BMJ Open

Advancing the development of real-world data for healthcare research in China: challenges and opportunities

Jia Zhong,[1] Jun Zhang,[2] Honghao Fang ,[1] Larry Liu,[3,4] Jipan Xie,[5] Eric Wu[6]

To cite: Zhong J, Zhang J, Fang H, *et al*. Advancing the development of real-world data for healthcare research in China: challenges and opportunities. *BMJ Open* 2022;**12**:e063139. doi:10.1136/bmjopen-2022-063139

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-063139).

John Cai, Can Chen, Erik Dasbach, Mengchun Gong, Yunfei Pei, and Yaoping Ruan presented in the workshop, which contributed to the manuscript.

Received 21 March 2022
Accepted 14 July 2022

[1]Analysis Group, Inc, Beijing, China
[2]MSD R&D (China) Co., Ltd, Beijing, China
[3]Merck & Co., Inc, Rahway, New Jersey, USA
[4]Weill Cornell Medical College, New York, New York, USA
[5]Analysis Group, Inc, Los Angeles, California, USA
[6]Analysis Group, Inc, Boston, Massachusetts, USA

Correspondence to
Dr Jia Zhong;
Jia.Zhong@analysisgroup.com

ABSTRACT

Various real-world data (RWD) sources have emerged in China with the intention of generating real-world evidence (RWE) that can be used in clinical and regulatory decision-making. Despite these efforts, significant barriers remain that hinder high-quality healthcare research. A workshop with 30 representatives from healthcare research agencies, technology companies focused on healthcare big data and pharmaceutical companies was held in December 2020 to identify strategies to overcome the barriers associated with the usability and quality of RWD in China. Across all sectors, examples of barriers identified included inconsistencies in terminology and non-standardised coding practices; the absence of longitudinal data; the absence of transparent data processing and validation practices; and the inability to access and share RWD. While cutting-edge technological innovations and data solutions provided powerful tools, the development of collaborative and synergistic research networks across multiple stakeholders is key to generate accessible, high-quality RWD in China. RWD has the potential to provide clinical, regulatory and reimbursement decision-makers with critical insights that can improve healthcare delivery in China. However, barriers to its access, collection and use must be addressed to generate RWE to guide healthcare stakeholders.

OVERVIEW

The collection and use of real-world data (RWD) in healthcare research is important to complement and supplement safety and efficacy data obtained from clinical trials, which often are not representative of the general patient population and may have limited follow-up time.[1] As such, real-world evidence (RWE) generated from RWD can provide invaluable information about long-term effectiveness and safety outcomes in large, heterogeneous, general patient populations of real-world clinical practice.[1]

Within the last decade, different types of RWD sources (eg, administrative claims and electronic heath records (EHR)) have emerged in China. Previously, we conducted a comprehensive review and evaluation of these secondary data sources in China and compared them with sources available in other countries.[2] When compared with other countries, there were two main limitations that we identified with secondary data sources that have emerged in China: data access and data quality.[2] To address these limitations, Analysis Group, Inc. invited 30 representatives from various healthcare sectors (ie, healthcare research agencies, technology companies focused on healthcare big data and pharmaceutical companies) to a workshop in December 2020. Representatives signed non-disclosure agreements and provided informed consent before attending the workshop. During the workshop, case studies with insights and experiences in RWD development were presented, followed by a round table discussion about the causes of barriers associated with the use of RWD in China, and ways to overcome the barriers. The four barriers we focused on addressing were as follows: (1) inconsistent terminology and non-standardised coding that impede linkage of data across different sources; (2) a lack of longitudinal data that prevents reconstruction of a patient's complete clinical pathway (eg, understanding treatment patterns, healthcare resource use and clinical outcomes); (3) a lack of transparency in data processing and validation, which has affected the quality of research conducted using RWD in China and (4) the inability to access and share data, which has restricted the use of government-sponsored data to academic researchers only and limited the widespread adoption of privately sponsored data. Following the workshop, secondary research with a literature review was conducted to summarise initiatives in overcoming the barriers and future directions.

The importance of RWE in clinical and regulatory decision making has been increasingly recognised in China, with policies and guidelines published in recent years. In January 2020, the National Medical Products Administration (NMPA) of China published

'Guidance on Real-World Evidence Supporting Drug Development and Review (Pilot)', which outlined the definition and sources of RWD and provided guidance on using RWE in supporting drug review, indication expansion, postapproval evaluation, and research and development of traditional Chinese medicine.[3] Following the publication of that guidance, a technical guideline on the development and review of drugs for children was released in September 2020 by the Centre for Drug Evaluation, an affiliated institution of the NMPA.[4] Besides drugs, RWD are also used in the clinical evaluation of medical devices, for which a technical guideline was published by the NMPA in November 2020.[5] Here, we summarise the key takeaways from the workshop and secondary research and discuss possible solutions to overcome barriers prohibiting the effective use of RWD in light of recent policies and guidelines.

INCONSISTENT TERMINOLOGY AND NON-STANDARDISED CODING

Variation in the ways that RWD are recorded and coded in China have contributed to inconsistencies in terminology and non-standardised data curation practices. Historically, standard coding practices were lacking when hospital information systems (HIS) or EHR were first established. Many hospitals relied on unstructured fields in patient records based on clinicians' descriptions of diagnoses and treatments. In recent upgrades of HIS, a diagnosis-related group pricing and payment schedule that requires physicians to classify hospital cases into distinct groups using established codes was introduced in China.[6] However, due to legacy issues within the HIS, there is a lack of adherence to the current coding systems, and thus, there is no standard system used across all hospitals. This issue may stem from broader inconsistencies, such as the existence of many clinical guidelines and expert-based consensus statements in China, even within the same disease,[7] and different hospitals may choose to follow different guidelines. Recent calls for standardisation include the use of the International Classification of Diseases, 10th revision (ICD-10) coding system, since it comprises significant updates that are not reflected in the 9th revision (ICD-9).[8] However, the adoption of the new coding system often requires infrastructural updates (eg, HIS), training for physicians, and substantial logistic support to ensure coding adherence. Therefore, the use of ICD-9 and ICD-10 codes varies substantially across hospitals. In general, ICD-10 are better adopted in tertiary hospitals compared with primary and secondary hospitals. The adoption time of the ICD-10 coding system also varies. For example, at Chongqing Cancer Hospital, the ICD-10 training was provided in 2022,[9] while at the First Affiliated Hospital Xi'an Jiaotong University, the training was provided in 2019.[10] As such, any RWD acquired must currently undergo an extensive data standardisation prior to analysis.

Establishing a common data model prior to data collection is a potential solution to address the challenges resulting from inconsistencies in data collection.[11] Using a common data model with predefined data forms/fields and entry restrictions would facilitate the generation of standardised data. The common data model has been applied to RWD in China across several projects, including a COVID-19 study conducted in Honghu City, Hubei Province in early 2020.[12 13] In this study, the Observational Medical Outcomes Partnership common data model developed in the Observational Health Data Sciences and Informatics (OHDSI) programme was used to collect and integrate heterogeneous data from a surveillance system, manual chart extraction, and the HIS.[14] However, it is worth noting that this application largely relied on manual data entry and quality checks, hence requiring extensive engagement of research staff. For greater efficiency and generalisability, multistakeholder collaboration is required to improve data coding and standardisation. For instance, in an effort to demonstrate the value of healthcare RWD using large-scale analytics, the OHDSI programme completed the initial steps to foster an international network with researchers from the pharmaceutical industry, healthcare providers and research agencies, as well as stakeholders from diverse backgrounds.[14] Collaboration with professional medical societies that may lead to the standardisation of coding systems for diseases and treatments, and with government agencies that may require their use at the policy level and provide funding to hospitals for implementation, may expedite the adoption of the standard coding systems in clinical practice and the implementation of the common data model, which will in turn form a consensus to guide data standardisation for healthcare research based on RWD.

LACK OF LONGITUDINAL DATA

In China, longitudinal healthcare data are scarce for several reasons. First, patients in China typically do not seek care at a single hospital, nor do they rely on a referral system across hospitals, especially in the outpatient setting. This results in scattered EHR data across multiple institutions that are unlikely to provide complete disease and treatment history. Therefore, reconstructing a patient's complete clinical pathway is challenging due to the absence of a healthcare referral system and the lack of collaboration across multiple sectors. Relatedly, while EHR systems are established in most hospitals, they are not often interoperable or used in the same manner across inpatient and outpatient settings. In the outpatient setting, very little information is recorded because of the large quantity of patient visits and the lack of standardised recording practices. Lastly, the recent popularity of online health counseling[15] and pharmacies[16] in China has exacerbated the data completeness issue, since drug costs and refill data in particular are not integrated into the systems in hospitals where patients often seek care.

Taken together, fragmented data across institutes, data gaps from outpatient settings, and the shift to online healthcare services are major obstacles in generating longitudinal RWD.

Several research institutes and technology companies have established initiatives to supplement inpatient data from EHR with data collected across multiple sources. For example, some databases have integrated patient-level data collected during hospitalizations and follow-up data, including patient-reported outcomes, into one system.[17] A similar approach that additionally incorporates collaboration across hospitals, regional governments and companies offering online health services may contribute to the compilation of significantly enriched longitudinal RWD in China. For example, Fudan University Shanghai Cancer Center collaborated with technology companies to establish a smart hospital information framework, which covers online counselling and mobile payments, data sharing and hospital management systems, and HIS, laboratory information system, radiology information system, picture archiving and communication system, etc. The establishment of such a framework enables integration of data across different systems and increases the potential of using data to support healthcare research. Such collaboration was encouraged by the National Health Commission in China and has been piloted in many hospitals across China.[18] To further advance multi-stakeholder collaboration and data integration, a supportive policy environment should be formed.

LACK OF TRANSPARENCY IN DATA PROCESSING AND VALIDATION

Substantial data processing and validation are often needed to transform raw data to analytical datasets for healthcare research. For example, natural language processing (NLP)-based algorithms have been used by technology companies to process data extracted from unstructured fields in EHR to derive treatment and response information. To ensure the accuracy of processed data, dedicated review teams were also established to manually review records and validate the NLP-generated results.[19–21] The data processing and validation have indeed increased the research value of RWD; however, as a proprietary process with limited information in the public domain, the quality of healthcare research based on processed data may be questioned. In addition, data curation rules may vary across technology companies, which result in limited comparability of data from different data sources.

Establishing a formal external quality assessment procedure is one approach that may examine and improve data quality. For example, Merck & Co, a multinational pharmaceutical company in Rahway, New Jersey in the USA, developed a multidimensional fit-for-purpose framework to rank available data sources and identify those that are appropriate for specific business needs.[22] Implementation of this framework has led to some success in the identification of quality issues that saved the industry from licensing low-quality data. This framework is particularly helpful in the validation of data elements that are essential in healthcare research. For example, treatment pattern and outcome analyses for metastatic cancers are often stratified by line of treatment. However, treatment line information is not always readily available in EHR and should be inferred based on data of treatment schedules, treatment gaps and treatment response, among others. Establishing standardised data curation rules for line of treatment can ensure the quality and integrity of data and enable relevant research studies based on multiple data sources. The long-term objectives of this framework are to improve the quality and reliability of RWD by facilitating cross-stakeholder collaboration and building partnerships with data providers to develop and maintain quality standards for RWD and RWE that are accepted by all healthcare decision-makers.

BARRIERS TO ACCESSING AND SHARING DATA: MULTISTAKEHOLDER COLLABORATION

Among the barriers hindering the use of RWD, the inability to access and share data are significant challenges in China. Several reasons may contribute to these barriers, including the lack of clear regulations governing data sharing and access to secondary databases. Without a central ethics review board in China, study applications must go through each participating hospital or medical centre individually. In addition, as per the Biosecurity Law of China and the Regulations on Administration of Human Genetic Resources, studies using healthcare RWD are required to go through a Human Genetic Resource Administration of China (HGRAC) approval process.[23 24] It is important for the researchers to incorporate the timeline of ethics review at individual hospitals or centres and HGRAC approval into the process of obtaining data access, as these steps are essential and required in China to ensure data security and privacy. As RWD evolves in China, a more time-efficient process of review and approval may be developed to facilitate data access and application.

In the short term, creative technical applications that foster collaboration and analytical integration across systems, while protecting data privacy and security, may be a solution to maximise the use of existing data needed to generate RWE. Federated analytics, a promising technology for data sharing without data transfer, is one such option. Instead of transferring data collected from various sources to a central repository, analytical approaches are developed centrally and divided based on the data source.[25] Federated analytics alleviates the data collection burden, avoids data privacy and security issues, and can be modified as needed based on the data accessed.

In the long term, development of a collaborative and synergistic research network is required to generate quality data, which can then be efficiently converted to RWE. A core feature of a collaborative and synergistic research network is that it requires collaboration across multiple

technology companies and stakeholders to provide the necessary infrastructure to support sharing of different data sources (eg, hospitals for EHR data, health security bureau for claims data, government, pharmaceutical industry and patient advocacy groups) on a joint network. Such programmes have been successfully started in other countries. For instance, Merck & Co in Rahway, New Jersey, USA recently spearheaded an initiative that aims to accelerate the generation of RWE through an innovative data platform called Real World Data Exchange (RWDEx).[26] In collaboration with renowned international research institutes, healthcare centres, and technology companies, the platform was developed to streamline data collection, data input and transformation, data storage, analytics and visualisation, to support healthcare decision-making and timely response to drug safety issues. Merck & Co is currently exploring opportunities to apply the collaboration model in China to advance the development of RWD for healthcare research.

High-quality RWD will eventually be used to generate RWE that can inform healthcare decision-making. To achieve this, it will be necessary to collaborate with industry experts to establish an RWE network that can reduce data access barriers. A successful example of this is the collaboration of the US Food and Drug Administration (FDA) with healthcare big data companies to monitor the safety of medical products (eg, FDA Sentinel Initiative)[27] and utilisation of RWD in drug review and regulatory approval (eg, RWE programme).[28] The benefits resulting from this collaboration are anticipated to improve medical development and efficient delivery of therapies to patients in need.

It is important to note that the development of a collaborative and synergistic research network should not solely focus on concurrent research needs, but also be innovative and have enough flexibility to adapt as necessary. Such a design often requires accurate temporal sequences of data, well-defined variables and a standardised procedure to minimise bias across multiple data sources and institutes. One such example is the collaborative effort across patient advocacy groups, business strategists in multiple pharmaceutical companies, and healthcare researchers in multiple countries to design a study of Duchenne muscular dystrophy (DMD).[29] As a rare disease, it is challenging to recruit patients with DMD for clinical trials or collect RWD on this disease by a single institute. The collaboration enabled the collection and sharing of real-world and clinical trial data on DMD from multiple data sources to construct patient cohorts, which facilitated the conduct of rapid, scalable, extensive and interoperable analytics, and the investigation of new drugs.[29]

CONCLUSIONS

Despite the barriers to RWD in healthcare research in China, the innovative solutions shared by participants during this workshop and the successful initiatives and programmes observed globally provide insight on the potential for overcoming the barriers. Technological innovations and multistakeholder collaboration are integral in the development of high-quality RWD in China. In the short run, well-designed common data models, data integration from multiple sources, and a balanced fit-for-purpose quality assessment framework are warranted to tackle technical challenges and enhance the value of RWD in healthcare research. In the long run, consensus on the RWD development (eg, data standardisation, data access, quality assessment) and application (eg, regulatory use, policy implications) need to be reached across multiple sectors to continue the strong evolvement of RWD in healthcare and advance policies and standards of high-quality RWE. As demonstrated by the successful collaborative efforts taken in other countries, multistakeholder interactions to enable efficient access to RWD will in turn further biomedical development and improve clinical quality. Additional focus on knowledge dissemination of high-quality RWE is instrumental for future biomedical research and drug development.

Acknowledgements Gloria DeWalt and Christine Tam, who are employees of Analysis Group, Inc., provided medical writing assistance.

Contributors JiZ, JuZ, HF, LL, JX and EW were involved in the conception and design of the workshop. JiZ, JuZ and HF were involved in the execution of the workshop and the drafting of the paper. JiZ, JuZ, HF, LL, JX and EW were involved in revising the paper critically for intellectual content and the final approval of the version to be published. JiZ, JuZ, HF, LL, JX and EW agree to be accountable for all aspects of the work.

Funding This work was supported by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Rahway, NJ, USA. The sponsor was involved in all stages of the workshop, study research and manuscript preparation.

Competing interests JiZ, HF and EW are employees of Analysis Group, Inc., a consulting company that has provided paid consulting services to Merck Sharp & Dohme, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA, which funded the development and conduct of this study. JX was an employee of Analysis Group, Inc. at the time of study conduct. JuZ is an employee of MSD R&D (China) Co., Ltd. LL is an employee of Merck Sharp & Dohme, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD

Honghao Fang http://orcid.org/0000-0003-1623-4314

REFERENCES

1 Katkade VB, Sanders KN, Zou KH. Real world data: an opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. *J Multidiscip Healthc* 2018;11:295–304.
2 Xie J, Wu E, Wang S. Real-World Data for Health Care Research in China: Call for Actions in press. *Value Health Reg Issues*;2021.
3 National Medical Products Administration. Guidance on real-world evidence supporting drug development and review (pilot), 2020. Available: https://www.nmpa.gov.cn/yaopin/ypggtg/ypqtgg/20200107151901190.html [Accessed 07 Dec 2021].
4 Center for Drug Evaluation of the National Medical Products Administration. Technical guideline on real-world studies in

supporting pediatric drug development and review (pilot), 2020. Available: https://www.nmpa.gov.cn/xxgk/ggtg/qtggtg/20200901104448101.html [Accessed 07 Dec 2021].
5 National Medical Products Administration. Technical guideline on real-world data in clinical evaluation of medical device (pilot), 2020. Available: https://www.nmpa.gov.cn/xxgk/ggtg/qtggtg/20201126090030150.html [Accessed 07 Dec 2021].
6 Jian W, Lu M, Chan KY, *et al*. The impact of a pilot reform on the diagnosis-related-groups payment system in China: a difference-in-difference study. *The Lancet* 2015;386:S26.
7 Chen Y, Wang C, Shang H, *et al*. Clinical practice guidelines in China. *BMJ* 2018;360:j5158.
8 Centers for Disease Control and Prevention. International classification of diseases. (ICD-10-CM/PCS) Transition - Background. Available: https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm [Accessed 17 Feb 2021].
9 Chongqing Cancer Hospital. Chongqing cancer Hospital completes the online training on ICD-10 coding, 2022. Available: https://www.cqch.cn/departments_baglk_ksdt/2022/xbo7pBbg.html [Accessed 31 May 2022].
10 The First Affiliated Hospital Xi'an Jiaotong University. Notice about a training on ICD-10 coding, 2019. Available: http://www.dyyy.xjtu.edu.cn/info/1780/33868.htm [Accessed May 31, 2022].
11 Lai EC-C, Ryan P, Zhang Y, *et al*. Applying a common data model to Asian databases for multinational pharmacoepidemiologic studies: opportunities and challenges. *Clin Epidemiol* 2018;10:875–85.
12 Digital China health. Available: http://www.dchealth.com/dch-en/#/ [Accessed 17 Feb 2021].
13 American Medical Informatics Association. Managing the global COVID-19 pandemic with health informatics. Available: https://www.amia.org/sites/default/files/AMIA-COVID19-Webinar-Series-Global-Health-East-Asia.pdf [Accessed 17 Feb 2021].
14 Observational Health Data Sciences and Informatics. OMOP common data model. Available: https://www.ohdsi.org/data-standardization/the-common-data-model/ [Accessed 17 Feb 2021].
15 Le W, Chang P-Y, Chang Y-W, *et al*. Why do patients move from online health platforms to hospitals? the perspectives of fairness theory and brand extension theory. *Int J Environ Res Public Health* 2019;16. doi:10.3390/ijerph16193755. [Epub ahead of print: 06 10 2019].
16 Gu Shanshan with Xinhua Silk Road. Online drug purchase accelerates dev. of China's internet economy amid epidemic, 2020. Available: https://en.imsilkroad.com/p/313379.html [Accessed 16 Feb 2021].
17 LinkDoc. 人工智能与医疗大**数据** 解决方案提供者 (artificial intelligence and medical big data solution provider) [content in Chinese]. Available: https://www.linkdoc.com/ [Accessed 17 Feb 2021].
18 The National Health Commission of the People's Republic of China. Strengthen the construction of smart hospitals with the support of informatization— —The National health Commission distributed materials at its regular press conference on March 21, 2019. Available: http://www.nhc.gov.cn/xcs/s7847/201903/c87c208841f14f76afcc0efa022d2126.shtml [Accessed 09 Jun 2022].
19 Happy Life Technology. Big data and medical AI technology. Available: https://www.hltpharma.com/en-index.html [Accessed 17 Feb 2021].
20 Chen L, Song L, Shao Y, *et al*. Using natural language processing to extract clinically useful information from Chinese electronic medical records. *Int J Med Inform* 2019;124:6–12.
21 Liu H, Xu Y, Zhang Z, *et al*. A natural language processing pipeline of Chinese Free-Text radiology reports for liver cancer diagnosis. *IEEE Access* 2020;8:159110–9.
22 Desai K, Chandwani S, Ru B, *et al*. Fit-For-Purpose real-world data assessments in oncology: a call for Cross-Stakeholder collaboration. *Value & Outcomes Spotlight* 2021;7:34–7.
23 State Council. The regulations on administration of human genetic resources, 2019. Available: http://www.gov.cn/zhengce/content/2019-06/10/content_5398829.htm [Accessed 31 May 2022].
24 The National People's Congress of the People's Republic of China. The biosecurity law of China, 2020. Available: http://www.npc.gov.cn/npc/c30834/202010/bb3bee5122854893a69acf4005a66059.shtml [Accessed 31 May 2022].
25 McMahan B, Ramage D. Federated learning: collaborative machine learning without centralized training data, 2017. Available: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html [Accessed 22 Jun 2021].
26 Merck Sharp & Dohme. Corporate responsibility report. Available: https://www.msdresponsibility.com/access-to-health/discovery-invention/product-patient-safety/ [Accessed 17 Feb 2021].
27 U.S. Food and Drug Administration. FDA's Sentinel Initiative, 2019. Available: https://www.fda.gov/safety/fdas-sentinel-initiative [Accessed 16 Feb 2021].
28 U.S. Food and Drug Administration. Framework for FDA's Real-World Evidence Program, 2018. Available: https://www.fda.gov/media/120060/download [Accessed 16 Feb 2021].
29 Collaborative trajectory analysis project. Available: https://www.ctap-duchenne.org/ [Accessed 16 Feb 2021].