Contents lists available at ScienceDirect

New Microbes and New Infections

journal homepage: www.journals.elsevier.com/new-microbes-and-new-infections

Original Article

Automated data collection tool for real-world cohort studies of chronic hepatitis B: Leveraging OCR and NLP technologies for improved efficiency

Xiaomei Zhou [a,1], Tao Zeng [b,1], Yibo Zhang [b,1], Yingying Liao [b], Jaime Smith [c], Lin Zhang [d], Chao Wang [d], Qinghai Li [e], Dongbo Wu [f,**], Yutian Chong [b,***], Xinhua Li [b,*]

[a] Information Center, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, 510630, China
[b] Department of Infectious Diseases, The Third Affiliated Hospital of Sun Yat-sen University, Guangdong Key Laboratory of Liver Disease, Guangzhou, 510630, China
[c] PAREXEL International, Billerica, MA, 01821, USA
[d] PAREXEL International, Shanghai, 200120, China
[e] KingPoint Data Technology Co., Ltd, Guangzhou, 510635, China
[f] West China Hospital of Sichuan University, Chengdu, 610041, China

ARTICLE INFO

Keywords:
Data collection
Hepatitis B
Natural language processing
Optical character recognition

ABSTRACT

Background: Collecting and standardizing clinical research data is a very tedious task. This study is to develop an intelligent data collection tool, named CHB-EDC, for real-world cohort studies of chronic hepatitis B (CHB), which can assist in standardized and efficient data collection.

Methods: CHB_EDC is capable of automatically processing various formats of data, including raw data in image format, using internationally recognized data standards, OCR, and NLP models. It can automatically populate the data into eCRFs designed in the REDCap system, supporting the integration of patient data from electronic medical record systems through commonly used web application interfaces. This tool enables intelligent extraction and aggregation of data, as well as secure and anonymous data sharing.

Results: For non-electronic data collection, the average accuracy of manual collection was 98.65 %, with an average time of 63.64 min to collect information for one patient. The average accuracy CHB-EDC was 98.66 %, with an average time of 3.57 min to collect information for one patient. In the same data collection task, CHB-EDC achieved a comparable average accuracy to manual collection. However, in terms of time, CHB-EDC significantly outperformed manual collection ($p < 0.05$). Our research has significantly reduced the required collection time and lowered the cost of data collection while ensuring accuracy.

Conclusion: The tool has significantly improved the efficiency of data collection while ensuring accuracy, enabling standardized collection of real-world data.

1. Introduction

Hepatitis B virus (HBV) infection is a global epidemic. According to the World Health Organization (WHO), the global prevalence of HBsAg in the general population is 3.8 %, with approximately 1.5 million new cases of HBV infection, 296 million cases of chronic infection, and 820,000 deaths due to HBV-related liver failure, cirrhosis, or hepatocellular carcinoma (HCC) in 2019 [1]. China has the highest number of HBV-infected individuals, accounting for one-third of the world's infected population, posing a serious threat to patients' lives and imposing a heavy burden on the healthcare system [2]. To achieve the World Health Organization's goal of "eliminating viral hepatitis as a public health threat by 2030," large-scale, multicenter studies on the clinical research of chronic HBV infection are essential.

In real-world cohort studies of chronic hepatitis B (CHB), structured patient data serves as the foundation for large-scale, multicenter analysis of real-world HBV cohort data, and it has significant clinical significance for the comprehensive prevention and treatment of hepatitis B.

[*] Corresponding author.
[**] Corresponding author.
[***] Corresponding author.
E-mail addresses: dongbohuaxi@scu.edu.cn (D. Wu), chongyt@mail.sysu.edu.cn (Y. Chong), lixinh8@mail.sysu.edu.cn (X. Li).
[1] These authors have contributed equally to this work.

https://doi.org/10.1016/j.nmni.2024.101469
Received 20 December 2023; Received in revised form 26 May 2024; Accepted 26 August 2024
Available online 28 August 2024

However, the traditional manual data collection is prone to human errors, and the quality and reliability of the data are insufficient, resulting in low efficiency and inability to meet the requirements of large-scale multicenter real world cohort studies. This challenge is mainly manifested in two aspects: (1) there is a large amount of unstructured data in electronic medical records [3], such as laboratory test reports, discharge summaries, and chief complaint, current medical history, past history, personal history, family history, and allergy history; (2) in multicenter real-world studies, there are differences in data among different centers, with some centers storing data in the form of images or paper documents, while others use electronic spreadsheet databases. To address this challenge, mature artificial intelligence (AI) technologies in the field of computer science, such as Optical Character Recognition (OCR) [4] and Natural Language Processing (NLP) [5,6], can be used for automated data extraction. These technologies offer intelligent proofreading functions, precise and efficient data collection, and significant cost reduction, providing guarantees for large-scale, multicenter data collection. OCR technology can recognize and extract text information from images and scanned documents, thereby achieving electronic data collection, and NLP technology can extract key information from unstructured data and convert it into structured data [4]. The application of AI technologies such as NLP and OCR in research data collection can significantly reduce the manual cost of data collection, cleaning, transformation, and standardization, while improving accuracy and efficiency [5].

In this study, we developed an AI-assisted data collection tool, named Chronic Hepatitis B Electronic Data Capture (CHB-EDC) System, for real-world data collection in the CHB research cohort as shown in Fig. 1. This tool can automatically process various formats of data, including raw data in image format, using internationally recognized data standards, OCR, and NLP models. It can automatically fill in the electronic case report forms (eCRFs) designed in the Research Electronic Data Capture (REDCap) system with data from image and other format sources [7]. Additionally, it intelligently extracts, summarizes, de-identifies, and facilitates data sharing, markedly enhancing the precision and efficiency of research data collection.

2. Methods

2.1. Requirements and technology selection

CHB-EDC system was designed in strict accordance with the "Implementation Specification for Real-World Research in Clinical Medicine in China" [8]. At the same time, the data definition and processing standards were referenced from the Clinical Data Interchange Standards Consortium (CDISC) standards to meet the requirements of multicenter research both domestically and internationally [9]. As shown in Fig. 1, our application development is divided into 4 stages. In the first stage, OCR and NLP tools were developed based on artificial intelligence technology. In the second stage, these tools were integrated into the REDCap system to achieve automated mapping of data into eCRF forms. In the third stage, integration with the electronic medical record (EMR) system was achieved through interfaces, enabling standardized data collection for both internal and external research purposes. In the fourth stage, we deployed our tool to the collaborating hospitals and validation.

2.2. Stage 1: research preparation and OCR/NLP models development

The study subjects must meet the definition of chronic hepatitis B prevention and treatment guidelines in China (2019 edition) [10] and also meet the following criteria. Inclusion criteria: (1) The time interval between the first positive Hepatitis B Surface Antigen (HBsAg) test and the last test is greater than 6 months; (2) The current examination data of the patients show normal Alanine Aminotransferase (ALT) levels; (3) Liver biopsy has been performed and corresponding pathological data are available; (4) Patients have not received antiviral treatment (including nucleoside (acid) or interferon) before liver biopsy. Exclusion criteria: (1) Patients with overlapping infections of Hepatitis A Virus (HAV), Hepatitis C Virus (HCV), Hepatitis D Virus (HDV), Hepatitis E Virus (HEV), or Human Immunodeficiency Virus (HIV); (2) Patients with other liver diseases (such as drug-induced, alcoholic, autoimmune, genetic metabolic liver diseases, etc.); (3) Patients who have undergone organ transplantation or are preparing for organ transplantation; (4) Other situations deemed unsuitable for inclusion by the researchers.

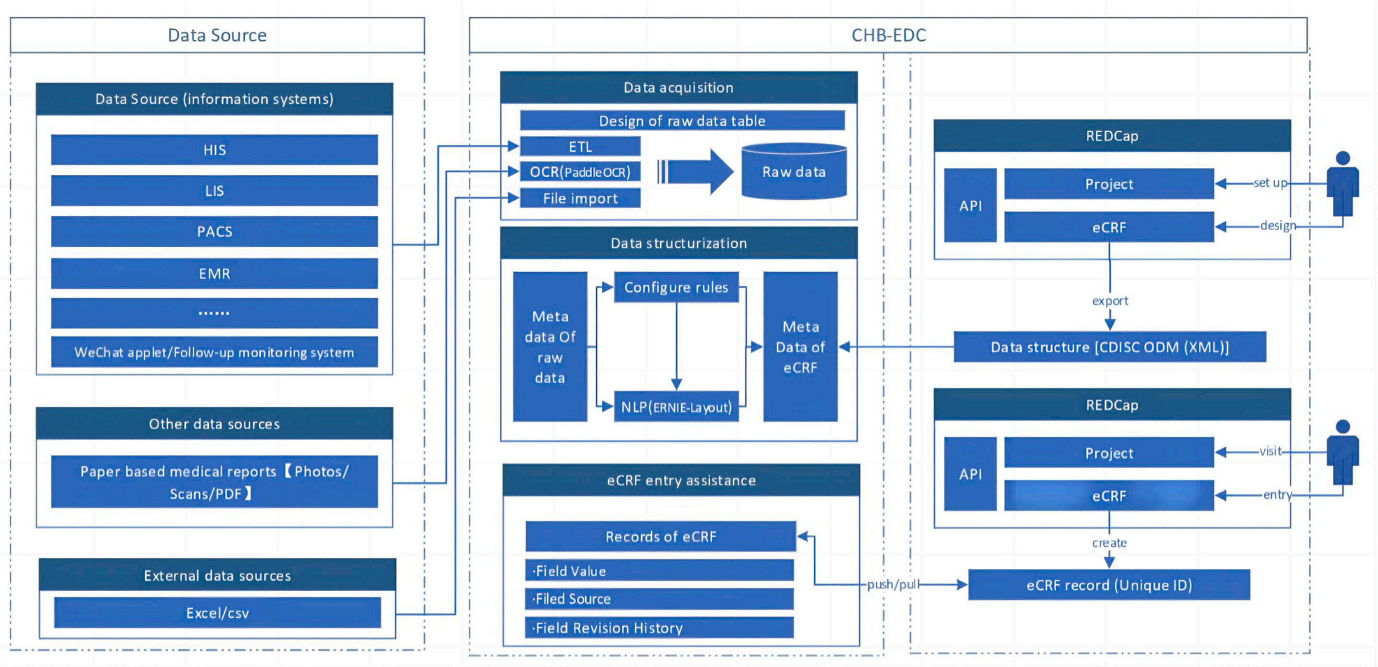


Fig. 1. CHB-EDC system design diagram. CHB-EDC supports data collection from various sources. HIS, hospital information system; LIS, laboratory information management system; PACS, picture archiving and communication system; EMR, electronic medical record system.

Experts in liver disease research discussed and formulated the fields and format of the eCRF. The planned collection of patient information includes demographic information, vital signs, complete blood count, liver function, hepatitis B virus markers, and pathology reports. The study subjects had to meet the definition of chronic hepatitis B outlined in the 2019 Chinese Guidelines for the Prevention and Treatment of Chronic Hepatitis B [11].

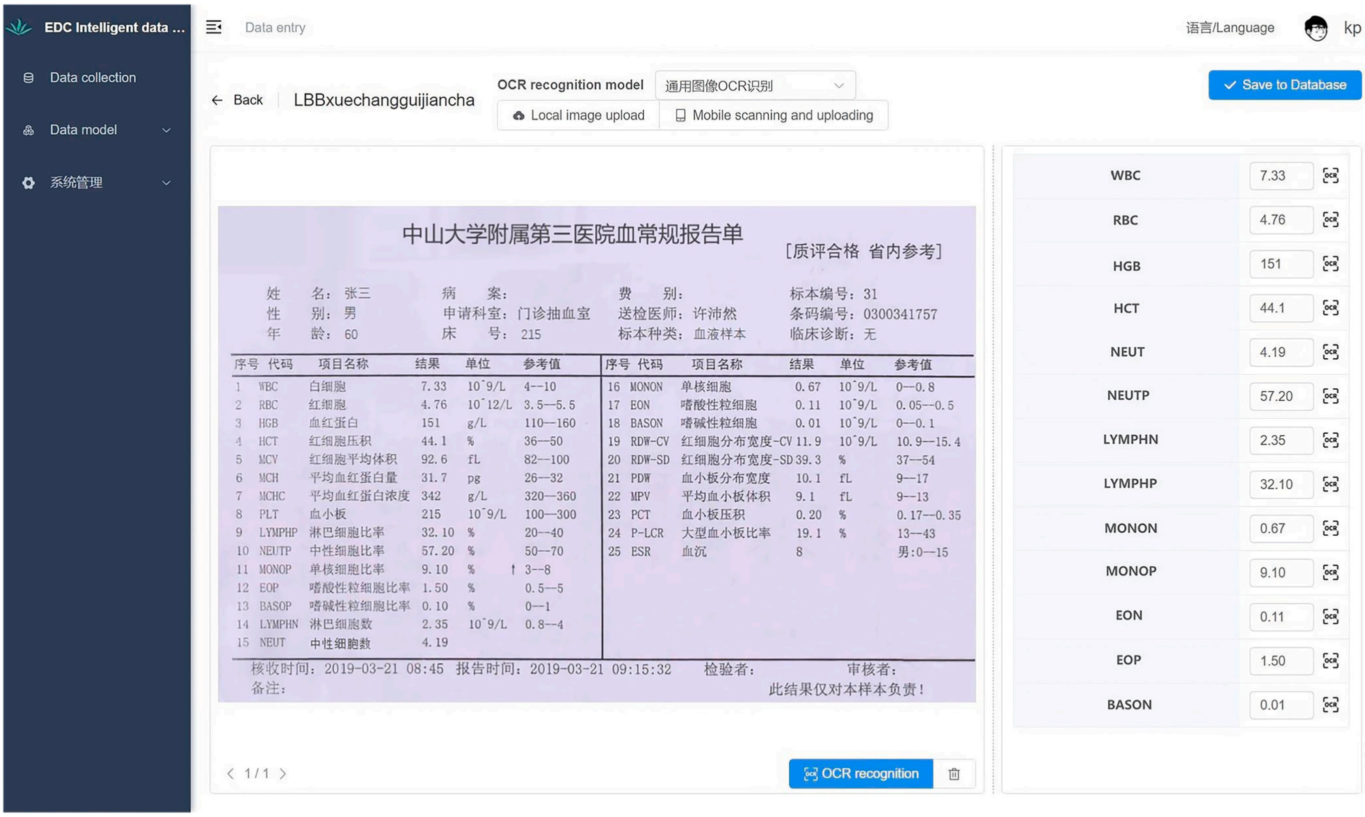
In order to digitize a large amount of external paper-based and image-format patient data, we utilized OCR technology. PaddleOCR is an open-source OCR tool developed based on Baidu's PaddlePaddle deep learning platform [12]. It supports accurate recognition of various types of text, including printed text, handwritten text, and scene text (Fig. 2). NLP models have been widely applied in the processing of clinical semi-structured and unstructured data. We used the ERNIE-Layout model for understanding and transforming clinical semi-structured and unstructured data [13]. ERNIE-Layout is an open-source model for cross-modal document understanding. It utilizes layout knowledge enhancement technology to integrate text, images, and layout information for joint modeling, enabling deep understanding and analysis of multi-modal documents such as document images and scans (Fig. 3). Our model's pretraining data was obtained from the Chinese Medical Entity Extraction dataset (CMeEE), which is part of the Chinese Biomedical Language Understanding Evaluation (CBLUE) benchmark 3.0 version. It consists of 15000 training data, 5000 validation data, and 3000 test data for Chinese medical named entity recognition.

2.3. Stage 2: integrated into REDCap

The next stage involved developing server programming to automatically extract data from OCR and NLP models and populate the electronic case report forms (eCRF) designed in the Research Electronic Data Capture (REDCap) system (Fig. 4). To ensure the security and privacy of patient data, CHB-EDC implemented strict user access restrictions and maintained an activity log tracking feature. To address security concerns during the data storage and sharing process, we designed and developed a data security protection tool. This tool is a web-based system with a B/S architecture that integrates methods such as data encryption and de-identification. It allows for encryption [14] and de-identification [15] of structured data, ensuring that locally sensitive files are uploaded in an encrypted form and stored in the database. During data exchange, the tool ensures that the data remains in a de-identified and encrypted state. Only after completing the data exchange can the data be decrypted, which ensures a comprehensive data protection process. CHB-EDC utilizes Node.js as the server platform, enabling seamless code reuse between the browser and server using JavaScript with minimal editing. This stage took place from June to August 2023, during which the application underwent further testing for browser compatibility.

2.4. Stage 3: integrated into EMR

The third stage involved integrating the CHB-EDC system into the hospital's EMR system. This integration was achieved through a data interface, which allowed for automatic collection of patient data from within the hospital to CHB-EDC using a single HTTPS interface. The application would periodically collect data from the EMR system for patients with chronic hepatitis B. Using the eCRF data format, the ERNIE-Layout model was employed to process unstructured and semi-structured data, automatically populating the eCRF in the REDCap system. This significantly reduced the workload for clinicians in collecting research data. For patient records brought from other medical institutions in the form of photocopies or images, the PaddleOCR model was used to digitize them, and then the ERNIE-Layout model was applied to map the data to the eCRF.



中山大学附属第三医院血常规报告单

[质评合格 省内参考]

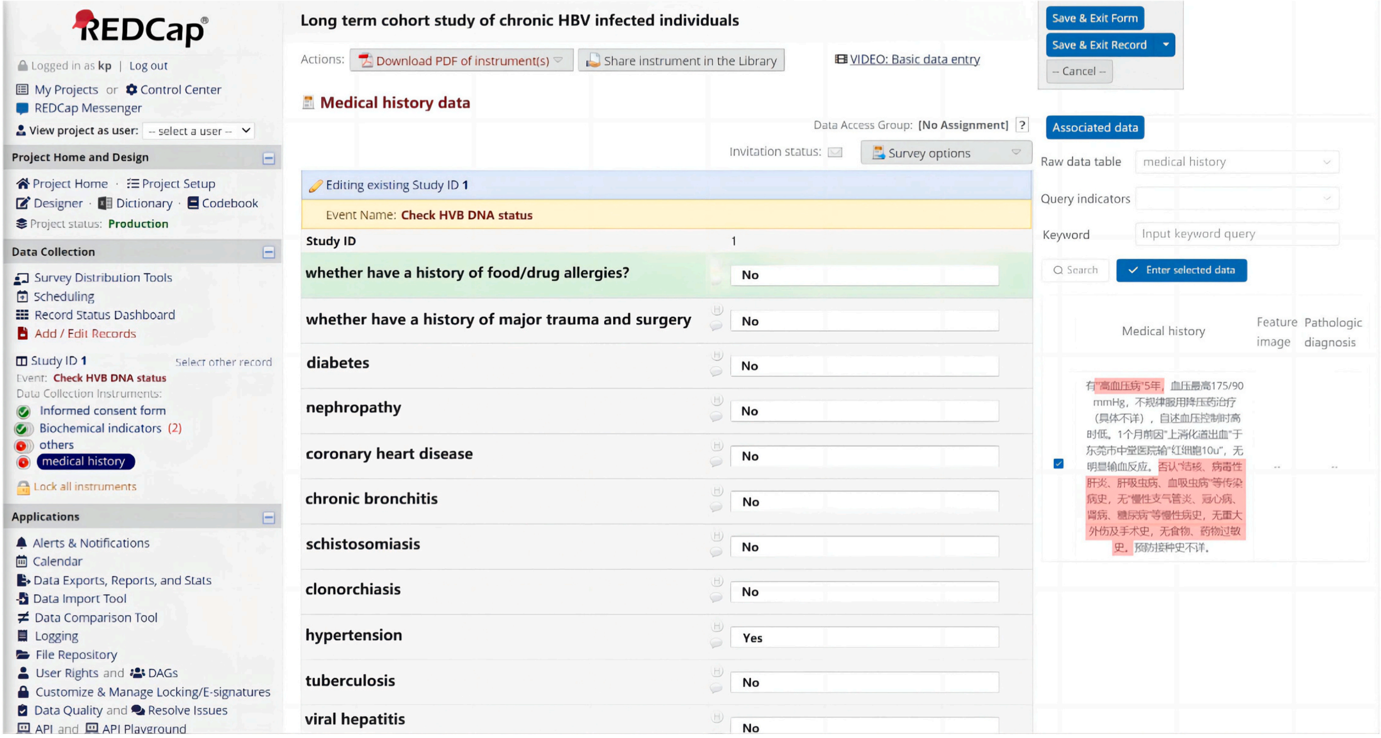
姓　名：张三　　病　案：　　费　别：　　标本编号：31
性　别：男　　申请科室：门诊抽血室　　送检医师：许沛然　　条码编号：0300341757
年　龄：60　　床　号：215　　标本种类：血液样本　　临床诊断：无

序号	代码	项目名称	结果	单位	参考值	序号	代码	项目名称	结果	单位	参考值
1	WBC	白细胞	7.33	10ˇ9/L	4--10	16	MONON	单核细胞	0.67	10ˇ9/L	0--0.8
2	RBC	红细胞	4.76	10ˇ12/L	3.5--5.5	17	EON	嗜酸性粒细胞	0.11	10ˇ9/L	0.05--0.5
3	HGB	血红蛋白	151	g/L	110--160	18	BASON	嗜碱性粒细胞	0.01	10ˇ9/L	0--0.1
4	HCT	红细胞压积	44.1	%	36--50	19	RDW-CV	红细胞分布宽度-CV	11.9	10ˇ9/L	10.9--15.4
5	MCV	红细胞平均体积	92.6	fL	82--100	20	RDW-SD	红细胞分布宽度-SD	39.3	%	37--54
6	MCH	平均血红蛋白量	31.7	pg	26--32	21	PDW	血小板分布宽度	10.1	fL	9--17
7	MCHC	平均血红蛋白浓度	342	g/L	320--360	22	MPV	平均血小板体积	9.1	fL	9--13
8	PLT	血小板	215	10ˇ9/L	100--300	23	PCT	血小板压积	0.20	%	0.17--0.35
9	LYMPHP	淋巴细胞比率	32.10	%	20--40	24	P-LCR	大型血小板比率	19.1	%	13--43
10	NEUTP	中性细胞比率	57.20	%	50--70	25	ESR	血沉	8		男:0--15
11	MONOP	单核细胞比率	9.10	% ↑	3--8						
12	EOP	嗜酸性粒细胞比率	1.50	%	0.5--5						
13	BASOP	嗜碱性粒细胞比率	0.10	%	0--1						
14	LYMPHN	淋巴细胞数	2.35	10ˇ9/L	0.8--4						
15	NEUT	中性细胞数	4.19								

核收时间：2019-03-21 08:45　报告时间：2019-03-21 09:15:32　检验者：　审核者：
备注：　此结果仅对本样本负责！

WBC	7.33
RBC	4.76
HGB	151
HCT	44.1
NEUT	4.19
NEUTP	57.20
LYMPHN	2.35
LYMPHP	32.10
MONON	0.67
MONOP	9.10
EON	0.11
EOP	1.50
BASON	0.01

Fig. 2. OCR tool text recognition interface. This tool supports the digitization and collection of data from paper documents, images, and scanned files.


REDCap
Logged in as kp | Log out
My Projects or Control Center
REDCap Messenger
View project as user: -- select a user --
Project Home and Design
Project Home · Project Setup
Designer · Dictionary · Codebook
Project status: Production
Data Collection
Survey Distribution Tools
Scheduling
Record Status Dashboard
Add / Edit Records
Study ID 1 Select other record
Event: Check HVB DNA status
Data Collection Instruments:
Informed consent form
Biochemical indicators (2)
others
medical history
Lock all instruments
Applications
Alerts & Notifications
Calendar
Data Exports, Reports, and Stats
Data Import Tool
Data Comparison Tool
Logging
File Repository
User Rights and DAGs
Customize & Manage Locking/E-signatures
Data Quality and Resolve Issues
API and API Playground
Actions: Download PDF of instrument(s)
Share instrument in the Library
VIDEO: Basic data entry
Save & Exit Form
Save & Exit Record
-- Cancel --
Associated data
Survey options
Search
Enter selected data

Long term cohort study of chronic HBV infected individuals

Medical history data

Data Access Group: [No Assignment]

Invitation status:

Editing existing Study ID 1

Event Name: Check HVB DNA status

Study ID	1
whether have a history of food/drug allergies?	No
whether have a history of major trauma and surgery	No
diabetes	No
nephropathy	No
coronary heart disease	No
chronic bronchitis	No
schistosomiasis	No
clonorchiasis	No
hypertension	Yes
tuberculosis	No
viral hepatitis	No

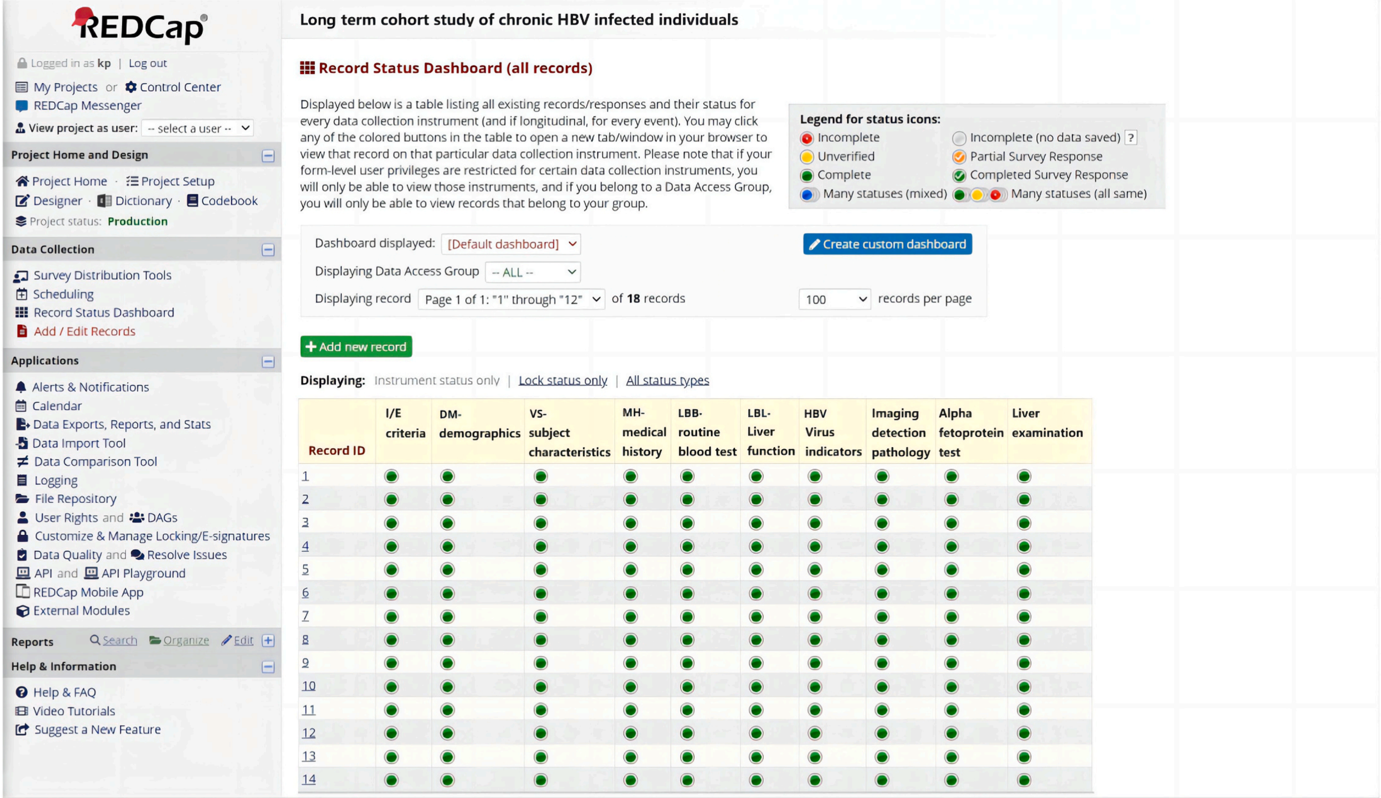
Raw data table: medical history

Query indicators:

Keyword: Input keyword query

	Medical history	Feature image	Pathologic diagnosis
☑	有"高血压病"5年，血压最高175/90mmHg，不规律服用降压药治疗（具体不详），自述血压控制时高时低。1个月前因"上消化道出血"于东莞市中堂医院输"红细胞10u"，无明显输血反应。否认"结核、病毒性肝炎、肝吸虫病、血吸虫病"等传染病史，无"慢性支气管炎、冠心病、肾病、糖尿病"等慢性病史，无重大外伤及手术史，无食物、药物过敏史。预防接种史不详。	--	--

Fig. 3. NLP tool data structuring interface. This tool can convert unstructured Chinese text into structured data and supports switching between Chinese and English.



Long term cohort study of chronic HBV infected individuals

Record Status Dashboard (all records)

Displayed below is a table listing all existing records/responses and their status for every data collection instrument (and if longitudinal, for every event). You may click any of the colored buttons in the table to open a new tab/window in your browser to view that record on that particular data collection instrument. Please note that if your form-level user privileges are restricted for certain data collection instruments, you will only be able to view those instruments, and if you belong to a Data Access Group, you will only be able to view records that belong to your group.

Legend for status icons:

🔴 Incomplete	○ Incomplete (no data saved)
🟡 Unverified	🟡 Partial Survey Response
🟢 Complete	✓ Completed Survey Response
🔵 Many statuses (mixed)	🟢🟡🔴 Many statuses (all same)

Dashboard displayed: [Default dashboard]

Displaying Data Access Group: -- ALL --

Displaying record: Page 1 of 1: "1" through "12" of 18 records; 100 records per page

Displaying: Instrument status only | Lock status only | All status types

Record ID	I/E criteria	DM-demographics	VS-subject characteristics	MH-medical history	LBB-routine blood test	LBL-Liver function	HBV Virus indicators	Imaging detection pathology	Alpha fetoprotein test	Liver examination
1	●	●	●	●	●	●	●	●	●	●
2	●	●	●	●	●	●	●	●	●	●
3	●	●	●	●	●	●	●	●	●	●
4	●	●	●	●	●	●	●	●	●	●
5	●	●	●	●	●	●	●	●	●	●
6	●	●	●	●	●	●	●	●	●	●
7	●	●	●	●	●	●	●	●	●	●
8	●	●	●	●	●	●	●	●	●	●
9	●	●	●	●	●	●	●	●	●	●
10	●	●	●	●	●	●	●	●	●	●
11	●	●	●	●	●	●	●	●	●	●
12	●	●	●	●	●	●	●	●	●	●
13	●	●	●	●	●	●	●	●	●	●
14	●	●	●	●	●	●	●	●	●	●

Fig. 4. Data collected by AI tools automatically mapped to eCRF in REDCap according to the set rules.

2.5. Stage 4: deployment and validation

Our CHB-EDC tool has been deployed at the Third Affiliated Hospital of Sun Yat-sen University and West China Hospital of Sichuan University. This study has collected clinical pathological and virological records of 4072 patients from the Third Affiliated Hospital of Sun Yat-sen University and West China Hospital of Sichuan University for evaluating the performance of our CHB-EDC tool in data collection. Four research assistants were recruited to annotate the data, and their analysis conclusions were compared. If there were significant differences in the analysis conclusions, a senior expert physician was invited to conduct further analysis, and the conclusions of the senior expert were considered final.

The evaluation metrics in this study included the average accuracy of data collection for paper medical records (calculated per field based on the eCRF design in this study), and the average time taken per record (in minutes), independent two-sample *t*-test was performed. The records were categorized into demographic information, vital signs, complete blood count, liver function, hepatitis B virus indicators, and pathological reports, totaling 6 categories.

3. Results

We conducted 10 experiments from October to December 2023, with the experimental group using CHB_EDC for data collection and the control group consisting of four test subjects for manual collection. Each experiment involved randomly sampling 100 paper medical records and comparing the accuracy and time taken for field recognition between manual collection and CHB_EDC collection. The experimental results showed that the average accuracy of manual collection was 98.65 %, with an average time of 63.64 min to collect information for one patient. The average accuracy of CHB_EDC collection was 98.66 %, with an average time of 3.57 min to collect information for one patient, as shown in Table 1. It is noteworthy that the time required for data collection using CHB_EDC consists of two parts: one part is the time used for model inference, and the other part is for manual verification of data quality and error correction. CHB_EDC supports highlighting the positions where field boxes are empty and their corresponding text, assisting staff in quickly verifying them. In the same data collection task, CHB_EDC achieved a comparable average accuracy to manual collection. However, in terms of time, CHB_EDC significantly outperformed manual collection ($p < 0.05$), saving an average of 60.07 min per medical record data collection. CHB_EDC, built upon the REDCap tool, is an OCR and NLP based tool specifically designed for non-electronic and non-structured data. It inherits the capabilities of data digitization and structured processing, greatly reducing the cost of data collection and preprocessing.

4. Discussion

This research is the first to develop a customized tool for the collection and standardization of clinical data for chronic hepatitis B. Previous research has applied OCR and NLP technologies to medical data processing with good results. Building on prior work, we evaluated the practical needs of collecting and standardizing clinical data for chronic hepatitis B, considering algorithm complexity (computational requirements), development challenges, cost, and efficiency. As a result, we developed an economically practical tool for data collection and structuring in chronic hepatitis B, known as CHB-EDC.

Table 1
Comparison of Accuracy between manual collection and CHB_EDC collection verification. SD: standard deviation.

Data Types	Manual model	CHB_EDC facilitated manual model
	Mean accuracy (SD) %	Mean accuracy (SD) %
Demographic information	98.99(0.98)	98.70(0.86)
Vital signs	98.64(0.90)	98.51(0.92)
Complete blood count	98.68(0.71)	98.61(0.78)
Liver function	98.53(1.06)	98.67(1.00)
Hepatitis B virus markers	98.80(0.85)	98.74(0.92
Pathological reports	98.28(1.09)	98.72(0.90)
Total	98.65(0.95)	98.66(0.90)

AI holds tremendous potential in the field of liver disease research, but currently, it is primarily focused on disease diagnosis and prognosis prediction [16,17]. Most studies do not address issues related to data collection and organization [18]. These predictive models require substantial manual data standardization from the modeling stage to the application and inference stage, which is not conducive to the subsequent research and use of the models.

In clinical research on chronic hepatitis B, our CHB-EDC tool serves as an auxiliary tool for early-stage data collection and standardization. Considering requirements such as cost and accuracy, we have chosen to optimize existing mature, open-source OCR and NLP models. Medical large language models have great potential in the field of clinical text data standardization processing. However, these models have a very large number of parameters and have not yet been adapted for the application of data collection and standardization in the context of chronic hepatitis B. The cost of development and training is very high, so large language models are not being used. For example, Med-PaLM performs exceptionally well in answering medical questions [19], PubMedGPT incorporates millions of full-text articles and abstracts in its training to interpret biomedical research language and assist in medical research [20]. GatorTron, developed by the University of Florida, is an electronic health record (EHR) big data model with 8.9 billion parameters, although much smaller than Med-PaLM. Currently, the application of large language models in medicine is still being explored and does not yet support specific fields, such as the application of data collection and standardization for chronic hepatitis B [21]. The high development and maintenance costs are also difficult for the majority of medical research teams to bear, so we need to choose traditional NLP solutions. The traditional NLP models include the BERT model and its variant models, with ERNIE-Layout being one of them. ERNIE-Layout incorporates a large amount of Chinese information material in the training corpus, and we only need to fine-tune it to achieve the structured and standardized clinical text data for chronic hepatitis B. The OCR model in CHB-EDC is used for the electronic collection of printed data from patient paper medical records, examination reports, and other documents. The features of these documents are white background with black text, and they hardly present challenges such as blurry images, complex backgrounds, messy handwriting, or complex semantic characters, which are common in cutting-edge OCR research [22]. Therefore, the recognition difficulty is not significant. Lianchi Zheng et al. compared the performance of Bert without additional coordinate information and LayoutXLM with additional coordinate information to improve medical OCR information extraction [23]. The open-source PaddleOCR model has achieved an accuracy rate of 99.5 %, and it can be directly invoked without the need for re-modeling and retraining.

In the management of hepatitis B patients, timely and accurate data is crucial for developing appropriate treatment plans. We have developed the CHB-EDC tool, which incorporates OCR and NLP technologies to enable the digitization and structured extraction of paper medical records and laboratory reports. This greatly simplifies the data collection and organization process. Traditional manual data collection requires personnel to enter, modify, and verify each data point, while CHB_EDC only requires partial manual data collection for model training and optimization. This allows for the creation of a pipeline for batch processing, significantly reducing the time required for data collection. Compared to traditional manual data collection, the average time for CHB_EDC data collection and verification is 3.57 min, which is 17.83 times faster than the traditional manual collection method, which takes an average of 63.64 min. Once successfully deployed, CHB_EDC can be reused multiple times, supporting batch processing for large-scale multicenter data collection needs. This significantly reduces time costs and lowers the difficulty of data collection in clinical research involving a large number of multicenter patient data.

In terms of accuracy, we compared the error rates of six different data types and found that when CHB_EDC data collection was combined with manual review, the accuracy was comparable to manual data collection. This indicates that CHB_EDC can achieve an average accuracy rate of 98.66 %, which is comparable to the accuracy of manual data collection at 98.65 %. This offers the potential to replace traditional manual operations and reduce repetitive labor for researchers. During the modeling and validation process, we found some methods that could improve accuracy. For example, (1) incorporating synonym mapping to address the issue of slight variations in vocabulary used in medical records from different hospitals, (2) including data from multiple hospitals in the training dataset to prevent potential overfitting issues from single-center data, and (3) supplementing with manual review for data collection, which can assist in data quality audit by highlighting text and field boxes. In this study, a total of 4072 medical records from Sun Yat-sen University Third Affiliated Hospital and Sichuan University West China Hospital were included for modeling and validation testing. The CHB-EDC tool effectively supported data collection for chronic hepatitis B patients from both hospitals.

On the other hand, by digitizing paper-based data through OCR technology, we preserve the integrity of the original data and use NLP technology to structure unstructured data, making it easier to manage and trace. This provides a reliable data foundation for long-term follow-up and treatment evaluation of hepatitis B patients. To protect sensitive data involving patient privacy, we have introduced encryption and de-identification algorithms to defend against potential privacy and security risks, ensuring that patient information is not misused.

However, this study also has certain limitations. Firstly, the use of AI technologies, such as OCR and NLP, requires high hardware and software requirements during model training and optimization. The technical capabilities and equipment differences among different medical institutions may affect the generalization and application of the methods. Additionally, although OCR and NLP technologies have made significant progress, there are still error rates when dealing with complex and diverse data in the medical field. Specifically, when processing unstructured data, the model's adaptability to specific domains needs further optimization. Next, we plan to optimize the performance and expand the functionality of our data collection tool from two aspects. Firstly, we have signed contracts with eight hospitals in China, and CHB-EDC will collect 200 medical records from each of these hospitals, totaling 1600 medical records for model optimization and validation. Finally, CHB-EDC will be deployed in the partner hospitals for subsequent joint research project data collection. On the other hand, we will use the same technical approach to develop an intelligent data collection tool for hepatitis B-related liver cirrhosis and liver cancer, in response to the research team's demand for data collection on other liver diseases. We have uploaded the project code to the GitHub open-source community (https://github.com/Clara10086/CHB-EDC) for reference by future researchers and staff involved in chronic hepatitis B data collection.

5. Conclusion

Based on the Chinese Chronic Hepatitis B Prevention and Treatment Expert Guidelines and the international CDISC data standards, we have developed a CHB-EDC system that incorporates OCR and NLP technologies to expand the functionality of the REDCap non-electronic data and non-structured data collection, transformation, and management tool. CHB-EDC is capable of automatically processing various formats of data, including raw data in image format, through CDISC standards, OCR, and NLP models. It can automatically populate data from image and other format source data into eCRF, while the system also includes built-in de-identification and encryption tools to protect data privacy and security. In the context of a large number of hepatitis B patients and complex disease information, the introduction of artificial intelligence technology makes it more efficient, secure, and accurate to extract key information from a large amount of unstructured data. This study provides an innovative means of data collection and processing for hepatitis B patient management and related research, offering a feasible solution for multicenter data integration and serving as a reference for other chronic disease research.

Ethics statement

This study was supported by the 5010 Cultivation Program of Clinical Research of Sun Yat-Sen University (2018024), Guangdong Province Rural Science and Technology Assistance (Special Commissioner) Project (No. KTPYJ2022012).

CRediT authorship contribution statement

Xiaomei Zhou: Writing – original draft, Investigation, Formal analysis. **Tao Zeng:** Writing – original draft. **Yibo Zhang:** Writing – original draft. **Yingying Liao:** Methodology. **Jaime Smith:** Methodology. **Lin Zhang:** Software, Formal analysis. **Chao Wang:** Software. **Qinghai Li:** Methodology. **Dongbo Wu:** Conceptualization. **Yutian Chong:** Conceptualization. **Xinhua Li:** Project administration, Conceptualization.

Declaration of competing interest

The authors of this article entitled "Automated Data Collection Tool for Real-World Cohort Studies of Chronic Hepatitis B: Leveraging OCR and NLP Technologies for Improved Efficiency" declare that they have no conflicts of interest to disclose.

References

[1] Global progress. report on HIV, viral hepatitis and sexually transmitted infections 2021. https://www.who.int/publications/i/item/9789240027077.

[2] Su S, Wong WC, Zou ZR, Cheng DD, Ong JJ, Chan PL, Ji FP, Yuen MF, Zhuang GH, Seto WK, et al. Cost-effectiveness of universal screening for chronic hepatitis B virus infection in China: an economic evaluation. Lancet Global Health 2022;10(2): E278–87.

[3] Yousefi A, Mastouri N, Sartipi K. Scenario-oriented information extraction from electronic health records. In: 2009 22nd IEEE international symposium on computer-based medical systems; 2-5 Aug. 2009 2009. p. 1–5.

[4] Karthikeyan S, Herrera AGSd, Doctor F, Mirza A. An OCR post-correction approach using deep learning for processing medical reports. IEEE Trans Circ Syst Video Technol 2022;32(5):2574–81.

[5] Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, Soni S, Wang Q, Wei Q, Xiang Y, et al. Deep learning in clinical natural language processing: a methodical review. J Am Med Inf Assoc 2020;27(3):457–70.

[6] Dudchenko A, Dudchenko P, Ganzinger M, Kopanitsa G. Extraction from medical records. Stud Health Technol Inf 2019;261:62–7.

[7] Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, McLeod L, Delacqua G, Delacqua F, Kirby J, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inf 2019;95: 103208.

[8] AcoCPAoRWSoC Medicine. Chinese practice algorithm on real world study of clinical medicine. Chinese Journal of Experimental and Clinical Infectious ((Electronic Edition) 2017;11(6):5.

[9] Kuchinke W, Aerts J, Semler SC, Ohmann C. CDISC standard-based electronic archiving of clinical trials. Methods Inf Med 2009;48(5):408–13.

[10] Chinese Society of Hepatology CMACSoID, Chinese Medical Association. Zhonghua Gan Zang Bing Za Zhi 2022;30(12):1309–31 [Guidelines for the prevention and treatment of chronic hepatitis B (version 2022)].

[11] [The guidelines of prevention and treatment for chronic hepatitis B (2019 version)]. Zhonghua Gan Zang Bing Za Zhi 2019;27(12):938–61.

[12] Du Y, Li C, Guo R, Yin X, Liu W, Zhou J, Bai Y, Yu Z, Yang Y, Dang Q, et al. PP-OCR: a practical ultra lightweight OCR system. 2020.

[13] Peng Q, Pan Y, Wang W, Luo B, Zhang Z, Zhengjie H, Hu T, Yin W, Chen Y, Zhang Y, et al. ERNIE-layout: layout knowledge enhanced pre-training for visually-rich document understanding. 2022.

[14] Adedeji KB, Nwulu NI, Aigbavboa C, Gbadamosi SL. Ieee: assessment of encryption and decryption schemes for secure data transmission in healthcare systems. In: IEEE AFRICON conference - powering africa s sustainable energy for AD agenda - the role ofICT and engineering: sep 25-27 2019; accra, Ghana. NEW YORK: Ieee; 2019.

[15] Sun ZK, Wang YL, Shu ML, Liu RX, Zhao HQ. Differential privacy for data and model publishing of medical data. IEEE Access 2019;7:152103–14.

[16] Kalapala R, Rughwani H, Reddy DN. Artificial intelligence in hepatology- ready for the primetime. J Clin Exp Hepatol 2023;13(1):149–61.

[17] Su TH, Wu CH, Kao JH. Artificial intelligence in precision medicine in hepatology. J Gastroenterol Hepatol 2021;36(3):569–80.

[18] Nam D, Chapiro J, Paradis V, Seraphin TP, Kather JN. Artificial intelligence in liver diseases: improving diagnostics, prognostics and response prediction. JHEP Rep 2022;4(4):100443.

[19] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172–80.

[20] Bolton E, Venigalla A, Yasunaga M, Hall D, Xiong B, Lee T, Daneshjou R, Frankle J, Liang P, Carbin M, et al. BioMedLM: a 2.7B parameter language model trained on biomedical text. ArXiv 2024:18421. abs/2403.

[21] Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Costa AB, Flores MG, et al. A large language model for electronic health records. NPJ Digit Med 2022;5(1):194.

[22] Memon J, Sami M, Khan R, Uddin M. Handwritten optical character recognition (OCR): a comprehensive systematic literature review (SLR). IEEE Access 2020. 1-1.

[23] Lianchi Z, Liu X, Sun Z, He Y. Improving medical OCR information extraction with integrated Bert and LayoutXLM models. 2023. p. 164–73.