http://www.cde.org.cn/attachmentout.do?mothed=list&id=0e208b282a448d40

用于产生真实世界证据的真实世界数据

指导原则

（试行）

April 2021

Attachment: "Guiding Principles of Real World Data Used to Generate Real World Evidence (Trial)"

I. Overview

Real-world evidence is an important part of the evidence chain of drug effectiveness and safety evaluation, and its related concepts and applications can be found in The Guiding Principles of Real-World Evidence in Support of Drug Development and Review (Trial). Real-world data is the basis for producing real-world evidence, without high-quality applicable real-world data support, real-world evidence is impossible to talk about.

Real-world data are data collected on a daily basis related to a patient's health and/or diagnosis and health care. Not all real-world data can be analyzed to produce real-world evidence, and real-world evidence is possible only after proper and adequate analysis of

real-world data that meets applicability. At present, the real world data recording, collection, storage and other processes lack strict quality control, there may be incomplete data, data standards, data models and description methods are not uniform and other issues, the effective use of real world data has formed an obstacle. Therefore, how to make the collected real-world data available or, after governance, the analytical data needed to meet clinical research objectives, and how to assess whether real-world data are appropriate for generating real-world evidence, is a key issue in using real-world data to form real-world evidence to support drug regulatory decisions.

This Guiding Principle complements the Guiding Principles (Trial) of Real World Evidence in Support of Drug Development and Review, which will give specific requirements and guidance to real-world data in terms of definition, source, evaluation, governance, standards, safety compliance, quality assurance, applicability, etc., to help bidders better manage their data, assess the suitability of real-world data, and prepare for the production of valid real-world evidence.

II. Real-world data sources and current conditions

Real-world data on drug development include recorded data (e.g. electronic medical records) of treatment processes in real medical settings, as well as various observational research data. Such data could have been collected prior to real-world research
data, which can also be newly collected for real-world research.

(I) The main source of real-world data

The sources of real-world data in China can be divided into hospital information system data, medical insurance payment data, registration research data, drug safety active monitoring, natural population queue data, etc., the following are common real-world data sources classified according to data function type.

1. Hospital information system data

Hospital information system data include structured and unstructured digital or non-digital patient records, such as the patient's demographic characteristics, clinical characteristics, diagnosis, treatment, laboratory examination, safety and clinical outcomes, which are usually stored in different information systems such as electronic medical records/electronic health files,

laboratory information management systems, medical image records and communication systems, and radiation information management systems. Some medical institutions set up hospital-level scientific research data platforms on the basis of data integration platforms or clinical data centers, integrate patient outpatient, inpatient, follow-up and other information to form data directly for clinical research. Some regional medical databases, using a relatively centralized physical environment for the storage and processing of clinical data across medical institutions, have the characteristics of large amounts of storage, many types, and can also be used as a potential source of real-world data.

Hospital information system data is based on the records of clinical practice processes, covering clinical outcomes and drug exposure, especially electronic medical records are widely used in real-world research.

2. Health insurance payment data

There are two main sources of medical insurance payment data in China, one is the basic medical insurance system established by the government and medical institutions, the establishment and unified management of the medical insurance payment database, which contains data on the basic information of patients, medical service utilization, prescription, settlement, medical claims and other structured fields; As a real-world data source, the health insurance system is more used for health technology evaluation and pharmaceutical economics research.

3. Register research data

Registered research data is collected through an organized system using observational research methods to collect data from clinical and other sources that can be used to evaluate the clinical outcomes of specific diseases, specific health conditions, and exposed populations. Registration research according to the characteristics of the population defined by the study mainly include medical product registration research, disease registration research and health service registration research three categories, China's registration research is mainly the first two categories. Among them, medical institutions and enterprises to support the drug registration study, the observation of patients using a certain drug, focusing on the use of drugs for different indications of clinical efficacy or monitoring adverse reactions.

The advantage of registering a research database is that it focuses on specific patients as the study population in combination with clinical diagnosis and treatment, medical insurance payment and other data sources, data collection is more standardized, generally including

patient self-reporting data and long-term follow-up data, observational outcome indicators are usually relatively rich, with high accuracy, strong structure and other advantages, for the evaluation of drug effectiveness, safety, economy and compliance has a better applicability, but also can be used in the natural history of disease and prognosis research.

4. Active monitoring data for drug safety

Active drug safety monitoring data are mainly used to carry out drug safety research and drug epidemiology research, through national or regional drug safety monitoring networks, from medical institutions, pharmaceutical companies, medical literature, online media, patient reporting outcomes and other channels, data collection. In addition, safety monitoring databases for their own medicines established by medical institutions and enterprises themselves may form part of such data sources.

5. Natural crowd queue data

Natural population queue data refers to a variety of data obtained from long-term forward-looking dynamic tracking observations of healthy and/or patient populations. Natural population queue data with uniform standards, information sharing, long time span and large sample size characteristics, such real-world data can help build common disease risk models, can support the drug research and development target population accurate positioning.

6. Histological data

Histological data as an important support for precision medicine, including genome,Epigenetics, transcriptomes, proteomics, and metabolic groups, which depict the characteristics of patients in genetics, physiology, biology, etc. from a system biology perspective. Often histological data need to be combined with clinical data to become suitable real-world data.

7. Death registration data

The death registration of a person is a continuous and complete collection and recording of the death information of its nationals. At present, china has four systems for collecting information on deaths, respectively, under the National Center for Disease Control and Prevention, the National Health Commission, the Ministry of Public Security and the Ministry of Civil Affairs. The population death registration data contain all the information in the medical

certificate of death, records the detailed cause of death and the time of death, and can be used as a data source for the population to divide the cause of death and the clinical outcome of major diseases.

8. Patients report outcome data

Patient report outcome is an index from the patient's own measurement and evaluation of disease outcome, including symptoms, physiology, psychology, medical service satisfaction, etc. There are two forms of recording, paper and electronic, the latter called electronic patient reporting outcomes, its rise and application, so that patient reporting outcomes and electronic medical records system docking and forming a complete flow of data at the patient level is possible.

9. Individual health monitoring data from mobile devices

Personal health monitoring data captures individual physiological signs in real time from mobile devices such as smartphones and wearables. These data often arise from the self-health management of the general population, the monitoring of patients with chronic illness by medical institutions, the process of evaluating the health status of the insured population by medical insurance companies, and are usually stored in wearable enterprises, medical institution databases, and data systems of commercial insurance companies.
Because wearable devices have the advantages of convenience and immediacy in collecting physiological and sign data, connecting with electronic health data can lead to more complete real-world data.

10. Other specific feature data

（1) Public health monitoring data

China has established a series of databases on public health monitoring, such as infectious disease surveillance, vaccination adverse events monitoring, etc., the recorded data can be used for division

Analysis of the incidence of infectious diseases, the general reaction of vaccines and the incidence of abnormal reactions.

（2）Patient follow-up data

In a real-world clinical environment, electronic medical records in hospitals are often not available.

The method covers some important clinical indicators of patients, such as total survival, five-year survival rate, adverse reaction information, etc., which need to be supplemented with long-term follow-up data in order to form the applicable real world data. Patient follow-up data mainly refers to clinical research purposes, hospital follow-up departments or third-party authorized service providers to the hospital patients by letter, telephone, outpatient, SMS, network follow-up and other ways to carry out clinical end points, rehabilitation guidance, medication reminders, satisfaction surveys and other services, services collected in the hospital follow-up data system, usually stored in the hospital follow-up data system. Through the link with medical records data, the integration of multi-source clinical data is realized to explore clinical research problems such as disease occurrence mechanism, development law, treatment methods and prognosis-related factors.

（3) Patient medication data

The drug use data of patient diagnosis and treatment process include patient information, drug regulations, and information such as drug usage and adverse reactions is usually stored in hospital drug management information system, pharmaceutical e-commerce platform, pharmaceutical enterprise product traceability and drug safety information database, as well as drug use monitoring platform. With the popularity of tele-diagnosis and Internet-slow disease management models, the number of out-of-hospital drug use data stored on prescription circulation platforms or pharmaceutical e-commerce platforms is increasing, and the effective use or stitching of such data can be used as a real-world data source for patient dimensional diagnosis and treatment process records.

With the continuous development of medical information technology, new real-world data types and sources will appear, but its specific application also depends on the clinical solution research issues, and the applicability of the data to produce real-world evidence.

(ii) The main challenges facing real-world data applications

From the data source, compared with randomized controlled Trial, RCT data, real-world data in most cases lack of strict quality control of its recording, collection, storage and other processes, will create incomplete data, missing key variables, inaccurate records and other

problems, the quality of these data defects, will greatly affect the subsequent data governance and application, and even affect the traceability of data, researchers also difficult to find the problems and check and revise. Due to changes in patient's course, place of consultation and time and space, it may lead to the loss of information on patient's disease status and related factors, which poses a challenge to the systematic evaluation of disease status and outcome in clinical research. Selective data collection, especially registered research data, is a potential risk that leads to bias.

Data fragmentation and information silos are prominent due to the relative independence and closure of real-world data sources, the wide variety of data management systems, the fragmentation and inconsistencies in data storage and data standards, and the difficulty of horizontal integration and exchange of data. For electronic medical records data, due to its high sensitivity, the system is generally closed management, the use of them may be limited. Electronic medical records may also affect the objective evaluation of clinical outcomes due to subjective descriptions of text types and differences in recorders. In addition, in the absence of uniform standards, data types are more diverse, from structured data, but also text, pictures, video and other unstructured and semi-structured data, in the process of data recording, collection, storage, will also lead to data redundancy and duplication, resulting in more difficult data processing.

III Evaluation of the applicability of real-world data

The evaluation of the suitability of real-world data should be based on specific research purposes and regulatory decision-making purposes.

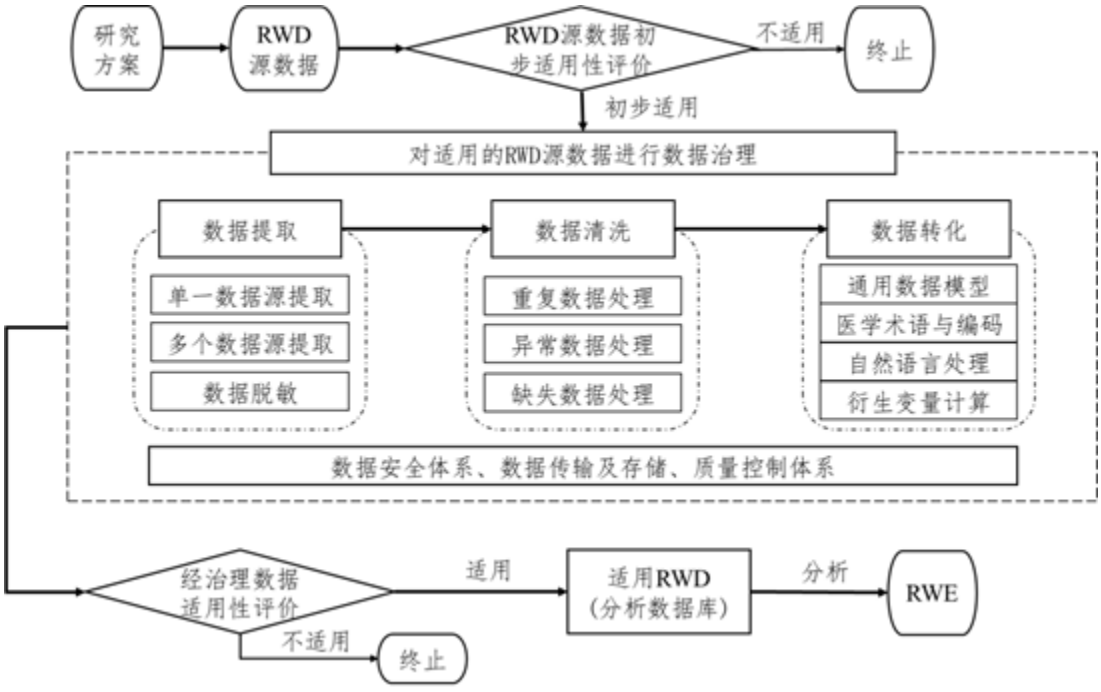
(I) Data governance and data management for real-world data

Real-world data can be obtained in retrospective and forward-looking ways, depending on when the study was conducted. Retrospective data collection usually requires data governance, mainly from retrospective observational research, prospective observational research, retrospective forward-looking observational research, etc. Forward-looking data collected is subject to data management, mainly from prospective observational studies, or practical clinical trials, because such data is similar to RCT data collection, that is, based on research programs to establish databases and collect data through electronic data collection systems, is forward-looking, planned, structured and standardized data. If a study takes advantage of past

data and collects future data, for example, retrospective forward-looking studies that begin immediately, the data collected retrospectively needs to be treated with data, while the data collected in the foresight is managed in a data management manner, and the key issue to note here is that the previously managed database should match the forward-looking database. For one-arm clinical trials with external control, if historical control, external data need to be used governance means, and if parallel control, external data can be used data management means.

The applicability evaluation of real-world data is mainly aimed at retrospective collection data, but also guidance for forward-looking data collection. The suitability evaluation can be divided into two stages, the first stage is from the dimensions of accessibility, ethics, compliance, representativeness, key variable integrity, sample size and source data activity status, the source data is evaluated and selected to determine whether it meets the basic analysis requirements of the research program; In the case of forward-looking collection of real-world data, there is no need for a preliminary application evaluation of the first phase.

Figure 1 A diagram of the applicability evaluation and data governance process for real-world data



(ii) Evaluation of the suitability of the source data

Source data that meets basic analysis requirements should have at least the following criteria:

1. The database is active and data is available

The database shall be continuously active during the study period and the recorded data shall be accessible, i.e. have access to the data and may be evaluated by third parties, in particular regulatory authorities.

2. Data usage meets ethical and security requirements

The use of source data should comply with ethical review regulations and with relevant data security and privacy requirements.

3. The coverage of key variables

Source data is usually incomplete, but should have some coverage, including at least outcome variables, exposure/intervention variables, demographic variables, and important covariates related to the purpose of the study.

4. The sample size is sufficient

The significant reduction in the number of source data cases after data governance should be fully considered and prejudged in order to ensure the sample size required for statistical analysis.

(iii) Evaluation of the applicability of governance data

The applicability evaluation of governed real world data is based mainly on data relevance and reliability.

1. Relevance evaluation

Relevance evaluations are designed to assess whether real-world data is closely related to the clinical issues of concern, focusing on coverage of key variables, accuracy of exposure/intervention and definition of clinical outcomes, representation of target populations, and fusion of multi-source heterogeneous data.

（1） Coverage of key variables and information Real-world data should contain important variables and information related to clinical outcomes, such as drug

use, patient demographics and clinical characteristics, covariates, outcome variables, follow-up time, potential safety information, and so on. If some of the above variables are missing, it is necessary to fully assess whether they can be filled using reliable statistical methods and the possible impact on causal inference results.

（2） The accuracy of the definition of exposure/intervention of the clinical outcome

Selecting and accurately defining clinically significant outcomes and defining exposure/interventions accurately is critical to real-world research and should be consistent with the clinical significance or theoretical basis of the study. The definition of clinical outcome should include the diagnostic criteria on which the diagnosis is based, the measurement method and its quality control (if any), the measuring tool (such as the use of the scale), the calculation method, the measurement time point, the variable type, the conversion of the variable type (e.g. from quantitative to qualitative), the endpoint event evaluation mechanism (such as the operational mechanism of the endpoint event determination committee), etc. When the definition of clinical outcome is inconsistent between different data sources, a unified clinical outcome should be defined and a reliable conversion method should be adopted. The definition of exposure/intervention should take into account the reasonableness of its time window.

（3) Representation of the target population

One of the advantages of real-world research over traditional RCT is the representation of a wider target population. Therefore, the criteria for inclusion and exclusion should be developed in such a way as possible to meet the target population in a real-world environment.

（4） The fusion of multi-source heterogeneous data

Because of the nature of real-world data, heterogeneous data, which in many cases belongs to multiple sources, requires the linking, fusion and isomorphic processing of data from different sources at the individual level. Therefore, accurate links at the individual level should be made through identifiers to support

the integration of key variables in the data source using common data models or data standards.

2. Reliability evaluation

The reliability of real-world data is evaluated mainly from the aspects of data integrity, accuracy, transparency, quality control and quality assurance.

(1) Integrity

Integrity refers to the degree to which data information is missing, including the loss and variation of variables The missing measure. For different studies, the degree of data loss, missing distribution, missing causes and missing mechanisms of variable values are not the same, should be described in detail.

When the proportion of data missing from a particular study is significantly higher than that of similar studies, the uncertainty of the study's conclusions is increased, and careful consideration needs to be given to whether that data can be used as data to support the generation of real-world evidence. A detailed analysis of the causes of the absence can help to make a comprehensive judgement of data reliability. If the problem of filling the missing data is involved, the appropriate filling method should be adopted according to the reasonable assumption of the missing mechanism.

(2) Accuracy

Accuracy refers to whether the data is consistent with the objective characteristics described, including whether the source data is accurate, whether the data value domain is within a reasonable range, whether the trend of the outcome variable changes over time is reasonable, whether the coding mapping relationship corresponds to and is unique, and so on. The accuracy of the data needs to be identified and verified against more authoritative references, for example, whether the endpoint event has been judged by an independent endpoint event determination committee.

（3) Transparency

Transparency of real-world data refers to the clarity of governance programs and governance processes for real-world data, ensuring that key exposure/intervention variables, covariates, and outcome variables in the analytical data can be traced back to the source data and reflect the process of data extraction, cleaning, conversion, and

standardization. Whether using manual data processing or automated processing, data governance standardized operating procedures and validation confirmation files should be clearly recorded and archived, especially reflecting data credibility issues such as data deletion, variable value fields, derived variable calculation methods, and mapping relationships. Data governance programmes should be developed in advance for research purposes and should ensure that the data governance process is consistent with governance programmes. Transparency of data also includes data accessibility, information sharing between databases, and transparency of methods to protect patient privacy. If you use algorithms to define research queues, the development of algorithms and their validation should also be transparent.

（4) Quality control

Quality control refers to the technology and activities implemented to verify that all aspects of data governance meet quality requirements. Quality control evaluation includes, but is not limited to: data extraction, safe processing, cleaning, structured, and subsequent storage, transmission, analysis and delivery links have quality control to ensure that all data is reliable, the data processing process is correct, whether to follow a complete, standardized, reliable data governance program and plan, and rely on the corresponding data quality verification and system verification procedures, to ensure that the data governance system in normal and stable conditions, to ensure the accuracy and reliability of real-world data.

(5) Quality assurance

Quality assurance refers to systematic measures to prevent, detect and correct data errors or problems that arise during research. Quality assurance of real-world data is closely related to regulatory compliance and should run through every aspect of data governance, taking into account, but not limited to, whether to establish research plans, programmes and statistical analysis plans related to real-world data, whether there are appropriate standard operating procedures, whether clear processes and qualified personnel are used for data collection, whether a common definition framework, the data dictionary, and compliance with the collection of critical data The common time frame for variables, whether the technical methods used for data element capture conform to

pre-specified technical specifications and operating procedures, including the integration of data from various sources, records of drug use and laboratory inspection data, follow-up records, links to other databases, whether data input is timely and secure transmission, and whether the requirements for on-site verification and access to source data, source documents, etc. are met by regulators.

4. Real-world data governance

Data governance refers to the specific clinical research problems, in order to achieve the applicable The management of raw data includes, but is not limited to, data security processing, data extraction (including multiple data sources), data cleaning (logical verification and abnormal data processing, data deletion processing), data transformation (data standards, common data models, normalization, natural language processing, medical coding, derived variable calculation), data transmission and storage, data quality control and so on.

(i) Personal information protection and data security processing

Real-world research related to the protection of personal information should follow the national information security technical norms, medical big data security management regulations, personal sensitive information should be de-identified processing, to ensure that according to the data can not be matched and restored personal sensitive information, through technical and management measures to prevent the leakage, destruction, loss, tampering of personal information. Data security processing should be based on the types of data involved in the study.

Quantity, nature and content, especially for sensitive personal information, establish data encryption technology requirements, risk assessment and emergency disposal procedures for all aspects of data governance, and conduct security measures effectiveness audit.

(ii) Data extraction

Depending on the storage format of the source data, whether it is electronic data, whether it contains unstructured data and other factors to choose the appropriate way to extract data, the following principles should be observed when extracting data:

The method of data extraction should be verified to ensure that the extracted data meet the requirements of the research program. The extract should ensure the consistency between the raw data extracted and the source data, and timestamp management should be performed on the extracted raw data and the source data.

Using extract tools that can be interoperable or integrated with source data systems reduces errors in data transcription, improving data accuracy and the quality and efficiency of data collection in clinical studies.

(iii) Data cleaning

Data cleaning refers to the removal of duplicate or redundant data from extracted raw data, logical verification of variable values and processing of outliers, and the management of data loss. It is important to note that data should not be modified to ensure the authenticity of the data if it cannot be traced back to the signature confirmation of the principal researcher or source data responsible party.

First, remove duplicate and irrelevant data while maintaining data integrity. In the process of merging different data sources, duplicate data may occur and needs to be removed. At the same time, because of the inaccurate mapping relationship between the data source and the common data model, data that is not relevant to the research objectives may be collected, and removing unwanted observations from the dataset can reduce unnecessary work.

Then carry out logical verification and exception data processing. Through logical verification, errors in raw data or data extraction can be found, such as discharge time earlier than admission time, birth year and month by age projection, laboratory test results do not conform to reality, qualitative judgment results and the criteria defined in the scheme are inconsistent. Handle exception data with great care to avoid the resulting bias. Error and exception data found should be further verified before the data can be changed, and changes to the data should be kept recorded.

Finally, the data loss is processed in statistical analysis, and for different studies, the degree of data loss, the reason for the deletion and the missing mechanism of variable values are different. If the problem of filling the missing data is involved, the appropriate filling method should be adopted according to the reasonable assumption of the missing mechanism.

(iv) Data conversion

Data transformation is the process of transforming the data format standard, medical terminology, coding standard and derivative variable calculation of raw data after data cleaning into real-world data according to the corresponding standard in the analysis database.

A reliable natural language processing algorithm can be used for the transformation of free text data, which can improve the conversion efficiency under the premise of ensuring that the data transformation is accurate and traceable.

In the calculation of derivative variables, the definition of the original data variables and variable values, calculation methods and derivative variables used for calculation should be clearly defined, and timestamp management should be carried out to ensure the accuracy and traceability of the data.

(v) Data transfer and storage

The transmission and storage of real-world data should be controlled throughout the life cycle of data collection, processing, analysis and destruction, based on a trusted network security environment. Encryption protection should be in place during data transfer and storage. In addition, operational setup approval processes, role permission control, and minimal authorization access control policies should be established to encourage the establishment of automated audit systems to monitor the processing and access activities of recorded data.

(vi) Data quality control

Data quality control is key to ensuring the integrity, accuracy, and transparency of research data. Data quality control requires the establishment of a sound real-world data quality management system and standard operating procedures, the recommended principles include:

1. Ensure the accuracy and authenticity of the source data

If electronic medical records are a key data source, there should be quality control standards for medical records to meet the requirements of analysis. The description, diagnosis and drug use information from the clinic need to be corroborated by the relevant chain of evidence. For any modifications in the entry process, the owner's confirmation and signature are required, and the reason for the modification is provided to ensure that the complete audit trail is left.

2. Take data integrity into account when extracting data

Evaluate and establish extraction fields and develop appropriate verification rules and database architecture.

3. Develop a comprehensive data quality management plan

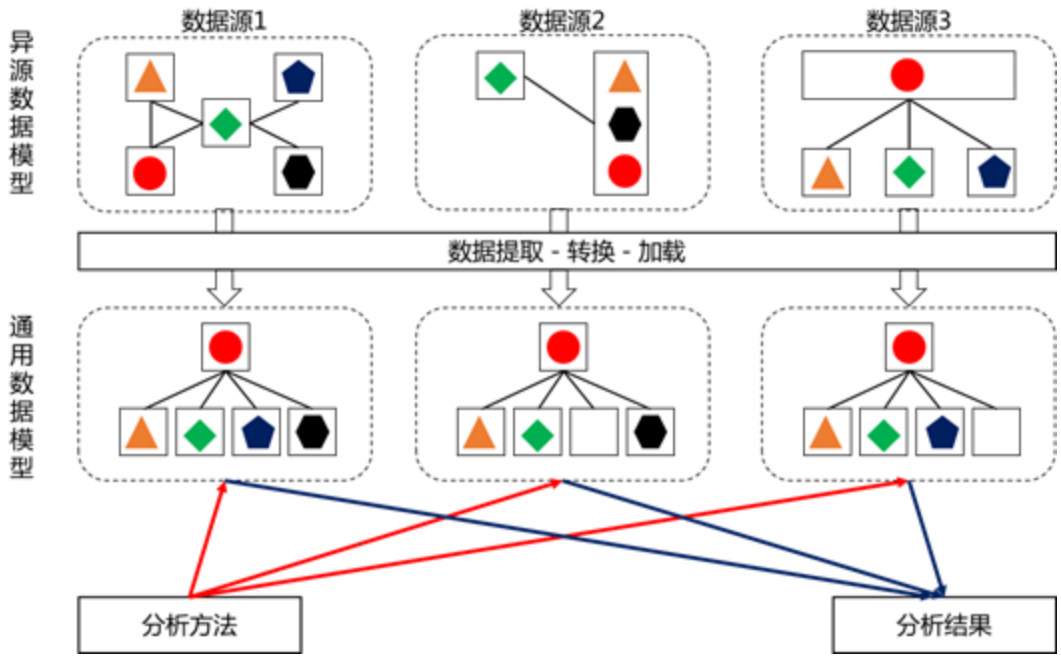
Develop system quality control and manual quality control plans to ensure the accuracy and completeness of data. For key variables, a comprehensive verification and source document review should be carried out, and other variables can be sampled according to the actual situation, for example, for demographic information, numerical variable thresholds, coding mapping relationship, etc., can be sampled in proportion to verify its accuracy and rationality.

(vii) A common data model

The common data model is a data model that is quickly centralized and standardized for multi-source heterogeneous data in a multidisciplinary collaboration model, and its main function is to convert different standard source data into a unified structure, format, and terminology for data consolidation across databases/datasets.

Due to the complexity of the structure and type of multi-source data, the differences in sample size and standards, the source data needs to be extracted, transformed and loaded in the overall process of transforming the source data into a common data model, and the source data should be consistent in both semantics and semantics with the structure and terminology of the target analysis database, as shown in Figure 2.

Figure 2 A diagram ideal for transforming a heterogeneous data model into a common data model should follow the following principles:



1. A common data model can be defined as a data governance mechanism that standardizes source data into common structures, formats, and terminology, allowing data consolidation across multiple databases/datasets. The common data model should have the ability to access the source data, is a data model that can be dynamically expanded and continuously improved, and has version control

2. The definition, measurement, merge, record and corresponding validation of variables in the common data model should be transparent, and the data transformation of multiple databases should have clear and consistent rules

3. Common variables or general ideas related to safety and effectiveness should be mapped to a common data model for different clinical research issues and compared to recognized or known studies.

(viii) Real-world data governance proposal

Real-world data governance proposals should be developed in advance and synchronized with the entire project research program. If the governance proposal needs to be revised during the study, it should be communicated with the review body and an updated governance proposal submitted. The proposal should describe the purpose for which real-world data is used for regulatory decision-making, the research design of the use of real-world data, and should also describe real-world data source data, including, but not limited to, the types of

real-world data source data/source files, such as health information system data, disease registration research data, health care data, etc.Past applications justify adoption, governance of real world data, i.e. the governance process from real world data sources to analytical databases, data models and data standards used, processing of missing data, measures taken to reduce or control potential bias resulting from the use of real-world data, quality control and quality assurance, and assessment of the applicability of real-world data.

5. Compliance, security and quality management systems for real-world data

(i) Data compliance

Real-world data comes from data from a variety of sources, including individual patient care, The collection, processing and use of data involves ethical and patient privacy issues. In order to fully protect the safety and interests of patients, access to and use of real-world data for real-world research is subject to review and approval by the Ethics Committee. Relevant personnel involved in real-world data governance must strictly abide by the requirements of relevant laws and regulations, and bidders should strictly implement and fulfill their protection and management obligations.

(ii) Data security management

Data security management should be done in accordance with national laws and regulations, industry regulatory requirements, etc., and the necessary security protection should be carried out for information systems and network facilities and cloud platforms carrying health and medical data. Data security should cover all life cycles, including data collection, data extraction, data transmission, data storage, data exchange, data destruction, and so on. Where encryption technology is used to ensure the integrity, confidentiality and traceability of data in the process of collection, extraction, transmission and storage, the media shall be controlled. Different protection measures are applied to the data forms of different media, and corresponding access control mechanisms are established to audit, register, archive and audit access records.

Data audit and related operating procedures for data collection, extraction, transmission,maintenance, storage, sharing, use, etc. to provide records and basis, should

include personnel audit, management audit, technical audit, the development and deployment of medical information system activity audit policy and appropriate standard operating procedures. The contents of the audit should include any action in the status of the data, including the act of logging in, creating, modifying, and deleting records, and should automatically generate time-stamped audit records, including, but not limited to, information such as authorization information, operational time, reason for operation, operation content, operator and signature, and be available for audit. Audit records should be securely stored and access control policies established.

(iii) Quality management system

A complete quality management system should be established to standardize the processing of real-world data and continuously optimize and perfect it in practice. Basic quality elements should be covered: to ensure the quality of real-world data, should establish the operation process covering the whole life cycle management of real-world data, computerized system functions should meet the management needs of real-world data, in line with the relevant requirements of relevant regulations on computerized systems, establish a sound personnel management system, data collection, governance, analysts should be appropriate Training to meet the requirements of responsibilities and competencies, standardized management of personnel's authority, establish risk management processes from data collection to data submission, and develop standard information and document management practices (paper, electronic media) to ensure that real-world data processing process records are complete, accurate, transparent, and protect data security and compliance.

Communication with regulatory bodies

In order to ensure that the quality of real-world data meets regulatory requirements, applicants are encouraged to communicate with regulatory authorities in a timely manner. Prior to the official start of real-world research, based on the overall research and development strategy and specific research programs, the real world data support the generation of real-world evidence, including the accessibility of real-world data, whether the sample size is large enough, whether the data governance plan is reasonable and feasible, data quality can be guaranteed, and so on. In the course of the study, if the data governance plan is adjusted according to the changes in the implementation of the study, the applicant is required to measure the potential

impact of the data governance plan adjustment on the pilot objectives, explain the full reasons for the adjustment to the regulatory authorities, and obtain their consent, and submit an updated research programme and data governance plan. After the completion of the study and before the submission of information, the applicant may consult with the regulatory authorities to submit the information and database for communication.

Bibliography

[1] Cai Wei , Zhan Siyan Thinking about speeding up the construction of china's active vaccine safety monitoring system

Chinese Journal of Preventive Medicine . 2019, 53 (7): 664-667

[2] National Health Commission, State Drug Administration Drug Clinical Trial Quality Tube

Code ofPractice. 2020.07.01.

[3] The Drug Review Center of the State Drug Administration Clinical Trial Data Management Techniques

Guide to technology. 2016.07.27.

[4] The State Drug Administration Real-world evidence supports drug development and review

Principles of Guidance (Trial). 2020.01.07.

[5] Hou Yongfang , Song Haibo , Liu Hongliang , etc. Based on the Chinese hospital drug alert system

Practice and discussion of active monitoring . China's drug alert . 2019, 16(4): 212-214.
Zhou Li, Ouyang Wenwei, Li Wei, etc. Analysis of the current situation of china registration research . . .
Journal of Medical Medicine . 2019, 19 (6): 702-707

[7] Berger M, Daniel G, Frank K, et al. A framework for regulatory use of real world evidence. https://healthpolicy.duke.edu/sites/default /files/ atoms/files/rwe_white_paper_2017.09.06.pdf.

[8] Booth CM, Karim S, Mackillop WJ. Real-world data: towards achieving the achievable in cancer care[J]. Nat Rev Clin Oncol. 2019,16(5): 312-325.

[9] Duke-Margolis Center for Health Policy. Characterizing RWD Quality and Relevancy for Regulatory Purposes. https://healthpolicy.duke.edu/ publications.

[10] Duke-Margolis Center for Health Policy. Determining Real-World Data's Fitness for Use and the Role of Reliability. https://healthpolicy. duke.edu/ publications.

[11] EMA. Reflection paper on expectations for electronic source data and data transcribed to electronic data collection tools in clinical trials. https://www.ema.europa.eu/en/documents/regulatory-proceduralguideline/reflection-paper-expectations-electronic-source-data-datatranscribed-electronic-data-collection_en.pdf.

[12] EMA. A Common Data Model for Europe – Why? Which? How? https://www.ema.europa.eu/en/documents/report/common-data-modeleurope-why-which-how-workshop-report_en.pdf.

[13] Khozin S, Abernethy AP, Nussbaum NC, et al. Characteristics of real-world metastatic non-small cell lung cancer patients treated with nivolumab and pembrolizumab during the year following approval [J]. Oncologist. 2018, 23: 328–336.

[14] OHDSI – Observational Health Data Sciences and Informatics, https://www.ohdsi.org.

[15] Ong TC, Kahn MG, Kwan BM, et al. Dynamic ETL: a hybrid approach for health data extraction transformation and loading [J]. BMC Medical Informatics and Decision Making 2017, 17(1) : 134.

Appendix 1 Vocabulary

Electronic Medical **Records (EMR):** Electronic records of health-related information about individual patients created, collected, managed, and accessed by authorized clinical professionals in medical facilities.

Electronic Health **Record**, **EHR:**Meets nationally recognized interoperability standards and is capable of creating, managing, and consulting electronic records of health-related information about individual patients created, managed, and consulted by authorized clinical professionals in multiple health facilities.

Observational Study:A study that explores the causal relationship between exposure/treatment and outcomes without active intervention, targeting natural or clinical populations, based on specific research issues.

Patient-Reported **Outcome, PRO:**is an indicator from the patient's own measurement and evaluation of the outcome of the disease, including symptoms, physiology, psychology, satisfaction with medical services, etc. It is recorded in both paper and electronic forms, the latter called electronic patient reporting outcomes(ePRO).

Logical Check: A review of the validity of clinical research data entered into a computer system, focusing on the input data and its expected numerical logic

Whether there are logical errors in numeric ranges or numeric attributes, etc.

Data Standard:A set of rules about how to build, define, format, or exchange specific types of data between computer systems. Data standards can make submissions predictable and consistent and in the form in which information technology systems or scientific tools can be used.

Data Cleaning: Data cleaning is designed to identify and correct noise in your data to minimize the impact of noise on the results of your data analysis. Noise in data mainly includes incomplete data, redundant data, conflicting data and incorrect data.

Data Fusion: Combines, correlates, and combines data and information from multiple sources to form a unified data set.

Data Element:A single observation of subjects recorded in a clinical study, such as date of birth, white blood cell count, severity of pain, and other clinical observations.

Data Governance: For specific clinical research issues, the governance of raw data for statistical analysis includes at least extracts (including multiple data sources), data security processing, and data cleaning

(logical verification and anomaly data processing, data integrity processing), data transformation (common data model, normalization, natural language processing, medical coding, derived variable calculation), data quality control, data transmission and storage, etc.

Common Data **Model**, **CDM**is a data model that is quickly centralized and standardized for multi-source heterogeneous data in a multidisciplinary collaboration model, whose primary function is to transform source data from different data standards into a unified structure, format, and terminology for data consolidation across databases/datasets.

Source **Data:**All information on clinical symptoms, observations, and original records and certified copies of other activities used to reconstruct and evaluate the study in clinical studies. The source data is contained in the source file, including the original record or its valid copy.

Real-World Data, RWD: Comes from a variety of data collected on a daily basis related to patient health and/or care and care. Not all real-world data can be analyzed as real-world evidence, and real-world evidence can only be produced if it meets the applicability of real-world data.

Real-World **Research/Study, RWR/RWS:**The process of collecting data**(real-world** data) or aggregated data derived from the health and treatment and health ofthe subjects in a real-world environment for clinical research issues, and obtaining the value and potential benefits of drug use through analysis - clinical evidence of risk (real-world evidence). Real-World Evidence , **RWE:**Drugs obtained through proper and adequate analysis of applicable real-world data Use and potential benefits - clinical evidence of risk.

Appendix 2 A comparison of Chinese and English vocabulary

A comparison of Chinese and English vocabulary

Chinese	English

	Adverse Events Following
Vaccination adverse events	
	Immunization, AEFI
A common data model	Common Data Model, CDM
Case reporting form	Case Report Form, CRF
Data governance	Data Curation
Case registration	Patient Registry
Electronic data collection	Electronic Data Capture, EDC
Electronic medical records	Electronic Medical Record, EMR
Electronic health file	Electronic Health Record, EHR
Electronic patient reporting outcomes	electronic Patient-Reported Outcome, ePRO
Observational research	Observational Study
The patient reports the outcome	Patient Reported Outcome, PRO

The outcome variable	Outcome Variable
Traceability	Traceability
Logical verification	Edit Check
Data standards	Data Standard
Data cleaning	Data Cleaning
Data governance	Data Curation
A common data model	Common Data Model, CDM
Hospital information system	Hospital Information System, HIS
Derivative variables	Derived Variable
The source data	Source Data
Real-world data	Real World Data, RWD
Real-world research	Real World Research/Study, RWR/RWS
Real-world evidence	Real World Evidence, RWE