Journal of Multidisciplinary Healthcare

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

LSTM-Based Prediction Model for Tuberculosis Among HIV-Infected Patients Using Structured Electronic Medical Records: A Retrospective Machine Learning Study

Jingfang Chen[1,2], Linlin Liu[3], Junxiong Huang[1], Youli Jiang[4], Chengliang Yin[1], Lukun Zhang[5], Zhihuan Li[1], Hongzhou Lu[1,5]

[1]Faculty of Medicine, Macau University of Science and Technology, Macau, 999078, People's Republic of China; [2]Department of Research and Teaching, The Third People's Hospital of Shenzhen, Shenzhen, 518112, People's Republic of China; [3]Hengyang Medical School, School of Nursing, University of South China, Hengyang, 421001, People's Republic of China; [4]Department of Neurology, The People's Hospital of Longhua, Shenzhen, 518109, People's Republic of China; [5]Department of Infectious Diseases, National Clinical Research Center for Infectious Diseases, The Third People's Hospital of Shenzhen, Shenzhen, 518112, People's Republic of China

Correspondence: : Zhihuan Li; Hongzhou Lu, Faculty of Medicine, Macau University of Science and Technology, Macau, 999078, People's Republic of China, Email baopullon@163.com; luhongzhou@szsy.sustech.edu.cn

Background: Both HIV and TB are chronic infectious diseases requiring long-term treatment and follow-up, resulting in extensive electronic medical records. With the exponential growth of health and medical big data, effectively extracting and analyzing these data has become the research hotspot. As a fundamental aspect of artificial intelligence, machine learning has been extensively applied in medical research, encompassing diagnosis, treatment, patient monitoring, drug development, and epidemiological investigations. This significantly enhances medical information systems and facilitates the interoperability of medical data.

Methods: In our study, we analyzed longitudinal data from the electronic health records of 4540 patients, gathered from the National Clinical Research Center for Infectious Diseases in Shenzhen, China, spanning from 2017 to 2021. Initially, we employed the fine-tuned ChatGLM to structure the electronic medical records. Subsequently, we utilized a multi-layer perceptron to classify each patient and determined the presence of tuberculosis in HIV patients. Using machine learning-based natural language processing, we structured these records to build a specialized database for HIV and TB co-infection. We studied the epidemiological characteristics, focusing on incidence patterns, patient characteristics, and influencing factors, to uncover the transmission characteristics of these diseases in Shenzhen. Additionally, we used Long Short-Term Memory to create a predictive model for TB co-infection among HIV patients, based on their medical records. This model predicted the risk of TB co-infection, providing scientific evidence for clinical decision-making and enabling early detection and precise intervention.

Results: Based on the refined ChatGLM model tailored for structured electronic health records, the accuracy of symptom extraction consistently surpassed 0.95 precision. Key symptoms such as diarrhea and normal showed precision rates exceeding 0.90. High scores were also achieved in recall and F1 scores. Among 4540 HIV patients, 758 were diagnosed with concurrent tuberculosis, indicating a 16.7% co-infection rate, while syphilis co-infection affected 25.1%, underscoring the prevalence of concurrent infections among HIV patients. Utilizing electronic health records, a Multilayer Perceptron classifier was developed as a benchmark against Long Short-Term Memory to predict high-risk groups for HIV and tuberculosis co-infections. The Multilayer Perceptron classifier demonstrated predictive ability with AUROC values ranging from 0.616 to 0.682 on the test set, suggesting opportunities for further optimization and generalization despite its accuracy in identifying HIV-TB co-infections. In tuberculosis intelligent diagnosis based on laboratory results, the Long Short-Term Memory showed consistent performance across 5-fold cross-validation, with AUROC values ranging from 0.827 to 0.850, indicating reliability and consistency in tuberculosis prediction. Furthermore, by optimizing classification thresholds, the model achieved an overall accuracy of 81.18% in distinguishing HIV co-infected tuberculosis from simple HIV infection.

Conclusion: Combining the Multilayer Perceptron classifier with Long Short-Term Memory represented an advanced approach for effectively extracting electronic health records and utilizing it for disease prediction. This underscored the superior performance of

Received: 8 March 2024
Accepted: 17 July 2024
Published: 23 July 2024

deep learning techniques in managing both structured and unstructured medical data. Models leveraging laboratory time-series data demonstrated notably better performance compared to those relying solely on electronic health records for predicting tuberculosis incidence. This emphasized the benefits of deep learning in handling intricate medical data and provided valuable insights for healthcare providers exploring the use of deep learning in disease prediction and management.

Keywords: Prediction models, HIV, Tuberculosis, Machine Learning, Artificial Intelligence

Introduction

Acquired Immunodeficiency Syndrome (AIDS), characterized by the progressive weakening of the immune system, frequently results in opportunistic infections like tuberculosis (TB).[1] Notably, TB is the leading cause of mortality among individuals with Human Immunodeficiency Virus (HIV), where the combined prevalence and interaction of HIV and TB epidemics significantly contribute to acute illness and elevated global mortality rates. Qi et al[2] conducted a meta-analysis in 2023 and determined that the pooled prevalence of HIV/TB co-infection in China was 6.0%. For people living with HIV (PLHIV), TB represents a critical cause of death, accounting for approximately one-quarter of all fatalities.[3] Nonetheless, there has been a consistent decline in TB-related deaths among PLHIV; in 2022, the estimated figure stood at 167,000 (95% UI: 139,000–198,000).[4] The epidemiological investigation unveiled that the combined mortality rate among individuals concomitantly affected by HIV and TB in China was recorded at 15.92%.[5]

Various methods are utilized for screening TB, encompassing symptom assessment, chest imaging, C-reactive protein (CRP) testing, laboratory tests, and rapid molecular biology testing. During each HIV/AIDS follow-up, screening for symptoms of TB is recommended. PLHIV who exhibit symptoms should undergo either chest X-ray or CRP testing. CRP testing stands out as a straightforward, cost-effective, and immediate diagnostic method. Its accuracy in identifying active TB among HIV/AIDS patients surpasses that of symptom screening. For TB screening in HIV/AIDS patients, using a cutoff of 5 mg/L demonstrates higher sensitivity compared to a cutoff of 10 mg/L. Furthermore, these individuals should undergo an annual chest X-ray examination.

With the rapid expansion of antiretroviral therapy (ART) in developing country, a pressing issue remains the persistently high mortality rates in patients co-infected with TB and HIV, even with the availability of effective treatments for both conditions.[6] Autopsy studies have disclosed a significant prevalence of undiagnosed TB in individuals positive for HIV-1, suggesting that the ramifications of co-infection may have been substantially underestimated.[7] Moreover, the concurrent presence of TB and HIV poses intricate clinical challenges, including diagnostic complexities, drug interactions, and increased adverse reactions to treatments. Despite extensive research efforts, accurately predicting TB development in PLHIV continues to be a formidable task, highlighting the necessity for sophisticated predictive models that can adapt to the evolving nature of these diseases.

Electronic medical records (EMRs) contain structured data, unstructured data, and time series data. Traditional statistical models are limited to handling structured data and cannot effectively analyze unstructured and time series data. In contrast, Machine Learning (ML) can simultaneously process and integrate these multi-dimensional and heterogeneous data types, providing more comprehensive and accurate analytical results. Traditional statistical models rely on manual feature engineering, requiring experts to select and extract features based on their experience. This process is time-consuming and prone to missing critical information. ML automates feature extraction by autonomously learning features and patterns through multi-layer neural networks. This capability enables the identification of complex non-linear relationships and hidden patterns, thereby improving predictive performance and accuracy. As medical informatics advances, the volume of EMR data continues to grow. ML, enhanced by big data technologies and distributed computing frameworks, can efficiently process and analyze massive datasets to extract valuable insights. However, traditional statistical models often face limitations in computing resources and time when dealing with large-scale data, making real-time analysis and prediction challenging. Specifically, ML endows systems with the ability to learn from data, enabling them to accomplish targeted tasks. Practically, this involves training models with large datasets to solve specific problems. ML excels at processing data, analyzing images, and identifying features without subjective interference, thereby providing more accurate anomaly detection, enhancing diagnostic accuracy, and predicting disease progression and prognosis.

There were several clinical prediction models designed for TB screening in PLHIV,[8–12] these models exhibit certain limitations. Some have been displayed sub-optimal performance during external validation, lack extensive external validation, or remain unassessed for clinical utility. Meanwhile, the unstructured nature of EHRs presents significant obstacles for data mining and reuse.[13] Recent advancements, such as ChatGLM, have enhanced natural language processing capabilities for structuring EHRs. ChatGLM shows promising accuracy in comprehending medical text.[14] But it lacks customization for institutional EHR quirks. Fine-tuning on local EHRs can adapt the model to local vocabulary and note patterns,[15] improving generalizability. Our approach involved fine-tuning ChatGLM using anonymized EHRs from the National Clinical Research Center for Infectious Diseases. The modified model, when tested on an annotated dataset, showed superior F1 scores in identifying medication, symptom, and diagnosis entities than the original version. It also generated more coherent key-value pairs. Implementing this refined ChatGLM for structuring EHR in downstream tasks such as cohort selection and clinical decision-making, while upholding patient privacy. Overall, fine-tuning large language models on local EHRs shows potential for unlocking EHR data. The convergence of AI and public health is paving new pathways for the prediction, management, and comprehension of complex diseases.

LSTM (Long Short-Term Memory) is an advanced type of Recurrent Neural Network (RNN) particularly suited for handling and predicting tasks involving time series data, designed to address the issue of vanishing gradients. Unlike traditional RNNs, LSTM introduces specialized units for storing and managing memory, implemented through a meticulously designed structure of gates, including the input gate, output gate, and forget gate. Each LSTM unit contains a memory cell responsible for maintaining the network's long-term state over time series. These gating mechanisms enable LSTM networks to more effectively control the flow of information within the network, thereby mitigating the vanishing gradient problem and capturing dependencies over extended sequences. These enhancements significantly boost the performance of RNNs, making them crucial for tackling more complex and demanding tasks involving long-sequence data.

The strength of LSTM lies in its ability to capture and utilize long-term dependencies within time series data, thereby improving the understanding and prediction of patients' health statuses. Additionally, LSTM is widely used for predicting medical events, such as when a patient might need intensive care or forecasting the health trends of chronic diseases. By learning from and modeling patients' historical data, LSTM can provide valuable predictive information, offering timely and accurate support for medical decision-making. Thus, the application prospects of LSTM in the medical field are promising, potentially bringing significant advancements and improvements to medical research and clinical practice.

In this study, by integrating MLP and LSTM, we combined the analysis of static and dynamic data, thereby improving the accuracy of predicting tuberculosis incidence among HIV/AIDS patients. The use of MLP allows us to delve into patients' basic biochemical indicators, while the incorporation of LSTM enables us to account for the time-dependence of disease progression. This complementary approach not only enhanced the predictive capability of the model but also underscored the importance of combining different types of machine learning models to tackle complex medical prediction problems. Our research results highlighted the potential of employing multi-model approaches in medical research and clinical practice, providing valuable insights for future studies on similar issues.

The study aimed to analyze the epidemiological characteristics and risk factors of tuberculosis infection in HIV/AIDS patients and to construct the LSTM-based predictive model to accurately forecast the incidence of TB. The application of this method will provide effective risk assessment tools in clinical practice, aiding healthcare providers in accurately identifying high-risk populations, optimizing resource allocation, and formulating more targeted prevention and treatment strategies. Ultimately, this will help alleviate the public health burden posed by the dual infection of HIV and TB. By integrating Named Entity Recognition (NER) and LSTM, our research not only offered new perspectives for EMR analysis but also provided support and solutions for managing and preventing HIV and TB co-infection.

Methods

Study design and population

In this study, we utilized EMR data, a diverse dataset that encompassed patient medical information sourced from National Clinical Research Center for Infectious Diseases, Shenzhen. The hospital has emerged as a leading institution in

the field of infectious disease research, with a particular emphasis on HIV and TB. Over the period spanning from January 1, 2017, to December 31, 2021, we amassed a cohort of 6426 individuals suspected or confirmed to have HIV. Given the substantial size and diverse nature of the dataset, rigorous data cleaning was imperative to ensure the quality and reliability of the analysis. The data cleaning process involved several critical steps, starting with the removal of cases where individuals had only undergone preliminary HIV testing without subsequent confirmatory tests. Confirmatory testing is crucial for accurate HIV diagnosis, and its absence may lead to data inaccuracies. Additionally, we excluded individuals who, despite being initially suspected of HIV, maintained regular medical visits up to December 31, 2021, without receiving a definitive HIV diagnosis. This exclusion helped to refine the dataset by focusing solely on confirmed cases, thereby enhancing the specificity of our research findings. Moreover, records with incomplete patient identification data were also removed. Accurate patient identification is pivotal in longitudinal health studies to track patient history and treatment outcomes effectively. Incomplete data can lead to duplication of records or misattribution of medical information, which could skew the results and lead to erroneous conclusions. After these exclusions, we retained a total of 4540 HIV patients for inclusion in this study. These records encompass a wealth of information, including patient identification, comprehensive medical histories, records of clinical visits, results from diagnostic tests and medical imaging, as well as the treatment plans meticulously crafted by attending physicians. The research received ethical approval from the Ethics Review Committee of the Third People's Hospital of Shenzhen (Approval Number: [2022–027]). Due to the study's methodology, which did not entail direct patient involvement, the Ethics Committee of The Third People's Hospital of Shenzhen, China, sanctioned the study's protocol and exempted the requirement for acquiring informed consent from participants. The research was conducted in strict conformity with pertinent ethical standards and legal mandates. The study complied with the Declaration of Helsinki.

The precise identification of patients is crucial in ensuring accurate patient recognition. The medical histories go beyond an account of individual past illness and surgeries to encompass familial medical backgrounds, thereby yielding valuable insights into hereditary or familial diseases. Clinical records provide a detailed reflection of both outpatient and inpatient treatment processes, while test and imaging results serve as vital pieces of evidence for diagnoses. The treatment plans offer a comprehensive outline of the planning and execution of therapeutic interventions, while nursing records monitor any changes in the patient's daily health status and vital signs. Prescription information duly documents the therapeutic guidance and recommendations of the physicians. This extensive compilation of data coalesces into a comprehensive information repository that not only underpins clinical decision-making but also provides a valuable database for medical research endeavors. Electronic medical records, with their expansive scope, offer large-scale real-world clinical data that can be instrumental in the development of clinical support decision systems.

Blood examinations encompass a wide array of tests, such as complete blood counts, platelet analysis, liver and renal function panels, cardiac enzymes, lipid profiles, glycemic indicators, thyroid assessments, infectious disease screenings, cancer diagnostics, autoimmune markers, hormonal levels, and genetic disease indicators. These tests play a crucial role in diagnosing various conditions, tracking disease progression, and evaluating treatment efficacy. For example, white blood cell counts and differentials can signal infections or blood disorders; liver enzymes, such as glutamic pyruvic transaminase and glutamic oxaloacetic transaminase, shed light on liver health; creatinine and urea measurements indicate kidney function; lipid profiles, including cholesterol and triglycerides, are critical for cardiovascular risk assessment; thyroid tests are vital for identifying thyroid issues; and markers for infectious diseases like HIV, Hepatitis B Virus, and Hepatitis C Virus are crucial in detecting these viral infections.

LSTM model

LSTM are a special type of recurrent neural network that possess the ability to learn long-term dependencies. They were introduced by *Hochreiter & Schmidhuber* (1997)[16] and refined and popularized in subsequent works, solidifying their status as highly effective solutions across a myriad of problems. Widely used in modern times, LSTMs have emerged as an important force in the field of artificial intelligence.[17]

The primary task of LSTM is to utilize a patient's historical health records to predict their likelihood of developing a disease at a future point in time. This system functions akin to an experienced physician who can forecast future health risks by analyzing a patient's past health conditions. Input (analogous to information gathered by a doctor): Patient's

blood sample data——These data are derived from blood tests conducted at various times in the patient's past, such as blood sugar levels, cholesterol levels, etc., similar to how a doctor obtains health information through blood tests. Patient's structured medical history data——This includes records of the patient's medical history, such as past illnesses, treatment records, etc., analogous to the process by which a doctor understands a patient's medical history. Output (doctor's diagnostic prediction): Likelihood of illness ——Based on the patient's historical health data, the system calculates the probability of the patient developing a certain disease at a future point in time, much like how a doctor predicts future health conditions based on examination results and medical history.

The core feature of LSTMs is the cell state, symbolized as the horizontal line in Figure 1. This cell state functions like a conveyor belt, smoothly carrying information across the network with minimal linear interference, allowing data to be transferred unaltered. LSTMs can delete or add information to the cell state, carefully regulated by structures called gates. Gates are a way to selectively let information through, composed of a sigmoid neural net layer and a pointwise multiplication operation. The sigmoid layer outputs numbers from 0 to 1 describing how much each component should be let through. A value of 0 represents "let nothing through" while a value of 1 represents "let everything through". LSTMs possess three such gates to protect and control the cell state. The output of the LSTM fuses contextual information from X, making it particularly well-suited for time series data analysis.[18] Simply adding a basic MLP on top of the LSTM output vector completes the design of a LSTM-based classifier, which is widely applied in intelligent diagnosis applications utilizing sequential follow-up data.[19]

MLP Classifier

The MLP Classifier implements a MLP architecture for regression tasks. The model consists of fully-connected neural network layers with a sigmoid output activation function.

The model starts by taking an input feature vector, denoted as x, of any size. This vector is then processed through two hidden layers, each using ReLU activations, followed by a final sigmoid output layer. In our study, the first hidden layer projects the input into a 30-dimensional representation, while the second layer maps this into a 10-dimensional embedding before the final regression output. By varying the input and output sizes, number of layers, and hidden dimensions, this model can be adapted for different regression problems. The modular implementation allows flexibility in model architecture. The use of fully-connected layers and stacked nonlinear activation gives the model the ability to learn complex mappings between input features and target variables. The ReLU activations introduce nonlinearity, while the final sigmoid squashes outputs to (0,1) for probabilistic regression.

In this study, we utilized longitudinal electronic medical records (EMRs) and detailed laboratory test data from individual patients to predict the probability of disease onset at a future predetermined time point. Our analysis focused on historical health data spanning several years, carefully extracting patterns and trends indicative of disease progression.

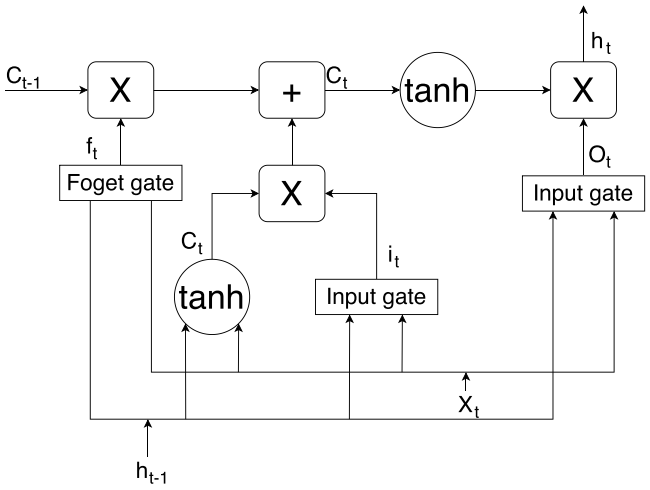

Figure 1 Diagram illustrating the structure of an LSTM unit.

Specifically, the laboratory test data included a comprehensive set of biomarkers such as ID, A/G, Glu, PDW, Cr, PCT, PLT, GGT, TG, AMY, HDL, IG%, HGB, MPV, DB, TB, GLO, EO#, MCH, LDL, HCT, EO%, IG#, TP, ALB, ALT, AST, MCV, Urea, NEUT#, LYMPH%, TH/TS, RBC, RDW-CV, Th-Cell, Ts-Cell, Th-Count, P-LCR, MONO#, NRBC#, RDW-SD, NRBC%, T-CELL, Ts-Count, MONO%, NEUT%, U/C, CHOL, MCHC, eGFR, BASO#, BASO%, TBA, WBC, LYMPH#, AST/ALT, Tc-Count. These biomarkers were instrumental in identifying biochemical and hematological changes that may signify early stages of disease development.

The EMRs data incorporated critical elements such as chief complaint, history of present illness, physical examination, diagnostic tests, and initial diagnosis. By integrating these EMRs with the laboratory test results, our predictive model leveraged temporal data sequences to forecast potential health outcomes.

This approach allowed us to identify at-risk individuals early on by interpreting subtle longitudinal changes in their health data, enabling proactive and personalized early intervention strategies. Our model was built using advanced machine learning algorithms that processed these vast and varied datasets to accurately predict the probability of disease occurrence at future time points. This predictive capability is crucial for implementing timely healthcare interventions that could potentially mitigate or even prevent the onset of disease, thus significantly improving patient outcomes and reducing healthcare costs.

Training sample generation

In the training of our MLP and LSTM models, we adopted a five-fold cross-validation strategy to prevent overfitting. This technique involves dividing the entire dataset into five distinct subsets. Throughout the training phase, each subset is systematically used once as a validation set while the remaining four subsets are utilized as the training data. This iterative process not only allows every segment of the dataset to be used for both training and validation but also significantly enhances the generalizability of the models. Employing such a rigorous validation method is essential to ensuring that the MLP and LSTM models maintain robust performance when exposed to new and diverse datasets.

The structured data of patients is aggregated by patient ID to obtain the follow-up time series of patients. The follow-up data of each patient were sorted in ascending order according to the follow-up time, and the LSTM training samples were constructed by using the method of permutation and combination. For example, if a patient has 5 follow-up data, then take the 1, 2, and 3 follow-up data as the input sequence x_1, x_2, x_3 of LSTM, and take the third diagnosis result as the expected output y_3 of LSTM. In this way, we can generate 9 training samples from 5 follow-up data (Figure 2).

Follow-up number	Characteristics					Diagnosis
1	chief complaint	history of present disease	physical examination	laboratory examination	imaging examination	HIV
2	chief complaint	history of present disease	physical examination	laboratory examination	imaging examination	HIV
3	chief complaint	history of present disease	physical examination	laboratory examination	imaging examination	HIV
4	chief complaint	history of present disease	physical examination	laboratory examination	imaging examination	HIV
5	chief complaint	history of present disease	physical examination	laboratory examination	imaging examination	HIV&TB

Figure 2 Training sample generation of HIV patients.

Results

Baseline characteristics

The study cohort comprised 4540 individuals between January 1, 2017 and December 31, 2021. As detailed in Table 1, the majority of participants were male (3876, representing 85.4%), with an average age of 39.5 years. Notably, 758 individuals (17%) were concurrently afflicted with HIV and TB. The median number of follow-up appointments per participant was 10, ranging from 1 to 152. The median interval between follow-ups was 88 days (ranging from 1 to 1465 days), and the median duration of follow-up per participant (from the last to the first visit) was 782 days (ranging from 1 to 1825 days).

In summary, our cohort covered a sizable HIV-infected population with longitudinal outpatient records, enabling modeling of disease progression. The male predominance aligns with the known HIV epidemiology in our context.[20] However, future studies should focus on enhancing the representation of underrepresented groups, such as women and adolescents.[21] The workflow diagram of the AI diagnosis framework was illustrated in Figure 3.

Table 1 Baseline Characteristics of the Cohorts

Variables	Classification	Number	$(\%)/\left(\bar{x}\pm SD\right)$
Sex	Male	3876	85.4
	Female	664	14.6
Ethics	the Han nationality	4364	96.1
	Others	176	3.9
Education level	Primary school	435	9.6
	Middle school	1167	25.7
	High school	1372	30.2
	University and above	1566	34.5
Married status	Married	1948	42.9
	Unmarried	2592	57.1
Payment Methods	Medical insurance	3444	75.9
	Self-pay	1096	24.1
Smoking	Yes	708	15.6
	No	3832	84.4
Drinking	Yes	196	4.3
	No	4344	95.7
Household registration	Shenzhen	318	7.0
	Non Shenzhen	4222	93.0
BMI group	<18.5	775	17.1
	18.5~	3005	66.2
	≥24	760	16.7
BMI	/	4540	21.28±3.108
Age	<18	45	1.0
	18~50	3665	80.7
	>50	830	18.3
Age group	/	4540	39.53±12.589
Height (cm)	/	4540	169.02±6.635
Weight (kg)	/	4540	61.05±10.585
Chronic disease	Diabetes	178	3.9
	Hypertension	246	5.4
	Coronary artery disease	35	0.8
	Cerebral infarction	68	1.5
	Chronic kidney disease	39	0.9

(*Continued*)

Table 1 (Continued).

Variables	Classification	Number	$(\%)/\left(\bar{x}\pm SD\right)$
Infectious diseases	Tuberculosis	758	16.70
	Pulmonary tuberculosis	619	13.6
	Multidrug-resistant tuberculosis	13	0.3
	Extra-pulmonary tuberculosis	327	7.2
	Syphilis	1141	25.1
	Viral Hepatitis	659	14.5
	HBV	283	6.2
	HCV	89	2.0
	Tuberculous Meningitis	67	1.5
	Malniferous Basketball Bacteria	135	3.0
Pulmonary infection	Yes	998	22.0
	No	3542	78.0
Tumor	Yes	669	14.7
	No	3871	85.3
Cancer	Yes	90	2.0
	No	4450	98.0
Calculus	Yes	248	5.5
	No	4292	94.5
Polyp	Yes	145	3.2
	No	4395	96.8
Admission Pathway	Emergency treatment	3122	68.8
	Outpatient	1418	31.2
Route of admission	Outpatient	1120	24.7
	Hospitalization	3420	75.3
Psychological status	Normal	3303	72.8
	Anxiety	1207	27.2
Risk of malnutrition	Yes	730	16.1
	No	3810	83.9
ICU history	Yes	463	10.2
	No	4077	89.8
Allergy history	Yes	195	4.3
	No	4345	95.7
Resuscitation history	Yes	617	13.6
	No	3923	86.4
Herpes zoster	Yes	112	2.5
	No	4428	97.5
Drug induced liver damage	Yes	221	4.9
	No	4319	95.1
Drug induced dermatitis	Yes	95	2.1
	No	4445	97.9
Condyloma	Yes	81	1.8
	No	4459	98.2
Anal Abnormalities	Yes	586	12.9
	No	3954	87.1
Perianal abscess	Yes	129	2.8
	No	4411	97.2
Fistula	Yes	222	4.9
	No	4318	95.1
Mixed hemorrhoids	Yes	142	3.1
	No	4398	96.9

(*Continued*)

Table 1 (Continued).

Variables	Classification	Number	(%)/$(\bar{x} \pm SD)$
Pregnancy	Yes	122	2.7
	No	4418	97.3
Premature rupture of membranes	Yes	12	0.3
	No	4528	99.7
Termination of Pregnancy	Yes	51	1.1
	No	4489	98.9
Chest Pain	Yes	174	3.8
	No	4366	96.2
Fever	Yes	1331	29.3
	No	3209	70.7
Cough	Yes	1120	24.7
	No	3420	75.3
Decreased white blood cells	Yes	504	11.1
	No	4036	88.9
NRS 2002 Score	/	/	1.20±1.263
Self-care score	/	/	92.26±17.856
Length of hospitalization	/	/	19.72±11.611

Abbreviation: SD, Standard Deviation.

Structuring EMR Results with ChatGLM

ChatGLM-6B[14] is a significant stride in the realm of open-source, bilingual (Chinese and English) conversational language models. Built upon the General Language Model (GLM) framework, it boasts an impressive 6.2 billion parameters. Leveraging model quantization technology, ChatGLM-6B can be locally deployed on consumer-grade graphics cards, requiring only 6GB of memory under INT4 quantization level.

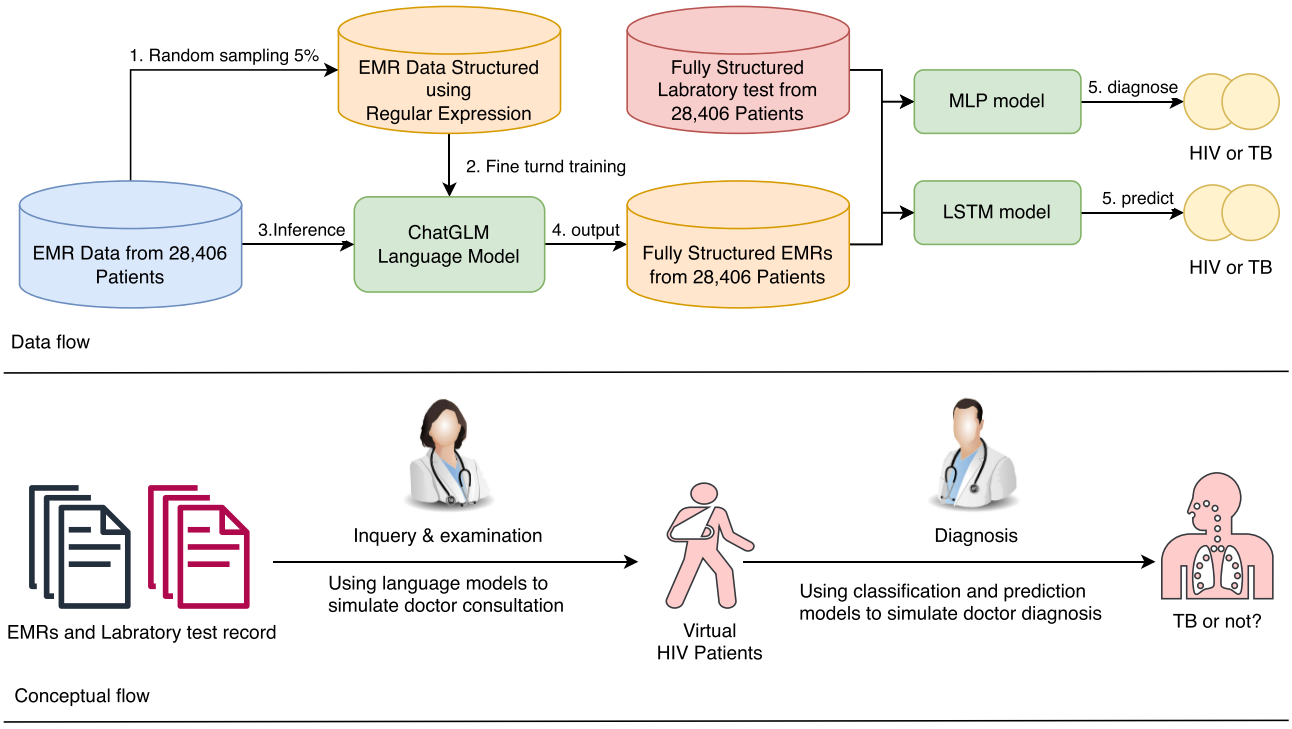

Figure 3 Workflow diagram of our AI pediatric diagnosis framework.

The unstructured nature of EHRs limit computational reuse.[22] We fine-tuned the ChatGLM language model on EHR notes to extract structured information. Our training data set comprises 4540 EMRs, which are randomly divided into 70% as training set, 20% as validation set and as 10% test set. The training labels are generated using self-developed regular expressions to structure part of the EHRs.

The disadvantage of regular expressions is that when there is new data or other description methods, regular expressions cannot accurately match the description content, thus marking errors. Therefore, more generalized models are often required to structure electronic case data. Here we chose the large language model ChatGLM for fine-tuning. Before fine-tuning, It can be seen that the initial ChatGLM didn't have the ability to structure EMRs. Fine-tuning significantly improved its extraction of clinical entities from free text. On the test set, it achieved high accuracy, precision, recall and F1 scores for common entities like bowel movements and mental status (Table 2).

In summary, fine-tuning improved ChatGLM's EHR structuring ability despite imperfect regular expression labels. Our approach enhances accessibility of unstructured EHR data. Further iterative training on expanded notes could improve generalization. Overall, large language models show promise in unlocking EHR data for clinical research and decision support.

Intelligent diagnosis of HIV and HIV/TB based on structured EMRs

Distinguishing HIV patients from those co-infected with TB using EHRs has the potential to significantly improve clinical decision making. We have investigated the performance of MLP models to accurately classify patients into those with HIV alone versus those with HIV/TB, using structured EHR inputs. EHR features including chief complaints, physical exam findings, lab tests and medications were extracted using a ChatGLM model fine-tuned on our EHR corpus. This transformed free text notes into structured inputs amenable for LSTM classification. We compiled a dataset of 4540

Table 2 Structuring Electronic Medical Records with ChatGLM

Symptoms	Accuracy	Precision	Recall	F1 Score
Stool-hematochezia	0.96352413	0.604651163	0.847826087	0.705882353
Stool-normal	0.952861953	0.910284464	0.997601918	0.95194508
Pharynx-congestion	0.982603816	0.64556962	0.944444444	0.766917293
Sane-clear	0.982603816	0.977755308	0.990778689	0.984223919
Abdomen-tenderness	0.988776655	0.529411765	0.818181818	0.642857143
Abdomen-soft	0.98372615	0.972098214	0.995428571	0.983625071
Anus	0.970819304	0.8	0.8	0.8
Abdomen-normal	0.98989899	0.902173913	0.902173913	0.902173913
Abdomen	0.976992144	0.434782609	0.588235294	0.5
Vulva-hyperemia	0.992143659	0.720930233	0.939393939	0.815789474
Anus-phyma	0.97979798	0.747368421	0.855421687	0.797752809
Pulmonary-infection	0.990460157	0.851351351	0.913043478	0.881118881
Pulmonary	0.976992144	0.768292683	0.940298507	0.845637584
Vulva	0.992143659	0.714285714	0.909090909	0.8
Body size-moderate	1	1	1	1
limbs-swollen	0.995510662	0.6875	0.785714286	0.733333333
Tonsil-swollen	0.997755331	0.870967742	1	0.931034483
Skin mucosa-normal	0.995510662	0.948905109	0.992366412	0.970149254
Skin mucosa	0.992704826	0.941605839	0.984732824	0.962686567
Intestinal-normal	0.994388328	0.967567568	0.978142077	0.972826087
Intestinal-intestinal gurgling sound normal	0.997194164	0.994318182	0.977653631	0.985915493
Intestinal	0.996071829	0.333333333	0.5	0.4
Kidney-pain	0.998877666	1	0.866666667	0.928571429
Pharyngeal	1	1	1	1
Average	0.987139918	0.805131385	0.896966465	0.84426834

Abbreviation: ChatGLM, Chat General Language Model.

de-identified patient encounters, with classes balanced via oversampling minority (HIV/TB) examples. 80% were randomly selected for LSTM training, with the rest for testing. Five-fold cross-validation was used to reduce variability. The MLP classifier achieved test set area under the receiver operating characteristic curve (AUROC) of 0.682–0.616 in predicting HIV/TB status based on EHR features (Figure 4).

In summary, MLP showed promising differentiation between HIV and HIV/TB patients given structured EHR data. Further refinements in feature selection and model optimization could improve generalizability. Our approach balancing prediction performance and scaling potential could help translate EHR data into actionable clinical insights.

Intelligent prediction of TB based on structured EMRs

Predicting future TB infection in HIV patients using longitudinal records has the potential to optimize screening and disease management. To address this issue, we developed LSTM models to forecast TB onset from structured EHRs. Per patient's EHRs were aggregated and sorted chronologically. Input sequences contained preceding visits, with the target label being TB status at the subsequent visit. We employed five-fold cross-validation to reduce model overfitting and evaluate model performance. The LSTM models achieved modest performance, with AUROCs of 0.503–0.688 for TB classification on held-out visits (Figure 5). Several factors likely contributed: (1) EHRs lacked sufficient quality and depth, and subjective descriptions varied across patients with the same diagnoses. (2) EHRs capture limited superficial data unlike more definitive lab tests.

However, our EHR structuring using a fine-tuned ChatGLM model enabled high-throughput feature extraction from free text. With higher-quality EHR data, the LSTM models could likely improve. To better determine if patient trajectories can predict impending TB onset, we trained LSTM models on structured laboratory tests results. This richer physiological data could better capture latent TB progression.

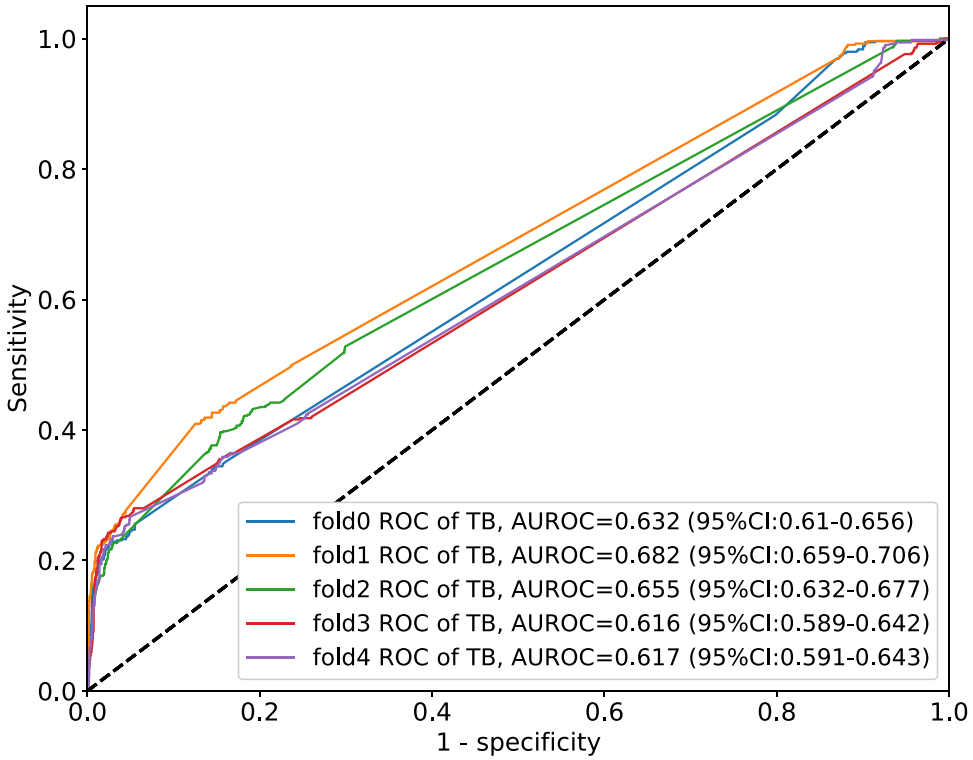

Figure 4 Five-fold cross-validation of the MLP predicting HIV-TB status based on EHR features.

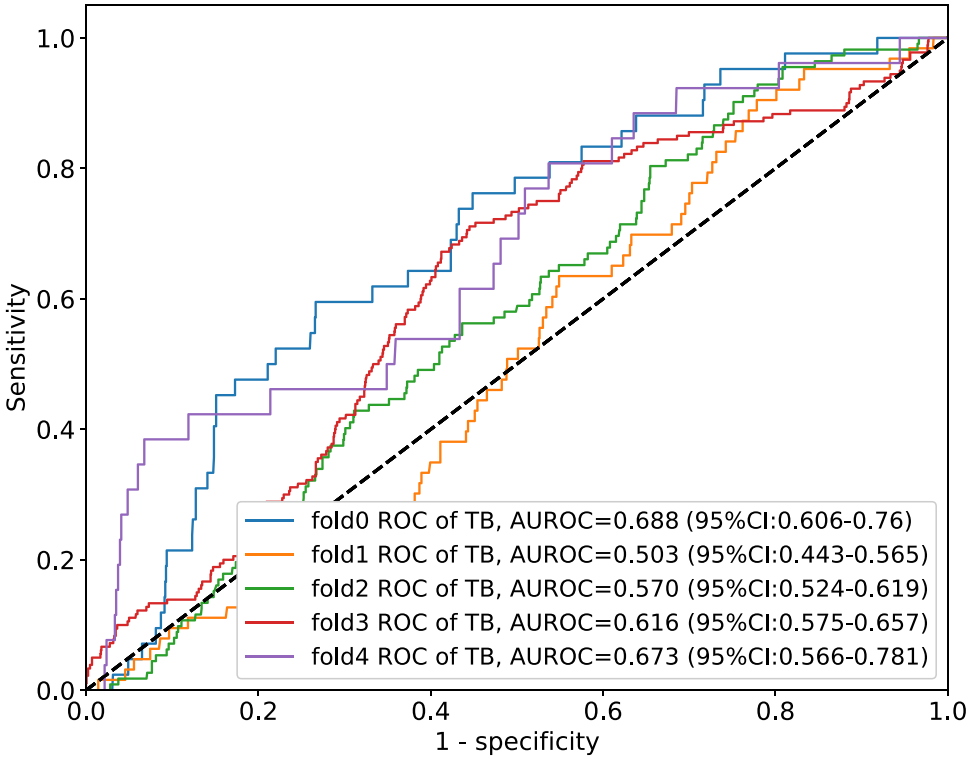

Figure 5 Five-fold cross-validation of the LSTM predicting TB based on EHR features.

Intelligent identification of HIV and TB based on inspection and inspection indicators

We have evaluated the performance of LSTM models in categorizing HIV/TB status using laboratory test features. Laboratory features including complete blood count, liver function tests, lipid profiles and others were extracted from our clinical data warehouse, without the erythrocyte sedimentation rate (ESR), interferon gamma release assays (IGRAs) and other TB-specific indicators. Values were z-score normalized before model input. We used the same 5-fold cross-validation approach as our prior EHR study, with an 80/20 train/test split. The LSTM classifier achieved an AUROC of 0.823–0.850 in predicting HIV/TB status based solely on lab tests (Figure 6). This performance surpassed its performance using EHR features.

The confusion matrix represented the performance of a classification model that differentiates between HIV and TB based on electronic medical record data. In this matrix, 2975 HIV cases were correctly identified, while 1051 were incorrectly classified as TB. For TB, 84 cases were correctly identified, and 14 were misclassified as HIV. The overall accuracy of the model is 74.18%. While the model is quite accurate in identifying HIV, as indicated by a high number of true positives, it showed lower precision in correctly classifying TB, suggesting areas for improvement in future model adjustments.

We evaluated LSTM models for forecasting TB onset in HIV patients using structured blood test data similarly to our EHR study. Longitudinal test results were sequenced for model input. Five-fold cross-validation yielded AUROCs of 0.869–0.644 for predicting future TB infection from earlier lab results (Figure 7). This significantly outperformed EHR-based models. The predictive models accurately screened individuals warranting closer follow-up and diagnostic workup. This could enable early case detection and treatment to improve outcomes and reduce transmission.[23]

The confusion matrix depicted the performance of a classification model distinguishing between HIV and TB from electronic medical record data. In this model, 6188 cases of HIV were correctly classified, while 1426 cases were incorrectly identified as TB. For TB, 513 cases were accurately classified, with 128 cases incorrectly identified as HIV. The overall accuracy of the model is noted as 81.18%. This matrix suggests a good performance in identifying HIV cases.

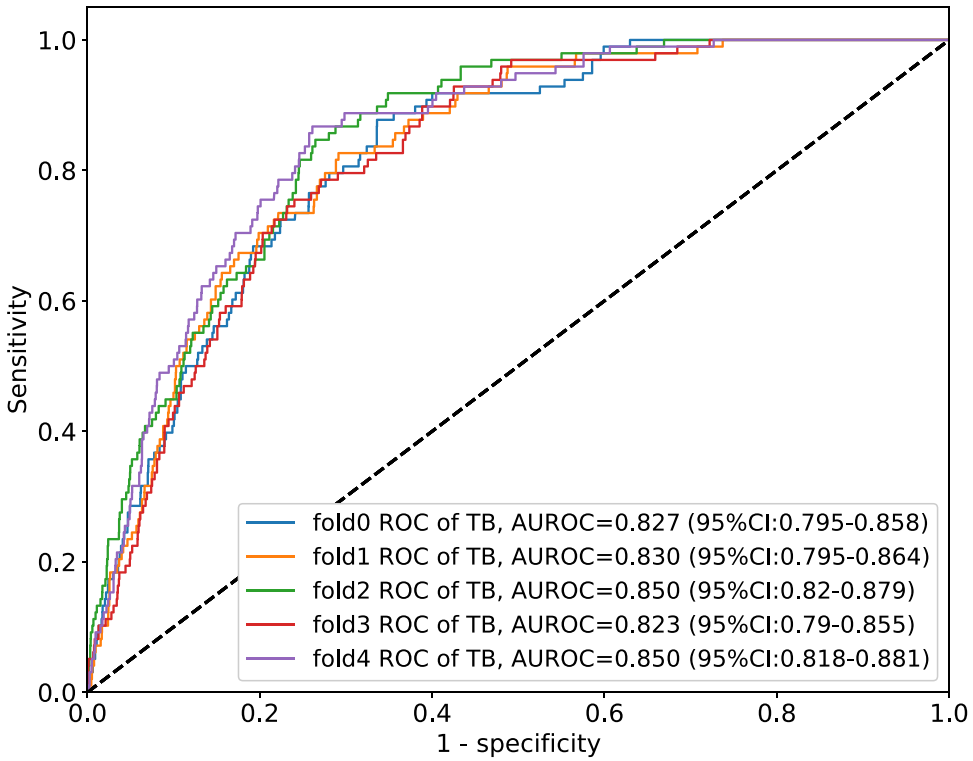

Figure 6 Five-fold cross-validation of the LSTM predicting TB based on lab tests.

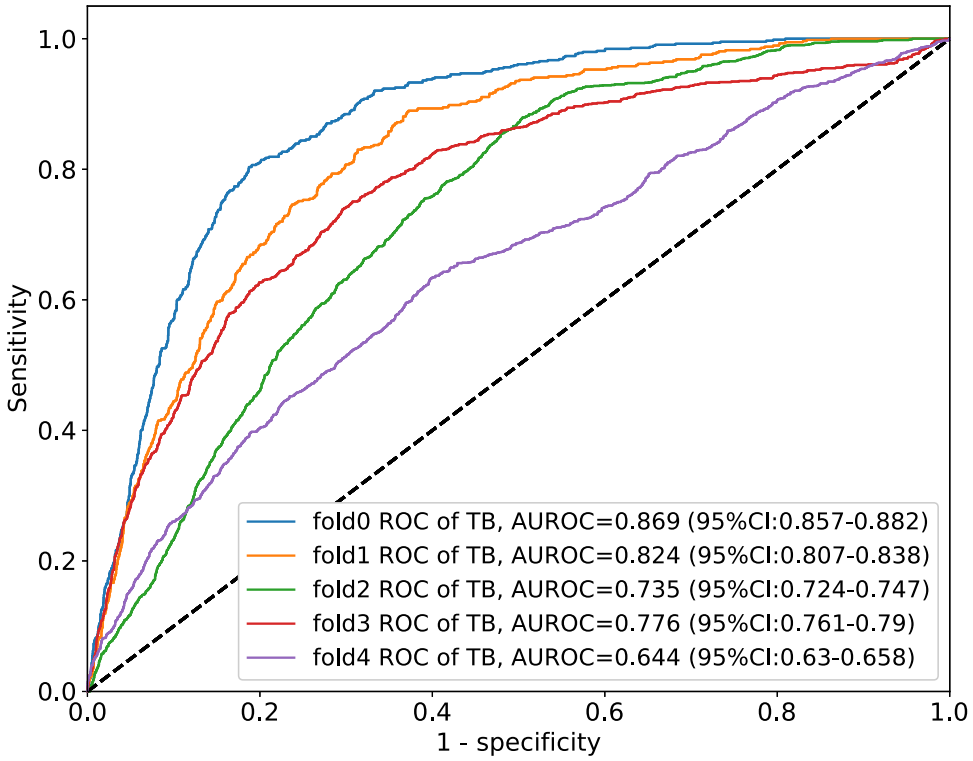

Figure 7 Five-fold cross-validation of the LSTM predicting TB based on structured blood test data.

Discussion

In this study, we analyzed the epidemiological characteristics of 4540 HIV patients admitted to the National Clinical Research Center for Infectious Diseases between 2017 and 2021. The findings revealed that a significant 16.7% of HIV patients were co-infected with TB. This statistic not only highlighted the prevalence of TB co-infection in the management of HIV but also underscored the urgent need for more effective prevention methods among patients with HIV. The study was conducted at the National Clinical Research Center for Infectious Diseases in Shenzhen, China, encompassing a broad population from the Pearl River Delta. This offered a wide-ranging perspective and a robust data foundation for our research. Additionally, Shenzhen, known for its substantial migrant population, boasts a diverse demographic structure representative of the entire country. This unique characteristic of population mobility makes Shenzhen particularly significant for studying the epidemiological features of infectious diseases. Collectively, these factors enhanced the representativeness and applicability of our findings.

In terms of patient demographics, the average age was 39.5 years, with males accounting for 85.4%, indicating a gender imbalance. This phenomenon aligned with global research findings, which show a higher rate of HIV infection among males, particularly among sexual minority groups. Regarding education levels, 34.5% of the patients had received college education or higher, challenging conventional assumptions about the socioeconomic status of HIV patients and showing that HIV crosses different educational backgrounds. Unmarried patients accounted for 57.1%, possibly related to the main transmission routes of HIV, especially in the high-risk sexual behavior. The presence of smoking and drinking behaviors, although not prevalent, still requires attention in clinical management, as these behaviors could exacerbate the health impact of HIV infection. Population mobility is another factor worth noting, with 93.0% of patients from other provinces, highlighting the importance of inter-regional cooperation in HIV epidemic monitoring and resource allocation.

In the field of EHR research, a major challenge is the unstructured nature of data, which significantly limits the reusability of EHR data for computational processing. This study enhanced the ability to extract structured information from EHR notes by meticulously fine-tuning the ChatGLM. After optimization, the model demonstrated significant improvements in identifying and extracting clinical entities from free text. Although regular expressions have limitations in data annotation, the finely tuned and optimized ChatGLM showed notable improvements in processing structured electronic health records. Furthermore, by iteratively training the model on larger datasets, its generalizability across different data types and application scenarios is expected to further improve.

We evaluated the performance of LSTM using laboratory test features for classifying HIV and HIV/TB co-infection. Notably, during the dataset construction, we deliberately excluded ESR, IGRAs, and other highly specific TB tests to explore the model's predictive performance without relying on these particular indicators. To ensure consistency and comparability of the data before inputting it into the model, all laboratory test values were standardized using z-score normalization. When relying solely on laboratory test data, the LSTM classifier achieved AUROC values between 0.823 and 0.850 in predicting HIV and HIV/TB co-infection, significantly outperforming the results obtained using only EHR data for TB prediction. Although we did not directly analyze the model weights to determine which input features were most crucial for the prediction results, we recognized the importance of such analysis for understanding the model's decision-making process and enhancing its applicability and interpretability in clinical settings. Future research could incorporate model interpretability techniques, such as SHAP value analysis or feature importance evaluation, to explore and validate which specific input data features are most critical for predicting the co-infection status of HIV and TB. This not only helped uncover the key biological and clinical factors behind the model's decisions but also provided clues for discovering new interactions between diseases. Moreover, such in-depth analysis can help identify and optimize the model's application to specific patient groups, thereby offering more personalized and precise support for clinical decision-making. Healthcare professionals can utilize the model established in this study to conduct personalized predictions for patients. Based on the results, they can provide health education and formulate targeted intervention measures, thereby enhancing patients' understanding of their conditions and improving treatment adherence.

We employed a method similar to EHR research to conduct an in-depth analysis of structured blood test data. We evaluated the performance of the LSTM model in predicting TB infection among HIV patients (AUROC: 0.869–0.644). This result significantly outperformed previous models based on EHR data. Utilizing this advanced

predictive model, we can accurately identify individuals who require closer follow-up and further diagnostic examinations. This not only facilitated early case detection but also significantly improved treatment outcomes through timely interventions, effectively reducing disease transmission. The research confirmed the efficacy of the LSTM model in predicting TB infection among HIV patients and highlighted the immense potential of integrating structured blood test data with machine learning techniques to enhance early disease diagnosis and intervention. Additionally, we also recognized that enhancing the model's interpretability in clinical applications is crucial for promoting broader adoption of these technologies in real-world clinical settings. Therefore, in future work, we plan to extensively utilize visualization and explanation tools such as T-SNE, LIME, and SHAP to clearly elucidate the basis of the model's decisions, thereby increasing transparency and trust in clinical environments.

However, data quality and accessibility are critical challenges in clinical practice. High-quality, complete medical records are essential for model training and prediction, but in clinical practice, data often suffer from missing values, inconsistencies, and noise, which can impact model accuracy and reliability. Therefore, establishing standardized data collection and processing procedures is necessary to ensure data quality and consistency. Additionally, it is crucial to strictly adhere to relevant regulations and ethical standards to ensure data privacy and security. Hospitals and healthcare institutions need the appropriate technical infrastructure to deploy and maintain these complex prediction models, including high-performance computing resources, data storage and management systems, and professional technical teams.

This study had several limitations. Prior research have indicated that TB/HIV co-infection may be linked to various factors, such as the route of HIV transmission, a history of TB exposure, CD4+ T cell counts, and the use of isoniazid preventive therapy.[24–26] However, due to the constraints of retrospective data, some factors were not included in a epidemiological analysis. Given the model's training on HIV-positive patients, challenges in feature generalization may arise when extrapolating to HIV-negative populations. Clinical predictions specifically tailored for high-risk groups or for diagnosing symptomatic individuals seeking medical care may not be as effective in identifying primarily subclinical tuberculosis cases in largely healthy individuals who are not actively seeking medical attention within a community setting.

Conclusion

This study underscored the necessity of integrating deep learning techniques with electronic health data, showcasing the immense potential of artificial intelligence in the public health. The model based on laboratory time-series data significantly outperformed those relying solely on electronic health records in predicting tuberculosis incidence. This finding not only highlighted the advantages of deep learning in handling complex medical data but also provided valuable insights for healthcare providers on exploring the application of deep learning in disease prediction and management. Combining deep learning techniques with electronic health records can substantially improve the accuracy of disease diagnosis and the personalization of treatment, bringing revolutionary breakthroughs to the field of public health. In the future, the study will progress by broadening the research parameters and conducting a multicenter prospective cohort study to identify and validate additional influencing factors, thereby enhancing the model's performance.

Data Sharing Statement

The datasets generated and/or analysed during the current study are not publicly available due [The Data Security Law of the People's Republic of China] but are available from the corresponding author on reasonable request.

Ethics Approval and Consent to Participate

The study protocol received was approved by the Ethics Review Committee of a infectious disease hospital from Shenzhen (protocol no: [2022-027]). Due to the study design, which did not require direct patient participation, the formal process of obtaining informed consent was not applicable. Furthermore, all research methodologies rigorously complied with relevant ethical standards and legal requirements.

Consent for Publication

All authors have read the journal policies and submit this manuscript in accordance with those policies. All authors consent for publication.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This work was supported by Shenzhen High-level Hospital Construction Fund (G2022006). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Nunn P, Williams B, Floyd K, et al. Tuberculosis control in the era of HIV. *Nat Rev Immunol*. 2005;5(10):819–826. doi:10.1038/nri1704
2. Qi -C-C, Xu L-R, Zhao C-J, et al. Prevalence and risk factors of tuberculosis among people living with HIV/AIDS in China: a systematic review and meta-analysis. *BMC Infect Dis*. 2023;23(1):584. doi:10.1186/s12879-023-08575-4
3. Straetemans M, Bierrenbach AL, Nagelkerke N, Glaziou P, van der Werf MJ. The effect of tuberculosis on mortality in HIV positive people: a meta-analysis. *PLoS One*. 2010;5(12):e15241. doi:10.1371/journal.pone.0015241
4. *WHO*. *Global Tuberculosis Report 2023*. Geneva; 2023.
5. Xiao J, Du S, Tian Y, et al. Causes of Death Among Patients Infected with HIV at a Tertiary Care Hospital in China: An Observational Cohort Study. *AIDS Res Hum Retroviruses*. 2016;32(8):782–790. doi:10.1089/aid.2015.0271
6. Bisson GP, Zetola N, Collman RG. Persistent high mortality in advanced HIV/TB despite appropriate antiretroviral and antitubercular therapy: an emerging challenge. *Curr HIV/AIDS Rep*. 2015;12(1):107–116. doi:10.1007/s11904-015-0256-x
7. Gupta RK, Lucas SB, Fielding KL, Lawn SD. Prevalence of tuberculosis in post-mortem studies of HIV-infected adults and children in resource-limited settings: a systematic review and meta-analysis. *AIDS*. 2015;29(15):1987–2002. doi:10.1097/QAD.0000000000000802
8. Auld AF, Kerkhoff AD, Hanifa Y, et al. Derivation and external validation of a risk score for predicting HIV-associated tuberculosis to support case finding and preventive therapy scale-up: a cohort study. *PLoS Med*. 2021;18(9):e1003739. doi:10.1371/journal.pmed.1003739
9. Hanifa Y, Fielding KL, Chihota VN, et al. A clinical scoring system to prioritise investigation for tuberculosis among adults attending HIV clinics in South Africa. *PLoS One*. 2017;12(8):e0181519. doi:10.1371/journal.pone.0181519
10. Balcha TT, Skogmar S, Sturegård E, et al. A Clinical scoring algorithm for determination of the risk of tuberculosis in hiv-infected adults: a cohort study performed at Ethiopian health centers. *Open Forum Infect Dis*. 2014;1(3):ofu095. doi:10.1093/ofid/ofu095
11. Aunsborg JW, Hønge BL, Jespersen S, et al. A clinical score has utility in tuberculosis case-finding among patients with HIV: a feasibility study from Bissau. *Int J Infect Dis*. 2020;92S:S78–S84. doi:10.1016/j.ijid.2020.03.012
12. Boyles TH, Nduna M, Pitsi T, et al. A clinical prediction score including trial of antibiotics and c-reactive protein to improve the diagnosis of tuberculosis in ambulatory people with HIV. *Open Forum Infect Dis*. 2020;7(2):ofz543. doi:10.1093/ofid/ofz543
13. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008;17:128–144.
14. Zeng A, Liu X, Du Z,et al. 'Glm-130b: an open bilingual pre-trained model.' *arXiv preprint arXiv*. 2022.
15. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234–1240. doi:10.1093/bioinformatics/btz682
16. Graves A, Graves A 'Long short-term memory.' *Supervised sequence labelling with recurrent neural networks*: (2012) 37–45.
17. Greff K, Srivastava RK, Koutnik J, Steunebrink BR, Schmidhuber J. LSTM: A Search Space Odyssey. *IEEE Trans Neural Netw Learn Syst*. 2017;28(10):2222–2232. doi:10.1109/TNNLS.2016.2582924
18. Malhotra P, Ramakrishnan A, Anand G, Vig L, Agarwal P, Shroff G.'LSTM-based encoder-decoder for multi-sensor anomaly detection.' *arXiv preprint arXiv:1607.00148*; (2016).
19. Lipton ZC, Kale DC, Elkan C, Wetzel R. 'Learning to diagnose with LSTM recurrent neural networks.' *arXiv preprint arXiv:1511.03677*: (2015).
20. He N, Detels R. The HIV epidemic in China: History, response, and challenge. *Cell Res*. 2005;15(11–12):825–832. doi:10.1038/sj.cr.7290354
21. Haber N, Tanser F, Bor J, et al. From HIV infection to therapeutic response: a population-based longitudinal HIV cascade-of-care study in KwaZulu-Natal, South Africa. *Lancet HIV*. 2017;4(5):e223–e230. doi:10.1016/S2352-3018(16)30224-7
22. Sheikhalishahi S, Miotto R, Dudley JT, et al. Natural Language Processing of Clinical Notes on Chronic Diseases: systematic Review. *JMIR Med Inform*. 2019;7(2):e12239. doi:10.2196/12239
23. Eang MT, Satha P, Yadav RP, et al. Early detection of tuberculosis through community-based active case finding in Cambodia. *BMC Public Health*. 2012;12(1):469. doi:10.1186/1471-2458-12-469

24. Sharan R, Bucşan AN, Ganatra S, et al. Chronic Immune Activation in TB/HIV Co-infection. *Trends Microbiol*. 2020;28(8):619–632. doi:10.1016/j.tim.2020.03.015
25. Alemu A, Bitew ZW, Yesuf A, Zerihun B, Getu M. The Effect of Long-Term HAART on the Incidence of Tuberculosis Among People Living with HIV in Addis Ababa, Ethiopia: A Matched Nested Case-Control Study. *Infect Drug Resist*. 2021;14:5189–5198. doi:10.2147/IDR.S345080
26. Wondmeneh TG, Mekonnen AT. The incidence rate of tuberculosis and its associated factors among HIV-positive persons in Sub-Saharan Africa: a systematic review and meta-analysis. *BMC Infect Dis*. 2023;23(1):613. doi:10.1186/s12879-023-08533-0