




BMJ Open Existing barriers and recommendations of real-world data standardisation for clinical research in China: a qualitative study

Junkai Lai ¹, Xiwen Liao,¹ Chen Yao,^{1,2} Feifei Jin ³, Bin Wang ¹, Chen Li,⁴ Jun Zhang,⁵ Larry Liu^{6,7}

To cite: Lai J, Liao X, Yao C, *et al.* Existing barriers and recommendations of real-world data standardisation for clinical research in China: a qualitative study. *BMJ Open* 2022;**12**:e059029. doi:10.1136/bmjopen-2021-059029

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2021-059029>).

Received 09 November 2021
Accepted 07 July 2022



© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Peking University Clinical Research Institute, Peking University First Hospital, Beijing, China

²Hainan Institute of Real World Data, Qionghai, Hainan, China

³National Center for Trauma Medicine, Peking University People's Hospital, Beijing, China

⁴Department of Health Statistics, School of Preventive Medicine, Fourth Military Medical

University, Xi'an, Shaanxi, China

⁵CORE, MSD China Ltd, Beijing, China

⁶Merck & Co Inc, Rahway, New Jersey, USA

⁷Weill Cornell Medical College, New York City, New York, USA

Correspondence to

Chen Yao;
yaochen@hsc.pku.edu.cn

ABSTRACT

Objective To investigate the existing barriers and recommendations of real-world data (RWD) standardisation for clinical research through a qualitative study on different stakeholders.

Design This qualitative study involved five types of stakeholders based on five interview outlines. The data analysis was performed using the constructivist grounded theory analysis process.

Setting Eight hospitals, four hospital system vendors, three big data companies, six medical products companies and four regulatory institutions were included.

Participants In total, 62 participants from 25 institutions were interviewed through purposive sampling.

Results The findings showed that the lack of clinical applicability in existing terminology standards, lack of generalisability in existing research databases, and lack of transparency in existing data standardisation process were the barriers of data standardisation of RWD for clinical research. Enhancing terminology standards by incorporating locally used clinical terminology, reducing burden in the usage of terminology standards, improving generalisability of RWD for research by using clinical data models, and improving traceability to source data for transparency might be feasible suggestions for solving the current problems.

Conclusions Efficient and reliable data standardisation of RWD for clinical research can help generate better evidence used to support regulatory evaluation of medical products. This research suggested enhancing terminology standards by incorporating locally used clinical terminology, reducing burden in the usage of terminology standards, improving generalisability of RWD for research by using clinical data models, and improving traceability to source data for transparency to guide efforts in data standardisation in the future.

INTRODUCTION

Real-world data (RWD) are data relating to patient health status or the delivery of health-care collected from a variety of sources such as electronic health records (EHRs).^{1–4} Internationally, especially in the USA and in China, RWD have become increasingly used to support regulatory decision making for drugs

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ Wide variety of relevant stakeholders on the subject.
- ⇒ Qualitative understanding of a major industry bottleneck.
- ⇒ Important recommendations that can guide the direction of the future of the subject.
- ⇒ Due to COVID-19, a portion of the interviews were not done in person and might limit the ability to read into the participants response for further exploration of the subject.
- ⇒ Recruitment of participants were limited to those that were already exploring the subject, which could result in selection bias.

and medical devices.^{1 5} In September 2019, China's National Medical Products Administration (NMPA) proposed to accelerate the approval process for advanced medical products listed abroad through the collection of RWD from patients using these products in Boao Lecheng Pilot Zone.^{6 7} The proposal has prompted medical products companies to conduct clinical research in Boao Lecheng using RWD, specifically the patient visit data collected in electronic medical records (EMRs), as real-world evidence for domestic product approval. An example of the first products to leverage the approval process included Johnson & Johnson's femtosecond ophthalmic surgical medical devices which started data collection in October 2019 and subsequently gained approval after 14 months.⁸ As more products being introduced into Boao Lecheng, there is an imminent need to efficiently translate the data within EMRs to clinical research data.

A current problem in China is that EMRs constitute a separate system that is not able to be directly connected to electronic data capture (EDC) system used for clinical research data collection, leading to the duplicative and manual transcription of EMR data

into the EDC system.^{9 10} The inefficient process results in poorer data quality due to the likelihood of human error and insufficient source data verification.¹¹ Solutions to the issue have been explored by the US Food and Drug Administration (FDA), which includes promoting the direct usage of electronic source data (eSource) from RWD systems for clinical research.^{12 13} In the eSource guidance, a key recommendation is to use data standards for the exchange of data to increase interoperability between EHR and EDC systems. In addition, initiatives led by the FDA promoted collaboration between standards organisations, such as Health Level Seven (HL7) and Clinical Data Interchange Standards Consortium, which produced solutions harmonising the differences between EHR data standards and clinical research data standards.¹⁴

However, these solutions are not directly translatable to China's clinical research context due to differences in the developed data standards. The data standards in China were developed by the Statistical Information Centre of the National Health Commission and used to evaluate the interoperability of hospital information systems.^{15–18} The first qualitative study on the problem of the gap between RWD and clinical research found several key problems, which included the lack of data standards usage, prevalence of unstructured data and data security concerns.¹⁹ Similarly, a literature review in China revealed that meaningful usage of RWD for clinical research is deterred by weak regulatory implementation of semantic level data standards, prevalence of unstructured data, and difficult hospital data access.²⁰ It is urgently important to address the standardisation of RWD for clinical research in China. However, limited literature and stakeholder opinion on the issue exist and have yet to be explored in China. Therefore, our research aimed to explore the barriers and recommendations regarding the standardisation of RWD for clinical research in China through a qualitative study conducted on industry-wide stakeholders.

METHODS

Design

Qualitative research allows us to understand a participant's experience through qualitative methods of capturing data such as the usage of interviews. Grounded theory is a qualitative research method used in research areas that are unexplored or under explored to inductively generate theory from data grounded in the perceptions of the participant.²¹ The method's extensive usage in healthcare research can be attributed to its systematic process of coding and analysis that allows important themes to emerge from the data, regarding the problems faced by participants and their resolutions toward these problems.²² Constructivist grounded theory (CGT) assumes that data are coconstructed through the researcher–participant interaction, and the product of analysis is influenced by the interaction of the researcher with the data.^{23 24} This study aimed to examine

an underexplored subject, the barriers experienced by stakeholders in the standardisation of RWD for clinical research and their recommendations in the context of China. Therefore, a qualitative research strategy guided by CGT was employed.

The research team conducted in-depth interviews with participants. The interviews were conducted between September and November 2021. The study is reported according to the Consolidated Criteria for Reporting Qualitative Research guidelines.²⁵

Participants selection

The selection of participants was based on the type of stakeholders involved in the construction of the regional data platform in Boao Lecheng, which aimed at the standardisation of RWD for clinical research. The type of stakeholders included participants from hospitals that generated RWD, hospital system vendors that installed EMRs, big data companies that centralised RWD onto a data platform, medical product companies that accessed RWD for clinical research and regulatory departments that evaluated the RWD used in clinical research. The type of stakeholders was categorised into three general categories: stakeholders that mainly affected the source data, stakeholders that mainly affected the standardisation of source data for clinical research and stakeholders that mainly affected the validity of RWD used for regulated clinical research. Hospital and hospital system vendors represented the first category, big data companies represented the second category, and medical products and regulatory departments represented the third category.

A stratified purposive sampling method was used to select representatives from each of the five stakeholder roles.^{26 27} Simultaneous data collection and analysis were conducted to determine when there was no longer new coding information generated for each role and the interviewing of participants stopped.²⁸ The resulting number of participants interviewed in the study at information saturation included 25 institutions with a total of 62 participants, which included no participant dropouts. YC and JL contacted the interviewees and briefed them on the subject matter of the investigation before the participants agreed to be arranged for an interview. Interviewees represented their own opinions based on their experience working at the institution and do not represent the institution. The number of participants interviewed for each type of stakeholder is shown in [table 1](#).

Table 1 Demographics of the participants

Type of stakeholder (# of institutions)	Total no of participants
Hospital (8)	16
Hospital system vendor (4)	10
Big data company (3)	15
Pharmaceutical (6)	12
Regulatory (4)	9

Detailed list of institutions for each type of stakeholder is included in see online supplemental appendix 1.

The inclusion criteria of the interviewees were as follows

Inclusion criteria

1. Participants who had extensive experience as a staff member at stakeholder's institution.
2. Participants who had experience evaluating RWD for clinical research for the institution.

Exclusion criteria

1. Participants who could not sign informed consent form.
2. Participants who could not provide at least 45 min for an interview.

Setting

The research team with training and experience in qualitative methods conducted interviews using a phone or in person. A quiet meeting room was chosen for each interview to allow for better recording of the study data. Each interview included only the participant and researchers.

Data collection

Semistructured interviews were recorded either over the phone or in person through a phone application with the ability to transcribe audio into text files.^{29 30} Field notes were taken to summarise important findings during the interview process, which helped guide later coding. A focus group interview was arranged instead of one-on-one interviews to promote discussion and communication for certain participants.³¹ Focus groups were used often for hospital and big data teams given the collaborative nature of the work and the tight schedules. Up to three people were involved in a single focus group. Each interview allowed 60 min, and basic information, including the interview time, place and interviewee, was collected at the beginning of the interview. Five sets of interview guides, designed for the five types of stakeholder roles, were pilot tested beforehand with similar participants that were not included in the study to make the flow of questioning better. Full interview guides are included in the appendix along with general categories that motivated these questions (see online supplemental appendices 2 and 3). The general categories of questions used for each role focused on how the stakeholders affected the data standardisation process at the source, from the source to research data and during evaluation at the research data. The interview questions guided the interviewer in exploring the subject with the participant. Further discussion on the questions or repeated interviews were allowed to explore deeper into the topic or for better clarification. Simultaneous data collection and analysis were determined when information saturation had occurred for each role, which implied that the interviewing of participants ended.

The interviewers were four doctoral students. JL (male) and XL (female) were mainly responsible for the interviews. BW (male) and FJ (female) played supportive roles and were mainly responsible for the recording of

interviews. The interviewers were trained in a qualitative research course and had previous experience conducting interviews.

Analysis

All interviews were transcribed to text using the automated transcription software and double checked by the two interviewers (JL and XL). Coding and memoing were done by three researchers (JL, XL, FJ) who drew on the techniques of CGT when they analysed the data. QSR NVivo V.12 software was used for coding. The team developed a structured coding tree based on the interviews that started with inductive open coding. Once the core categories emerged, deductive selective coding was performed. Memos were used to assist the researchers during the entire analysis process to help them understand the data, critique the codes, and identify the theoretical categories that the data represented. Open coding was performed independently by two researchers, and the derived core categories were compared in multiple rounds of discussions until all three research members (JL, XL and FJ) agreed. Participants did not provide feedback on the findings.

Patient and public involvement

There was no patient or public involvement in this research.

RESULTS

Barriers and recommendations in the standardisation of RWD for clinical research

The CGT framework generated from the three stages of coding and the 62 participants' responses were summarised in the flow chart (figure 1). The study found three main barriers and four main suggestions. The barriers included lack of clinical applicability in existing terminology standards, lack of common data elements in existing databases and lack of transparency in existing data standardisation processes. The recommendations included enhancing terminology standards by incorporating locally used clinical terminology, reducing burden in the usage of terminology standards, improving applicability of databases using clinical data models (CDM), and improving traceability to source data for transparency.

Causes

Lack of clinical applicability in existing terminology standards

The findings showed that hospital and hospital system participants have expressed the lack of applicability of terminology standards in the clinical setting. Clinicians expressed that terminology standards such as International Classification of Diseases, Tenth Revision (ICD-10) are not granular enough to reflect the diagnosis that they want to make. In addition, they expressed that terminology standards often use technical expressions that are not commonly used by physicians, making the search process for terminology burdensome. Therefore,

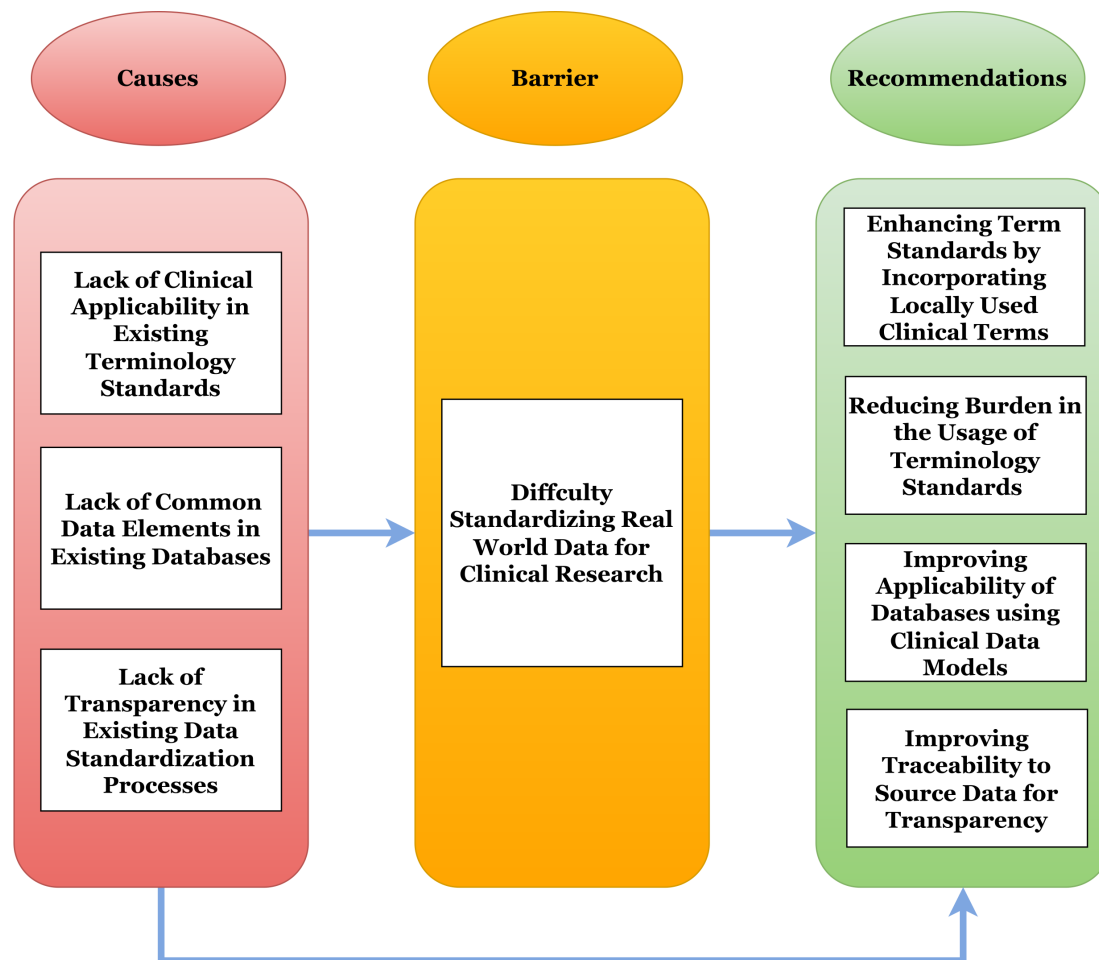


Figure 1 Caption barrier and suggestions in data standardisation of real-world data for clinical research.

clinicians expressed that they often use the ‘other’ option to input their own answers. Hospital system participants expressed that they often must implement custom made terminology lists created by the hospital instead of using default terminology standards to improve the usability of the system.

‘We give our clients default standards to use, but they may feel that the standards do not match their needs and will ask us to perform more customizations’—Hospital Information System Vendor Participant 1

‘When implementing standard terminology for the diagnosis field, doctors often just fill in their own answers in the ‘other’ option’—Hospital Participant 8

Lack of common data elements in existing databases

The findings showed that medical product companies and regulatory departments expressed that the existing RWD databases such as disease specialty databases formed by hospitals are standardised to specific research questions and not generalisable to others. Medical product participants expressed that there is substantial variation in the type of available data even when standardised. This resulted in the inability to leverage multiple databases together to answer a specific clinical research question due to differences in available data and their definitions. Regulatory department participants also expressed

similar views regarding the applicability of the existing RWD databases to support regulatory decision making regarding medical products. Currently, the existing data were not organised in a way that could be combined into a generalisable research database used to address multiple regulatory questions by different departments.

‘For feasibility studies, we may look at disease specialty databases. Although data are standardized for clinical research, the data elements in these databases are usually very different from each other, and we may have to focus on data elements that are more widely available to conduct our studies’—Medical Products Participant 7

‘Beside our department, other departments are also using RWD in specific datasets. There is currently no general platform that can organize RWD to be used by multiple departments to support regulatory decision making. Developing such a platform may be in our interest’—Regulatory Participant 3

Lack of transparency in existing data standardisation process

The findings showed that hospital and medical product participants expressed that the data standardisation process from RWD to clinical research data lacks transparency. Medical product participants expressed that they can use data completeness as well as other metrics to determine the quality of the data, but the exact methods

used for data standardisation are not transparent. In addition, they had concerns over the interpretability of standardisation methods such as natural language processing algorithms in extracting relevant research data and the determination of whether regulatory institutions would accept these methods. Hospital participants also expressed that inaccurate data produced by external vendors are difficult to correct or target due to the unknown methods used to transform the data. As the producers of research data, big data participants expressed that the standardisation process requires many steps and teams involved, which can reduce its transparency.

‘The exact methods used for data standardisation in producing research databases from RWD are not very transparent. My concerns for the usage of hard to interpret artificial intelligence algorithms for the extraction and standardisation of data are whether regulatory institutions will accept them’—Medical Products Participant 4

‘When vendors standardise our data into research data, the produced data may sometimes be inaccurate. We are not able to understand the methods used in standardization and find the reasons why the data may be incorrect’—Hospital Participant 9

‘Data standardization may require many teams and communication between many systems, which can lead to reduced transparency in the process and make the methods used hard to document comprehensively’—Big Data Participant 5

Suggestions

Enhancing terminology standards by incorporating locally used clinical terminology

The findings showed that big data companies and hospital information system participants suggested that the incorporation of their collection of locally used clinical terminology can improve the coverage of the existing terminology standards. Big data participants expressed the need to add and map RWD terminology found in their databases to standard terminology to enhance current terminology standards. Hospital system participants expressed that they have collected practical terminology lists from different hospitals that are used instead of default standard terminology lists. In addition, they expressed that the choice to use local lists in a clinical setting is to improve better departmental communication and may be a key component in the revision of terminology standards.

‘When working to develop different research databases, our team has incorporated medical experts that help us aggregate common terminologies that are synonyms with standard terminology into a library. Using the library will help search for relevant RWD’—Big Data Participant 15

‘Standards will get adopted if they can be easily used by our clients. Through our experience working with hospitals, we have collected terminology lists that are used often instead of standard terminology lists due to its ability to improve communication within hospitals’—Hospital Information System Vendor Participant 4

Reduce burden in the usage of terminology standards

The findings showed that hospital participants expressed that the efficiency of the usage of data standards can be improved by using more automatic methods of terminology standardisation. Hospital participants expressed various methods used to automatically standardise terminology before and after the documentation phase. Before the documentation phase, hospital participants suggested that terminology standards can be pre-coordinated with more familiar terminologies before usage. After the documentation phase, terminology standards can be post-coordinated through natural language processing algorithms that can match local terminologies with standard terminology.

‘To facilitate the usage of standards during medical documentation, we may recommend more familiar terminologies used to display the terminology standards before documentation’—Hospital Participant 9

‘Doctors are unfamiliar with the different standards. We will usually work with companies that can use better technology such as terminology matching to help us standardize the data after documentation’—Hospital Participant 13

Improving applicability of databases using CDM

The findings showed that hospital system and big data participants expressed that the usage of CDM standards to organise RWD can improve the applicability of RWD to different clinical research questions or services. Hospital system participants expressed that the usage of HL7 RIM data model can facilitate the reuse of data for different services including clinical decision support services. Big data participants suggested the usage of the OHDSI data model to organise RWD for the answering of different clinical research questions. In addition, they suggested that research in different disease areas may require a further extension of the models by analysing where these models fail to capture specific types of data.

‘Learning from Huawei’s and Alibaba’s approach to organize their services, we are starting to apply the HL7 RIM (Health Level HL7 Reference Information Model) model to build a middle layer in which our different hospital systems can create their services. Eventually, we would like to use it to support clinical decision support systems’—Hospital Information System Vendor Participant 1

‘When we participate in more clinical studies, we find that the usage of data models such as OHDSI data model can be used to help organize data to answer multiple research questions. However, we may need to extend the data models for more specific diseases by analyzing gap between our schema and the sponsors research case report forms’—Big Data Participant 5

Improving traceability to source data for transparency

The findings showed that regulatory department and medical product participants suggested the improvement in the traceability to source data for better transparency in

the data standardisation process. Regulatory departments recommended that clinical research involving RWD should adhere to the Good Clinical Practice (GCP) principles which require that research data are traceable to its source data. In addition, aspects of a clinical trial management workflow to authenticate and monitor the quality of the data should be used to increase the confidence in the research data obtained. Medical product company participants suggested the usage of eSource methods that can standardise the transmission of source data and help meet regulatory expectations in terms of auditing the quality of source data used for clinical research.

‘The GCP principles should be upheld similarly when using RWD for clinical research. Applying aspects of the clinical trial workflow may be needed to raise the confidence in the quality of RWD collection.’ Regulatory Institution Participant 2

‘We have been searching for eSource tools/companies that can help us collect reliable source data for clinical research that can be easily audited and used as evidence for regulatory approval’—Medical Products Participant 7

DISCUSSION

The barriers and recommendations in the standardisation of RWD for clinical research are the research questions central to the current qualitative study. Through a CGT approach, the study found three main barriers and four main suggestions. The barriers included lack of clinical applicability in existing terminology standards, lack of common data elements in existing databases and lack of transparency in the existing data standardisation process. The recommendations included enhancing terminology standards by incorporating locally used clinical terminology, reducing burden in the usage of terminology standards, improving applicability of databases using CDM, and improving traceability to source data for transparency. The grounded theory used in the paper was applied to address a specific problem regarding the difficulty in RWD standardisation for clinical research. The use of the methods in grounded theory was to find the barriers and recommendation to the research problem, with the goal of applying the recommendations found to the barriers that similar stakeholders may face in China.

In this study, the first reason identified was the lack of clinical applicability of current China terminology standards. The current terminology standards do not fit the expressions commonly used by physicians in China and may be burdensome to use. Thus, it is important to enhance terminology standards by adding locally used clinical terminology as well as reduce the burden associated with using terminology standards. Internationally, the problem is addressed in many countries through the usage of SNOMED-CT as a comprehensive terminology for clinical application.³² The deficiencies of China’s EMR standards include its emphasis on the standardisation of data elements and limited focus on terminology standards, preventing meaningful exchange of information

at the semantic level.²⁰ Thus, researchers believed that the localisation and implementation of a comprehensive international terminology standard such as SNOMED-CT within EHRs could help represent clinically relevant information comprehensively in China.³³ However, previous translation of SNOMED-CT had been insufficient without the collection of terminology synonyms, since physicians did not follow the precise expressions in terminologies.³⁴ In contrast, local terminology datasets in China showed its ability to cover 74.8% of commonly terms used within EHRs.³⁵ Therefore, the recommendations to collect local terminology is particularly important to increase the clinical applicability of current terminology standards.

The other issue regarding clinical applicability of existing terminology standards is the burden associated with its usage. A literature review studying the impact of EHR data structures, such as coding systems, on clinical efficiency found conflicting results with some studies suggesting that structured data made work processes easier while other studies suggesting that coding and entering structured data was slower.³⁶ The study further explained that the perceived difficulties might be due to the lack of familiarity with the coding systems. Participants in our study suggested leveraging pre-coordination and post-coordination methods to use terminology standards without depending on a clinician’s familiarity with terminology standards. Pre-coordination is a strategy that constrains and maps coding systems to existing local terminology lists, allowing for the usage of local terminology lists without familiarity with external coding systems. A successful implementation of pre-coordination was demonstrated in Hong Kong by binding local terminology, the Hong Kong Clinical Terminology Table, to international terminology standards with the outcome of not influencing regular clinical workflow.³⁷ Post-coordination can be applied to existing terminology lists, but here the emphasis is its application to free text by using natural language processing algorithms to extract terms and match them with coding systems. Recent improvements in using NLP showed a 90% accuracy in the extraction and matching of Chinese clinical text terms to SNOMED-CT.³⁸ The success of these methods in their respective studies has demonstrated the capability of improving the efficiency of using terminology standards without impacting normal clinical workflow.

The second reason identified was the lack of generalisability in existing research databases. The lack of generalisability of databases can lead to the limited usage of RWD even after standardisation since the databases only address a specific question. Thus, the usage of CDM can improve the generalisability of databases by organising RWD in a consistent and research relevant way to enable the answering of research questions. In the USA, the same problem was first discovered in 2008 when met with the technical challenge surrounding the detection of 10 outcomes in 10 drug classes in a network of multiple databases in the Observational Medical Outcomes Partnership (OMOP) research network. The result was the

development of a generalisable common CDM that each database could conform, allowing for the efficient answering of clinical research questions.^{39 40} In 2021, HL7 and OHDSI (previously OMOP) collectively announced their initiative to create a CDM that integrated data standards common to EHRs with the goal of better organising EHR data into a clinical research data model.⁴¹ Although the usage of common data models in China has not been pushed by the government, the growing usage among big data companies and other research organisations is evident. Confirming the experiences of the participant in this study, research teams in China have found that even if the same clinical problem is studied, the heterogeneity of cohort studies in terms of variable definition and data collection hinders the integration and sharing of data for clinical research.⁴² The problem has been a motivating factor in the review of a suitable international CDM that can be used to address the heterogeneity in databases.⁴² Application of the OHDSI CDM in China in its first application to study chronic diseases at a single site has now expanded to its usage domestically to answer COVID-19 treatment questions using country-wide databases.^{43 44} In addition to the application of common data models, translational research and the development of tools to transform related domestic RWD standards, such as HL7 CDA, to common data models, such as OHDSI CDM, are ongoing in Korean and China.^{45 46}

The final reason was the lack of transparency in the existing data standardisation process. The lack of well-documented and understandable methods used in the data standardisation process can compromise the reliability of the data for clinical research. Thus, improving traceability of research data to the source data can help evaluate the quality of the standardise data, increase transparency, and meet regulatory expectations. Despite the importance of traceability requirements for regulated clinical research, it remains as a top data standard issue identified by the US FDA in the successful review of submitted data.⁴⁷ In response, the US FDA has promoted the use of electronic source data (eSource) including EHRs to enhance the traceability of research data and reduce errors in transcription in several guidance.^{12 13} The implementation of eSource has been researched by the Society of Clinical Data Management to satisfy regulatory expectations regarding data integrity principles.⁴⁸ Among the expectations is the emphasis on GCP ALCOA principles including the declaration of source data, usage of standards, real-time capture of data and automatic data quality checks. Further, the TransCelerate eSource initiative examined the slow adoption of eSource and found that the main reasons included the lack of standards usage and interoperability between EHRs and EDC systems.⁴⁹ In China, researchers have highlighted the need to increase the transparency of the data standardisation process through source data sharing and statistical analysis protocol publishing.⁵⁰ In addition, source data verification, which checks consistency between the research data and source data, is promoted with great

emphasis by the NMPA, where extreme deviations of the source data with research data may lead to legal repercussions.⁵¹ To address these issues, suggestions in China were made to develop and use an independent eSource platform for the storage and transmission of research source data to guard data integrity and increase transparency. The development and usage of such a platform was tested using RWD collected from the Catalys Precision Laser System medical device real world study in Boao Lecheng and showed great promise in its ability to efficiently transform data while guarding data integrity.^{52 53} In 2021, the National Health Commission of China solidified the need for the usage of a research source data management platform at medical institutions as a requirement for the conduct of clinical research.⁵⁴

The strength of the study was the selection of a wide and comprehensive range of stakeholder that better represented the issue in China. Several limitations of this study warranted attention. The participants included specific institutions that were selected to represent the perspective of different stakeholder roles. The unselected companies may have different views, which could result in selection bias. To minimise selection bias, stratified purposive sampling methods were used. Various key institutions were included, and information saturation was assumed to be achieved. In addition, the cultural background and experience of the authors may have influenced the interpretation of the data, although the interviewers had experience and training in conducting qualitative research.

CONCLUSION

The qualitative study investigated the barriers in RWD standardisation for clinical research based on CGT. This study found barriers including lack of clinical applicability in existing terminology standards, lack of common data elements in existing databases and lack of transparency in existing data standardisation process. Enhancing terminology standards by incorporating locally used clinical terminology, reducing burden in the usage of terminology standards, improving applicability of databases using CDM and improving traceability to source data for transparency may be feasible suggestions for solving the current problems. The findings can be used to promote the development of efficient and reliable methods for the data standardisation of RWD for clinical research. Furthermore, the contributions of the study can guide the usage of standards, support the implementation of eSource methods and facilitate the development of real-world evidence. In the future, we aim to use the suggestions in our study to develop and evaluate eSource tools in China that can standardise RWD for clinical research with efficiency and reliability. Second, we aim to use the themes discovered to improve communication among relevant stakeholder groups as well as use their collaborative opinion to improve the development of data standards that can facilitate the standardisation of RWD for clinical research.

Acknowledgements We thank all individuals who took the time to participate in our interviews.

Contributors JL, CY and CL designed the study. JL and XL collected the data. CY and JL contacted the respondents. JL, XL, FJ and BW analysed the data. JL and XL wrote the first draft of the manuscript. FJ, CY and CL revised the manuscript. JZ contributed to the concept and protocol development, implementation of study and review and revised the manuscript. LL contributed to concept development, implementation of study and review and revised the manuscript. All authors contributed to the interpretation of the data and editing of the manuscript and approved the final manuscript. CY had full access to all data in the study and had final responsibility for the decision to submit for publication. JL and CY are the guarantors.

Funding This work was supported by Department of Science, Technology and International Cooperation, National Medical Products Administration (no award/grant number) and Merck Sharp & Dohme, a subsidiary of Merck & Co, Rahway, NJ, USA (no award/grant number).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Consent obtained directly from patient(s)

Ethics approval Ethical approval was obtained from Peking University Institutional Review board (No. IRB00001052-21081). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Data are available upon reasonable request. Study protocol and original data are available on request by emailing the corresponding author.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Junkai Lai <http://orcid.org/0000-0002-2272-3870>

Feifei Jin <http://orcid.org/0000-0001-6500-0800>

Bin Wang <http://orcid.org/0000-0003-0012-9835>

REFERENCES

- US Food And Drug Administration. Use of real-world evidence to support regulatory decision-making for medical devices guidance for industry and food and drug administration staff; 2017.
- US Food And Drug Administration. Real-World evidence program; 2018.
- Corrigan-Curay J, Sacks L, Woodcock J. Real-World evidence and real-world data for evaluating drug safety and effectiveness. *JAMA* 2018;320:867–8.
- Sun X, Tan J, Tang L, *et al*. Real world evidence: experience and lessons from China. *BMJ* 2018;360:j5262.
- National Medical Product Association. Real world evidence supports the guiding principles of drug development and review; 2020.
- National Development And Reform Commission, National Health Commission. State administration of traditional Chinese medicine. implementation measures on supporting the construction of Boao Le Cheng international medical tourism pilot area; 2021.
- National Medical Products Association. The State Food and Drug Administration launched the scientific action plan for China's drug supervision; 2021.
- Johnson Johnson Surgical Vision. A real-world evidence study in China of the Catalys precision laser system; 2020.
- Blitz R, Dugas M, Design C. Conceptual design, implementation, and evaluation of generic and Standard-Compliant data transfer into electronic health records. *Appl Clin Inform* 2020;11:374–86.
- Matsumura Y, Hattori A, Manabe S, *et al*. Interconnection of electronic medical record with clinical data management system by CDISC ODM. *Stud Health Technol Inform* 2014;205:868–72.
- Wu Y, Yin D, Abbasi K. China's medical research revolution. *BMJ* 2018;360:k547.
- US Food And Drug Administration. Electronic source data in clinical investigations; 2013.
- US Food And Drug Administration. Use of electronic health record data in clinical investigations; 2018.
- The Office Of The National Coordination For Health Information Technology. Harmonization of various common data models and open standards for evidence generation to support patient-centered outcomes research; 2020.
- National Health Commission of the People's Republic of China. Electronic medical record sharing document specification; 2020.
- National Health Commission of the People's Republic of China. Administrative measures for hierarchical evaluation of application level of electronic medical record system; 2018.
- National Health Commission of the People's Republic of China. Standardized maturity evaluation scheme for hospital information interconnection; 2020.
- National Health Committee Of The People'S Republic Of China. Basic architecture and data standard of electronic medical record; 2009.
- Jin F, Yao C, Yan X, *et al*. Gap between real-world data and clinical research within hospitals in China: a qualitative study. *BMJ Open* 2020;10:e38375.
- Xie J, Wu EQ, Wang S, *et al*. Real-World data for healthcare research in China: call for actions. *Value Health Reg Issues* 2022;27:72–81.
- Byrne M. Grounded theory as a qualitative research methodology. *Aorn J* 2001;73:1155–6.
- Chapman AL, Hadfield M, Chapman CJ. Qualitative research in healthcare: an introduction to grounded theory using thematic analysis. *J R Coll Physicians Edinb* 2015;45:201–5.
- Metelski FK, Santos JLG, Cechinel-Peiter C, *et al*. Constructivist Grounded theory: characteristics and operational aspects for nursing research. *Rev Esc Enferm USP* 2021;55:e3776.
- Mills J, Bonner A, Francis K. Adopting a constructivist approach to grounded theory: implications for research design. *Int J Nurs Pract* 2006;12:8–13.
- Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care* 2007;19:349–57.
- Moser A, Korstjens I. Series: practical guidance to qualitative research. Part 3: sampling, data collection and analysis. *Eur J Gen Pract* 2018;24:9–18.
- Setia MS. Methodology series module 5: sampling strategies. *Indian J Dermatol* 2016;61:505–9.
- Sutton J, Austin Z. Qualitative research: data collection, analysis, and management. *Can J Hosp Pharm* 2015;68:226–31.
- Peters K, Halcomb E. Interviews in qualitative research. *Nurse Res* 2015;22:6–7.
- Whiting LS. Semi-structured interviews: guidance for novice researchers. *Nurs Stand* 2008;22:35–40.
- Britten N. Qualitative interviews in medical research. *BMJ* 1995;311:251–3.
- Bodenreider O, Cornet R, Vreeman DJ. Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. *Yearb Med Inform* 2018;27:129–39.
- Zhu Y, Pan H, Zhou L, *et al*. Translation and localization of SNOMED CT in China: a pilot study. *Artif Intell Med* 2012;54:147–9.
- Zhang R, Liu J, Huang Y, *et al*. Enriching the International clinical nomenclature with Chinese daily used synonyms and concept recognition in physician notes. *BMC Med Inform Decis Mak* 2017;17:54.
- Cheng Y, Jiang T, Deng L, *et al*. Research on the coverage of standard Chinese medical terminology to practical application. *Chinese Journal of Health Informatics and Management* 2020.
- Forsvik H, Voipio V, Lamminen J, *et al*. Literature review of patient record structures from the physician's perspective. *J Med Syst* 2017;41:29.
- Hung K, Lau M, Fung V. Successful implementation of terminology binding in Hong Kong Hospital authority. *Stud Health Technol Inform* 2019;264:1486–7.

- 38 Chen Y, Hu D, Li M, *et al*. Automatic SNOMED CT coding of Chinese clinical terms via attention-based semantic matching. *Int J Med Inform* 2022;159:104676.
- 39 Overhage JM, Ryan PB, Reich CG, *et al*. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* 2012;19:54–60.
- 40 Stang PE, Ryan PB, Racoosin JA, *et al*. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Ann Intern Med* 2010;153:600–6.
- 41 Observational Health Data Sciences Informatics. HL7 international and OHDSI announce collaboration to provide single common data model for sharing information in clinical care and observational research; 2021.
- 42 Yue HX, Zhan YL, Bian F, *et al*. [Data standard and data sharing in clinical cohort studies]. *Zhonghua Liu Xing Bing Xue Za Zhi* 2021;42:1299–305.
- 43 Prats-Urbe A, Sena AG, Lai LYH, *et al*. Use of repurposed and adjuvant drugs in hospital patients with covid-19: multinational network cohort study. *BMJ* 2021;373:n1038.
- 44 Zhang X, Wang L, Miao S, *et al*. Analysis of treatment pathways for three chronic diseases using OMOP CDM. *J Med Syst* 2018;42:260.
- 45 Ji H, Kim S, Yi S, *et al*. Converting clinical document architecture documents to the common data model for incorporating health information exchange data in observational health studies: CDA to CDM. *J Biomed Inform* 2020;107:103459.
- 46 OMAHA. Mapping with OMOP CDM; 2021.
- 47 Hume S, Sarnikar S, Becnel L, *et al*. Visualizing and validating metadata traceability within the CDISC standards. *AMIA Jt Summits Transl Sci Proc* 2017;2017:158–65.
- 48 Society for Clinical Data Management. eSource implementation in clinical research: a data management perspective; 2014.
- 49 Kellar E, Bornstein SM, Caban A, *et al*. Optimizing the use of electronic data sources in clinical trials: the landscape, part 1. *Ther Innov Regul Sci* 2016;50:682–96.
- 50 Xinyao J, Wenke Z, Junhua Z. Promote transparency in real-world study. *World Chinese Medicine* 2019.
- 51 Yan X, Dong C, Yao C. Protecting the accuracy of clinical trial data in China. *BMJ Opinion* 2018.
- 52 Dong C, Yan X, Tian R, *et al*. Strengthen the process report of clinical trials, promote full transparency of clinical trials. *Chinese Journal of Evidence-Based Medicine* 2018.
- 53 Jin F, Yao C, Ma J. Explore efficient and feasible clinical real world data collection mode in Hainan Boao Lecheng international medical tourism pilot zone. *China Food & Drug Administration Magazine* 2021.
- 54 National Health Commission of the People's Republic of China. Measures for the administration of clinical research initiated by researchers in medical and health institutions; 2021.