenized textual laboratory text to pre-train the LLMs. In our first submodel pipeline, we developed an LLM pre-training unimodal method to extract text feature embedding from the EHR corpus as shown in Fig. 1b (top).We used primary language unimodal methods such as BERT[41], RoBERTa[42], BiomedBERT[43], Flan-T5[44], and GPT-2[45] for various tokenization and pre-training techniques on our FEMH corpus, allowing the model to comprehend a significant amount of domain-specific clinical knowledge and contextual semantics.

Large quantitative feature encoding

For our feature selection, we selected representative laboratory test items associated with T2DM as our second submodel input, as shown in "Appendix". In clinical terms, this approach allows LLMMs to identify groups at similar risk for T2DM. We first address missing values in each blood test by imputing them with mean values. Then, the data is normalized using Z-scores. During training, a simple deep neural network (DNN) is employed to extract key blood test characteristics, as illustrated in Fig. 1b (bottom panel). Subsequently, these extracted features are combined with the latent features of text already learned by the unimodal language models. This submodel is then integrated with the LLMs, which fuse the combined features within a latent space, incorporating both the extracted blood test characteristics and the semantic information from the T2DM corpus.

Multi-head attention fusion

We designed an attention module to calculate two domain embeddings for the attention score and to improve the individual unimodal contributions to the overall model prediction. We concatenated two embeddings, the text representation from LLM encoders and the blood representation from the DNN outputs. Thus, we used a multi-head attention module to facilitate an improved fusion of features from the two domains in the latent space. This attention mechanism allows us to perform a dot-product operation on text and blood vectors. We concatenated embedding vectors as query, key, and value for the attention module to generate attention-weighted matrices. By comparing the relevance of a query and key, attention weights determine the importance of each value in answering the current query, where a higher attention weight indicates a greater significance of the value for the query's resolution. Next, to enhance latent feature fusion, we used the final concatenated encoded features from multi-head attention embedding with LLMs and DNN output vectors for the final fully connected layers. Furthermore, to visualize interpretable contextual text and provide corresponding importance, we calculated

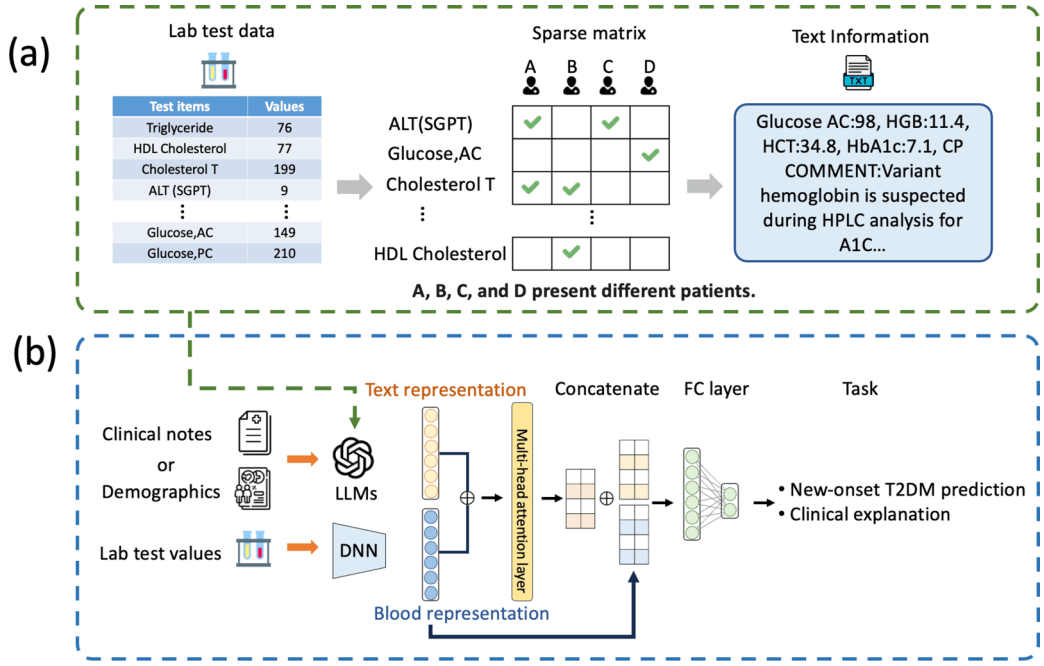

Fig. 1. The overall framework of the LLMMs. Panel (**a**) shows the values-to-text for training language models after textualizing laboratory values. Panel (**b**) demonstrates that LLMMs propose unimodal language models and DNN modules to extract and embed features from clinical notes and laboratory values. Then, multi-head attention modules are used for final feature fusion for downstream classification tasks.

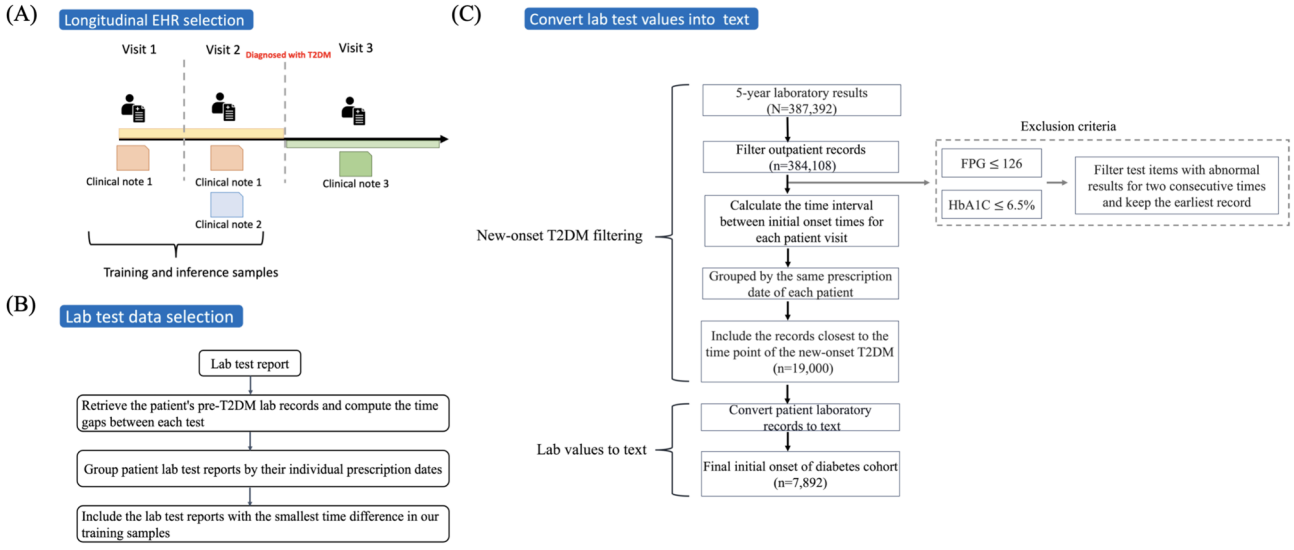

Fig. 2. The research methodology encompasses the study procedures for T2DM in EHR and laboratory tests. Panel (**A**) illustrates the process of filtering longitudinal EHR data for new-onset T2DM patients. Panel (**B**) delineates the procedure for grouping patients' laboratory test data by selecting the shortest time interval between consecutive visits. Panel (**C**) presents the overall data preprocessing steps, which involve filtering for new-onset T2DM cases and converting the final values into text format.

Shapley values based on the attention weight outputs of LLMMs, offering a comprehensive interpretability of our contextualized corpus in "Interpretable attention in textual laboratory results" section.

Conversion of laboratory values to text

Traditionally, disease modeling has primarily relied on numerical laboratory values. However, for most blood test items, there exist variations where different patients have different measurement items, or measurements are taken at different time points, leading to a problem of laboratory value sparsity, as illustrated in Fig. 1a. Recent

Demographics	Diabetes (n = 6929)	Non-diabetes (n = 23,665)	Significant
Age	67.08	55.09	$t = -52.45$, $p < 0.05$
Sex/gender (M/F)	3867/3062	11668/11997	$\chi^2 = 90.45$, $p < 0.05$
FPG	138.87 (13.60)	132.85 (9.43)	$t = -41.80$, $p < 0.05$
HbA1c	7.03 (0.35)	6.89 (0.18)	$t = -44.59$, $p < 0.05$

Table 1. Comparison of demographics between Diabetes and Non-diabetes Groups.

Modality	Data type (Task code)	Description	Example
Corpus	Textual laboratory values (A)	The laboratory examination section comprises the text description of laboratory items and the corresponding value	Free T4:1.42, TSH:1.450, HDL Cholesterol:57, BUN:22, Cholesterol T:146, Estimated GFR(MDRD):60, Glucose AC:148, ALT (SGPT):29, Uric Acid:6.5, Creatinine:1.22, K:4.5, Triglyceride:66, LDL Cholesterol:88
Corpus	Clinical note (B)	The notes document the patient's chief complaint, present illness, past medical and surgical history, physical examination findings, and assessment	This is a 56-year-old male patient with underlying hypertension and gout, with ophthalmic history of left eye retinal detachment status post pars plana vitrectomy with encircling buckling in 2016. This time, he complained of right eye with floaters since 2020.10, and he came to us and retinal breaks of right eye were noted. Therefore, a focal laser was applied. However, the patient complained of progressive visual field defect and blurred vision of right eye in recent days. Upon examination, decreased right eye vision (0.1) was found. Fundus examination showed retinal detachment from the upper parts with breaks. Pars plana vitrectomy was suggested. He received surgery on 2020.12.24. After surgery, he was admitted for further treatment
Numerical data	Laboratory values (C)	Record the biochemical data and test indicators	

Table 2. An overview of the LLMMs' training format, including corpus and modality details.

studies have utilized the manual insertion of identical templates as pseudo-notes (e.g., "Given the vitals: pulse is {value}...") into textual laboratory values [46,47]. Inspired by this perspective, we consider more nuanced approaches in our numerical-to-textual conversion process, we calculated the time difference between all records prior to the onset of diabetes and the point of recent diabetes occurrence for each patient in the training data. Subsequently, we grouped these records based on their temporal proximity to the diabetes onset. Our training dataset includes objective information extracted from the SOAP (Subjective, Objective, Assessment, Plan) components of nursing reports. These reports include entries such as, "*blood pressure/pulse measurement upload data–>BP: mmHg; PR: 72 /min [OU], No apparent diabetic retinopathy (No DR)*". These unstructured data contain vital signs recorded by professional nurses and laboratory test results, thereby preserving crucial symptomatic information about the patients.

To address this challenge, we first followed the process outlined to extract data on laboratory values from patients, then performed non-text serialization encoding and generated encoder-to-text embedding. This facilitated more direct corpus encoding for our LLMMs. This approach mitigates the issue of sparse data in testing items and overcomes the limitation of LLMs in predicting textual outcomes from solely numerical features. In addition, it helps to address the scenario in which patients have missing data for most of their laboratory test items.

Results

Single modality methods comparison

In our study, we initially validated the quantitative metrics of single modality using traditional machine learning methods. We selected common machine learning algorithms as the baseline for evaluating single modality laboratory test values, including Logistic Regression, K-Nearest Neighbors (KNN), K-Means, SVM, Random Forest, XGBoost, CatBoost and DNN. In Table 3, we can observe that the linear classifier such as logistic regression only achieved an accuracy of 0.79, while KNN and K-Means showed improved performance. Tree-based methods such as Random Forest, XGBoost, and CatBoost were able to effectively achieve accuracies above 0.85 under different metric measurements. Finally, we compared the performance of a three-layer DNN on quantitative metrics and found that it had lower performance in precision and recall. It's worth noting that the experimental results above demonstrate that while using ML methods on a single modality has some predictive power, it is limited by making predictions based on unstructured data.

Modality data for early T2DM prediction

To improve early prediction and risk of T2DM before the appearance of clinical symptoms, this task used predictive models based on relevant clinical notes or laboratory values, formulated as a binary classification problem. We evaluated the combination of three data formats as inputs for LLMMs training: (A) textual laboratory values, (B) clinical notes, and (C) laboratory values. Afterward, we evaluated early T2DM prediction based on either unimodal language models with different NLP frameworks or LLMMs architectures, as shown in Table 3.

Architecture	Modality	Model	Accuracy	Recall	Precision	F1-score
Machine Learning Classifier	C	Logistc Regeression	0.79	0.79	0.79	0.73
		KNN	0.83	0.83	0.81	0.81
		K-Means	0.83	0.83	0.82	0.81
		SVM	0.78	0.78	0.80	0.71
		Random Forest	0.86	0.86	0.85	0.85
		XGboost	0.86	0.86	0.85	0.85
		CatBoost	0.86	0.86	0.85	0.85
		DNN (Three-layer)	0.85	0.53	0.74	0.61
Unimodal	A	BiomedBERT	0.65	0.65	0.66	0.65
		ClinicalBERT	0.61	0.61	0.61	0.61
		SciFive	0.66	0.66	0.66	0.66
		RoBERTa	0.65	0.65	0.65	0.65
		Flan-T5-base-220M	0.62	0.62	0.67	0.62
		Flan-T5-large-770M	0.79	0.79	0.78	0.79
		BERT	0.66	0.66	0.66	0.66
		GPT-2	0.78	0.78	0.77	0.77
Unimodal	A+B	BiomedBERT	0.82	0.82	0.82	0.82
		ClinicalBERT	0.78	0.78	0.79	0.77
		SciFive	0.76	0.76	0.76	0.76
		RoBERTa	0.83	0.83	0.83	0.83
		Flan-T5-base-220M	0.81	0.81	0.82	0.81
		Flan-T5-large-770M	0.93	0.93	0.93	0.93
		BERT	0.79	0.79	0.79	0.79
		GPT-2	0.93	0.93	0.93	0.93
LLMMs	A+C	BiomedBERT	0.81	0.81	0.80	0.80
		ClinicalBERT	0.81	0.81	0.80	0.80
		SciFive	0.81	0.81	0.80	0.80
		RoBERTa	0.80	0.80	0.79	0.80
		Flan-T5-base-220M	0.80	0.80	0.80	0.80
		Flan-T5-large-770M	0.81	0.81	0.81	0.81
		BERT	0.82	0.82	0.80	0.81
		GPT-2	0.81	0.81	0.79	0.81
LLMMs	(A+B) + C	BiomedBERT	0.93	0.93	0.93	0.93
		ClinicalBERT	0.90	0.90	0.90	0.91
		SciFive	0.93	0.93	0.93	0.93
		RoBERTa	0.92	0.92	0.92	0.92
		Flan-T5-base-220M	0.93	0.93	0.93	0.93
		Flan-T5-large-770M	0.92	0.92	0.93	0.92
		BERT	0.93	0.93	0.93	0.93
		GPT-2	0.92	0.92	0.92	0.92

Table 3. Performance comparison of unimodal and LLMMs in new-onset T2DM prediction using different modality combinations of A, B, and C.

First, we evaluated early T2DM prediction based on unimodal language model predictions using only textual laboratory data. The majority of unimodal language model predictions could only be made using laboratory terms in non-sequential semantic order, such as "*K:4.1, HGB:14.1, Platelet:260, ALT (SGPT):12.*". Within the unimodal GPT-2 framework, we identified laboratory sequence patterns in groups undergoing the same disease evaluations, achieving an accuracy of 0.78 and precision and F1-scores of 0.77. We can assert that groups undergoing blood tests for the same disease (such as hypertension or heart disease) exhibit identical test items and sequences in textual information. Then, our analysis indicated that integrating clinical notes with predictions based on textual laboratory data significantly enhances the predictive capability of the unimodal model in the early detection of T2DM. Finally, in our proposed LLMMs, we incorporated quantifiable laboratory values with unimodal language models to perform attention fusion. The experimental results revealed that even using only textual laboratory and laboratory values, performance was enhanced compared to the unimodal effect. We also contemplated utilizing unimodal language models for clinical notes, textual laboratory values, and laboratory values in LLMMs, which can achieve a performance score of over 0.90 across different LLMs architectures.

Longitudinal T2DM risk prediction based on textual laboratory corpus

Current practice in most clinical settings relies on blood tests to confirm a preliminary diagnosis of T2DM. To proactively predict the risk of developing T2DM, we estimated the likelihood of new-onset T2DM at time intervals of 90, 180, 270, and 365 days. We then evaluated the model's performance using AUC and areas under precision-recall curves (AUPRC) metrics. To train and evaluate our model, we first selected patients with T2DM onset records from the textual laboratory data and combined them with an equal number of randomly sampled negative samples.

Figure 3 illustrates the performance of various unimodal language models in predicting the onset of T2DM at different timeframes (T days). Notably, textual post-processing of laboratory reports helps mitigate the challenge of imbalanced data samples in our training dataset. Furthermore, the experiments demonstrated consistent and stable prediction performance across different prediction timeframes for various LLMMs. Interestingly, some models even exhibited performance improvements as the prediction window increased. For example, Biomed-BERT achieved an AUC and AUPRC of 0.72 for predictions made 365 days in advance. Similarly, the larger Flan-T5 model maintained an AUC and AUPRC above 0.70 across all prediction stages.

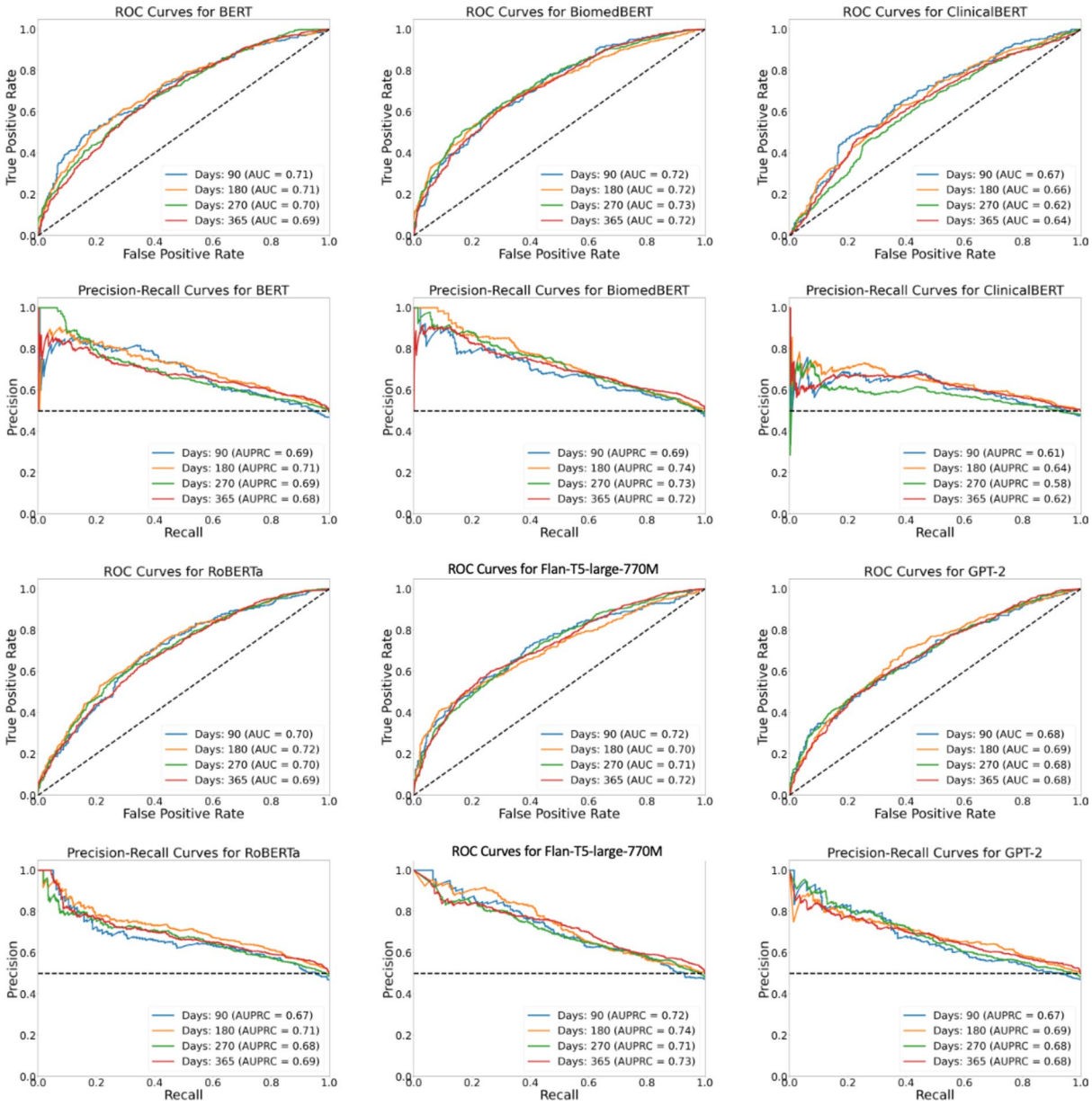

Fig. 3. Evaluating the performance of different unimodal LLMs in predicting early T2DM T Days in advance trained on textual laboratory values.

Interpretable attention in textual laboratory results

The SHAP was developed from game theory and generates Shapley values to explain the importance of features. Particularly, SHAP-based approaches have been employed as a baseline for feature importance interpretation in hemorrhagic stroke data (e.g., time-series vital signs) [48]. In SHAP-based interpretability research for EHR, the most significant clinical features for predicting various diseases were identified through SHAP analysis. This analysis incorporated pre-trained Word2Vec embeddings and either a Bidirectional Gated Recurrent Unit (BiGRU) architecture [49] or a multimodal transformer for clinical notes visualization and interpretation [50]. Given LLMs' superior ability in contextual understanding, we propose an interpretable approach for analyzing complex corpora of textual lab values and clinical notes. Our method leverages this contextual strength to provide meaningful explanations. We began by pre-training the LLMs to emphasize word positioning during the encoding process, which allows for the computation of attention scores. Subsequently, we utilized the SHAP values to analyze the combined corpus of clinical notes and textual laboratory data. This visualization tool helps us understand the individual contributions of words within the corpus. By highlighting each word's positive or negative influence on predicting specific clinical terms from the LLMMs output, SHAP values enhanced the model's clinical interpretability.

Figure 4 showcases a sample analysis of a non-diabetic patient using SHAP values. Red highlights indicate words associated with a higher risk of disease onset, while blue highlights indicate lower risk factors specific to this case. This visualization reveals the complex interplay between clinical indicators and predicted outcomes. In Fig. 4a, focusing on the non-diabetic case from the laboratory textual data, lighter colors represent low-risk test items. These include key indicators such as glucose and A1C levels, which can provide early warning signs of potential T2DM. Conversely, we analyze the effectiveness of SHAP values for feature importance using the diabetes patient case shown in Fig. 4b. This case depicts a diabetic patient with multiple chronic disease histories and a blood sugar level of 220 mg/dL. Our analysis reveals that the most influential feature is T2DM, while other significant features contributing substantially to the prediction outcome include various health conditions such as hypertension, depression, and dementia. By identifying and analyzing critical keywords within the narrative, the model unveils the intricate relationship between textual data and T2DM outcomes, providing comprehensive insight into the prediction process.

Discussion

The global trend of EHR data collection presents hospitals with significant challenges, particularly when sample sizes exceed one hundred thousand patients. Traditional machine learning algorithm packages like Synthetic Minority Over-sampling Technique (SMOTE) [51] often struggle with computational and memory constraints at this scale. In our study, we evaluated the effectiveness of mean and median imputation methods for handling missing data in new-onset cases, as depicted in Fig. 5. Our findings revealed that the choice between these two imputation techniques had minimal impact on the performance of most models. Notably, ClinicalBERT and BioBERT, which leverage specialized domain knowledge and serve as the foundation for LLMs, demonstrated superior performance. These models achieved comparatively high AUC and Accuracy (ACC) scores in predicting new-onset diabetes cases.

For complex unstructured EHRs, we provide an approach as a reference to enhance the clinical interpretability of SHAP values in a multimodal corpus derived from our proposed LLMMs. In our case study as shown in Fig. 6A, highlights significant SHAP values in the clinical report. Additionally, we utilize the feature importance

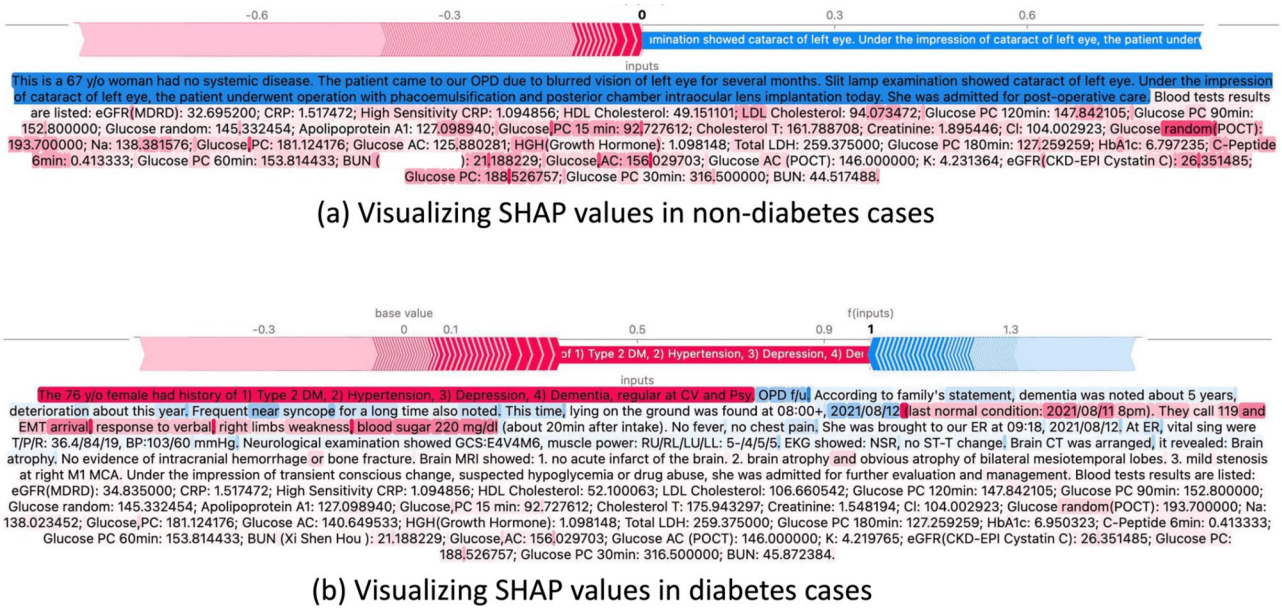

Fig. 4. The comparison and visualization of the interpretation sample of diabetics and non-diabetics with SHAP values.

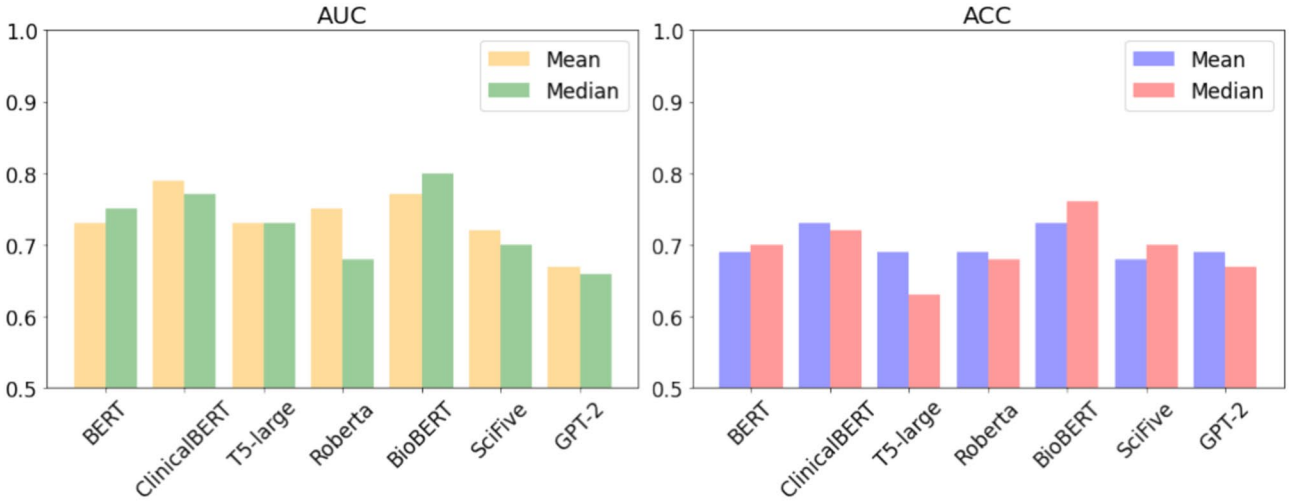

Fig. 5. Performance of new-onset T2DM prediction in LLMMs based on mean and meidan imputation methods in various backbone models.

Case Study ID: 18

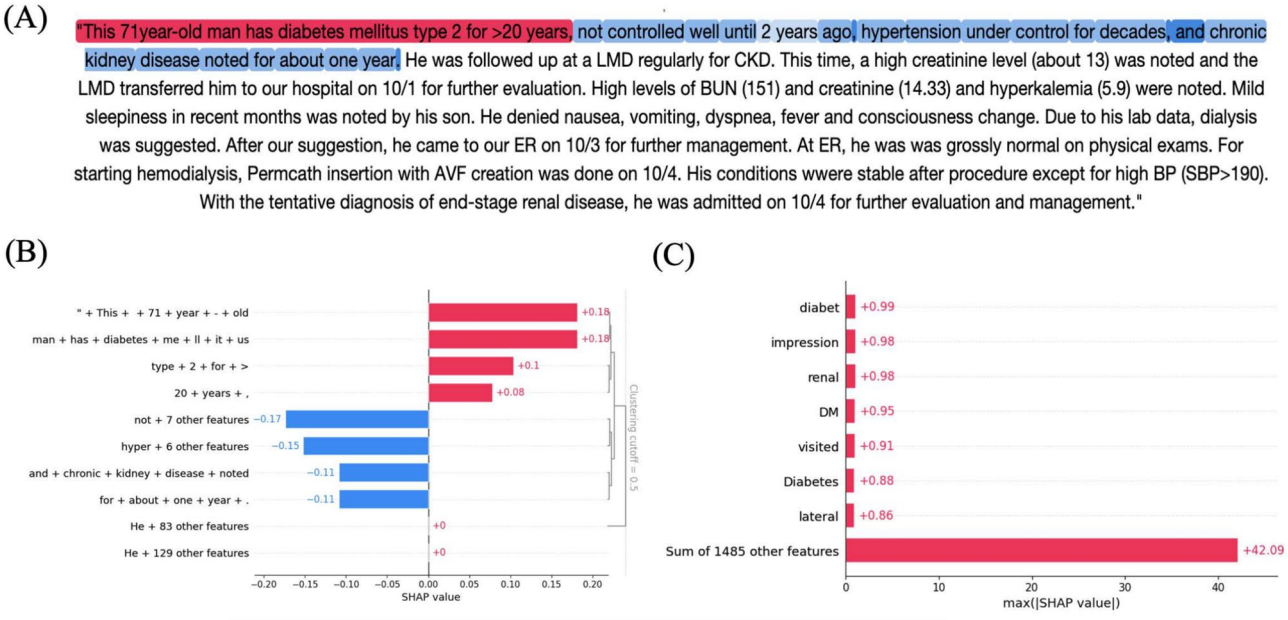

Fig. 6. Visualization of important interpretability using SHAP values in the case study of T2DM. Panel (**A**) shows the predicted SHAP values after the LLMMs, highlighting the key points of a T2DM case. Panel (**B**) represents the degree of importance (red) and negative impact (blue) of clinical token text based on SHAP values. Panel (**C**) shows the summarized importance of features by sorting SHAP values in descending order.

predicted by LLMMs in our analysis of the clinical report. In Fig. 6B, the horizontal axis represents SHAP values, extending from 0 as the baseline. Each row depicts a distinct feature, color-coded to indicate its impact. Red sections denote a positive influence on the prediction, while blue sections indicate a negative influence. Color intensity corresponds to the magnitude of the feature value. We can observe that being 71 years old and having Type 2 diabetes positively influence the model's attention, particularly corresponding well with a diabetes history of over 20 years. This provides good explanatory power for these factors. Furthermore, Fig. 6C identifies important hidden chronic disease-related medical terms such as "*impression*", "*renal*", and "*visited*". These observations substantiate that the multimodal SHAP approach offers valuable clinical reference points.

In our study limitations, two key points emerge. Firstly, while LLMs demonstrate proficiency in addressing single-modality text problems, there remains substantial scope for extended research in the realm of unstructured and structured EHRs. For instance, LLMs could potentially be employed for imputation of missing values, or

to enhance the fusion of predictions and explanations across different modalities. Secondly, the training samples derived from a single hospital's database, subject to privacy constraints, may limit the model's capacity for generalized inference.

Conclusions

In conclusion, this study explored the potential of LLMs with attention mechanisms for integrating clinical notes and laboratory data. We introduced a novel approach using textual laboratory data, demonstrating that the selection of pre-trained LLMs significantly enhances the performance of T2DM classification. Our experiments yielded promising results, with both AUC and AUPRC exceeding new-onset T2DM prediction scores of 0.70 when using LLMs on textual laboratory values. Furthermore, we investigated LLMs equipped with attention modules. By applying Shapley values to textual lab values, we enabled these LLMs to provide interpretable insights from clinical notes. In future research, we can focus on developing models that serve as real-time, effective risk alert systems for clinicians and patients.

Data availibility

The data for this study was collected from the research database of Far Eastern Memorial Hospital in Taiwan with permission. Due to patient privacy protection, the availability of the data is restricted and not publicly accessible. We will release the relevant research code (https://github.com/Ding1119/LLMMs_FEMH/tree/main) to ensure the reproducibility of our experiments. For further research and data access, please contact the corresponding author on reasonable request.

Appendix

- *Blood test items*: eGFR (MDRD), CRP, High Sensitivity CRP, HDL Cholesterol, LDL Cholesterol, Glucose PC 120min, Glucose PC 90min, Glucose random, Apolipoprotein A1, Glucose, PC 15 min, Cholesterol T, Creatinine, Glucose random (POCT), Na, Glucose PC, Glucose AC, HGH (Growth Hormone), Total LDH, Glucose PC 180min, HbA1c, C-Peptide 6min, Glucose PC 60min, BUN, Glucose AC (POCT), K, eGFR (CKD-EPI Cystatin C), Glucose PC 30min, Triglyceride, ALT (SGPT), AST (SGOT), Creatinine (POCT).

Received: 13 May 2024; Accepted: 23 August 2024
Published online: 06 September 2024

References

1. (WHO), W. H. O. The top 10 causes of death (2020).
2. Chew, N. W. *et al.* The global burden of metabolic disease: Data from 2000 to 2019. *Cell Metab.* **35**, 414–428 (2023).
3. Gregg, E. W. *et al.* Improving health outcomes of people with diabetes: Target setting for the who global diabetes compact. *Lancet* **401**, 1302–1312 (2023).
4. Health Promotion Administration, M. O. H. & Welfare. Statistical yearbook of health promotion. https://www.hpa.gov.tw/EngPages/Detail.aspx?nodeid=3850 &pid=17613 (2021).
5. Bernardini, M., Romeo, L., Misericordia, P. & Frontoni, E. Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine. *IEEE J. Biomed. Health Inform.* **24**, 235–246 (2019).
6. Bisercic, A. *et al.* Interpretable medical diagnostics with structured data extraction by large language models (2023). CoRR https://doi.org/10.48550/ARXIV.2306.05052
7. Lee, R. Y. *et al.* Assessment of natural language processing of electronic health records to measure goals-of-care discussions as a clinical trial outcome. *JAMA Netw. Open* **6**, e231204–e231204 (2023).
8. Jiang, L. Y. *et al.* Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
9. Cahan, N. *et al.* Multimodal fusion models for pulmonary embolism mortality prediction. *Sci. Rep.* **13**, 7544 (2023).
10. Guazzo, A. *et al.* Deep-learning-based natural-language-processing models to identify cardiovascular disease hospitalisations of patients with diabetes from routine visitsâ€™ text. *Sci. Rep.* **13**, 19132 (2023).
11. Juhn, Y. & Liu, H. Artificial intelligence approaches using natural language processing to advance ehr-based clinical research. *J. Allergy Clin. Immunol.* **145**, 463–469 (2020).
12. Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J. & Eisenstein, J. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 3 (eds Walker, M. *et al.*) 1101–1111 (Association for Computational Linguistics, 2018). https://doi.org/10.18653/v1/N18-1100.
13. Liu, J., Zhang, Z. & Razavian, N. Deep ehr: Chronic disease prediction using medical notes. In *Machine Learning for Healthcare Conference* 440–464 (PMLR, 2018).
14. Zhao, W. X. *et al.* A survey of large language models. arXiv:2303.18223
15. Yang, X. *et al.* A large language model for electronic health records. *NPJ Digit. Med.* **5**, 194 (2022).
16. Johnson, A. E. *et al.* Mimic-iii, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
17. Belyaeva, A. *et al.* Multimodal llms for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data* 86–102 (Springer, 2023).
18. Bycroft, C. *et al.* The UK biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
19. Minot, J. R. *et al.* Interpretable bias mitigation for textual data: Reducing genderization in patient notes while maintaining classification performance. *ACM Trans. Comput. Healthcare*https://doi.org/10.1145/3524887 *(2022).*
20. Wu, J., Gan, W., Chen, Z., Wan, S. & Yu, P. S. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)* 2247–2256 (2023). https://doi.org/10.1109/BigData59044.2023.10386743
21. Gasparetto, A., Marcuzzo, M., Zangari, A. & Albarelli, A. A survey on text classification algorithms: From text to predictions. *Information (Basel)* **13**, 83 (2022).
22. Sun, X. *et al.* Text classification via large language models. In *Conference on Empirical Methods in Natural Language Processing* 8990–9005. (Association for Computational Linguistics, Singapore, 2023).

23. Zhang, L., Tashiro, S., Mukaino, M. & Yamada, S. Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: A comparative test case. *J. Rehabil. Med.* **55**, 13373 (2023).
24. Steinberg, E. *et al.* Language models are an effective representation learning technique for electronic health record data. *J. Biomed. Inform.* **113**, 103637 (2021).
25. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 35 (2017).
26. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining* 785–794 (2016).
27. Laupacis, A. *et al.* Clinical prediction rules: A review and suggested modifications of methodological standards. *JAMA* **277**, 488–494 (1997).
28. Hosmer, D. W. & Lemeshow, S. *Applied Logistic Regression*, 2 edn. Wiley Series in Probability and Statistics-Applied Probability and Statistics Section (Wiley-Interscience, New York, 2013)
29. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
30. Yu, K. & Xie, X. Predicting hospital readmission: A joint ensemble-learning model. *IEEE J. Biomed. Health Inform.* **24**, 447–456 (2019).
31. Han, J. M. *et al.* Unsupervised neural machine translation with generative language models only (2021). arXiv:2110.05448
32. Cascella, M., Montomoli, J., Bellini, V. & Bignami, E. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. *J. Med. Syst.* **47**, 33 (2023).
33. Chen, S. *et al.* Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation (2023). CoRR https://doi.org/10.48550/ARXIV.2305.13614
34. Huang, H. *et al.* ChatGPT for shaping the future of dentistry: The potential of multi-modal large language model. *Int. J. Oral Sci.* **15**, 29 (2023).
35. Kleesiek, J., Wu, Y., Stiglic, G., Egger, J. & Bian, J. An opinion on ChatGPT in health care—Written by humans only. *J. Nucl. Med. Soc. Nucl. Med.* **64**, 701–703 (2023).
36. Chirino, A. *et al.* High consistency between recommendations by a pulmonary specialist and ChatGPT for the management of a patient with non-resolving pneumonia. *Nort. Healthc. Med. J.* **8**, 9. https://doi.org/10.59541/001c.75456 (2023).
37. Wang, S., Zhao, Z., Ouyang, X., Wang, Q. & Shen, D. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv:2302.07257
38. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* **4**, 86 (2020).
39. Yan, A. *et al.* Radbert: Adapting transformer-based language models to radiology. *Radiol. Artif. Intell.* **4**, e210258. https://doi.org/10.1148/ryai.210258 (2022).
40. Kather, J. N. Artificial intelligence in oncology: Chances and pitfalls. *J. Cancer Res. Clin. Oncol.* **149**, 7995–7996. https://doi.org/10.1007/s00432-023-04666-6 (2023).
41. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding (2018). arxiv:1810.04805
42. Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach (2019). arXiv:1907.11692
43. Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare (HEALTH)* **3**, 1–23 (2021).
44. Chung, H. W. *et al.* Scaling instruction-finetuned language models (2022). arXiv:2210.11416
45. Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).
46. Hegselmann, S. *et al.* Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics* 5549–5581 (PMLR, 2023).
47. Lee, S. A. *et al.* Multimodal clinical pseudo-notes for emergency department prediction tasks using multiple embedding model for ehr (meme) (2024). arXiv:2402.00160
48. Feng, Q. *et al.* Can attention be used to explain ehr-based mortality prediction tasks: A case study on hemorrhagic stroke. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* 1–6 (2023).
49. Grout, R. *et al.* Predicting disease onset from electronic health records for population health management: A scalable and explainable deep learning approach. *Front. Artif. Intell.* **6**, 1287541 (2024).
50. Lyu, W. *et al.* A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction. In *AMIA Annual Symposium Proceedings*, vol. 2022, 719 (American Medical Informatics Association, 2022).
51. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

Author contributions

Jun-En Ding and Phan Nguyen Minh Thao were the co-first authors, assisting in the complete programming, experimental design, and manuscript writing. Professor Wen-Chih Peng, Dr. Fang-Ming Hung, and Professors Ling Che, Dongsheng Luo, and Feng Liu provided guidance and supervision to students in their research. Jian-Zhe Wang, Chun-Cheng Chug, Min-Chen Hsieh, Yun-Chien Tseng, and Chenwei Wu assisted in organizing five years of EHR data and programming. Dr. Chi-Te Wang, Chih-Ho Hsu, Yi-Tui Chen, and Pei-fu Chen provided clinical advice and guidance on the manuscript.

Competing interests

The authors declare no competing interests.

Ethical approval

The FEMH Research Ethics Review Committee Board granted ethical approval for this study protocol (Reference FEMH No.: 112086-F). The research database was approved and acquired for this study on July 13, 2023, and the research period extends until December 31, 2025. Informed consent was not required because the study involved anonymized, retrospective data. The informed consent was waived by the Institutional Review Board (IRB) of FEMH.

Additional information

Correspondence and requests for materials should be addressed to F.-M.H.

Reprints and permissions information is available at www.nature.com/reprints.