BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (http://bmjopen.bmj.com).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

BMJ Open

Existing Barriers and Recommendations of Real-World Data Standardization for Clinical Research in China: A Qualitative Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-059029
Article Type:	Original research
Date Submitted by the Author:	09-Nov-2021
Complete List of Authors:	Lai, Junkai; Peking University First Hospital, Peking University Clinical Research Institute, Liao, Xiwen; Peking University First Hospital, Department of Biostatistics Jin, Feifei; Peking University People's Hospital, Department of Trauma Medicine Wang, Bin; Peking University First Hospital, Department of Biostatistics Yao, Chen; Peking University First Hospital, Department of Biostatistics; Peking University Clinical Research Institute Li, Chen; Fourth Military Medical University, Department of Health Statistics; School of Preventive Medicine
Keywords:	QUALITATIVE RESEARCH, Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

SCHOLARONE™
Manuscripts

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

For peer review only

Existing Barriers and Recommendations of Real-World Data Standardization for Clinical Research in China: A Qualitative Study

Junkai Lai[1], Xiwen Liao[1], Feifei Jin[2], Bin Wang[1], Chen Yao[134], Chen Li[5]

1 Department of Biostatistics, Peking University First Hospital, Beijing, Beijing, China.
2 Peking University People's Hospital, Beijing, Beijing, China.
3 Peking University Clinical Research Institute, Beijing, Beijing, China.
4 Hainan Institute of Real World Data, Qionghai, Hainan, China.
5 Department of Health Statistics, School of Preventive Medicine, Fourth Military University, Xi'an, Shaanxi, China.

Correspondence to
Chen Yao
Peking University First Hospital, Xicheng District, Beijing 100034, China
Tel +86 18610640562
Email yaochen@hsc.pku.edu.cn

Word Count
3996

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Existing Barriers and Recommendations of Real-World Data Standardization for Clinical Research in China: A Qualitative Study

Abstract

Objective

To investigate the existing barriers and recommendations of real-world data standardization 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

8 Hospitals, 4 Hospital System Vendors, 3 Big Data Companies, 6 Pharmaceutical Companies, and 4 Regulatory Institutions were included.

Participants

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

Results

Real world data is difficult to standardize for clinical research in China. The main causes were difficulty integrating standards in routine clinical process, lack of generalizable research datasets, and difficulty auditing the standardization process. The main suggestions were better feedback loop for standards, better separation of standards and documentation, promoting generalizable clinical research data models, and improving traceability to source data.

Conclusions

Determining the barriers and recommendations could contribute to the development of clinical research in China. The key findings in the study make a clear point that data standardization relies on the consensus of many working groups. Data standards cannot be integrated at the source data unless it is customized for practical usage and can only be secondarily applied to source data adequately when the meaning of the source data can be clearly understood.

Keywords: real world data; data standards; data standardization; clinical research; qualitative research; China;

Strengths and Limitations of this study:

- Strength: Wide variety of relevant stakeholders on the subject
- Strength: Qualitative understanding of a major industry bottleneck

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

- Strength: Important recommendations that can guide the direction of the future of the subject
- Limitations: Due to COVID-19, a portion of the interviews were not done in person and may limit the ability to read into the participants response for further exploration of the subject
- Limitations: Recruitment of participants were limited by those that were already exploring the subject and cannot be generalized to participants that may not be familiar with the subject.

Introduction

Real World Data are data relating to patient health status or the delivery of health care routinely collected from a variety of sources.[1-4] Data standards related to real-world data sources (such as EMR data in China) have gradually evolved from the basic guidelines of clinical documentation to standard terminology usage and clinical data models.[5-7] Data standard usage is meant to improve the data quality of real-world data so that it can be further used for secondary purposes such as clinical decision support and clinical research.

Since 2008, the Informatization Leading Group of the Ministry of Health and other departments have organized relevant experts to write the basic architecture and data standard specification of the Electronic Medical Record (EMR).[6] By considering the basic guides for case writing that integrates Chinese and Western medicine and a survey on the common medical functions, the institution published a guidance for basic standards associated with EMRs.[7] Through several national strategies to promote EMR meaningful use and data sharing, China has gradually improved the standardization of real-world data.

In 2018, a performance rubric for the meaningful usage of EMRs were created to rate hospitals;[8] The policy pushed for all tertiary hospitals to reach grade 4 or above (hospital wide information sharing and primary medical decision support) and secondary hospitals to reach grade 3 or above (inter-departmental data exchange) by 2020 which has been overall achieved. Later in 2020 when EMRs have matured at hospitals, the Statistical Information Center of the National Health Commission further pushed for the data sharing capability of hospitals through another performance rubric.[9] The main guidance on data standards currently is based on a localized version of HL7 CDA (Health Level 7 Clinical Document Architecture) and international terminology standards such as ICD-10 (International Classification of Disease version 10).[6-7]

The future of clinical research will benefit greatly if EMR data is standardized and able to be shared. However, even if there are existing guidance on industry health data standards, it is unclear what the barriers or best methods are for real-world data standardization for clinical research. Drawing on our previous qualitative study of the gap between real world data and clinical research, data standards are one of the components that causes this gap.[10] This study seeks to address this question through a qualitative study by interviewing different stakeholders seeking to conduct clinical research using real-world data to benefit their decision-making process.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Methods

Design

Qualitative research allows us to understand a participants experience through qualitative methods of capturing data often through interviews. Constructivist grounded theory (CGT) provides a way for theory construction from qualitative data and is of the view that researchers do not discover theory but rather construct it through interaction and interpretation of the participant.[11-12] Exploring why real-world data is difficult to standardize for usage in clinical research is the concern of this study. Therefore, a qualitative research strategy guided by CGT was employed.

The research team conducted in-depth interviews with different institutions representing 3 categories of key stakeholders: stakeholders affecting the source data, stakeholders affecting the standardization of source data for research, and stakeholders affecting the validity of data for clinical research. The interviews were conducted between August and October 2021. The study is reported following the guidelines of the Consolidated Criteria for Reporting Qualitative Research (COREQ) guidelines.[13]

Participants Selection

We intended to interview three categories of stakeholders: stakeholders affecting the source data, stakeholders affecting the standardization of source data for research, and stakeholders affecting the validity of data for clinical research. Hospital and Hospital System Vendors represented the first category, Big Data Companies represented the second category, and Pharmaceutical and Regulatory Institutions represented the third category. Purposive sampling was used for the selection of participants which represented institutions that had a team working with real world data for clinical research and recommended experts.[14-15] The participants recruited for the study included 25 institutions with a total of 62 participants with no refusal in participation or dropped out. YC and JL contacted the interviewee through phone and briefed the subject matter and the objective of investigation before the participants agreed to be arranged for an interview. Interviewees represents their own opinions based on their experience working at the institution and do not represent the institution. The number of participants by the type of stakeholder is shown in shown in **Table 1.** Detailed list of institutions by type of stakeholder is included in the Appendix.

The inclusion criteria for the interviewees were as follows

Inclusion criteria

1. Participant had extensive experience as a staff member at stakeholder's institution
2. Participant has experience evaluating real world data for clinical research for the institution

Exclusion criteria

1. Participant could not sign informed consent form
2. Participant could not provide at least 45 minutes for an interview

Setting

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

The research team with training and experience in qualitative methods conducted interviews using a telephone or in person. A quiet meeting room was chosen for each interview to allow for better recording of the study data and exclusion of non-participants.

Data Collection

Semi-structured interviews were recorded either over the phone or in person through an iPhone app with the ability to transcribe audio into text files.[16-18] Field notes were taken to summarize important findings during the interview process that would help guide later coding. Depending on the stakeholder's time and willingness, a focus group interview was arranged instead of one-on-one interviews to promote discussion and communication.[19] The allowed time for interviews must not exceed 60 minutes in respect to the participants daily schedule. After reading a confidentiality and privacy statement and informing participants that interview will be recorded, the researcher will conduct interviews in accordance with steps already described in this protocol. In each interview, basic information of interview time, place and interviewee will be collected. Five sets of interview guides were designed for the five types of stakeholders of the interviewees and pilot tested before hand with similar participants not included in the study to make the flow of questioning better. Full interview guides are included in the Appendix. The interview questions guide the interviewer in exploring the subject with the participant. Further discussion on the questions or repeat interviews were allowed to explored deeper into the topic or for better clarification but in the study none occurred. Transcripts were not returned to the participants for correction. The collection of data was not limited by data saturation and finished once all participants were interviewed.

The interviewers were four doctoral students. JL (Male) and XL (Female) were mainly responsible for the interviews, and BW (Male) and FJ (Female) played supportive roles and were mainly responsible for the recordings. The interviewers were trained in a qualitative research course and had experience conducting interviews.

Sample Interview Guideline for the Hospitals

- This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
- Describe your role and experience facilitating clinical research at the Hospital?
- What are the motivating goals of clinical research?
- How do you determine the research dataset that are needed for your research?
- Do you think that electronic medical records or routine care data at the hospital are enough to accomplish your research?
- How do you aggregate and store all data from different hospital systems?
- How do you implement standards on your data?
- What areas in your clinical research process do you have to rely on external vendors to help you?
- How does data sharing happen for medical records inside and outside of the hospital?

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

- Beyond clinical research, have you standardized your data for other purposes?

Analysis

All interviews were transcribed to text using the automated transcription software and double checked for each recording by the two interviewers (JL and XL). Coding and memoing were done by the four researchers (JL, XL, FJ, BW) whom drew on the techniques of constructivist grounded theory while analyzing the data.[16-18] 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. Open coding was performed independently by the two researchers, and the derived core categories were compared in multiple rounds of discussions until all four research members (JL, XL, FJ, BW) agreed. Participants did not provide feedback on the findings.

Patient and public involvement

There was no patient or public involvement in this research.

Results

Barrier and Suggestions in Data Standardization of Real-World Data for Clinical Research

The CGT framework generated from the three stages of coding and the 62 participants' responses are summarized in the flow chart (figure 1). We found three causes that create the barrier in data standardization of real-world data, including difficulty integrating standards with routine clinical process, lack of generalizable research datasets, and difficulty auditing the standardization process. The main suggestions relating to the causes were found to be better feedback loop for standards, separation of data standards and documentation, promoting clinical research data models, and improving traceability to source data.

Causes

Hard to Integrate Standards with the Routine Clinical Process

Standard terminology libraries do not map well to the expressions used by physicians to describe the patient's conditions even when translated since they do not incorporate easily searchable colloquial terms. Standardized lists given as options will often lead to physicians using the "other" option to fill in their answers. Data standards are rarely utilized for the communication of data within hospitals and only standardized later on for research or regulatory data submission purposes; In addition, there are lots of isolated systems that may not be able to communicate with each other. Even if hospital system vendors want to use default standards incorporated into the system, hospitals will often reject them in favor of what they are familiar with. For research, specialty departments with hospitals rely on external vendors to create databases for them with standardized data; Physicians find it dissatisfying to give a search criterion and retrieve datasets that find patients that do not match their initial inclusion criteria due to the

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

granularity of the standards.

"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 System Participant 1

"When implementing standard answers for the diagnosis field, doctors often just fill in their own answers in the "other" option" – Hospital Participant 8

Lack of Generalizable Research Dataset

Pharmaceutical companies hoping to use real world data as part of their product development process find it to be a struggle. The usage of existing real-world data is often limited to lab and demographic data that may provide limited use. The developed hospital disease specialty databases by different big data vendors may differ greatly in terms of data definition and cannot be externally validated or aggregated to conduct studies. Regulatory institutions also express that the data shared by hospitals from data sharing policies are limited in usage and cannot be directly applied for decision making. Real world data currently lacks generalized data models that can be used for different therapeutic areas as well as regulatory evaluations.

"For feasibility studies, we may look into disease specialty databases. 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" - Pharmaceutical Participant 7

"Beside our department, other departments are also using real world data. Although data quality may not be great, there may still be important signals that can support the evaluation process. Developing a platform that can be used by multiple departments may be in our interest." – Regulatory Participant 3

Hard to Audit Data Standardization Process

Using existing databases, pharmaceutical companies may find it difficult to evaluate the quality of data based on plausibility measures alone. Without a way to guarantee traceability of data back to the source data, real world data is limited to conducting feasibility studies and hard to incorporate into clinical studies such using real world data as an external control group. Pharmaceutical companies also notice the usage of advanced algorithms for the transformation of real-world data to research datasets and find it concerning. The main concern is that new methodologies that risk data integrity may be rejected by regulatory institutions during evaluation. Big data companies find that advanced algorithms may produce unexpected and hard to explain variations in new data sources and are striving to reduce the complexity in data transformation.

"From our experience, our team has had to reduce the amount steps for the transformation of data from one model to another model and even resorted to merging

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

groups of IT staff working on different problems to clear up the standardization process" – Big Data Participant 5

"My concerns for the usage of artificial intelligence algorithms for the extraction and standardization of data are whether regulatory institutions will accept them." – Pharmaceutical Participant 4

Suggestions

Better Feedback Loop for Standards

Employees of hospital systems vendors will attend workshops and certification programs by HL7 China to help implement more standardized data communication between hospital systems. They find localization and extension of international standards to be the primary focus in their efforts and require a joint effort between participants from different institutions guided by standards organizations. Big data companies and hospital system vendors also have rich experience customizing hospital systems and standardizing local data and feel that they can enrich the development of a more suitable local standard.

"When working to develop different research databases, our team has incorporated medical experts that help us aggregate terminology libraries and understand the best way to search for data." – Big Data Participant 15

"Standards will get adopted if they make our systems communicate better with each other and services well-liked by our clients." – Hospital System Participant 4

Better Separation of Data Standards and Documentation

Good balance in separation of data standardization and routine documentation of data is an important strategy to reduce the burden of physicians during documentation. Hospitals will negotiate with vendors on mapping locally used terminology and standard terminologies in the backend of the system to alleviate differences while ensuring data can be standardized. In addition, hospital specialty departments may choose to implement recommended text or other methods that simplify documentation while enhancing consistency in data capture. Finally, hospitals will usually consult external vendors that can leverage technology such as natural language processing that deals better with standardizing electronic text data.

"Sometimes it is necessary for there to be some standardization during data collection because doctors who are busy will often elaborate very little about the patient. For some specialty departments, we may implement text recommendations to help standardize the documentation" – Hospital Participant 9

"Doctors are unfamiliar with the different standards. We will usually work with companies that can use better technology to help us standardize the data." – Hospital

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Participant 13

Promoting Clinical Research Data Models

Data is often needed for many projects or services in a given company and may benefit from having a data model that will reduce the repetition of work done to gather data. Big data and hospital system vendors have been developing their own data models inspired by approaches from international data models such as HL7, OHDSI (Observational Health Data Sciences and Informatics), CDISC (Clinical Data Interchange Standards Consortium), and other organizations to meet this goal. As more research is done using real world data, big companies have been able to gather core datasets needed for different types of clinical studies that they can use to increase the generalizability of their data model.

"Learning from Huawei's and Alibaba's approach to organize their services, we are starting to apply the HL7 RIM (Health Level 7 Reference Information Model) model to build a middle layer in which our different hospital system can create their services. Eventually we would like to use it to support clinical decision support systems" – Hospital System Participant 1

"When we participate in more clinical studies, we find better overlap between our schema and their research case report forms. Sometimes a therapeutic area may be very specific and we may need to extend our data model." – Big Data Participant 5

Improving Traceability to Source Data

Regulatory institutions recommend that real world data standardization for clinical research should adhere to GCP (Good Clinical Practice) principles of data integrity in which data traceability is a key focus. Furthermore, they suggest that better integration between the collection of real-world data and clinical trial data management processes will lead to better regulatory acceptance. Prompted by the concerns for data traceability, pharmaceutical companies are exploring methods of data capture that can meet regulatory expectation while reducing data collection efforts.

"GCP principles should be upheld similarly when using real world data for clinical research. Applying aspects of the clinical trial workflow may be needed to raise the confidence in real world data collection." – Regulatory Institution Participant 2

"We have been searching for eSource capability that can help us collect reliable data that can be easily audited and used as evidence for regulatory approval" - Pharmaceutical Participant 7

Discussion

This study investigated the existing problems that prevent the data standardization of real-world data for clinical research and further recommendation on each of these problems. Qualitative interviews were conducted on five types of stakeholders which

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

included hospitals, hospital information system vendors, big data companies, pharmaceutical companies, and regulatory institutions. The wide range of stakeholders were meant to better gauge industry views on the usage of real-world data for clinical research and their thoughts on data standardization.

Difficulty integrating standards to routine clinical process arise from the conflicts between clinical research and routine care source data collection. Data collected routinely is not meant to be expressed without variations or based on a research protocol for data collection; This raises an important question: what is the proper context for the implementation of data standards? The recommendations which include having a better feedback loop for standards and better separation of data standards and documentation may give a signal for the right context of data standardization. If data standards can gradually assist the physician to clarify, summarize, and guide their clinical decision-making process, the chance of data becoming misclassified can be reduced and data quality can be increased. Furthermore, improvements in artificial intelligence and usability of technology can better separate the natural workflow of the two processes by ensuring that standardization can still be done even if there are variations in data input.

The problem surrounding the generalizability of research datasets are often the result of differences in the needs for data by the stakeholders. Often, if the data is standardized to fulfill the needs of only a single stakeholder, the data is transformed into a data silo that cannot be reused for other purposes. The recommendation to promote common research data models is to provide a place of cooperation and negotiation between different stakeholders that allows them the room to analyze the data themselves according to their expertise and data access. For retrospective clinical studies, pharmaceutical stakeholders have worked with big data companies utilizing common research data models to conduct feasibility studies without accessing the data; The data quality was usually audited using plausibility measures based on baseline characteristics in previous studies. However, pharmaceutical companies want to eventually increase usage of real-world data as source data for prospective clinical studies which may require more access and better traceability. In studies that have these requirements, recommendations were made to better integrate electronic data capture and real-world data systems to enforce good clinical practice principles that would help alleviate regulatory evaluation and hospital security concerns. The integration would provide for more direct traceability to the source data and eventually lead to the development of interfaces that can audit the data more efficiently.

Recently published FDA (Food Drug Administration) guidance on “Data Standards for Drug and Biological Product Submissions containing Real World Data” have also mentioned similar challenges with real world data standardization.[20] Notable challenges including differences in source data standards, data exchange formats, and business processes may often result from the difficulty integrating standards with the routine clinical process leading to variations in standards usage. Similarly, research data sets do not readily exist and may go through complicated data transformation processes that can reduce transparency. FDA further mentions real world data sources may have inconsistent data source formats and wide range of methods and algorithms used to

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

create these datasets that can be hard to audit. To increase auditing capability, FDA suggests that sponsors should try to conform to FDA-supported data standards, such as CDISC, and document the process of data transformation of source data to these standards which may include audit trail, data curation, differences in data definition, and quality control processes. Our study has also found that stakeholders in China are moving towards a similar direction in leveraging more broadly supported data models and working on creating more transparency and traceability in the data transformation process.

Conclusion

The qualitative study investigated the barriers that prevent real world data standardization for clinical research based on constructivist grounded theory. A wide range of stakeholders involved in the source data, standardization process, and validity of data for clinical research were interviewed. The main causes were found to be difficulty integrating standards in routine clinical process, lack of generalizable research datasets, and difficulty auditing the standardization process. The main suggestions for these barriers were found to be better feedback loop for standards, better separation of standards and documentation, promoting generalizable clinical research data models, and improving traceability to source data. The key findings in the study make a clear point that data standardization relies on the consensus of many working groups. Data standards cannot be integrated at the source data unless it is customized for practical usage and can only be secondarily applied to source data adequately when the meaning of the source data can be clearly understood.

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 WB analyzed the data. JL and XL wrote the first draft of the manuscript. FJ, CY, and CL 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.

Funding Statement Merck Sharp and Dohme ECT2109014314.

Ethics Approval Ethical approval was obtained from Peking University Institutional Review board (No. IRB00001052-21081).

Figure 1 Caption Barrier and Suggestions in Data Standardization of Real-World Data for Clinical Research

References

1. Administration USFaD. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices - Guidance for Industry and Food and Drug Administration Staff: Food and Drug Administration, 2017.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices

2. Administration USFaD . Framework for FDA's Real-World Evidence Program: Food and Drug Administration, 2018. Available: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
3. Sun X, Tan J, Tang L, et al. . Real world evidence: experience and lessons from China. BMJ 2018;360:j5262. 10.1136/bmj.j5262
4. 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. 10.1001/jama.2018.10136
5. Statistical Information Center of National Health Commission. Conceptual data model of national health and population information [EB / OL] (2020-05-22) http://www.nhc.gov.cn/wjw/s9497/202006/7e895db195394b36ac4c806f2748cce3.shtml?tdsourcetag=s_ pctim_ aiomsg
6. General Office of the Ministry of Health and the State Administration of Traditional Chinese Medicine. The basic architecture and data standard of electronic medical record (Trial Implementation) [EB / OL] (2009-12-31) http://www.nhc.gov.cn/cms-search/xxgk/getManuscriptXxgk.htm?id=45414.
7. Statistical Information Center of National Health Commission. Specification for electronic medical record sharing document part 1 (2020 version) [EB / OL] (2016-09-29) http://www.nhc.gov.cn/wjw/s9497/201609/f725ee2635c74ed3a728cd2350953bff.shtml
8. General Office of the National Health Commission. Administrative measures for hierarchical evaluation of the application level of electronic medical record system (for Trial Implementation) and evaluation standards (for Trial Implementation) [EB / OL] (2018-12-07) http://www.nhc.gov.cn/yzygj/s7659/201812/3cae6834a65d48e9bfd783f3c7d54745.shtml.
9. Statistical Information Center of the National Health Commission. Standardized maturity evaluation scheme for hospital information interconnection (2020 version) [EB / OL] (2020-08-06) http://www.nhc.gov.cn/mohwsbwstjxxzx/s8553/202008/e80dafa1334c44c38f644602406a4973.shtml.
10. 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(12):e038375. Published 2020 Dec 29. doi:10.1136/bmjopen-2020-038375
11. C K. Constructing Grounded Theory: A practical guide through qualitative analysis Kathy Charmaz Constructing Grounded Theory: A practical guide through qualitative analysis Sage 224 £19.99 0761973532 0761973532 [Formula: see text]. Nurse Res 2006;13:84. 10.7748/nr.13.4.84.s4
12. Woods P, Gapp R, King MA. Generating or developing grounded theory:

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

methods to understand health and illness. Int J Clin Pharm 2016;38:663–70. 10.1007/s11096-016-0260-2
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. 10.1093/intqhc/mzm042
14. Setia MS. Methodology series module 5: sampling strategies. Indian J Dermatol 2016;61:505–9. 10.4103/0019-5154.190118
15. 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. 10.1080/13814788.2017.1375091
16. Whiting LS. Semi-structured interviews: guidance for novice researchers. Nurs Stand 2008;22:35–40. 10.7748/ns2008.02.22.23.35.c6420
17. Peters K, Halcomb E. Interviews in qualitative research. Nurse Res 2015; 22:6–7. 10.7748/nr.22.4.6.s2
18. Britten N. Qualitative interviews in medical research. BMJ 1995;311:251–3. 10.1136/bmj.311.6999.251
19. Rabiee F. Focus-group interview and data analysis. Proc Nutr Soc 2004;63:655–60. 10.1079/PNS2004399
20. Administration USFaD. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices - Guidance for Industry and Food and Drug Administration Staff: Food and Drug Administration, 2017. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices

Table 1 Demographics of the participants

Type of Stakeholder (# of Institutions)	Total Number of Participants
Hospital (8)	16
Hospital System Vendor (4)	10
Big Data Company (3)	15
Pharmaceutical (6)	12
Regulatory (4)	9

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

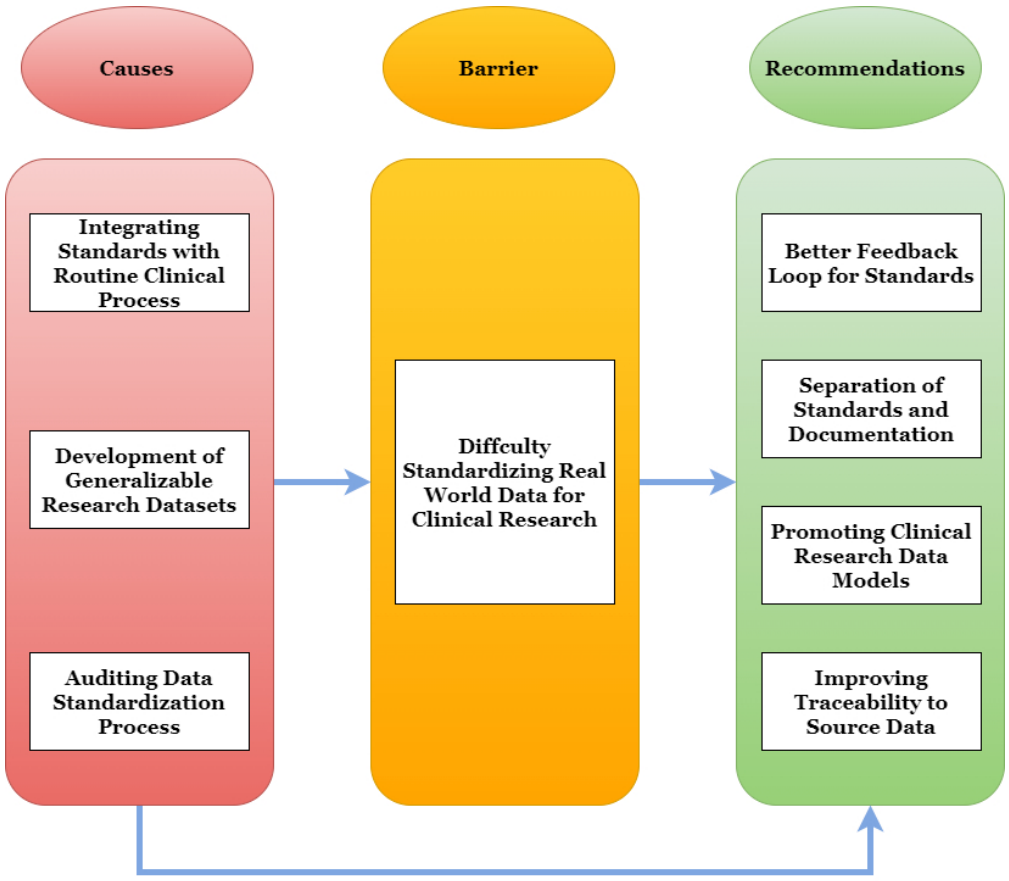


Figure 1 Barrier and Suggestions in Data Standardization of Real-World Data for Clinical Research

289x254mm (72 x 72 DPI)

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

Consolidated criteria for reporting qualitative studies (COREQ): 32-item checklist

Please indicate in which section each item has been reported in your manuscript. If you do not feel an item applies to your manuscript, please enter N/A.

For further information about the COREQ guidelines, please see Tong *et al.*, 2017: https://doi.org/10.1093/intqhc/mzm042

No.	Item	Description	Section #
Domain 1: Research team and reflexivity			
Personal characteristics			
1.	Interviewer/facilitator	Which author/s conducted the interview or focus group?	
2.	Credentials	What were the researcher's credentials? *E.g. PhD, MD*	
3.	Occupation	What was their occupation at the time of the study?	
4.	Gender	Was the researcher male or female?	
5.	Experience and training	What experience or training did the researcher have?	
Relationship with participants			
6.	Relationship established	Was a relationship established prior to study commencement?	
7.	Participant knowledge of the interviewer	What did the participants know about the researcher? *E.g. Personal goals, reasons for doing the research*	
8.	Interviewer characteristics	What characteristics were reported about the interviewer/facilitator? *E.g. Bias, assumptions, reasons and interests in the research topic*	
Domain 2: Study design			
Theoretical framework			
9.	Methodological orientation and theory	What methodological orientation was stated to underpin the study? *E.g. grounded theory, discourse analysis, ethnography, phenomenology, content analysis*	
Participant selection			
10.	Sampling	How were participants selected? *E.g. purposive, convenience, consecutive, snowball*	
11.	Method of approach	How were participants approached? *E.g. face-to-face, telephone, mail, email*	
12.	Sample size	How many participants were in the study?	
13.	Non-participation	How many people refused to participate or dropped out? What were the reasons for this?	
Setting			
14.	Setting of data collection	Where was the data collected? *E.g. home, clinic, workplace*	
15.	Presence of non-participants	Was anyone else present besides the participants and researchers?	

For peer review only

16.	Description of sample	What are the important characteristics of the sample? *E.g. demographic data, date*	
Data collection			
17.	Interview guide	Were questions, prompts, guides provided by the authors? Was it pilot tested?	
18.	Repeat interviews	Were repeat interviews carried out? If yes, how many?	
19.	Audio/visual recording	Did the research use audio or visual recording to collect the data?	
20.	Field notes	Were field notes made during and/or after the interview or focus group?	
21.	Duration	What was the duration of the interviews or focus group?	
22.	Data saturation	Was data saturation discussed?	
23.	Transcripts returned	Were transcripts returned to participants for comment and/or correction?	
Domain 3: analysis and findings			
Data analysis			
24.	Number of data coders	How many data coders coded the data?	
25.	Description of the coding tree	Did authors provide a description of the coding tree?	
26.	Derivation of themes	Were themes identified in advance or derived from the data?	
27.	Software	What software, if applicable, was used to manage the data?	
28.	Participant checking	Did participants provide feedback on the findings?	
Reporting			
29.	Quotations presented	Were participant quotations presented to illustrate the themes / findings? Was each quotation identified? *E.g. Participant number*	
30.	Data and findings consistent	Was there consistency between the data presented and the findings?	
31.	Clarity of major themes	Were major themes clearly presented in the findings?	
32.	Clarity of minor themes	Is there a description of diverse cases or discussion of minor themes?	

When submitting your manuscript via the online submission form, please upload the completed checklist as a Figure/supplementary file.

If you would like this checklist to be included alongside your article, we ask that you upload the completed checklist to an online repository and include the guideline type, name of the repository, DOI and license in the *Data availability* section of your manuscript.

Developed from: Allison Tong, Peter Sainsbury, Jonathan Craig, Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups, International Journal for Quality in Health Care, Volume 19, Issue 6, December 2007, Pages 349–357, https://doi.org/10.1093/intqhc/mzm042

List of Institutions:

Hospitals:

1. Peking University People's Hospital, Beijing, Beijing, China
2. Peking University First Hospital, Beijing, Beijing, China
3. First Teaching Hospital of Tianjin University of Traditional Chinese Medicine, Tianjin, Tianjin, China
4. Hainan General Hospital, Haikou, Hainan, China
5. Boao Evergrande International Hospital, Boao, Hainan, China
6. Boao Super Hospital, Boao, Hainan, China
7. Boao Yiling Lifecare Center, Hainan, China
8. Boao Worldlight Hospital, Hainan, China

Hospital System Vendors:

1. Haitai International
2. Goodwill
3. Winning Health
4. Orion Health Rhapsody

Big Data Companies:

1. Yiducloud
2. Digital Health China Technologies
3. Inspur

Pharmaceutical Companies:

1. Pfizer
2. Tigermed
3. AstraZeneca
4. Bristol-Meyers Squibb
5. Johnson & Johnson
6. BeiGene

Regulatory Institutions:

1. China National Health Development Research Center
2. National Medical Products Administration
3. China Center for Food and Drug International Exchange
4. Hainan Boao Lecheng International Medical Tourism Pilot Zone Administration

Questionnaire:

Hospital:

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience facilitating clinical research at the Hospital?

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

3. What are the motivating goals of clinical research?
4. How do you determine the research dataset that are needed for your research?
5. Do you think that electronic medical records or routine care data at the hospital are enough to accomplish your research?
6. How do you aggregate and store all data from different hospital systems?
7. How do you implement standards on your data?
8. What areas in your clinical research process do you have to rely on external vendors to help you?
9. How does data sharing happen for medical records inside and outside of the hospital?
10. Beyond clinical research, have you standardized your data for other purposes?

Big Data:

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience handling data at the company?
3. Describe your interaction with clients that want to utilize your service for clinical research?
4. How do you organize source data and any standards that you use to do so?
5. How do you manage the variety of standards that are published?
6. How do you transform the data to fit these standards?
7. How do you manage the differences between source data and standards?
8. How do you track the data transformation process?
9. How do you manage the different research projects that need to use real world data?
10. What type of standards are your clients required to fulfill?

Hospital System:

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience developing hospital systems? What type of systems does your company produce?
3. What is behind the motivation for hospitals to use more standardized systems?
4. Describe how data is organized and presented for hospital systems and the standards used to create them?
5. How has the usage of standards improved your services?
6. What is the process of negotiating standards usage when customizing your product for clients?
7. How does customization of hospital systems affect standard usage?
8. How do you localize these standards for practical usage?
9. How do you improve communication between your systems or external systems using standards?
10. What areas of hospital systems rely most on existing international standards?

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

Pharmaceutical:

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience in using real world data for clinical research? What types of real-world data do you use as source data for your studies?
3. How do you obtain or access real world data?
4. How does the process of sourcing real world data differ from the traditional data collection for clinical research the most?
5. What standards are used for real world data?
6. What data standards would like to see used for real world data?
7. What standardization methods for real world data are used to produce research data?
8. How do you check whether the data is reliable and what types of data do you think are most reliable?
9. Does real world data meet your research needs?
10. What do you think the standards for real world data necessary to meet evidence requirement by regulatory institutions?

Regulatory:

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience in regulating real world studies? What types of real-world data do you see used as source data for these studies?
3. What are the common characteristics of clinical studies using real world data do you often see (study design, phase, purpose)?
4. How does the process of sourcing real world data differ from the traditional data collection for clinical research the most?
5. What are some ethical standards that may be violated in real world data usage for clinical studies?
6. What standards are used for real world data?
7. What data standards are necessary for real world data to meet evidence requirements?
8. What are the main considerations surrounding data standardization?
9. How does real world data meet other regulatory evaluation needs besides clinical trials?
10. How can real world data be better collected to establish a platform in China that can benefit more stakeholders?

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

BMJ Open

Existing Barriers and Recommendations of Real-World Data Standardization for Clinical Research in China: A Qualitative Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-059029.R1
Article Type:	Original research
Date Submitted by the Author:	07-Apr-2022
Complete List of Authors:	Lai, Junkai; Peking University First Hospital, Peking University Clinical Research Institute, Liao, Xiwen; Peking University First Hospital, Peking University Clinical Research Institute Jin, Feifei; Peking University People's Hospital, National Center for Trauma Medicine Wang, Bin; Peking University First Hospital, Peking University Clinical Research Institute Yao, Chen; Peking University First Hospital, Peking University Clinical Research Institute; Hainan Institute of Real World Data Li, Chen; Fourth Military Medical University, Department of Health Statistics; School of Preventive Medicine
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Qualitative research
Keywords:	QUALITATIVE RESEARCH, Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

SCHOLARONE™
Manuscripts

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

For peer review only

Existing Barriers and Recommendations of Real-World Data Standardization for Clinical Research in China: A Qualitative Study

Junkai Lai[1], Xiwen Liao[1], Chen Yao[13], Feifei Jin[2], Bin Wang[1], Chen Li[4]
1 Peking University Clinical Research Institute, Peking University First Hospital, Beijing, China
2 National Center for Trauma Medicine, Peking University People's Hospital, Beijing, China
3 Hainan Institute of Real World Data, Qionghai, Hainan, China.
4 Department of Health Statistics, School of Preventive Medicine, Fourth Military University, Xi'an, Shaanxi, China

Correspondence to
Chen Yao
Peking University First Hospital, Xicheng District, Beijing 100034, China
Tel +86 18610640562
Email yaochen@hsc.pku.edu.cn

Keywords real world data; data standards; data standardization; clinical research; qualitative research; China;

Word Count
5085

For peer review only

Existing Barriers and Recommendations of Real-World Data Standardization for Clinical Research in China: A Qualitative Study

Abstract

Objective To investigate the existing barriers and recommendations of real-world data (RWD) standardization 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 8 Hospitals, 4 Hospital System Vendors, 3 Big Data Companies, 6 Medical Products Companies, and 4 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 generalizability in existing research databases, and lack of transparency in existing data standardization process were the barriers of data standardization of RWD for clinical research. Expanding coverage of terminology through collecting common terminology, reducing burden in the usage of terminology standards, improving generalizability 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 standardization of RWD for clinical research can help generate better evidence used to support regulatory evaluation of medical products. This research suggests expanding coverage of terminology through collecting common terminology, reducing burden in the usage of terminology standards, improving generalizability of RWD for research by using clinical data models, and improving traceability to source data for transparency to guide efforts in data standardization in the future.

Strengths and Limitations of this study:

- Strength: Wide variety of relevant stakeholders on the subject
- Strength: Qualitative understanding of a major industry bottleneck
- Strength: Important recommendations that can guide the direction of the future of the subject
- Limitations: Due to COVID-19, a portion of the interviews were not done in person and may limit the ability to read into the participants response for further exploration of the subject
- Limitations: Recruitment of participants were limited by those that were already exploring the subject and could not be generalized to participants that may not be familiar with the subject. The unselected companies may have different views, which could result in selection bias.

For peer review only

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 in the United States (U.S.) and in China, RWD have become increasingly used to support regulatory decision making for drugs 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 electronic medical records (EMR) of patient visits, as evidence for domestic product approval. An example of the first products to leverage the approval process include Johnson & Johnson's femtosecond ophthalmic surgical medical devices which started data collection in October 2019 and was subsequently approved after 6 months[8]. As more products are introduced into Boao Lecheng, there is an imminent need to efficiently translate the data within EMRs to clinical research data.

The current problem in China is that EMRs constitute a separate system that is not directly connected to electronic data capture (EDC) systems, leading to duplicative and manual transcription of EMR data into the EDC system during clinical research[9-10]. The inefficient process results often in poor data quality due to human error and insufficient source data verification[11]. Solutions to the issue have been explored by the U.S. Food and Drug Administration (FDA), which includes promoting the direct usage of electronic source data (eSource) within real world data systems such as EHRs for clinical research[12-13]. In the eSource guidance, interoperability between EHR and EDC systems through the usage of data standards is emphasized. In addition to publishing guidance, initiatives led by the FDA promoted collaboration between standards organizations Health Level Seven (HL7) and Clinical Data Interchange Standards Consortium (CDISC), which have produced solutions that can translate EHR data standards to clinical research data standards[14].

However, these solutions are not directly translatable to China's context due to differences in the developed data standards and methods used to standardize data. The data standards in China were developed through the Statistical Information Center of the National Health Commission and pushed through government evaluation of hospital information system's meaningful usage[15-18]. The first qualitative study on the problem of the gap between RWD and clinical research, found several key domestic problems which included the lack of data standards usage, prevalence of unstructured data, and other data security concerns[19]. Similarly, a literature review in China reveals the deterrents of the meaningful usage of RWD for clinical research to include the lack of regulatory implementation of semantic level data standards, unstructured data, and data accesss[20]. Therefore, it is important to address the topic of the standardization of RWD for clinical research in China. However, limited literature has addressed the issue and opinions of stakeholders urgently needing to use RWD for clinical research have yet to be collected in China. Therefore, a qualitative study of the relevant stakeholders in China regarding the barriers and suggestions related to the topic is needed.

Methods

Design

For Peer Review Only

Qualitative research allows us to understand a participant's experience through qualitative methods of capturing data often through interviews. Grounded theory is a qualitative research method used in areas previously unexplored or under explored to inductively generate theory from data grounded in the perceptions and concerns of the participant[21]. The method's extensive history in healthcare research can be attributed to its systematic process of analysis and stages of coding that allows themes to emerge from the data regarding the problems faced by participants and their resolution to the problems[22]. Constructivist grounded theory (CGT) assumes that data are co-constructed through the researcher-participant interaction, and the product of analyses is influenced by the interaction of the researcher with the data[23-24]. Studying the underexplored barriers experienced by stakeholders and their resolutions in the process of standardizing RWD for clinical research in China is the central problem of the study. 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 following the guidelines of the Consolidated Criteria for Reporting Qualitative Research (COREQ) guidelines[25].

Participants Selection

The selection of participants was based on their relevance to the type of stakeholders involved in the construction of the regional data platform in Boao Lecheng which seeks to efficiently standardize RWD for clinical research. The stakeholders include hospitals generating RWD, hospital system vendors that install EMRs for hospitals, big data companies that aggregate the hospital EMRs onto a data platform, Medical Products companies that consume the data for clinical research, and regulatory departments in charge of evaluating the usage of RWD for research. The type of stakeholders can be categorized into 3 categories: stakeholders affecting the source data, stakeholders affecting the standardization of source data for research, and stakeholders affecting the validity of data 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 Institutions 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 done to determine when new coding information was no longer 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 with no dropouts. YC and JL contacted the different interviewee and briefed on the subject matter and the objective of 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 by the type of stakeholder is shown in **Table 1.** Detailed list of institutions by type of stakeholder is included in the Appendix.

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

For peer review only

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 minutes 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 and included only the participant and researchers.

Data Collection

Semi-structured 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 summarize 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]. Each interview allowed 60 minutes and basic information including the interview time, place, and interviewee was collected at the beginning. Five sets of interview guides were designed for the five types of stakeholder roles and pilot tested before hand with similar participants that were not included in the study to make the flow of questioning better. Full interview guides were included along with general categories that motivate the questions in the Appendix. The general categories of questions used for each role focus on how the stakeholders affects the data standardization process including at the source, during data standardization to research, and evaluation of research data. The interview questions guided the interviewer in exploring the subject with the participant. Further discussion on the questions or repeat interviews were allowed to explore deeper into the topic or for better clarification. Simultaneous data collection and analysis determined when information saturation had occurred for each role, implying the interviewing of participants ended.

The interviewers were four doctoral students. JL (Male) and XL (Female) were mainly responsible for the interviews, and BW (Male) and FJ (Female) played supportive roles and were mainly responsible for the recordings. The interviewers were trained in a qualitative research course and had experience conducting interviews.

Analysis

All interviews were transcribed to text using the automated transcription software and double checked for each recording by the two interviewers (JL and XL). Coding and memoing were done by three researchers (JL, XL, FJ) who drew on the techniques of constructivist grounded theory while analyzing 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 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, FJ) agreed. Participants did not provide feedback on the

For peer review only

findings.

Patient and public involvement

There was no patient or public involvement in this research.

Results

Barriers and Recommendations in the Standardization of RWD for Clinical Research

The CGT framework generated from the three stages of coding and the 62 participants' responses were summarized 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 standardization processes. The recommendations included expanding coverage of terminology by collecting common terminology, reducing burden in the usage of terminology standards, improving applicability of databases using clinical data models, 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 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, 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 Products companies and regulatory departments expressed that the existing RWD databases such as disease specialty databases formed by hospitals are standardized to specific research questions and not generalizable to others. Medical Products participants expressed that there is substantial variation in the type of available data even when standardized. 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.

For peer review only

Currently, the existing data were not organized in a way that could be combined into a generalizable 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 Standardization Process

The findings showed that hospital and Medical Products participants expressed that the data standardization process from RWD to clinical research data lacks transparency. Medical Products 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 standardization are not transparent. In addition, they had concerns over the interpretability of standardization methods such as natural language processing algorithms in extracting relevant research data and determining whether regulatory institutions will accept these methods. Hospital participants also expressed that inaccurate data produced by external vendors are difficult to correct or target due to not knowing the exact methods used to transform the data. As the producers of research data, big data participants expressed that the standardization process requires many steps and teams involved, which can reduce its transparency.

"The exact methods used for data standardization 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 standardization of data are whether regulatory institutions will accept them." – Medical Products Participant 4

"When vendors standardize 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, making the methods used hard to document comprehensively" – Big Data Participant 5

Suggestions

Expanding Coverage of Terminology by Collecting Common Terminology

The findings showed that big data companies and hospital information system participants suggested that the incorporation of their collection of local terminology can improve the coverage of the existing terminology standards. Big data participants expressed using RWD to find and

For peer review only

aggregate colloquial terminology used by clinicians to improve the coverage of terminologies used in RWD. Hospital system participants expressed that they have collected practical terminology lists from different hospitals, instead of standard terminology lists. In addition, they expressed that these local lists are more likely to be used in a clinical setting, because it improved communication and could be key to the adoption of terminologies in a clinical setting.

"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, instead of standard terminology lists, which improves 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 standardization. Hospital participants expressed various methods used to automatically standardize 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 Clinical Data Models

The findings showed that hospital system and big data participants expressed that the usage of data model standards to organize 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 efficiency of reusing data for different services including clinical decision support services. Big data participants suggested the usage of the OHDSI data model to organize their data for the reusage of data to answer different research questions. In addition, they suggested that research in different disease areas may require a further extension of the models by analyzing 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

For peer review only

the HL7 RIM (Health Level 7 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 Products participants suggested the improvement in the traceability to source data for better transparency in the data standardization 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 Products company participants suggested the usage of eSource methods that meet regulatory expectations in terms of auditing the source data to determine the quality of the collected data.

"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 standardization of RWD for clinical research is the research question central to the current qualitative study. Through a constructivist grounded theory 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 standardization process. The recommendations included expanding coverage of terminology by collecting common terminology, reducing burden in the usage of terminology standards, improving applicability of databases using clinical data models, 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 standardization for clinical research. The use of the methods in grounded theory were to find the barriers and recommendation to the research problem, with the goal of using 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

For peer review only

used by physicians in China and may be burdensome to use. Thus, it is important to promote the collection of common 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 standardization of data elements and limited focus on terminology standards, preventing meaningful exchange of information[20]. Thus, researchers believed that the localization 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 have shown 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 (HKCTT), 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 generalizability in existing research databases. The lack of generalizability of databases can lead to the limited usage of RWD even after standardization since the databases only address a specific question. Thus, the usage of clinical data models can improve the generalizability of databases by organizing RWD in a consistent and research relevant way to enable the answering of research questions. In the US, 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 generalizable common data model (CDM) that each database could conform to that would allow for the efficient answering of clinical

research questions[39-40]. In 2021, HL7 and OHDSI (previously OMOP) collectively announced their initiative to create a common data model that integrates data standards common to EHRs with the goal of better organizing EHR data into a clinical research data model[41]. Although the usage of common data models in China have not been pushed by the government, the growing usage among big data companies and other research organizations is evident. Confirming the experiences of the participant in the current 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 clinical data model 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 have now expanded its usage domestically and have also been internationally used to answer COVID-19 treatment questions[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 is the lack of transparency in the existing data standardization process. The lack of well-documented and understandable methods used in the data standardization 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 standardize 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. However, 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 in the data standardization process by through source data sharing and statistical analysis protocol publishing to increase transparency of standardization methods used[50]. In addition, source data verification, which checks consistency between the data recorded in the database with 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 utilize an independent eSource platform for the transferring and storing of research source data to guard data integrity and increase transparency. The development and usage of such a platform was tested using real world data 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 development and usage of a research source data management platform at medical institutions when they conduct clinical research[54].

For peer review only

The strength of the study is the selection of a wide and comprehensive range of stakeholder that can better represent the issue in China. Several limitations of this study warrant 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 minimize 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 standardization for clinical research based on constructivist grounded theory. 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 standardization process. Expanding coverage of terminology through collecting common terminology, reducing burden in the usage of terminology standards, improving applicability of databases using clinical data models, 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 standardization 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 standardize RWD for clinical research with efficiency and reliability.

Figure 1 Caption Barrier and Suggestions in Data Standardization of Real-World Data for Clinical Research

Table 1 Demographics of the participants

Type of Stakeholder (# of Institutions)	Total Number of Participants
Hospital (8)	16
Hospital System Vendor (4)	10
Big Data Company (3)	15
Pharmaceutical (6)	12
Regulatory (4)	9

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 WB analyzed the data. JL and XL wrote the first draft of the manuscript. FJ, CY, and CL 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.
Funding Statement This work was supported by the Department of Science, Technology, and International Cooperation, National Medical Products Administration.

For peer review only

Competing Interests None declared.

Patient consent for publication Not required

Ethics Approval Ethical approval was obtained from Peking University Institutional Review board (No. IRB00001052-21081).

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

References

1 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. In, 2017.

2 US Food And Drug Administration. REAL-WORLD EVIDENCE PROGRAM. In, 2018.

3 J Corrigan-Curay, L Sacks, J Woodcock. Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. *JAMA* 2018;320(9):867-68.

4 X Sun, J Tan, L Tang, JJ Guo, X Li. Real world evidence: experience and lessons from China. *BMJ* 2018;360:j5262.

5 National Medical Product Association. Real world evidence supports the guiding principles of drug development and review. In, 2020.

6 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. In, 2021.

7 National Medical Products Association. The State Food and Drug Administration launched the scientific action plan for China's drug supervision. In, 2021.

8 Johnson Johnson Surgical Vision. A Real-World Evidence Study in China of the Catalys Precision Laser System. In, 2020.

9 R Blitz, M Dugas. Conceptual Design, Implementation, and Evaluation of Generic and Standard-Compliant Data Transfer into Electronic Health Records. *Appl Clin Inform* 2020;11(3):374-86.

10 Y Matsumura, A Hattori, S Manabe, et al. Interconnection of electronic medical record with clinical data management system by CDISC ODM. *Stud Health Technol Inform* 2014;205:868-72.

11 Y Wu, D Yin, K Abbasi. China's medical research revolution. *BMJ* 2018;360:k547.

12 US Food And Drug Administration. Electronic Source Data in Clinical Investigations. In, 2013.

13 US Food And Drug Administration. Use of Electronic Health Record Data in Clinical Investigations. In, 2018.

14 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. In, 2020.

15 National Health Commission of the People's Republic of China. Electronic medical record sharing document specification. In, 2020.

16 National Health Commission of the People's Republic of China. Administrative measures for hierarchical evaluation of application level of electronic medical record system. In, 2018.

17 National Health Commission of the People's Republic of China. Standardized maturity evaluation scheme for hospital information interconnection. In, 2020.

18 National Health Committee Of The People'S Republic Of China. Basic architecture and data standard of electronic medical record. In, 2009.

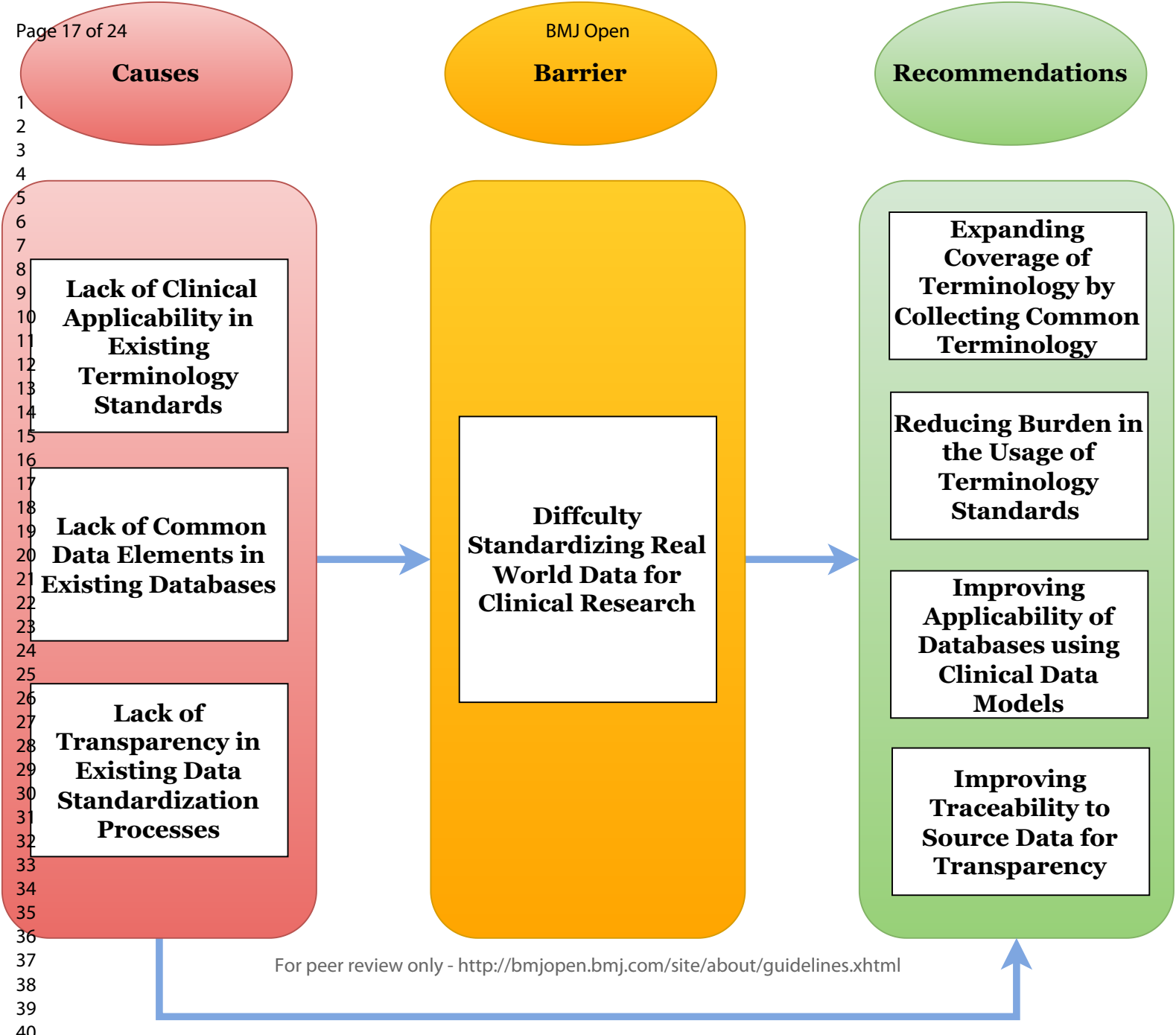
For peer review only

19 F Jin, C Yao, X Yan, et al. Gap between real-world data and clinical research within hospitals in China: a qualitative study. *Bmj Open* 2020;10(12):e38375.
20 J Xie, EQ Wu, S Wang, et al. Real-World Data for Healthcare Research in China: Call for Actions. *Value Health Reg Issues* 2022;27:72-81.
21 M Byrne. Grounded theory as a qualitative research methodology. *Aorn J* 2001;73(6):1155-56.
22 AL Chapman, M Hadfield, CJ Chapman. Qualitative research in healthcare: an introduction to grounded theory using thematic analysis. *J R Coll Physicians Edinb* 2015;45(3):201-05.
23 FK Metelski, J Santos, C Cechinel-Peiter, et al. Constructivist Grounded Theory: characteristics and operational aspects for nursing research. *Rev Esc Enferm Usp* 2021;55:e3776.
24 J Mills, A Bonner, K Francis. Adopting a constructivist approach to grounded theory: Implications for research design. *Int J Nurs Pract* 2006;12(1):8-13.
25 A Tong, P Sainsbury, J Craig. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care* 2007;19(6):349-57.
26 A Moser, I Korstjens. Series: Practical guidance to qualitative research. Part 3: Sampling, data collection and analysis. *Eur J Gen Pract* 2018;24(1):9-18.
27 MS Setia. Methodology Series Module 5: Sampling Strategies. *Indian J Dermatol* 2016;61(5):505-09.
28 J Sutton, Z Austin. Qualitative Research: Data Collection, Analysis, and Management. *Can J Hosp Pharm* 2015;68(3):226-31.
29 K Peters, E Halcomb. Interviews in qualitative research. *Nurse Res* 2015;22(4):6-07.
30 LS Whiting. Semi-structured interviews: guidance for novice researchers. *Nursing standard* 2008;22(23):35-40.
31 N Britten. Qualitative interviews in medical research. *BMJ* 1995;311(6999):251-53.
32 O Bodenreider, R Cornet, DJ Vreeman. Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. *Yearb Med Inform* 2018;27(1):129-39.
33 Y Zhu, H Pan, L Zhou, et al. Translation and localization of SNOMED CT in China: a pilot study. *Artif Intell Med* 2012;54(2):147-49.
34 R Zhang, J Liu, Y Huang, 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(1):54.
35 Y Cheng, T Jiang, L Deng, L Chen, J Ming. Research on the Coverage of Standard Chinese Medical Terminology to Practical Application. *Chinese Journal of Health Informatics and Management* 2020.
36 H Forsvik, V Voipio, J Lamminen, et al. Literature Review of Patient Record Structures from the Physician's Perspective. *J Med Syst* 2017;41(2):29.
37 K Hung, M Lau, V Fung. Successful Implementation of Terminology Binding in Hong Kong Hospital Authority. *Stud Health Technol Inform* 2019;264:1486-87.
38 Y Chen, D Hu, M Li, H Duan, X Lu. Automatic SNOMED CT coding of Chinese clinical terms via attention-based semantic matching. *Int J Med Inform* 2022;159:104676.
39 JM Overhage, PB Ryan, CG Reich, AG Hartzema, PE Stang. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* 2012;19(1):54-60.
40 PE Stang, PB Ryan, JA Racoosin, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med* 2010;153(9):600-06.
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

For Peer Review Only

Observational Research. In, 2021.
42 HX Yue, YL Zhan, F Bian, et al. [Data standard and data sharing in clinical cohort studies]. *Zhonghua Liu Xing Bing Xue Za Zhi* 2021;42(7):1299-305.
43 A Prats-Uribe, AG Sena, L Lai, et al. Use of repurposed and adjuvant drugs in hospital patients with covid-19: multinational network cohort study. *BMJ* 2021;373:n1038.
44 X Zhang, L Wang, S Miao, et al. Analysis of treatment pathways for three chronic diseases using OMOP CDM. *J Med Syst* 2018;42(12):260.
45 H Ji, S Kim, S Yi, 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. In, 2021.
47 S Hume, S Sarnikar, L Becnel, D Bennett. Visualizing and Validating Metadata Traceability within the CDISC Standards. *AMIA Jt Summits Transl Sci Proc* 2017;2017:158-65.
48 Society for Clincial Data Management. eSource Implementation in Clinical Research: A Data Management Perspective. In, 2014.
49 E Kellar, SM Bornstein, A Caban, et al. Optimizing the Use of Electronic Data Sources in Clinical Trials: The Landscape, Part 1. *Ther Innov Regul Sci* 2016;50(6):682-96.
50 J Xinyao, Z Wenke, Z Junhua, et al. Promote Transparency in Real-World Study. *World Chinese Medicine* 2019.
51 X Yan, C Dong, C Yao. Protecting the accuracy of clinical trial data in China. *BMJ Opinion* 2018.
52 C Dong, X Yan, R Tian, Z Bian, C YAO. Strengthen the process report of clinical trials, promote full transparency of clinical
trials. *Chinese Journal of Evidence-Based Medicine* 2018.
53 F Jin, C Yao, J Ma, et al. 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. In, 2021.

For peer review only



For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

List of Institutions:

Hospitals:

1. Peking University People's Hospital, Beijing, Beijing, China
2. Peking University First Hospital, Beijing, Beijing, China
3. First Teaching Hospital of Tianjin University of Traditional Chinese Medicine, Tianjin, Tianjin, China
4. Hainan General Hospital, Haikou, Hainan, China
5. Boao Evergrande International Hospital, Boao, Hainan, China
6. Boao Super Hospital, Boao, Hainan, China
7. Boao Yiling Lifecare Center, Hainan, China
8. Boao Worldlight Hospital, Hainan, China

Hospital System Vendors:

1. Haitai International
2. Goodwill
3. Winning Health
4. Orion Health Rhapsody

Big Data Companies:

1. Yiducloud
2. Digital Health China Technologies
3. Inspur

Pharmaceutical Companies:

1. Pfizer
2. Tigermed
3. AstraZeneca
4. Bristol-Meyers Squibb
5. Johnson & Johnson
6. BeiGene

Regulatory Institutions:

1. China National Health Development Research Center
2. National Medical Products Administration
3. China Center for Food and Drug International Exchange
4. Hainan Boao Lecheng International Medical Tourism Pilot Zone Administration

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

Interview Guide:

Hospital:

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience facilitating clinical research at the Hospital?
3. What are the motivating goals of clinical research?
4. How do you determine the data needed for your research? What are the barriers and recommendations?
5. Do you think that electronic medical records or routine care data at the hospital are enough to accomplish your research? What are the barriers and recommendations?
6. How do you use data standards to aggregate and store all data from different hospital systems? What are the barriers and recommendations?
7. What data standards are used and how do you implement data standards during routine data collection? What are the barriers and recommendations?
8. What areas in your clinical research process do you have to rely on external vendors to help you standardize the data and how have you evaluated their data standardization process? What are the barriers and recommendations?
9. How are data standards used to share medical records inside and outside of the hospital? What are the barriers and recommendations?
10. Beyond clinical research, have you standardized your data for other purposes?

Big Data:

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience handling data at the company?
3. Describe your interaction with clients that want to utilize your service for clinical research?
4. What type of standards are your clients required to fulfill?
5. How do you use data standards to organize and aggregate source data? What are the barriers and recommendations?
6. How do you use data standards when you transform source data into research datasets? What are the barriers and recommendations?
7. What methods are used to standard source data for clinical research? What are the barriers and recommendations?
8. How do you track and evaluate the quality of the data transformation process? What are the barriers and recommendations?
9. How do you manage the variety of standards that are published? What are the barriers and recommendations?
10. How do you manage the different research projects that need to use real world data? What are the barriers and recommendations?

Hospital System Vendor:

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience developing hospital systems? What type of systems does your company produce?
3. What is behind the motivation for hospitals to use data standards?
4. Describe the data standards that are used for hospital systems?
5. How are data standards implemented and customized for the hospital? What are the barriers and recommendations?
6. How do you use data standards to improve internal and external communication at the hospital? What are the barriers and recommendations?

Medical Products Company:

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience in using real world data for clinical research? What types of real-world data do you use as source data for your studies?
3. How do you obtain or access real world data?
4. How does the process of sourcing real world data differ from the traditional data collection for clinical research the most?
5. What standards are used for real world data?
6. What data standards would you like to see used for real world data?
7. What standardization methods for real world data are used to produce research data? What are the barriers and recommendations?
8. How do you check whether the data is reliable and what types of data do you think are most reliable? What are the barriers and recommendations?
9. Does real world data meet your research needs? What are the barriers and recommendations?
10. Given regulatory consideration for the usage of real-world data for clinical research, what do you see as the problems in the current methods used to standardize data? What are some recommendations?

Regulatory:

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience in regulating real world studies? What types of real-world data do you see used as source data for these studies?
3. What are the common characteristics of clinical studies using real world data do you often see (study design, phase, purpose)?
4. How is real world data used to support regulatory decision making?
5. How does the process of sourcing real world data differ from the traditional data collection for clinical research the most?

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

6. What standards are used for existing real-world data?
7. What data standards are necessary for the submission of real-world data used to gain product approval?
8. How can the standardization of real-world data for clinical research meet regulatory expectations? What are the barriers and recommendations?
9. How can we establish a real-world data platform that can be used for clinical research that can benefit the most stakeholders in China? What are the barriers and recommendations?

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

General Categories for Interview Questions:

General Category	Hospital	Hospital System Vendor	Big Data Company	Medical Products Company	Regulatory Department
Privacy and Information Consent Statement	1	1	1	1	1
Experience and aim when using RWD for clinical research	2,3,10	2,3	2,3	2,3,4	2,3,4,5
Relevant RWD Standards	7	4	4,9	5,6	6,7
RWD relevance for clinical research	4,5		10	8,9	9
Standardization of RWD at source	6,7,9	5,6	5		
Standardization of RWD for clinical research	8		6,7	7	

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

Reliability of RWD Standardization for clinical research	8		8	8,10	8

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

Consolidated criteria for reporting qualitative studies (COREQ): 32-item checklist

Please indicate in which section each item has been reported in your manuscript. If you do not feel an item applies to your manuscript, please enter N/A.

For further information about the COREQ guidelines, please see Tong *et al.*, 2017: https://doi.org/10.1093/intqhc/mzm042

No.	Item	Description	Section #
Domain 1: Research team and reflexivity			
Personal characteristics			
1.	Interviewer/facilitator	Which author/s conducted the interview or focus group?	
2.	Credentials	What were the researcher's credentials? *E.g. PhD, MD*	
3.	Occupation	What was their occupation at the time of the study?	
4.	Gender	Was the researcher male or female?	
5.	Experience and training	What experience or training did the researcher have?	
Relationship with participants			
6.	Relationship established	Was a relationship established prior to study commencement?	
7.	Participant knowledge of the interviewer	What did the participants know about the researcher? *E.g. Personal goals, reasons for doing the research*	
8.	Interviewer characteristics	What characteristics were reported about the interviewer/facilitator? *E.g. Bias, assumptions, reasons and interests in the research topic*	
Domain 2: Study design			
Theoretical framework			
9.	Methodological orientation and theory	What methodological orientation was stated to underpin the study? *E.g. grounded theory, discourse analysis, ethnography, phenomenology, content analysis*	
Participant selection			
10.	Sampling	How were participants selected? *E.g. purposive, convenience, consecutive, snowball*	
11.	Method of approach	How were participants approached? *E.g. face-to-face, telephone, mail, email*	
12.	Sample size	How many participants were in the study?	
13.	Non-participation	How many people refused to participate or dropped out? What were the reasons for this?	
Setting			
14.	Setting of data collection	Where was the data collected? *E.g. home, clinic, workplace*	
15.	Presence of non-participants	Was anyone else present besides the participants and researchers?	

For peer review only

16.	Description of sample	What are the important characteristics of the sample? *E.g. demographic data, date*	
Data collection			
17.	Interview guide	Were questions, prompts, guides provided by the authors? Was it pilot tested?	
18.	Repeat interviews	Were repeat interviews carried out? If yes, how many?	
19.	Audio/visual recording	Did the research use audio or visual recording to collect the data?	
20.	Field notes	Were field notes made during and/or after the interview or focus group?	
21.	Duration	What was the duration of the interviews or focus group?	
22.	Data saturation	Was data saturation discussed?	
23.	Transcripts returned	Were transcripts returned to participants for comment and/or correction?	
Domain 3: analysis and findings			
Data analysis			
24.	Number of data coders	How many data coders coded the data?	
25.	Description of the coding tree	Did authors provide a description of the coding tree?	
26.	Derivation of themes	Were themes identified in advance or derived from the data?	
27.	Software	What software, if applicable, was used to manage the data?	
28.	Participant checking	Did participants provide feedback on the findings?	
Reporting			
29.	Quotations presented	Were participant quotations presented to illustrate the themes / findings? Was each quotation identified? *E.g. Participant number*	
30.	Data and findings consistent	Was there consistency between the data presented and the findings?	
31.	Clarity of major themes	Were major themes clearly presented in the findings?	
32.	Clarity of minor themes	Is there a description of diverse cases or discussion of minor themes?	

When submitting your manuscript via the online submission form, please upload the completed checklist as a Figure/supplementary file.

If you would like this checklist to be included alongside your article, we ask that you upload the completed checklist to an online repository and include the guideline type, name of the repository, DOI and license in the *Data availability* section of your manuscript.

Developed from: Allison Tong, Peter Sainsbury, Jonathan Craig, Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups, International Journal for Quality in Health Care, Volume 19, Issue 6, December 2007, Pages 349–357, https://doi.org/10.1093/intqhc/mzm042

BMJ Open

Existing Barriers and Recommendations of Real-World Data Standardization for Clinical Research in China: A Qualitative Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-059029.R2
Article Type:	Original research
Date Submitted by the Author:	04-Jul-2022
Complete List of Authors:	Lai, Junkai; Peking University First Hospital, Peking University Clinical Research Institute, Liao, Xiwen; Peking University First Hospital, Peking University Clinical Research Institute Yao, Chen; Peking University First Hospital, Peking University Clinical Research Institute; Hainan Institute of Real World Data Jin, Feifei; Peking University People's Hospital, National Center for Trauma Medicine Wang, Bin; Peking University First Hospital, Peking University Clinical Research Institute Li, Chen; Fourth Military Medical University, Department of Health Statistics; School of Preventive Medicine Zhang, Jun; MSD China Ltd, CORE Liu, Larry; Merck & Co Inc; Weill Cornell Medical College
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Qualitative research
Keywords:	QUALITATIVE RESEARCH, Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

SCHOLARONE™
Manuscripts

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

For peer review only

Existing Barriers and Recommendations of Real-World Data Standardization for Clinical Research in China: A Qualitative Study

Junkai Lai[1], Xiwen Liao[1], Chen Yao[13], Feifei Jin[2], Bin Wang[1], Chen Li[4], Jun Zhang[5], Larry Liu[6,7]
1 Peking University Clinical Research Institute, Peking University First Hospital, Beijing, China
2 National Center for Trauma Medicine, Peking University People's Hospital, Beijing, China
3 Hainan Institute of Real World Data, Qionghai, Hainan, China.
4 Department of Health Statistics, School of Preventive Medicine, Fourth Military University, Xi'an, Shaanxi, China
5 MSD R&D (China) Co., Ltd.., Beijing, China
6 Merck & Co., Inc., Rahway, NJ, USA
7 Weill Cornell Medical College, New York, NY, USA

Correspondence to
Chen Yao
Peking University First Hospital, Xicheng District, Beijing 100034, China
Tel +86 18610640562
Email yaochen@hsc.pku.edu.cn

Keywords real world data; data standards; data standardization; clinical research; qualitative research; China;

Word Count
5238

For peer review only

Existing Barriers and Recommendations of Real-World Data Standardization for Clinical Research in China: A Qualitative Study

Abstract

Objective To investigate the existing barriers and recommendations of real-world data (RWD) standardization 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 8 Hospitals, 4 Hospital System Vendors, 3 Big Data Companies, 6 Medical Products Companies, and 4 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 generalizability in existing research databases, and lack of transparency in existing data standardization process were the barriers of data standardization of RWD for clinical research. Enhancing terminology standards by incorporating locally used clinical terminology, reducing burden in the usage of terminology standards, improving generalizability 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 standardization 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 generalizability of RWD for research by using clinical data models, and improving traceability to source data for transparency to guide efforts in data standardization in the future.

Strengths and Limitations of this study:

- Strength: Wide variety of relevant stakeholders on the subject
- Strength: Qualitative understanding of a major industry bottleneck
- Strength: Important recommendations that can guide the direction of the future of the subject
- Limitations: 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
- Limitations: Recruitment of participants were limited to those that were already exploring the subject, which could result in selection bias.

For peer review only

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 United States (U.S.) and in China, RWD have become increasingly used to support regulatory decision making for drugs 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 (EMR), 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 U.S. Food and Drug Administration (FDA), which includes promoting the direct usage of electronic source data (eSource) from real world data 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 organizations like Health Level Seven (HL7) and Clinical Data Interchange Standards Consortium (CDISC), which produced solutions harmonizing 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 Center 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 standardization 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 standardization of RWD for clinical research in China through a qualitative study conducted on industry-wide stakeholders.

Methods

For peer review only

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 co-constructed 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 standardization 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 (COREQ) 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 standardization 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 centralized 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 categorized into 3 general categories: stakeholders that mainly affected the source data, stakeholders that mainly affected the standardization 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.** Detailed list of institutions for each type of stakeholder is included in the (See Appendix 1).

The inclusion criteria of the interviewees were as follows

For peer review only

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 minutes 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

Semi-structured 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 summarize 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 minutes, 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 Appendix 2,3). The general categories of questions used for each role focused on how the stakeholders affected the data standardization 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 constructivist grounded theory when they analyzed the data. QSR NVivo V.12 software was used for coding. The team developed a structured coding tree

For peer review only

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, 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 Standardization of RWD for Clinical Research

The CGT framework generated from the three stages of coding and the 62 participants' responses were summarized 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 standardization 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, 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 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, 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 standardized

For peer review only

to specific research questions and not generalizable to others. Medical product participants expressed that there is substantial variation in the type of available data even when standardized. 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 organized in a way that could be combined into a generalizable 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 Standardization Process

The findings showed that hospital and medical product participants expressed that the data standardization 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 standardization are not transparent. In addition, they had concerns over the interpretability of standardization 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 standardization process requires many steps and teams involved, which can reduce its transparency.

"The exact methods used for data standardization 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 standardization of data are whether regulatory institutions will accept them." – Medical Products Participant 4

"When vendors standardize 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

For peer review only

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 standardization. Hospital participants expressed various methods used to automatically standardize 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 Clinical Data Models

The findings showed that hospital system and big data participants expressed that the usage of clinical data model standards to organize 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 reusage of data for different services including clinical decision

For peer review only

support services. Big data participants suggested the usage of the OHDSI data model to organize 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 analyzing 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 7 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 standardization 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 standardize 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 standardization of RWD for clinical research are the research questions central to the current qualitative study. Through a constructivist grounded theory 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 standardization 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 clinical data models, and improving traceability to source data for transparency. The grounded theory used in the paper was applied to address a specific problem

For peer review only

regarding the difficulty in RWD standardization 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 standardization of data elements and limited focus on terminology standards, preventing meaningful exchange of information at the semantic level [20]. Thus, researchers believed that the localization 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 (HKCTT), 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 generalizability in existing research databases. The

lack of generalizability of databases can lead to the limited usage of RWD even after standardization since the databases only address a specific question. Thus, the usage of clinical data models can improve the generalizability of databases by organizing RWD in a consistent and research relevant way to enable the answering of research questions. In the US, 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 generalizable common clinical data model (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 clinical data model that integrated data standards common to EHRs with the goal of better organizing 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 organizations is evident. Confirming the experiences of the participant in the current 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 clinical data model 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 standardization process. The lack of well-documented and understandable methods used in the data standardization 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 standardize 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 standardization 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,

For peer review only

suggestions in China were made to develop and utilize 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 real world data 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 minimize 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 standardization for clinical research based on constructivist grounded theory. 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 standardization process. 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, 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 standardization 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 standardize RWD for clinical research with efficiency and reliability. Secondly, 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 standardization of RWD for clinical research.

Figure 1 Caption Barrier and Suggestions in Data Standardization of Real-World Data 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 WB analyzed 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.

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

Ethics Approval Ethical approval was obtained from Peking University Institutional Review board (No. IRB00001052-21081).

Competing Interest Statement No, there are no competing interests for any author

Table 1 Demographics of the participants

Type of Stakeholder (# of Institutions)	Total Number of Participants
Hospital (8)	16
Hospital System Vendor (4)	10
Big Data Company (3)	15
Pharmaceutical (6)	12
Regulatory (4)	9

For peer review only

References

1 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. In, 2017.

2 US Food And Drug Administration. REAL-WORLD EVIDENCE PROGRAM. In, 2018.

3 J Corrigan-Curay, L Sacks, J Woodcock. Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. *JAMA* 2018;320(9):867-68.

4 X Sun, J Tan, L Tang, JJ Guo, X Li. Real world evidence: experience and lessons from China. *BMJ* 2018;360:j5262.

5 National Medical Product Association. Real world evidence supports the guiding principles of drug development and review. In, 2020.

6 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. In, 2021.

7 National Medical Products Association. The State Food and Drug Administration launched the scientific action plan for China's drug supervision. In, 2021.

8 Johnson Johnson Surgical Vision. A Real-World Evidence Study in China of the Catalys Precision Laser System. In, 2020.

9 R Blitz, M Dugas. Conceptual Design, Implementation, and Evaluation of Generic and Standard-Compliant Data Transfer into Electronic Health Records. *Appl Clin Inform* 2020;11(3):374-86.

10 Y Matsumura, A Hattori, S Manabe, et al. Interconnection of electronic medical record with clinical data management system by CDISC ODM. *Stud Health Technol Inform* 2014;205:868-72.

11 Y Wu, D Yin, K Abbasi. China's medical research revolution. *BMJ* 2018;360:k547.

12 US Food And Drug Administration. Electronic Source Data in Clinical Investigations. In, 2013.

13 US Food And Drug Administration. Use of Electronic Health Record Data in Clinical Investigations. In, 2018.

14 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. In, 2020.

15 National Health Commission of the People's Republic of China. Electronic medical record sharing document specification. In, 2020.

16 National Health Commission of the People's Republic of China. Administrative measures for hierarchical evaluation of application level of electronic medical record system. In, 2018.

17 National Health Commission of the People's Republic of China. Standardized maturity evaluation scheme for hospital information interconnection. In, 2020.

18 National Health Committee Of The People'S Republic Of China. Basic architecture and data standard of electronic medical record. In, 2009.

19 F Jin, C Yao, X Yan, et al. Gap between real-world data and clinical research within hospitals in China: a qualitative study. *Bmj Open* 2020;10(12):e38375.

20 J Xie, EQ Wu, S Wang, et al. Real-World Data for Healthcare Research in China: Call for Actions. *Value Health Reg Issues* 2022;27:72-81.

21 M Byrne. Grounded theory as a qualitative research methodology. *Aorn J* 2001;73(6):1155-56.

22 AL Chapman, M Hadfield, CJ Chapman. Qualitative research in healthcare: an introduction to

For peer review only

grounded theory using thematic analysis. *J R Coll Physicians Edinb* 2015;45(3):201-05.

23 FK Metelski, J Santos, C Cechinel-Peiter, et al. Constructivist Grounded Theory: characteristics and operational aspects for nursing research. *Rev Esc Enferm Usp* 2021;55:e3776.

24 J Mills, A Bonner, K Francis. Adopting a constructivist approach to grounded theory: Implications for research design. *Int J Nurs Pract* 2006;12(1):8-13.

25 A Tong, P Sainsbury, J Craig. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care* 2007;19(6):349-57.

26 A Moser, I Korstjens. Series: Practical guidance to qualitative research. Part 3: Sampling, data collection and analysis. *Eur J Gen Pract* 2018;24(1):9-18.

27 MS Setia. Methodology Series Module 5: Sampling Strategies. *Indian J Dermatol* 2016;61(5):505-09.

28 J Sutton, Z Austin. Qualitative Research: Data Collection, Analysis, and Management. *Can J Hosp Pharm* 2015;68(3):226-31.

29 K Peters, E Halcomb. Interviews in qualitative research. *Nurse Res* 2015;22(4):6-07.

30 LS Whiting. Semi-structured interviews: guidance for novice researchers. *Nursing standard* 2008;22(23):35-40.

31 N Britten. Qualitative interviews in medical research. *BMJ* 1995;311(6999):251-53.

32 O Bodenreider, R Cornet, DJ Vreeman. Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. *Yearb Med Inform* 2018;27(1):129-39.

33 Y Zhu, H Pan, L Zhou, et al. Translation and localization of SNOMED CT in China: a pilot study. *Artif Intell Med* 2012;54(2):147-49.

34 R Zhang, J Liu, Y Huang, 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(1):54.

35 Y Cheng, T Jiang, L Deng, L Chen, J Ming. Research on the Coverage of Standard Chinese Medical Terminology to Practical Application. *Chinese Journal of Health Informatics and Management* 2020.

36 H Forsvik, V Voipio, J Lamminen, et al. Literature Review of Patient Record Structures from the Physician's Perspective. *J Med Syst* 2017;41(2):29.

37 K Hung, M Lau, V Fung. Successful Implementation of Terminology Binding in Hong Kong Hospital Authority. *Stud Health Technol Inform* 2019;264:1486-87.

38 Y Chen, D Hu, M Li, H Duan, X Lu. Automatic SNOMED CT coding of Chinese clinical terms via attention-based semantic matching. *Int J Med Inform* 2022;159:104676.

39 JM Overhage, PB Ryan, CG Reich, AG Hartzema, PE Stang. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* 2012;19(1):54-60.

40 PE Stang, PB Ryan, JA Racoosin, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med* 2010;153(9):600-06.

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. In, 2021.

42 HX Yue, YL Zhan, F Bian, et al. [Data standard and data sharing in clinical cohort studies]. *Zhonghua Liu Xing Bing Xue Za Zhi* 2021;42(7):1299-305.

43 A Prats-Uribe, AG Sena, L Lai, et al. Use of repurposed and adjuvant drugs in hospital patients with covid-19: multinational network cohort study. *BMJ* 2021;373:n1038.

44 X Zhang, L Wang, S Miao, et al. Analysis of treatment pathways for three chronic diseases using

For peer review only

OMOP CDM. *J Med Syst* 2018;42(12):260.

45 H Ji, S Kim, S Yi, 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. In, 2021.

47 S Hume, S Sarnikar, L Becnel, D Bennett. Visualizing and Validating Metadata Traceability within the CDISC Standards. *AMIA Jt Summits Transl Sci Proc* 2017;2017:158-65.

48 Society for Clincial Data Management. eSource Implementation in Clinical Research: A Data Management Perspective. In, 2014.

49 E Kellar, SM Bornstein, A Caban, et al. Optimizing the Use of Electronic Data Sources in Clinical Trials: The Landscape, Part 1. *Ther Innov Regul Sci* 2016;50(6):682-96.

50 J Xinyao, Z Wenke, Z Junhua, et al. Promote Transparency in Real-World Study. *World Chinese Medicine* 2019.

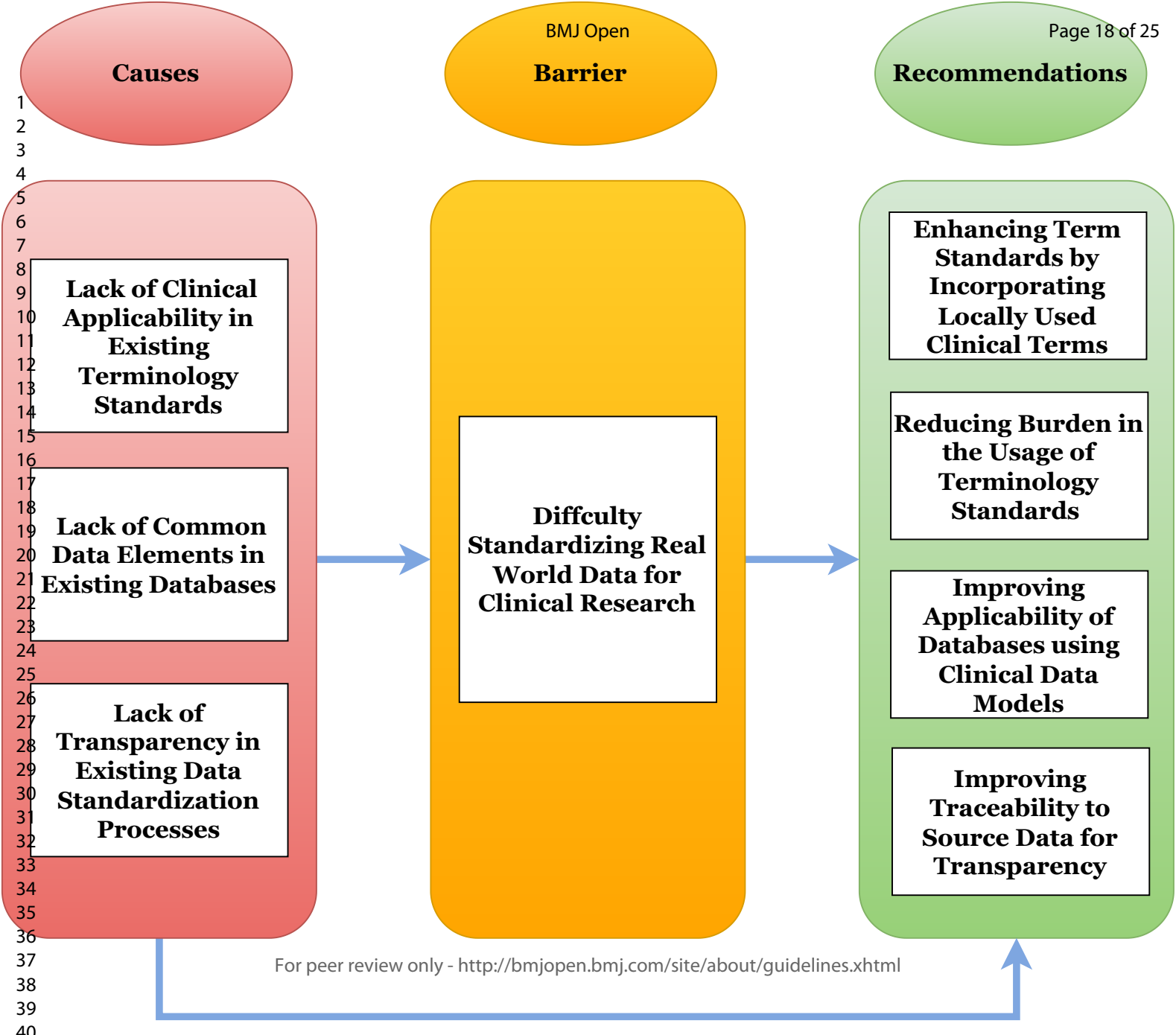
51 X Yan, C Dong, C Yao. Protecting the accuracy of clinical trial data in China. *BMJ Opinion* 2018.

52 C Dong, X Yan, R Tian, Z Bian, C YAO. Strengthen the process report of clinical trials, promote full transparency of clinical
trials. *Chinese Journal of Evidence-Based Medicine* 2018.

53 F Jin, C Yao, J Ma, et al. 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. In, 2021.

For peer review only



For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Appendix 1:
List of Institutions:
Hospitals:

1. Peking University People's Hospital, Beijing, Beijing, China
2. Peking University First Hospital, Beijing, Beijing, China
3. First Teaching Hospital of Tianjin University of Traditional Chinese Medicine, Tianjin, Tianjin, China
4. Hainan General Hospital, Haikou, Hainan, China
5. Boao Evergrande International Hospital, Boao, Hainan, China
6. Boao Super Hospital, Boao, Hainan, China
7. Boao Yiling Lifecare Center, Hainan, China
8. Boao Worldlight Hospital, Hainan, China

Hospital System Vendors:

1. Haitai International
2. Goodwill
3. Winning Health
4. Orion Health Rhapsody

Big Data Companies:

1. Yiducloud
2. Digital Health China Technologies
3. Inspur

Pharmaceutical Companies:

1. Pfizer
2. Tigermed
3. AstraZeneca
4. Bristol-Meyers Squibb
5. Johnson & Johnson
6. BeiGene

Regulatory Institutions:

1. China National Health Development Research Center
2. National Medical Products Administration
3. China Center for Food and Drug International Exchange
4. Hainan Boao Lecheng International Medical Tourism Pilot Zone Administration

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

Appendix 2:
Interview Guide:

Hospital:

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience facilitating clinical research at the Hospital?
3. What are the motivating goals of clinical research?
4. How do you determine the data needed for your research? What are the barriers and recommendations?
5. Do you think that electronic medical records or routine care data at the hospital are enough to accomplish your research? What are the barriers and recommendations?
6. How do you use data standards to aggregate and store all data from different hospital systems? What are the barriers and recommendations?
7. What data standards are used and how do you implement data standards during routine data collection? What are the barriers and recommendations?
8. What areas in your clinical research process do you have to rely on external vendors to help you standardize the data and how have you evaluated their data standardization process? What are the barriers and recommendations?
9. How are data standards used to share medical records inside and outside of the hospital? What are the barriers and recommendations?
10. Beyond clinical research, have you standardized your data for other purposes?

Big Data:

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience handling data at the company?
3. Describe your interaction with clients that want to utilize your service for clinical research?
4. What type of standards are your clients required to fulfill?
5. How do you use data standards to organize and aggregate source data? What are the barriers and recommendations?
6. How do you use data standards when you transform source data into research datasets? What are the barriers and recommendations?
7. What methods are used to standard source data for clinical research? What are the barriers and recommendations?
8. How do you track and evaluate the quality of the data transformation process? What are the barriers and recommendations?
9. How do you manage the variety of standards that are published? What are the barriers and recommendations?
10. How do you manage the different research projects that need to use real world data? What are the barriers and recommendations?

Hospital System Vendor:

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience developing hospital systems? What type of systems does your company produce?
3. What is behind the motivation for hospitals to use data standards?
4. Describe the data standards that are used for hospital systems?
5. How are data standards implemented and customized for the hospital? What are the barriers and recommendations?
6. How do you use data standards to improve internal and external communication at the hospital? What are the barriers and recommendations?

Medical Products Company:

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience in using real world data for clinical research? What types of real-world data do you use as source data for your studies?
3. How do you obtain or access real world data?
4. How does the process of sourcing real world data differ from the traditional data collection for clinical research the most?
5. What standards are used for real world data?
6. What data standards would you like to see used for real world data?
7. What standardization methods for real world data are used to produce research data? What are the barriers and recommendations?
8. How do you check whether the data is reliable and what types of data do you think are most reliable? What are the barriers and recommendations?
9. Does real world data meet your research needs? What are the barriers and recommendations?
10. Given regulatory consideration for the usage of real-world data for clinical research, what do you see as the problems in the current methods used to standardize data? What are some recommendations?

Regulatory:

1. This interview will be recorded and used in a qualitative study, your identity will be concealed to protect your privacy, do we have your full consent in this interview and have your signed information consent form?
2. Describe your role and experience in regulating real world studies? What types of real-world data do you see used as source data for these studies?
3. What are the common characteristics of clinical studies using real world data do you often see (study design, phase, purpose)?
4. How is real world data used to support regulatory decision making?

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

5. How does the process of sourcing real world data differ from the traditional data collection for clinical research the most?
6. What standards are used for existing real-world data?
7. What data standards are necessary for the submission of real-world data used to gain product approval?
8. How can the standardization of real-world data for clinical research meet regulatory expectations? What are the barriers and recommendations?
9. How can we establish a real-world data platform that can be used for clinical research that can benefit the most stakeholders in China? What are the barriers and recommendations?

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

Appendix 3:
General Categories for Interview Questions:

General Category	Hospital	Hospital System Vendor	Big Data Company	Medical Products Company	Regulatory Department
Privacy and Information Consent Statement	1	1	1	1	1
Experience and aim when using RWD for clinical research	2,3,10	2,3	2,3	2,3,4	2,3,4,5
Relevant RWD Standards	7	4	4,9	5,6	6,7
RWD relevance for clinical research	4,5		10	8,9	9
Standardization of RWD at source	6,7,9	5,6	5		
Standardization of RWD for clinical research	8		6,7	7	

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

Reliability of RWD Standardization for clinical research	8		8	8,10	8

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2021-059029 on 3 August 2022. Downloaded from http://bmjopen.bmj.com/ on April 17, 2024 by guest. Protected by copyright.

Consolidated criteria for reporting qualitative studies (COREQ): 32-item checklist

Please indicate in which section each item has been reported in your manuscript. If you do not feel an item applies to your manuscript, please enter N/A.

For further information about the COREQ guidelines, please see Tong *et al.*, 2017: https://doi.org/10.1093/intqhc/mzm042

No.	Item	Description	Section #
Domain 1: Research team and reflexivity			
Personal characteristics			
1.	Interviewer/facilitator	Which author/s conducted the interview or focus group?	
2.	Credentials	What were the researcher's credentials? *E.g. PhD, MD*	
3.	Occupation	What was their occupation at the time of the study?	
4.	Gender	Was the researcher male or female?	
5.	Experience and training	What experience or training did the researcher have?	
Relationship with participants			
6.	Relationship established	Was a relationship established prior to study commencement?	
7.	Participant knowledge of the interviewer	What did the participants know about the researcher? *E.g. Personal goals, reasons for doing the research*	
8.	Interviewer characteristics	What characteristics were reported about the interviewer/facilitator? *E.g. Bias, assumptions, reasons and interests in the research topic*	
Domain 2: Study design			
Theoretical framework			
9.	Methodological orientation and theory	What methodological orientation was stated to underpin the study? *E.g. grounded theory, discourse analysis, ethnography, phenomenology, content analysis*	
Participant selection			
10.	Sampling	How were participants selected? *E.g. purposive, convenience, consecutive, snowball*	
11.	Method of approach	How were participants approached? *E.g. face-to-face, telephone, mail, email*	
12.	Sample size	How many participants were in the study?	
13.	Non-participation	How many people refused to participate or dropped out? What were the reasons for this?	
Setting			
14.	Setting of data collection	Where was the data collected? *E.g. home, clinic, workplace*	
15.	Presence of non-participants	Was anyone else present besides the participants and researchers?	

For peer review only

16.	Description of sample	What are the important characteristics of the sample? *E.g. demographic data, date*	
Data collection			
17.	Interview guide	Were questions, prompts, guides provided by the authors? Was it pilot tested?	
18.	Repeat interviews	Were repeat interviews carried out? If yes, how many?	
19.	Audio/visual recording	Did the research use audio or visual recording to collect the data?	
20.	Field notes	Were field notes made during and/or after the interview or focus group?	
21.	Duration	What was the duration of the interviews or focus group?	
22.	Data saturation	Was data saturation discussed?	
23.	Transcripts returned	Were transcripts returned to participants for comment and/or correction?	
Domain 3: analysis and findings			
Data analysis			
24.	Number of data coders	How many data coders coded the data?	
25.	Description of the coding tree	Did authors provide a description of the coding tree?	
26.	Derivation of themes	Were themes identified in advance or derived from the data?	
27.	Software	What software, if applicable, was used to manage the data?	
28.	Participant checking	Did participants provide feedback on the findings?	
Reporting			
29.	Quotations presented	Were participant quotations presented to illustrate the themes / findings? Was each quotation identified? *E.g. Participant number*	
30.	Data and findings consistent	Was there consistency between the data presented and the findings?	
31.	Clarity of major themes	Were major themes clearly presented in the findings?	
32.	Clarity of minor themes	Is there a description of diverse cases or discussion of minor themes?	

When submitting your manuscript via the online submission form, please upload the completed checklist as a Figure/supplementary file.

If you would like this checklist to be included alongside your article, we ask that you upload the completed checklist to an online repository and include the guideline type, name of the repository, DOI and license in the *Data availability* section of your manuscript.

Developed from: Allison Tong, Peter Sainsbury, Jonathan Craig, Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups, International Journal for Quality in Health Care, Volume 19, Issue 6, December 2007, Pages 349–357, https://doi.org/10.1093/intqhc/mzm042