ESC
European Society
of Cardiology

European Heart Journal - Digital Health (2024) **5**, 712–724
https://doi.org/10.1093/ehjdh/ztae066

ORIGINAL ARTICLE
Artificial intelligence (machine learning, deep learning)

Enhancing the interoperability and transparency of real-world data extraction in clinical research: evaluating the feasibility and impact of a ChatGLM implementation in Chinese hospital settings

Bin Wang [1,†], Junkai Lai[2,3,†], Han Cao[4,†], Feifei Jin[5,6,7,†], Qiang Li[8], Mingkun Tang[4], Chen Yao[9,10,*], and Ping Zhang[11,*]

[1]School of Clinical Medicine, Tsinghua University, No. 30 Shuangqing Road, Haidian District, Beijing 100084, China; [2]Institute of Automation, Chinese Academy of Sciences, No. 95 Zhongguancun Road, Haidian District, Beijing 100080, China; [3]Hangzhou LionMed Medical Information Technology Co., Ltd, No.19 Jugong Road, Xixing Sub-District, Hangzhou 310000, China; [4]Medical Data Science Center, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, No. 168 Litang Road, Changping District, Beijing 102218, China; [5]Trauma Medicine Center, Peking University People's Hospital, No. 11 Xizhimen South Street, Xicheng District, Beijing 100044, China; [6]Key Laboratory of Trauma Treatment and Neural Regeneration, Peking University, Ministry of Education, No. 11 Xizhimen South Street, Xicheng District, Beijing 100044, China; [7]National Center for Trauma Medicine of China, No. 11 Xizhimen South Street, Xicheng District, Beijing 100044, China; [8]Department of Information Administration, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, No. 168 Litang Road, Changping District, Beijing 102218, China; [9]Peking University Clinical Research Institute, Peking University First Hospital, No. 8 Xishiku Street, Xicheng District, Beijing 100034, China; [10]Hainan Institute of Real-World Data, No. 32 Kangxiang Road, Qionghai 571437, China; and [11]Department of Cardiology, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, No. 168 Litang Road, Changping District, Beijing 102218, China

Received 14 May 2024; revised 4 July 2024; accepted 19 August 2024; online publish-ahead-of-print 12 September 2024

Aims

This study aims to assess the feasibility and impact of the implementation of the ChatGLM for real-world data (RWD) extraction in hospital settings. The primary focus of this research is on the effectiveness of ChatGLM-driven data extraction compared with that of manual processes associated with the electronic source data repository (ESDR) system.

Methods and results

The researchers developed the ESDR system, which integrates ChatGLM, electronic case report forms (eCRFs), and electronic health records. The LLaMA (Large Language Model Meta AI) model was also deployed to compare the extraction accuracy of ChatGLM in free-text forms. A single-centre retrospective cohort study served as a pilot case. Five eCRF forms of 63 subjects, including free-text forms and discharge medication, were evaluated. Data collection involved electronic medical and prescription records collected from 13 departments. The ChatGLM-assisted process was associated with an estimated efficiency improvement of 80.7% in the eCRF data transcription time. The initial manual input accuracy for free-text forms was 99.59%, the ChatGLM data extraction accuracy was 77.13%, and the LLaMA data extraction accuracy was 43.86%. The challenges associated with the use of ChatGLM focus on prompt design, prompt output consistency, prompt output verification, and integration with hospital information systems.

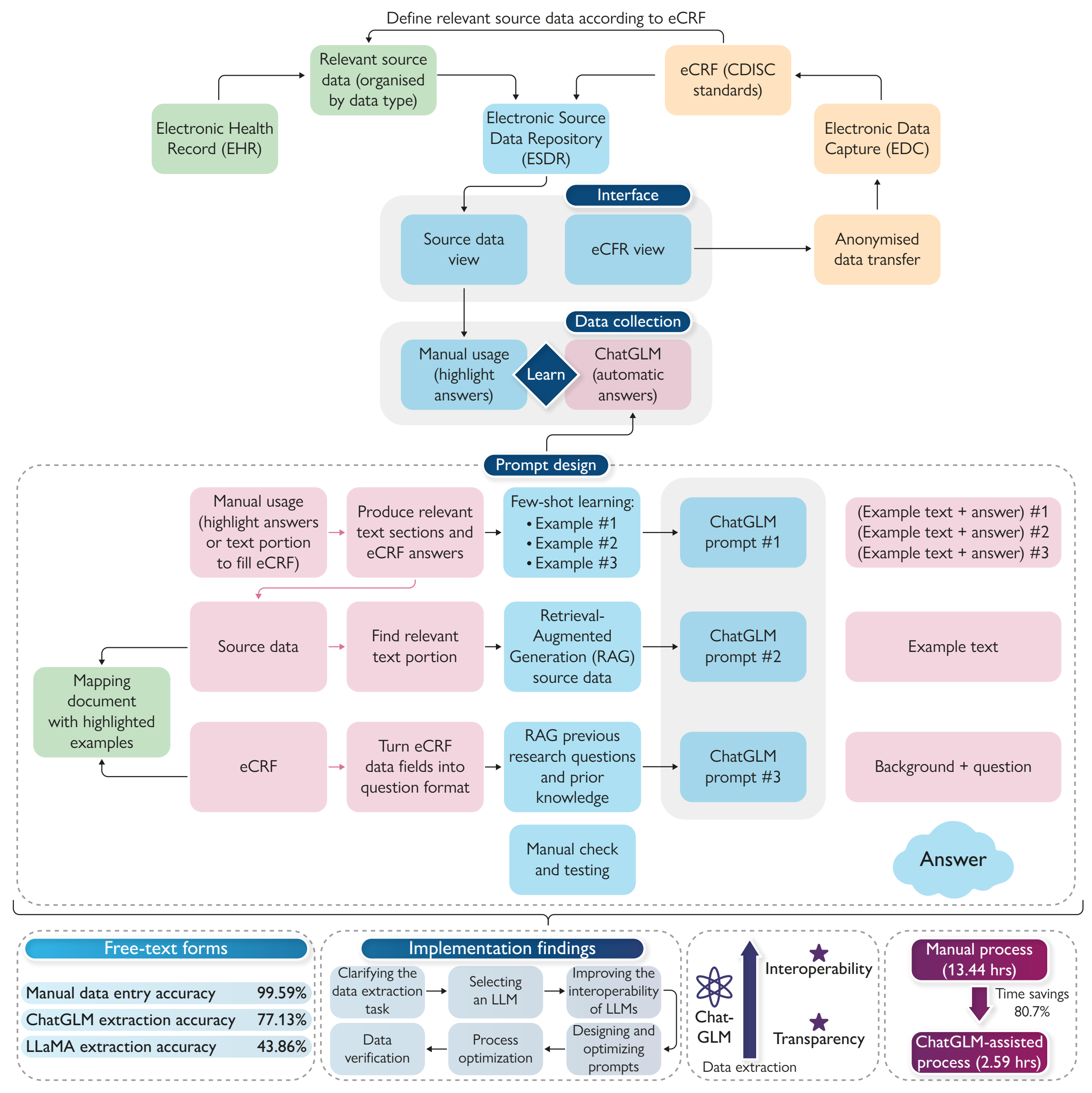
Conclusion

The main contribution of this study is to validate the use of ESDR tools to address the interoperability and transparency challenges of using ChatGLM for RWD extraction in Chinese hospital settings.

* Corresponding author. Tel: +86 01066551053, Email: yaochen@hsc.pku.edu.cn (C.Y.); Tel: +86 01056118899, Email: zhpdoc@126.com (P.Z.)

[†]These authors contributed equally to this work.

Graphical Abstract


Define relevant source data according to eCRF
Relevant source data (organised by data type)
eCRF (CDISC standards)
Electronic Health Record (EHR)
Electronic Source Data Repository (ESDR)
Electronic Data Capture (EDC)
Interface
Source data view
eCFR view
Anonymised data transfer
Data collection
Manual usage (highlight answers)
Learn
ChatGLM (automatic answers)
Prompt design
Manual usage (highlight answers or text portion to fill eCRF)
Produce relevant text sections and eCRF answers
Few-shot learning: • Example #1 • Example #2 • Example #3
ChatGLM prompt #1
(Example text + answer) #1 (Example text + answer) #2 (Example text + answer) #3
Source data
Find relevant text portion
Retrieval-Augmented Generation (RAG) source data
ChatGLM prompt #2
Example text
Mapping document with highlighted examples
eCRF
Turn eCRF data fields into question format
RAG previous research questions and prior knowledge
ChatGLM prompt #3
Background + question
Manual check and testing
Answer
Free-text forms
Manual data entry accuracy 99.59%
ChatGLM extraction accuracy 77.13%
LLaMA extraction accuracy 43.86%
Implementation findings
Clarifying the data extraction task
Selecting an LLM
Improving the interoperability of LLMs
Data verification
Process optimization
Designing and optimizing prompts
Chat-GLM
Interoperability
Transparency
Data extraction
Manual process (13.44 hrs)
Time savings 80.7%
ChatGLM-assisted process (2.59 hrs)

Keywords Challenge • Data extraction • Electronic health records • Interoperability • Large language models • Prompt

Introduction

The heightened interest in the application of artificial intelligence (AI) and large language models (LLMs) in the medical sector is the result of the potential of such models to enhance various facets of healthcare.[1] In the context of clinical research, prior methodologies in natural language processing (NLP) have predominantly concentrated on named entity recognition, such as by employing notable models such as bidirectional encoder representations from transformers for biomedical text mining (BioBERT), to identify entities pertinent to clinical research. Models such as BioBERT[2] represent domain-specific language models that are initially pretrained on extensive biomedical corpora. However, this methodology requires additional rule-based transformations to generate responses to specific clinical queries. The development of LLMs specifically designed for biomedical and clinical text mining has further enhanced the capabilities of NLP in this domain.[3] In contrast, contemporary LLMs exhibit a question–answer generative pretrained transformer (GPT) structure with supervised fine-tuning. At present, LLMs based on the GPT structure employ dialog for both input and output, in which context the input dialog consists of questions and the output dialog comprises direct answers. This approach eliminates the need for laborious rule-based transformations, thereby streamlining the process by directly providing responses that are more general and adaptable than the named entities previously used. Additionally, large clinical language models such as GatorTron,[4] which was trained on extensive clinical text, have been associated with promising results with respect to a variety of clinical NLP tasks, including extracting clinical concepts and answering medical questions.[5,6] These advancements highlight the opportunities for extracting data from medical-free text that LLMs offer.

Previous studies have extensively demonstrated the use of LLMs for information extraction tasks such as named entity recognition[7] and relationship extraction.[8] Integrating LLMs into hospital systems necessitates addressing data privacy, security, interoperability, data mapping, standardization, quality, and scalability issues to comply with healthcare regulations and meet clinical research requirements.[9–11] Healthcare systems are complex, with a variety of data sources and formats, as well as stringent privacy and security requirements.[12] The lack of interoperability with electronic health record (EHR) systems prevents LLMs from accessing raw medical documents directly. Similarly, the absence of interoperability with electronic data capture (EDC) systems means that the data extraction results from LLMs cannot be standardized within the structure of electronic case report forms (eCRFs), rendering them unsuitable for direct data analysis. Most importantly, the black-box nature of LLMs makes it impossible to verify the reliability of the data extraction process. These challenges significantly limit the feasibility of using LLMs for real-world data (RWD) extraction in actual healthcare settings.

The burden of inconsistencies, missing or incomplete observations, and the presence of noise and outliers in EHR data render them unsuitable for research purposes.[13] The use of EHRs for clinical research is associated with historical progress and current applications, and relevant efforts have focused on using the most recent standards and technologies to facilitate data transfer from EHR systems into clinical research databases, thereby improving data quality.[14] The primary challenge in translating data between EHR systems and clinical research databases, such as EDC systems, is the unstructured nature of EHR source data and a lack of methods for improving data interoperability between these systems. First, the majority of clinically relevant source data are documented as narrative text in EHR systems, which can be inconsistent and noisy due to time constraints and physician documentation practices. Second, in China, the standard for source data usage is limited primarily to the International Classification of Diseases codes for diagnoses and procedures, making it impossible to extract EHR source data directly from these coding systems within EDC systems. To improve interoperability between these systems, we created a framework that uses a digital product designed to connect EHR and EDC systems during the data collection process. This framework aims to increase transparency and interoperability. Within this framework, we set up ChatGLM to learn from the data collection process, allowing them to respond accurately to research queries on the basis of source data. The digital product is called the electronic source data repository (ESDR).[15–18] It can integrate the source data required for clinical studies and facilitate the electronic transfer of study data from the EHR system to the EDC system.

The primary objective of this study is to assess the feasibility and impact of the implementation of ChatGLM for RWD extraction within a Chinese hospital setting. Specifically, this research aims to evaluate the effectiveness of ChatGLM-driven data extraction and traceability functions compared with the manual processes associated with the ESDR. Additionally, the investigation seeks to identify and analyse the challenges encountered during the implementation of ChatGLM with the goal of obtaining insights from practical experiences in the field.

Methods

System design

The ESDR interface, as detailed in prior research,[15] integrates eCRFs and EHRs to promote enhanced traceability and eCRF field highlighting.[15] For further information, please refer to previous studies.[16–18] *Figure 1* shows the ChatGLM workflow description and the composition of the ESDR system. The clinical research data extraction pipeline is initialized when research data requirements are sent from the EDC system via the Clinical Data Interchange Standards Consortium (CDISC) standards to the ESDR. On the basis of source data samples sent from the EHR, ESDR will bind case report data fields to the most relevant type of data source for later usage in the ESDR interface. The interface will visualize eCRFs on the left alongside relevant patient source data on the right. Data collection is initialized through the manual entry method, where researchers can highlight relevant source data to be used as answers for eCRF data fields. The manual process is recorded as examples and used to optimize ChatGLM prompts via the few-shot learning method. For new data, the ChatGLM first finds relevant text sections from the source data and finds related content used to localize medical terminology and understanding via the retrieval-augmented generation (RAG) method and combines previous examples to increase inference ability. Supplementary material online, *file S1* includes several examples of prompts.

Pilot case selection

The pilot cases were selected on the basis of investigator-initiated clinical research projects that had been ethically approved in the cardiovascular medicine departments that participated in the pilot collaboration. These research projects required a formal clinical study protocol as well as an eCRF. To make the pilot more manageable, only ongoing retrospective studies were considered. The Paroxysmal Atrial Tachycardia Project, a single-centre retrospective cohort study, explores the correlations among paroxysmal atrial tachycardia, thromboembolic events, and atrial fibrillation. Baseline data, including sociodemographic information, medical history, medication history, laboratory results, electrocardiogram data, and other types of information, were drawn from the EHR system. The eCRF field that addresses data extraction from free-text sources was given the highest evaluation in this investigation (see Supplementary material online, *Table S1* and *file S2*). In addition, ChatGLM's capacity to identify drug terms as standardized terms was evaluated via structured discharge medication.

Ethical approval

This study was conducted in accordance with the principles of the Declaration of Helsinki and received approval from the Beijing Tsinghua Changgung Hospital Institutional Review Board (number 21440-4-03). The anonymization of patient data adhered to data safety standards.

Implementation process

Beginning on 1 August 2023, researchers deployed the ESDR system on the local area network of a hospital, which interfaced with the hospital information system (HIS) to obtain certified copies of patient data. LLM localization and deployment were performed by utilizing the Chinese open-source ChatGLM2-6B model.[19] The researchers designed questions for each eCRF variable, employing preset questions as prompts to encourage the ChatGLM to extract answers from free-text data. Through a 2-week optimization process involving three rounds of fine-tuning, technicians enhanced the ChatGLM instructions to ensure optimal data extraction. The LLaMA2-7B (Large Language Model Meta AI) model[20] was also deployed to compare and reference the accuracy of the ChatGLM extraction process in free text across different eCRF fields. The parameters of ChatGLM2-6B include the temperature = 0.3, top_p = 0.8, top_k = 50. The parameters of LLaMA2-7B consist of the temperature = 0.4, top_p = 0.8, top_k = 50. The web chat interface of HuatuoGPT-II[21] was used to conduct and demonstrate prompt query tests, and its performance was compared to that of ChatGLM2-6B.

Research design

The primary objective of this research was to evaluate the efficacy of the ChatGLM-driven or LLaMA-driven data extraction and traceability functions (the AI-assisted process) of ESDR software compared with those of the manual data collection and verification methods associated with the ESDR system (the ESDR manual process). The study focused on differences in the accuracy rates and time allocation associated with these approaches. Five eCRF forms (see Supplementary material online, *Table S1* and *file S2*),

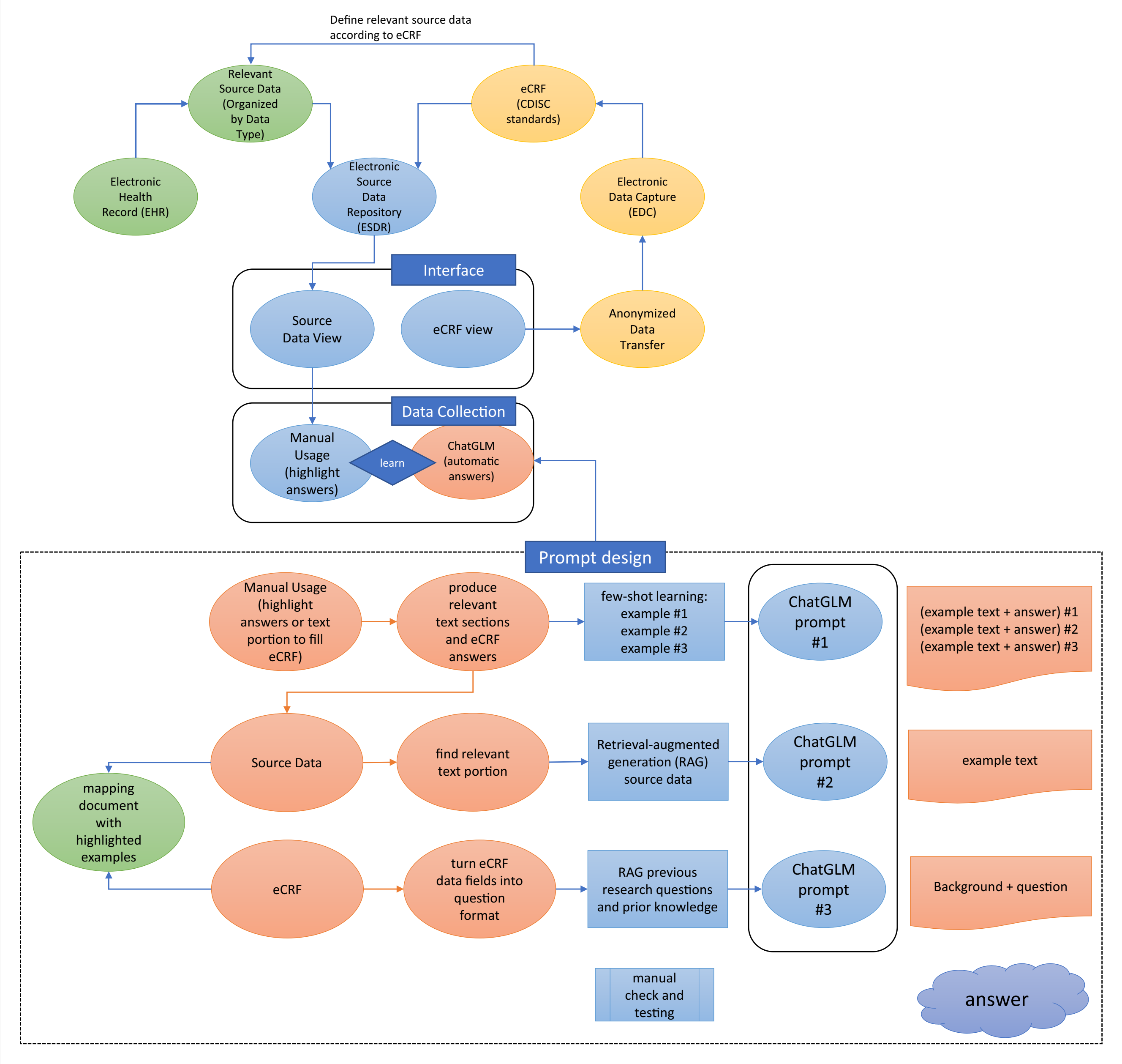

Define relevant source data according to eCRF
Relevant Source Data (Organized by Data Type)
eCRF (CDISC standards)
Electronic Health Record (EHR)
Electronic Source Data Repository (ESDR)
Electronic Data Capture (EDC)
Interface
Source Data View
eCRF view
Anonymized Data Transfer
Data Collection
Manual Usage (highlight answers)
learn
ChatGLM (automatic answers)
Prompt design
Manual Usage (highlight answers or text portion to fill eCRF)
produce relevant text sections and eCRF answers
few-shot learning: example #1 example #2 example #3
ChatGLM prompt #1
(example text + answer) #1
(example text + answer) #2
(example text + answer) #3
mapping document with highlighted examples
Source Data
find relevant text portion
Retrieval-augmented generation (RAG) source data
ChatGLM prompt #2
example text
eCRF
turn eCRF data fields into question format
RAG previous research questions and prior knowledge
ChatGLM prompt #3
Background + question
manual check and testing
answer

Figure 1 Composition of the electronic source data repository system and description of the prompt design. eCRF, electronic case report form; EDC, electronic data capture; EHRs, electronic health records; ESDR, electronic source data repository; RAG, retrieval-augmented generation; CDISC, Clinical Data Interchange Standards Consortium.

predominantly comprising free-text content, were evaluated; the relevant data focused on 630 subjects.

During the implementation phase, technicians utilized ChatGLM or LLaMA for batch processing to extract all patient data. To assess the extraction efficacy of ChatGLM or LLaMA, this study subsequently employed traditional manual methods for secondary data extraction. Given ChatGLM's considerable advantage over traditional manual processes, employing the latter for extracting data from all patients appears unnecessary. Instead, a subset comprising 10% of the samples (63 subjects) was selected for comparative assessments of extraction effectiveness.

Conceptually, three processes were considered. The traditional manual process refers to the method commonly employed by clinical researchers, the ESDR manual process that simplifies manual entry, and the AI-assisted process that automates data extraction via ChatGLM or LLaMA:

(A) Traditional manual process:
   (a) The EDC and EHR platforms were opened separately.
   (b) The clinical research coordinator reviewed EDCs for data fields and subsequently examined EHRs for relevant text, manually completing the EDC.

(B) ESDR manual process:
   (a) Manual Data Entry: Participants utilized ESDR software to input patient admission records manually into the ESDR eCRF form and filled in the eCRF forms directly.
   (b) Manual Verification: In reference to the ESDR records, the participants manually traced and verified the eCRF form, correcting any input errors. This process allowed for the simultaneous viewing of relevant EHRs and EDCs on a single platform, thus facilitating data comparison from left to right.

(C) AI-assisted process:

(a) AI Data Entry (Batch Processing): Batch processing codes for AI data extraction were configured, and AI data extraction was executed for all 63 subjects; the total runtime was recorded. This process filled in relevant eCRF fields directly.

(b) AI-Assisted Data Verification (Traceability Function): Researchers validated the accuracy of fields filled in by the ChatGLM or LLaMA within the ESDR software, manually correcting incorrectly entered fields via the AI-assisted source data location feature to facilitate swift tracing. This feature highlights the relevant medical text used to populate eCRF forms.

The traditional manual process (Process A) involves manual extraction and data entry from separate EDC and EHR platforms and thus represents a labour-intensive and error-prone procedure that contributes to inefficiencies. Once a standard practice, this method has become a bottleneck in research efficiency because of its resource-intensive requirements. Therefore, given the substantial time investment required for the implementation of the traditional manual process, we opted not to assess this method.

To mitigate bias in the assessment across various researchers, a clinician proficient in ESDR software employed two procedures to evaluate the sampled patient data. The participants utilized both the ESDR manual process and the AI-assisted process in two workflows to ensure a comprehensive assessment.

Data collection

The accuracy of eCRF data transcription refers to an assessment of whether the values entered into the eCRF are in line with the corresponding source data from the EHRs. After eliminating errors in the source data, if consistency is confirmed, the eCRF question is considered to have been completed correctly. Data completeness is defined by the presence of the required eCRF data within the EHR. If the response 'not mentioned' or 'unknown' is recorded in the verified eCRF field, incompleteness is indicated. To calculate the completeness rate of eCRF fields, the eCRF data are exported from the ESDR. ESDR employs audit trials to log all user actions related to eCRF completion and modifications, including users and timestamps, which can be retrieved to measure the time spent on eCRF completion.

Data analysis

Data analysis was guided by descriptive statistics; Python software (version 3.11.5) was employed, and plotting methods were selected on the basis of the characteristics of the data distribution.

Results

Distribution of data sources

Data were drawn from the EHRs and prescription records of 13 departments (*Figure 2*). The top three departments were cardiology (36.5%, 23), neurology (22.2%, 14), and the cardiac intensive care unit (19.0%, 12).

Data completeness

The discharge medication form achieved 100% data completeness for 123 eCRF fields. However, for free-text forms, which accounted for 27 eCRF fields, six fields presented data completeness rates below 20%; all of these fields were from the health status form (*Figure 3*).

Electronic case report forms data transcription time

For 63 subjects with 9450 eCRF fields, the ESDR manual process required ∼48 382 s (∼13.44 h) to complete, with an average of 5.12 s per field (*Table 1*). In contrast, the ChatGLM-assisted process required ∼22 126 s (∼6.15 h), with an average of 2.34 s per field (*Table 2*). Considering batch processing time, the actual amount of human resource time invested was only 9337 s (∼2.59 h), with an average of 0.99 s per field. The ChatGLM-assisted process was associated with an estimated efficiency improvement of 80.7%, thus indicating a significant reduction in human labour time. In all fields, including free-text forms and discharge medication, ChatGLM-assisted total time savings were highest (87.75%) in the respiratory medicine department and lowest (71.31%) in the internal medicine department (see Supplementary material online, *Figure S1*).

Electronic case report forms data transcription quality

For the ESDR manual process, the overall accuracy of the initial manual entry was 99.08%. In the ChatGLM-assisted process, the overall data extraction accuracy rate was 94.84%. The initial manual input accuracy for free-text forms was 99.59%, the ChatGLM data extraction accuracy was 77.13%, and the LLaMA data extraction accuracy was 43.86% (*Table 3*). *Figure 4* shows the mean number of characters in the admission notes for patients across departments. *Figures 5* and *6* illustrate the accuracy of free-text form extraction for the two models by field type and department category, respectively. The mean number of characters in admission notes for six common internal medicine departments (Respiratory Medicine, Cardiac Intensive Care Unit, Cardiology, Neurology, Internal Medicine, and Cardiac Surgery) was >1800 (*Figure 4*). However, the extraction accuracy for free-text forms was <78% for ChatGLM and <47% for LLaMA in the six departments listed above (*Figure 6*). The extraction accuracy of LLaMA was superior to that of ChatGLM in the hypertension, diabetes, and coronary heart disease fields. However, in terms of vital signs and family history, LLaMA performed worse than ChatGLM did (*Figure 5*).

Errors in manually entered fields primarily included numerical entry errors (such as inaccuracies in recording the number of cigarettes smoked, diastolic blood pressure, and frequency of alcohol consumption) and instances of missing responses (resulting from the failure to click on single-choice questions). The ChatGLM data extraction exhibited exceptionally high accuracy with respect to capturing information related to other family history, frequency of alcohol consumption, and various vital signs. However, some fields, particularly those related to specific diseases such as coronary heart disease, diabetes, and hypertension, were associated with lower accuracy rates, thus suggesting potential areas for improvement with respect to understanding localized Chinese clinical terminology and clinical contexts in the ChatGLM-assisted data extraction process (*Figure 5*). The extraction errors for discharge medication identified by ChatGLM were related primarily to the inability to accurately identify data with similar names for certain medications and reasoning errors in unit conversions for doses taken. The accuracy of the LLaMA in extracting vital signs from three fields (systolic blood pressure, diastolic blood pressure, and pulse) was <5%, and the errors were mainly in the form of null values returned without recognition.

Challenges and possible solutions pertaining to the use of ChatGLM

The challenges associated with the use of ChatGLM focus on prompt design, prompt output consistency, prompt output verification, and integration with HISs (see Supplementary material online, *Table S2*). To increase the interoperability of ChatGLM with health information systems, the development of tools such as the ESDR software used in this study is a recommended strategy.

Extension of the research findings to other regions

Although this study was conducted in a Chinese healthcare setting, the insights gained from this pilot study may be applicable to broader regions. Supplementary material online, *Table S3* provides a

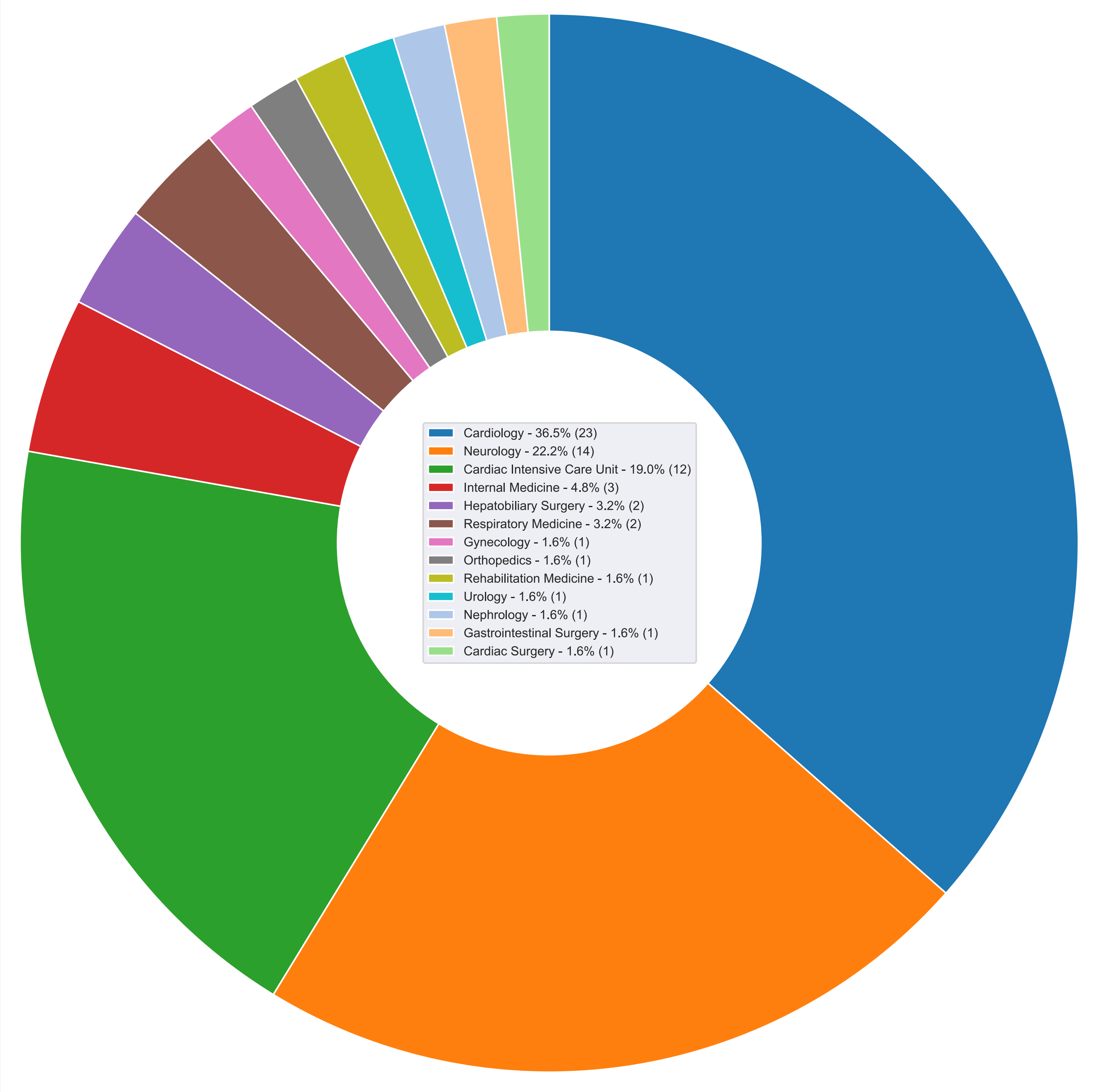


Figure 2 Distribution of departments from which the patient data were drawn. The percentages and numbers of patients in different departments are listed in the legend.

comprehensive summary of the implementation findings derived from this study. Supplementary material online, *Figure S2* uses the web chat interface of ChatGLM2-6B to demonstrate the prompt query test, while Supplementary material online, *Figure S3* uses HuatuoGPT-II.

Discussion

Interpretation of the research results

In this retrospective study, some study-related free-text elements were absent, reflecting inherent limitations in terms of data completeness. EHRs, which are designed for medical rather than research purposes, pose challenges for retrospective research in terms of data quality, as clinicians may not prioritize research elements in their record-keeping. Therefore, the discrepancy in the completeness of the eCRF fields arises from the initial medical records being incomplete. Since free-text medical records are transmitted intact to ChatGLM or LLaMA for extraction, the results of the data integrity assessment reflect the inherent deficiencies of the medical records rather than the data extraction process used.

The eCRF data transcription time results indicate an 80.7% improvement in the ChatGLM-assisted process, a finding that is in line with prior research.[16,17] The batch processing of ChatGLM data extraction requires significantly less time than does the labour-intensive traditional

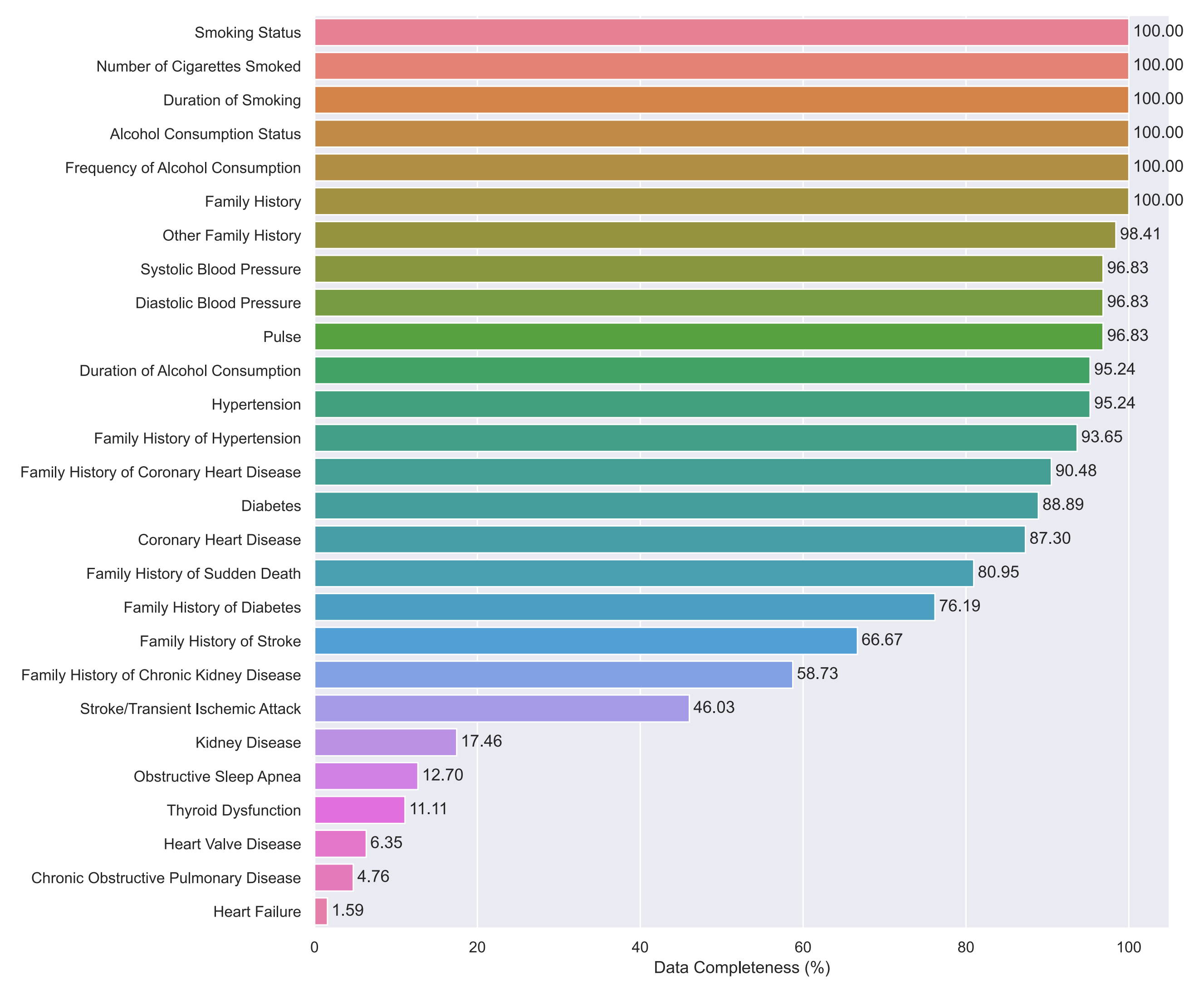


Figure 3 Data completeness across different fields in the free-text forms for all patients (Note: The phrase 'free-text forms' refers to the overall category, including four forms: lifestyle and behaviour, family history, health status, and vital signs.).

Table 1 Time consumption associated with the electronic source data repository manual process

Electronic case report form name	Manual data entry time (in seconds)	Manual verification time (in seconds)	Total (in seconds)
Discharge medication	12 847	8883	21 730
Free-text forms	15 031	11 621	26 652
Total	27 878	20 504	48 382

The phrase 'free-text forms' refers to the overall category, which includes four forms: lifestyle and behaviour, family history, health status, and vital signs.

manual process. As retrospective research projects become increasingly common, the scalability and cost-effectiveness of batch processing become evident, thus emphasizing the substantial value of this approach.

Table 2 Time consumption associated with the ChatGLM-assisted process

Process	Time (in seconds)
Batch processing (free-text forms + discharge medication)	12 789
ChatGLM-assisted data traceability (free-text forms)	6586
ChatGLM-assisted data traceability (discharge medication)	2751
Total	22 126

A comparison of the accuracy of free-text field extraction between LLaMA and ChatGLM revealed differences. LLaMA, an exemplary model with performance comparable to that of ChatGPT, exhibits greater accuracy in recognizing disease terms. However, it is less accurate than ChatGLM in recognizing fields such as family history and vital signs. The

Table 3 Comparison of accuracy rates with respect to initial data entry or extraction between the two processes

Form name	Single-person eCRF fields	Total fields	Manual entry verification (corrected fields)	Manual data entry accuracy (%)	ChatGLM-assisted data traceability (corrected fields)	ChatGLM extraction accuracy (%)	LLaMA-assisted data traceability (corrected fields)	LLaMA extraction accuracy (%)
Discharge medication	123	7749	80	98.97	99	98.72	—	—
Free-text forms	27	1701	7	99.59	389	77.13	955	43.86
Total	150	9450	87	99.08	488	94.84	—	—

eCRF, electronic case report form; LLaMA, Large Language Model Meta AI.

variation in accuracy can be attributed to the fact that the free-text admission records are in Chinese, which presents a challenge for the LLaMA model in adapting to the recording conventions of the Chinese medical language. Conversely, the localized Chinese model ChatGLM demonstrates superior accuracy. In the overall accuracy comparison of free-text extraction, the ChatGLM achieved 77.13%, whereas the LLaMA model reached only 43.86%. Owing to restrictions imposed on the transfer of Chinese medical data from hospitals, very few open-source medical LLMs trained on such data are available. HuatuoGPT-II,[21] a Chinese medical LLM, was primarily designed to assist in diagnosing diseases and formulating treatment plans by providing medical Q&A services. Extracting clinical research data from medical records requires the employed LLM to identify descriptions in the input medical text that are relevant to eCRF fields, subsequently outputting concise answers through basic reasoning. In this context, unlike general-purpose LLMs such as ChatGLM and ChatGPT, HuatuoGPT-II generates extensive medical knowledge texts related to the given query (see Supplementary material online, *Figure S3*). This necessitates the integration of an additional LLM into the task flow to further distil precise answers from the output of HuatuoGPT-II. This dual-LLM approach inherently increases the risk of inducing data extraction errors due to the complexity of and potential for misinterpretation between the two models. Consequently, ChatGLM was selected as the primary model for this study, as the differences in language, styles, and information systems preclude the deployment of high-performing open-source models such as LLaMA, which are unable to achieve optimal performance in the Chinese medical environment.

While LLM enhancements can improve their reasoning capabilities, the impact of this strategy on the resulting data extraction quality remains limited. This limitation arises because the extraction of data from medical records primarily requires only basic reasoning abilities. More critically, it depends on augmenting prompts with medically relevant knowledge to compensate for the general lack of medical text comprehension by generalized LLMs. As demonstrated in this study, the developed prompt design method can still yield favourable results even when using older models, such as ChatGLM2-6B.

A comparative analysis of the extraction accuracy of ChatGLM for free-text fields across different departments revealed a negative correlation between the length of the admission record and the extraction accuracy. This is because the admission records of inpatients in internal medicine departments are typically more complex and longer than those in surgical departments. The efficiency improvement of ChatGLM-assisted is determined primarily by the time spent on data traceability and correction via ESDR, resulting in an inverse relationship between the efficiency improvement and extraction accuracy.

Process improvement

The adoption of ChatGLM signifies a transformative shift in free-text data extraction, overcoming the constraints of conventional NLP methods. ChatGLM not only enhances the security of medical data but also leads to a significant increase in data extraction efficiency, thereby prioritizing clinical understanding over technical intricacies.

Regarding safety, conventional NLP methods pose security risks, as technicians handle patient medical records for the purpose of text annotation. In contrast, ChatGLM facilitates prompt design implementation without requiring direct interaction with EHRs, thus effectively mitigating the risk of data exposure and ensuring secure data extraction.

Unlike rule-based NLP, which relies on predefined patterns and manual rule creation, ChatGLM uses deep learning to understand the nuances and context of language, thereby offering a more adaptable and efficient approach. The primary advantage of this approach lies in the substantial reduction in the amount of human labour required for data extraction. Traditional NLP methods demand extensive manual effort for rule creation and maintenance, whereas ChatGLM autonomously learns from vast datasets, leading to significant cost savings and enabling human resources to be directed to more complex tasks.

Moreover, beyond the transformation of data extraction, ChatGLM plays a crucial role in enhancing safety and communication within the healthcare domain. By minimizing the need for meticulous rule creation and maintenance, ChatGLM facilitates collaboration between technical experts and healthcare professionals. In contrast to traditional NLP methods, in which context technical experts grapple with intricate rule sets, ChatGLM allows these actors to focus on correcting and fine-tuning the model. This ability to learn from extensive datasets reduces the likelihood of rule-based errors and establishes a more reliable process for data extraction from EHRs.

Efficacy of the electronic source data repository system with an integrated ChatGLM

In the past, employing traditional NLP models on a central processing unit occasionally required as long as 1 h for an eCRF to execute. However, the use of a graphics processing unit to run ChatGLM enables the same task to be completed within a few minutes. At present, technicians can simply configure the ChatGLM question prompts during project deployment, thus eliminating the need for extensive time investments in labelling text and supervising training. Furthermore, the adoption of ChatGLM enhances the reusability of ESDR across various research projects. This approach allows a limited number of technical personnel to support the development of multiple clinical research projects within the hospital efficiently.

Innovations of this study

The main contribution of this study is to validate the use of ESDR tools to address the interoperability and transparency challenges of using

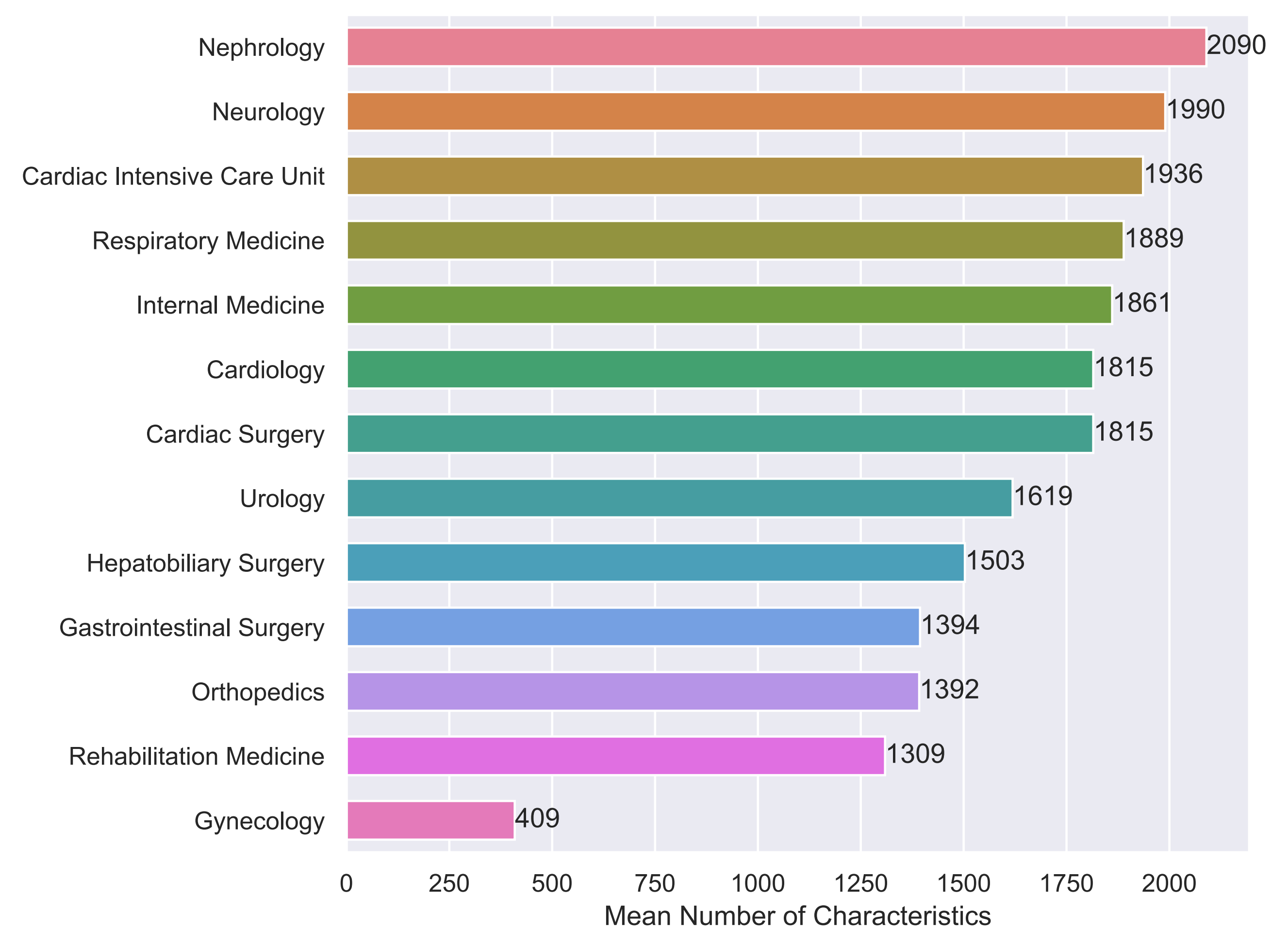


Figure 4 The mean number of characters in the patient admission notes acquired from different departments.

ChatGLM for RWD extraction in Chinese hospital settings. ESDR tools integrate EHR source data and eCRF study data interfaces to achieve interoperability between ChatGLM and both systems. ChatGLM extracts study data from EHR source data and electronically populates it into the eCRF for interoperability. By tracing study data back to the corresponding medical records, the ESDR tool addresses interoperability and transparency issues in assessing the reliability of ChatGLM extraction results.

Comparisons with similar work

Lee *et al.*[22] demonstrated that the performance of a privacy-preserving FastChat-T5 is promising with respect to the automatic answering of medical questions on the basis of an evaluation of 84 thyroid cancer surgical pathology reports. Chiang *et al.*[23] discussed GPT-2 models that had been finetuned on clinical notes regarding the accurate extraction of headache frequency from EHRs. The study by Ge *et al.*[24] compared the use of Versa GPT-4, implemented in a protected health information-compliant manner, with manual chart review to extract eight data elements from electronic medical records related to hepatocellular carcinoma. Notably, these investigations focused primarily on the stage of theoretical evaluation and testing, thus highlighting a substantial gap pertaining to the practical application of LLMs in clinical research projects.

Implications for future research

To improve the accuracy of ChatGLM-driven data extraction, researchers have proposed the use of a multiround processing approach. Multiple rounds of processing enable the ChatGLM to detect and correct its own errors, resulting in more consistent and accurate data extraction.[25] The effectiveness of ChatGLM could be improved by incorporating RAG techniques. The RAG algorithm combines the generative capabilities of ChatGLM with the precision of information retrieval, allowing the model to access a medical knowledge database and retrieve relevant information as needed.[26] This approach can address the model's medical knowledge gaps and improve its ability to handle localized clinical terms.

The success of the ESDR system with ChatGLM integration highlights possible avenues for future research on the optimization of data extraction and traceability in clinical settings. In the future, this study's insights can guide the use of ESDR tools featuring ChatGLM to conduct prospective clinical research, expanding the application scenarios of such tools to encompass clinical trials. Further studies could explore the scalability of the system across different healthcare institutions, assess its adaptability to various types of medical specialty research, and investigate the generalizability of ChatGLM-driven data extraction in diverse research scenarios.

Finally, the impact of long-term use of ChatGLM in healthcare settings on patient outcomes and data integrity needs to be evaluated in

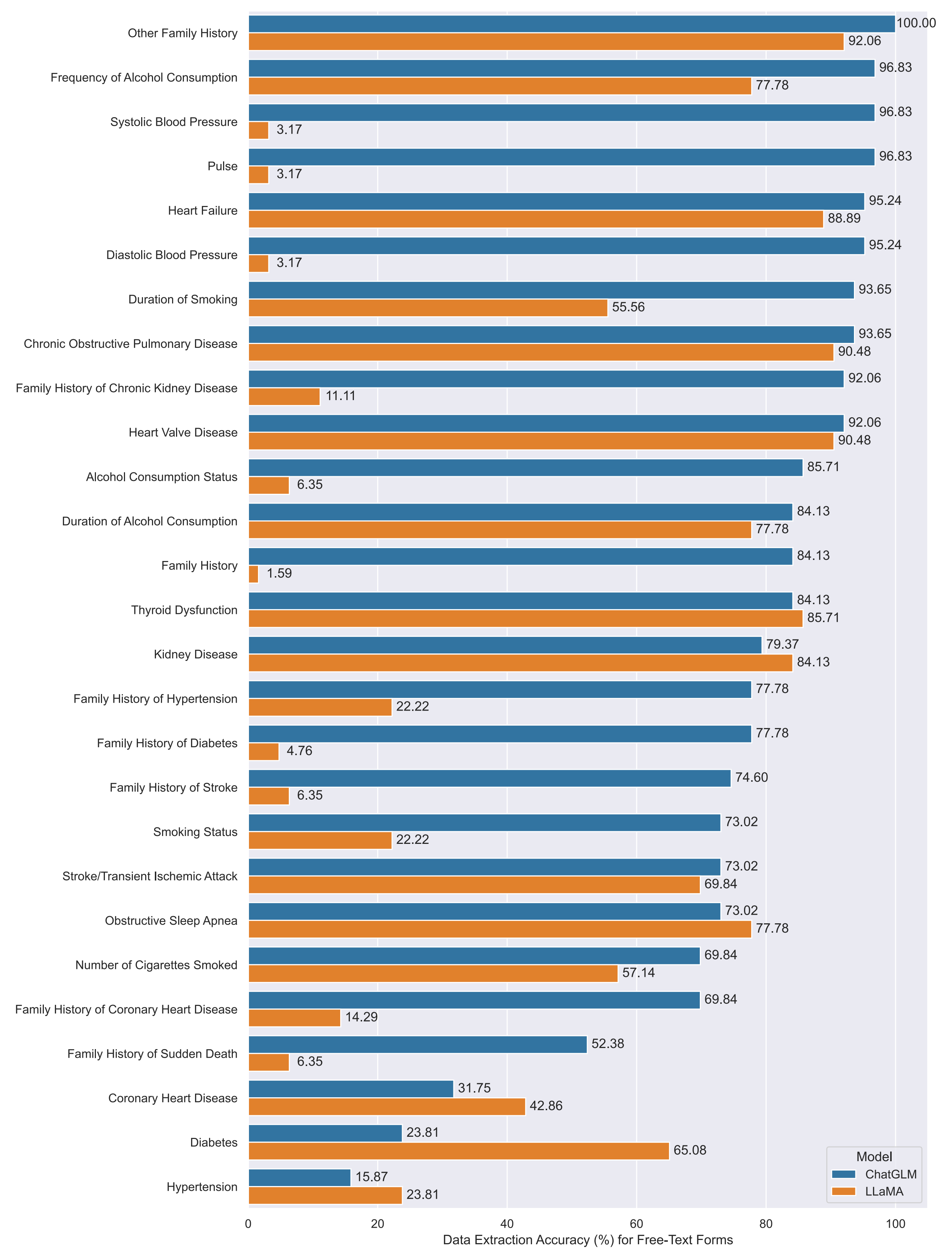


Figure 5 The accuracy of ChatGLM or LLaMA data extraction from different fields of the free-text forms for all patients.

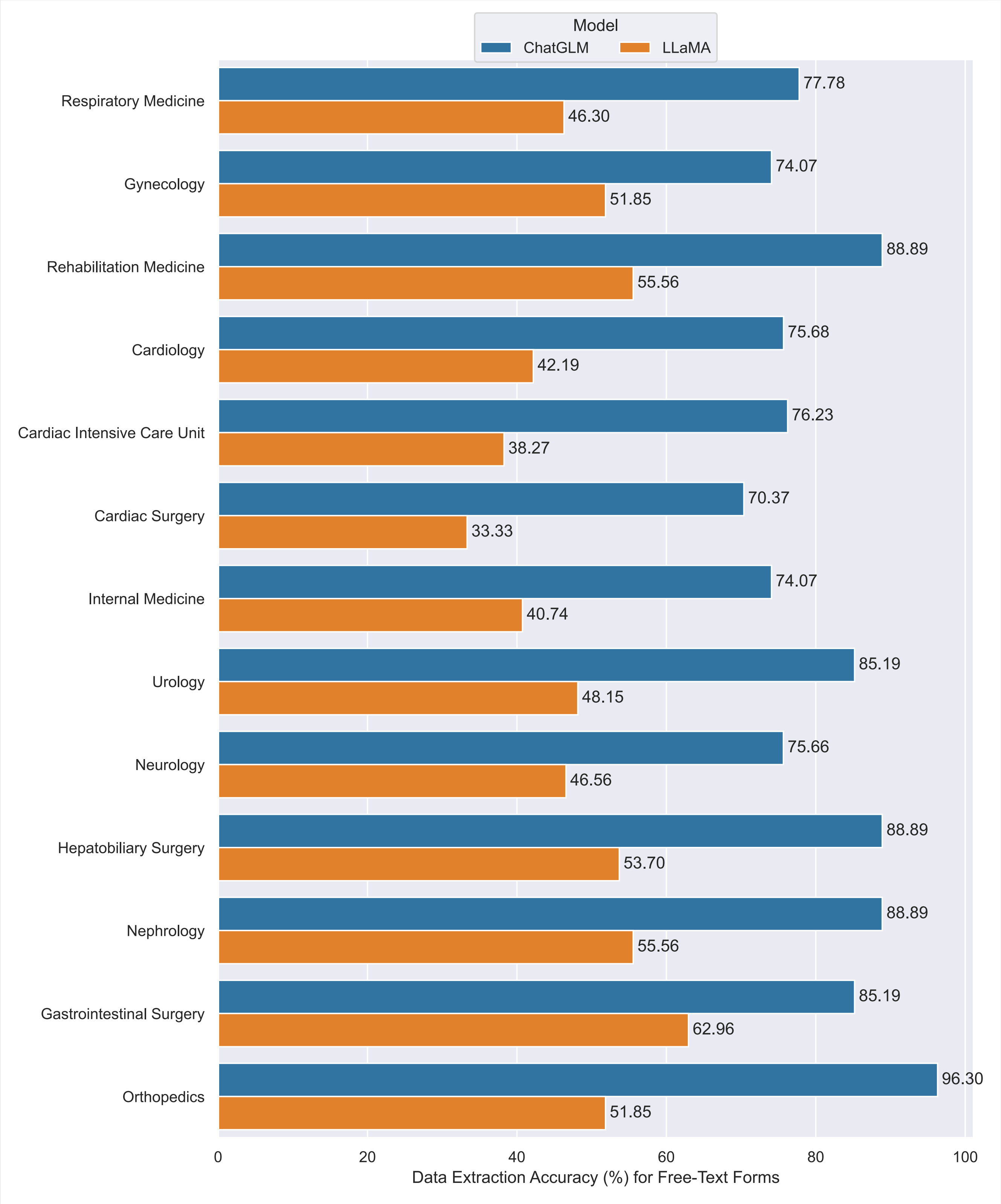


Figure 6 The mean accuracy of ChatGLM or LLaMA in extracting data from free-text forms of patients across different departments.

detail. The potential benefits of ChatGLM to healthcare systems are considerable, with the ability to guide initial medical record entries, enhance data integrity, and manage data quality at the source. By ensuring the accurate capture of information from the outset, ChatGLM could assist in reducing errors and inconsistencies in medical records, which are crucial for reliable research and patient care. Its ability to identify critical adverse events from medical records could facilitate early detection and timely implementation of interventions, thereby preventing complications and reducing hospital readmissions. Healthcare professionals could be supported in their decision-making by detailed data summaries provided by ChatGLM, which would highlight essential patient information and trends, facilitating more informed decisions.

Limitations

The study was conducted within a single centre. The impact of ChatGLM may vary across different healthcare settings. This study was further constrained by the scope of the knowledge input into the model, and as such, it did not comprehensively address the challenge of enhancing the prompt design process. Instead, this research utilized a limited set of terminology specific to the given hospital and specialty departments. To advance the model's understanding of the clinical context and terminology related to the specific hospital, it is crucial to integrate a broader and more comprehensive source of terminology drawn from the local unified medical language system,[27] such as the Chinese Hospital Information Management Association terminology dataset.[28]

Conclusions

The main contribution of this study is to validate the use of ESDR tools to address the interoperability and transparency challenges of using ChatGLM for RWD extraction in Chinese hospital settings. The overall data extraction accuracy of ChatGLM in free-text fields is greater than that of the LLaMA model, suggesting that a localized model of LLMs adapted to the language of medical records should be chosen for implementation. However, ChatGLM shows a decrease in data extraction accuracy and time savings as the complexity increases with the length of the medical-free text.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health*.

Acknowledgements

We sincerely thank Hangzhou LionMed Medical Information Technology Co., Ltd., for its strong support of our research.

Funding

This work was supported by the Real-World Study Project of the Hainan Boao Lecheng Pilot Zone (Real-World Study Base of NMPA) (no. HNLC2022RWS017).

Conflict of interest: none declared.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author upon reasonable request.

Lead author biography

Bin Wang, MD, is a postdoctoral fellow at the School of Clinical Medicine, Tsinghua University. He received an MD degree in Clinical Research Methodology from Peking University Health Science Center. His research interests focus on digital clinical research.

Junkai Lai, MD, is a postdoctoral fellow at the Institute of Automation, Chinese Academy of Sciences. He received an MD degree in Clinical Research Methodology from Peking University Health Science Center. His research interests focus on digital clinical research and artificial intelligence.

References

1. Arora A, Arora A. The promise of large language models in health care. *Lancet* 2023;**401**:641.
2. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–1240.
3. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, *et al.* Large language models encode clinical knowledge. *Nature* 2023;**620**:172–180.
4. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, *et al.* A large language model for electronic health records. *NPJ Digit Med* 2022;**5**:194.
5. Skalidis I, Cagnina A, Fournier S. Performance of artificial intelligence in answering cardiovascular textual questions. *Eur Heart J Digit Health* 2023;**4**:364–365.
6. Skalidis I, Cagnina A, Fournier S. Use of large language models for evidence-based cardiovascular medicine. *Eur Heart J Digit Health* 2023;**4**:368–369.
7. Wang S, Sun X, Li X, Ouyang R, Wu F, Zhang T, *et al.* Gpt-ner: named entity recognition via large language models. arXiv 2304.10428, https://doi.org/10.48550/arXiv.2304.10428, 7 October 2023, preprint: not peer reviewed.
8. Zhou H, Li M, Xiao Y, Yang H, Zhang R. LEAP: LLM instruction-example adaptive prompting framework for biomedical relation extraction. *J Am Med Inform Assoc* 2024;**31**:2010–2018.
9. Wiest IC, Ferber D, Zhu J, van Treeck M, Meyer SK, Juglan R, *et al.* From text to tables: a local privacy preserving large language model for structured information retrieval from medical documents. medRxiv 23299648, https://doi.org/10.1101/2023.12.07.23299648, 8 December 2023, preprint: not peer reviewed.
10. Chen S, Savova GK, Bitterman DS. Considerations for prompting large language models—reply. *JAMA Oncol* 2024;**10**:538–539.
11. Denecke K, May R, Rivera Romero O. Potential of large language models in health care: Delphi study. *J Med Internet Res* 2024;**26**:e52399.
12. Shahnaz A, Qamar U, Khalid A. Using blockchain for electronic health records. *IEEE Access* 2019;**7**:147782–147795.
13. Jin F, Yao C, Yan X, Dong C, Lai J, Li L, *et al.* Gap between real-world data and clinical research within hospitals in China: a qualitative study. *BMJ Open* 2020;**10**:e038375.
14. Nordo AH, Levaux HP, Becnel LB, Galvez J, Rao P, Stem K, *et al.* Use of EHRs data for clinical research: historical progress and current applications. *Learn Health Syst* 2019;**3**:e10076.
15. Wang B, Lai J, Jin F, Liao X, Zhu H, Yao C. Clinical source data production and quality control in real-world studies: proposal for development of the eSource record system. *JMIR Res Protoc* 2022;**11**:e42754.
16. Wang B, Lai J, Liu M, Jin F, Peng Y, Yao C. Electronic source data transcription for electronic case report forms in China: validation of the electronic source record tool in a real-world ophthalmology study. *JMIR Form Res* 2022;**6**:e43229.
17. Wang B, Hao X, Yan X, Lai J, Jin F, Liao X, *et al.* Evaluation of the clinical application effect of eSource record tools for clinical research. *BMC Med Inform Decis Mak* 2022;**22**:98.

18. Wang B, Lai J, Liao X, Jin F, Yao C. Challenges and solutions in implementing eSource technology for real-world studies in China: qualitative study among different stakeholders. *JMIR Form Res* 2023;**7**:e48363.
19. Zeng A, Liu X, Du Z, Wang Z, Lai H, Ding M, *et al.* Glm-130b: an open bilingual pretrained model. arXiv 2210.02414, https://doi.org/10.48550/arXiv.2210.02414, 5 October 2022, preprint: not peer reviewed.
20. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, *et al.* LLaMA: open and efficient foundation language models. arXiv 2302.13971, https://doi.org/10.48550/arXiv.2302.13971, 27 February 2023, preprint: not peer reviewed.
21. Chen J, Wang X, Gao A, Jiang F, Chen S, Zhang H, *et al.* HuatuoGPT-II, one-stage training for medical adaption of LLMs. arXiv 2311.09774, https://doi.org/10.48550/arXiv.2311.09774, 16 November 2023, preprint: not peer reviewed.
22. Lee DT, Vaid A, Menon KM, Freeman R, Matteson DS, Marin MP, *et al.* Development of a privacy preserving large language model for automated data extraction from thyroid cancer pathology reports. medRxiv 23298252, https://doi.org/10.1101/2023.11.08.23298252, 8 November 2023, preprint: not peer reviewed.
23. Chiang CC, Luo M, Dumkrieger G, Trivedi S, Chen YC, Chao CJ, *et al.* A large language model-based generative natural language processing framework fine-tuned on clinical notes accurately extracts headache frequency from electronic health records. *Headache* 2024;**64**:400–409.
24. Ge J, Li M, Delk MB, Lai JC. A comparison of a large language model vs manual chart review for the extraction of data elements from the electronic health record. *Gastroenterology* 2024;**166**:707–709.e703.
25. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, *et al.* Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta Radiol* 2023;**1**:100017.
26. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, *et al.* Retrieval-augmented generation for large language models: a survey. arXiv 2312.10997, https://doi.org/10.48550/arXiv.2312.10997, 18 December 2023, preprint: not peer reviewed.
27. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267–D270.
28. Zhang Y, Xu Y, Shang L, Rao K. An investigation into health informatics and related standards in China. *Int J Med Inform* 2007;**76**:614–620.