Guideline on Using Real-World Data to Generate Real-World Evidence

(Trial Version)

English Translation by:

Ligang Luo, Pingyan Chen, Jianing Di, Jun Wang, Chunquan Ou

Center for Drug Evaluation, NMPA

April, 2021

Contents

Guidance on Using Real-World Data to Generate Real-World Evidence

1. Introduction

Real-World Evidence (RWE) is an important component of evidence chain for evaluating the efficacy and safety. For its concept and implementation please refer to *Guidance on Using Real-World Evidence to Support Drug Development and Evaluation (for Trial Implementation)*. On the other hand, high-quality and applicable Real-World Data (RWD) are the basis for the generation of RWE.

RWD refer to a variety of data, collected through regular practice, that are related to an individual patient's health status and/or diagnosis, treatment and healthcare. Only those RWD that satisfy the fitness standard and after proper and sufficient analyses can be used to generate RWE. For RWD, at present, there is a lack of strict quality control in the process of data documentation, collection, and storage. There are also challenges in the completeness of data and the consistency of data standards, data model and description methods, which hinders the effective use of RWD. Therefore, how to convert the collected RWD, possibly after curation, into the analysis data needed for clinical research, and how to evaluate the fitness of RWD in generating RWE, remain as the key questions in how to use RWD generating RWE to support drug regulatory decision making.

As a complement to the *Guidance for Using Real-World Evidence to Support Drug Development and Evaluation (for Trial Implementation)*, this guidance provides specific requirements and advises in terms of key aspects including definition, sources, evaluation, curation, standards, security and compliance, quality assurance and fitness, etc., in order to help sponsors properly curate and evaluate the fitness of the RWD, and make sufficient preparations for the generation of RWE.

2. Sources and current status of RWD

The RWD related to drug development mostly consist of data recorded in regular medical environment (such as in Electronic Health Records [HER]) and in observational studies. Such data may have been already documented before conducting a Real-World Study (RWS) or collected for the purpose of conducting an RWS.

2.1 Common sources of RWD

Categorized by specific functionality, the sources of RWD in China include hospital information system data, medical insurance claim data, disease registry data, public health surveillance data and natural population cohort data.

3.3.2 Hospital information system data

Hospital information system data may include records that are structured or unstructured, digital or non-digital. They may contain demographic characteristics, clinical characteristics, diagnosis, treatment, laboratory tests, safety information，and clinical outcomes. The data are usually

stored in different information systems such as Electronic Medical Records (EMR/EHR), Laboratory Information Systems (LIS), Picture Archiving and Communication Systems (PACS), Radiological Information Systems (RIS), etc. Utilizing data integration platforms or clinical data repositories, some medical institutions have built hospital-level research data platforms that consolidate various outpatient and inpatient records, follow-up, etc. Data in such systems may be used directly in clinical research. With physically centralized environment clinical data storage and process across medical institutions, some regional medical databases with large storage and rich types of data could also be used as a potential source of RWD.

Hospital information system data capture the records collected in the course of routine clinical practice, and may include a wide range of clinical outcome and exposure variables. Such information especially the electronic medical records (EMR) is commonly used in RWS.

3.3.2 Medical insurance claim data

There are two main categories of medical insurance claim data in China: one is the essential medical insurance system established by government and medical institutions, used to construct and manage the medical insurance claim database, that includes structural data as individual information, utilization of medical services, prescriptions, settlements and medical claims; the other is the commercial healthcare insurance database established by insurance companies that includes information classified according to claim payment and duration of insurance. Data from this

category are often simpler in data dimensionality. As a source of RWD, the medical insurance system is often used for health technology assessment and pharmacoeconomic research.

3.3.2 Registry data

Registry data refer to clinical information and data of other kind collected through an organized system and based on observational research, and can be used to evaluate specific diseases, or clinical outcome of population with specific health condition and exposure. Depending on the characteristics of the population of interest, registry studies mainly include those for products, healthcare services, and diseases. Registries in China are mostly for specific disease and products. Among these, drug product registry studies supported by medical institutions and companies are designed for specific medical product to monitor adverse events or effectiveness for different indications.

Registry databases have several advantages: Targeting on specific patient population，integrating multiple data sources including clinical treatment and medical insurance payment, standard data collection including self-reported and long-term follow-up information, rich observed outcomes with high accuracy and clear structure, etc. Such data well fit the purpose of the evaluation of drug effectiveness, safety, cost and compliance，also can be used in the research of natural history and prognosis of disease.

3.3.2 Active drug safety surveillance data

Active drug safety surveillance data are mainly used to conduct drug

safety research and drug epidemiology research. Data are collected from medical institutions, pharmaceutical companies, medical literature, public media, patient reported outcomes and other channels through national or regional drug safety surveillance networks. In addition, medical institutions and companies with their own drug safety surveillance databases may also be sources of such data.

3.3.2 Natural population cohort data

Natural population cohort data refer to all kinds of data collected on people with health or unhealth condition through prospective and dynamic long-term tracking and observation. These data feature common standards, shared information, long time span and large sample size. RWD of this type can be used to build a common disease risk models or to support the identification of precise target population for drug development.

3.3.2 Omics data

Omics data as an important support for precision medicine, mainly include genomics, epigenetics, transcriptomes, proteomes and metabolomes. These data describe the characteristics of patients with respect to genetics, physiology and biology from a systematic biological perspective. Generally, omics data always combine with clinical data so as to satisfy RWS purposes.

3.3.2 Death registry data

Death registry is a country's continuous and complete collection and recording of death information of its nationals. There are currently four

systems for the collection of death information in China belonging to the National Center for Disease Control and Prevention, the National Health Commission, the Ministry of Public Security, and the Ministry of Civil Affairs, respectively. Death registry data contain all information in the death certificates, including the detailed cause and date of death. These databases provide a data source that produces cause-specific death rates and clinical outcomes of critical illness.

3.3.2 Patient reported outcome data

Patient Reported Outcomes (PRO) are measurements and evaluations of disease outcomes reported by patients themselves, including symptoms, physiological and psychological reactions, and medical service satisfaction. PROs have become increasingly important for drug evaluation. PROs can be recorded on paper or electronically, and the latter is referred to as electronic PRO (ePRO). The rise and application of ePRO make it possible for PRO to connect with EMR systems and form a comprehensive data flow at the patient level.

3.3.2 Personal health surveillance data from mobile devices

Personal health surveillance data can be collected through mobile devices (such as smartphone, wearable devices) in real time. Generally, such data are generated in the course of self-managed health programs, monitoring of patients with chronic diseases by medical institutions, or evaluation of insured individuals by medical insurance companies. These data are usually stored in the data systems of wearable device companies, the database of medical institutions, or the database of insurance companies.

Due to their convenience of use and immediacy in the collection of physiological and physical information by wearable devices, through connecting with electronic health data can generate fairly complete RWD.

3.3.2 Data for other specific purposes

1) Public health surveillance data

There is a series of databases that are related to the surveillance of public health, such as infectious disease monitoring, the Adverse Events Following Immunization (AEFI) monitoring system, etc. The information collected can be used to analyze the morbidity of infectious disease, the incidence of common and severe vaccination reactions.

2) Data from patient follow-up

In the real-world environment of clinical diagnosis and treatment, in-hospital EMRs are often insufficient in capturing all important clinical indicators, such as overall survival, 5-year survival rate, adverse reaction information, etc. EMR data may need to be supplemented with long-term follow-up data, in order to produce fitted RWD. Patient follow-up data mainly refer to out-of-hospital data collected via letter, telephone, outpatient service, SMS, online follow-up, etc., as authorized by the hospital or a third-party service provider for the purpose of clinical research. The follow-up service includes clinical endpoints collected, rehabilitation guidance, medication reminders, and satisfaction surveys, etc. Data are generally stored in hospital follow-up data systems. Through connecting with EMR data and merging clinical data from multiple sources, follow up data can be used to explore clinical research problems

such as disease occurrence mechanism, disease progression, treatment methods, prognostic factors, etc.

3) Patient medication data

Patient medication data collected during the course of diagnosis and treatment can include patient information, drug specifications, prescription information, and adverse reactions, etc. They are usually stored in hospital drug management information systems, pharmaceutical e-commerce platforms, pharmaceutical company databases for monitoring product safety information, and drug usage surveillance platforms. As telemedicine and internet chronic disease management become more popular, out-of-hospital medication data stored in prescription circulation platforms or pharmaceutical e-commerce platforms are increasing. Patient medication data could be used as a RWD source of recording the treatment process，through effective using.

With the continuous development of medical information technologies, new RWD types and sources will continue to emerge. Their specific application will depend on the clinical research questions to be answered and the fitness of such data that supports the RWE generation.

2.2 Main issues in the application of RWD

Compared with data from randomized clinical trials (RCT), RWD usually lack strict quality control over the process of recording, collection, and storage. This can lead to incomplete data, missing key variables, or inaccurate records. The presence of such quality defects will greatly affect subsequent data curation, applicability and even traceability of data.

These existing problems are also difficult for investigators to identify and correct. Information such as patients' disease status and related factors may be missing if there is a change in a patient's disease course, care provider, and treatment date, brining challenges to clinical research of disease status and systematic evaluation of disease outcome. Selective data collection, especially by registries, will result in potential risks of research result bias.

Data fragmentation and information isolation become pressing issues, because different RWD sources are relatively independent and closed, data storage from multiple data management systems are scattered and with inconsistent data standards, and challenges exist in data integration and exchange.

EMR data might be of limited use because of the closed management of the highly sensitive information system. Because electronic medical records are subjective records, and there are differences between recorders, these characteristics may affect the evaluation and objectivity of clinical outcomes. Additionally, in the absence of a unified standard, data structures can be diverse, ranging from structured data to non-structured or semi-structured data such as free texts, pictures and videos. It may lead to redundancy and duplication of data in the process of data recording, collection and storage, and in turn make data curation more difficult.

3. Fitness evaluation of RWD

Evaluation of the “fitness” of RWD should be based on specific research objectives and regulatory decision-making purposes.

3.1 Curation and management of RWD

Real-world data can be collected prospectively or retrospectively. Data collected retrospectively usually requires data curation. The data mainly comes from retrospective observational studies, prospective observational studies, ambispective observational studies, etc. that have been carried out in the past. For prospectively collected data, data management is required. The data mainly comes from prospective observational studies or practical clinical trials to be carried out. This type of data collection method is similar to RCT research, which is prospective, planned, structured and standardized, and establishes a database according to the research plan and collects data through EDC. If a study uses both past data and will collect future data, like an ambispective study, data curation is required for retrospectively collected data, and data management is used for prospectively collected data. The key issue is that the database of curated retrospective data should match the prospective database. For single-arm clinical trials with external controls, if it is a historical control, the external data needs to be curated. If it is a parallel control, the external data need be managed.

The fitness evaluation of real-world data mainly focuses on retrospective data, but it also has guiding significance for prospective data.

The fitness evaluation can be divided into two stages: Stage 1 is the preliminary evaluation and selection of source data in terms of

accessibility, ethical issues, compliance, representativeness, completeness of key variables, sample size and activity of source data so as to judge whether the data satisfy the basic requirements of the study analysis plan; Stage 2 is the evaluation of data relevance and reliability, and data curation mechanisms (data standards and common data models) being or to be adopted so as to judge whether the curated data are fit for the generation of RWE (see Figure 1). In an RWS, preliminary evaluation of fitness in Stage 1 can be ignored if RWD collected prospectively.

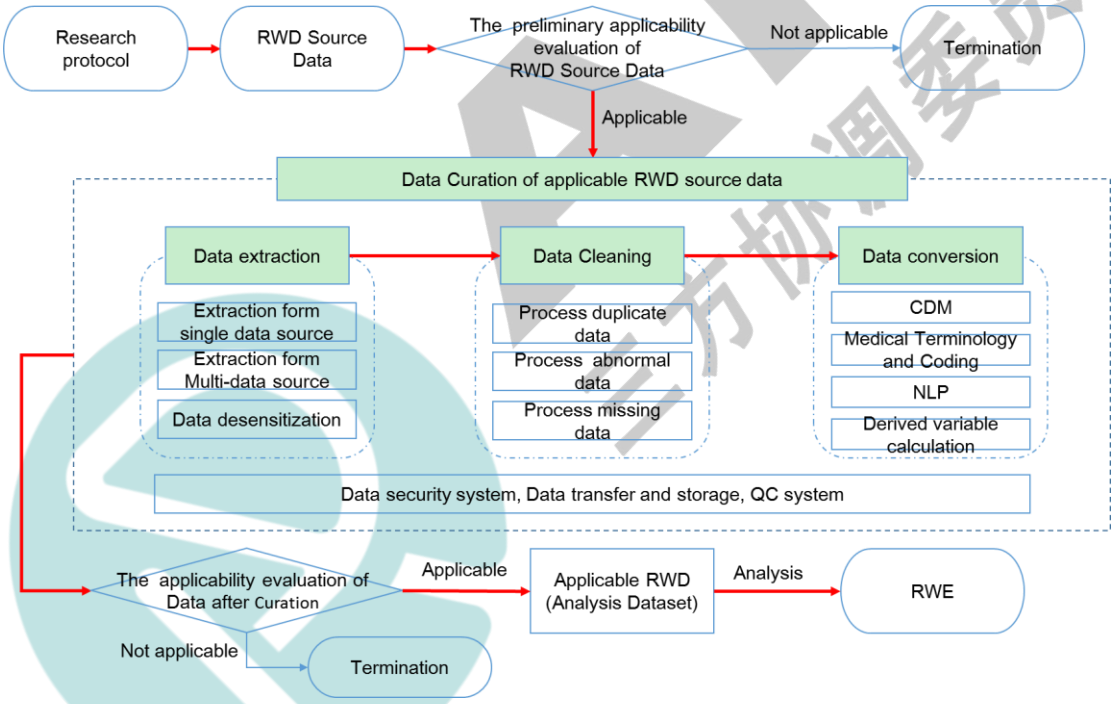

Figure 1 Fitness evaluation and data curation process of RWD

3.2 Fitness evaluation of source data

The source data satisfying basic requirements for analysis should at least meet the following conditions:

3.2.1 Active database and data accessibility

The database should be continuously active during the study period and all recorded data should be accessible, i.e. data can be obtained and evaluated by a third party especially regulatory agency.

3.2.2 Compliance with ethical and data security requirements

The use of source data should be subject to ethical approval and comply with data security requirements.

3.2.3 Data completeness

Although the source data may be incomplete, a minimum degree of completeness is required. At a minimum, outcome variable, exposure/ intervention variable, demographics variable, and important covariates should exist. In addition, the impact of missing data in analysis model to the robustness of the research conclusion needs to be considered.

3.2.4 Sufficient sample size

There needs thorough consideration of potentially significant loss of data records after data curation, in order to ensure sufficient sample size for statistical analyses.

3.3 Fitness evaluation of curated data

Fitness evaluation of curated RWD is mainly based on data relevance and reliability.

3.3.2 Relevance evaluation

The relevant evaluation aims at evaluating whether RWD is closely relevant to the clinical question of concern, with particular focus on coverage of key variables, accuracy of the definition of exposure/intervention and clinical outcome, representativeness of target

population and the integration of data with different structures.

1) Coverage of key variables and information.

Important variables and information relevant to clinical outcomes should be included in RWD, such as drug use, patient demographic and clinical characteristics, covariates, outcome variables, follow-up duration, potential safety information, etc. If missing data exist in some of these variables, it should be fully evaluated whether they can be resolved using reliable statistical methods, and the potential impact on the result of causal inference should be evaluated.

2) Accuracy of the definition of exposure/intervention and clinical outcome.

The selection and accurate definition of clinical outcomes and exposure/intervention are essential for RWS, and should be consistent with the clinical significance or theoretical basis of the research objectives. Clinical outcome definitions should contain diagnostic criteria, measuring methods and their quality control (if any), measurement tools (e.g. the use of questionnaire scales), calculation methods, measurement time points, variable types, transformation of variable types (e.g. from quantitative to qualitative variable), mechanism of endpoint event evaluation (e.g. the operation mechanism of the endpoint event committee), etc. When the clinical outcome definitions are different between various data sources, a unified clinical outcome definition should be established using a reliable transformation method. The definition of exposure/intervention should consider the reasonableness of its time

window.

3) Representativeness of target population.

One of the advantages of RWS over traditional RCTs is the broader representation of the target population. Therefore, the development of inclusion and exclusion criteria should conform to the target population under the real-world practice as much as possible.

4) Merging multi-source heterogeneous data.

In many cases, RWD come from multiple sources and with different data structures. Therefore, there is a need to link or merge them at the individual level, based on individual subject identifiers, in order to make the data support the integration of key variables by a common data model (CDM) or data standard.

3.3.2 Reliability evaluation

The reliability of RWD is mainly evaluated from several aspects including completeness, accuracy, transparency, quality control and quality assurance.

1) Completeness

Completeness refers to the degree of missing of data information, including missing variables and/or missing variable values. For different studies, the proportion, distribution, reason and mechanism of missing are not the same and should be described in detail. While the missing data problem cannot be avoided in RWD, there should be a limit on the proportion of missing. When the proportion of missing data in a particular study obviously exceeds that of similar studies, the uncertainty of the

study conclusions will increase. In this case, it is necessary to carefully consider whether the data can be used to support the generation of RWE. A detailed analysis of the reasons for missing data helps comprehensively judge the reliability of the data. If the imputation of missing data is considered, proper imputation methods should be used based on reasonable assumptions of the missing mechanism.

2) Accuracy

Accuracy (or plausibility) refers to whether the data are consistent with the objective characteristics described, including whether the source data are accurate, whether the data value range is reasonable, whether the trend in change of outcome variables over time is reasonable, and whether the mapping code is unique, etc. The accuracy of the data needs to be identified and verified based on established references with proper authority, for example, whether the outcome event is assessed by an independent outcome event committee.

3) Transparency

Transparency refers to the clear and transparent plan and process of curation of RWD. It should ensure that key exposure variables, covariates and outcome variables can be traced back to the source data, and reflects the process of data extraction, cleaning, conversion and standardization. Regardless of whether the data process is done manually or in automated fashion, the standardized operating procedures (SOP) of data curation and verification and confirmation documents should be clearly recorded and archived. Potential data credibility issues, such as degree of missing,

ranges of variables, and calculation method and mapping relationship of derived variables, should all be well documented. The data curation plan should be formulated in advance according to the study objectives, and the data curation process should be consistent with the curation plan. Transparency of data also includes accessibility of data, information sharing between databases, and protection for patient privacy. If an algorithm is used to define the research cohort, the development of the algorithm and its verification should also be transparent.

4) Quality control

Quality control refers to the technologies and activities implemented to confirm that all aspects of data curation meet the quality requirements. Quality control evaluation includes but is not limited to data extraction, security processing, cleaning, structuring, and subsequent storage, transmission, analysis, and submission. Quality control aims to ensure that all data are reliable, and that the data processing has been conducted correctly. Complete, standardized, and reliable data curation schemes and plans should be followed, and corresponding data quality inspection and system verification procedures used to ensure that the data curation system operates under normal and stable conditions to ensure the accuracy and reliability of the RWD.

5) Quality assurance

Quality assurance refers to systematic actions to prevent, detect, and correct data errors or problems that occur during research. Quality assurance of RWD is closely related to regulatory compliance and should

be implemented throughout the entire data curation process. Considerations include but are not limited to whether a related research plan, study protocol, and statistical analysis plan are established; whether corresponding SOPs exist; whether clear processes and qualified personnel for data collection are in place; whether a common definition framework, i.e. data dictionary, is used; whether a common time frame for collecting key variables is followed; whether the technical methods used for data element capture are sufficient, including the integration of data from various sources, the recording of drug use and laboratory data, follow-up records, and links to another database, etc.; whether data input is timely and the data transmission is safe; whether the relevant regulatory requirements of on-site inspection (e.g. review of source data and files) are met.

4. Curation of RWD

Data curation refers to the curation of the source data for the purpose of statistical analysis to investigate a specific research question. Data curation includes, but is not limited to the following aspects: data security processing, data extraction (including multiple data sources), data cleaning (edit checks, outliers processing, and data completeness processing), data conversion (data standard, CDM, normalization, natural language processing, medical coding, derived variable calculation), data transmission and storage, data quality control, etc.

4.2Personal information protection and data safety processing

Real world studies require the protection of personal information. Therefore, according to the national information security technical specifications and medical big data security management regulations, de-identification should be applied to sensitive personal information to ensure this information cannot be restored in the data. It aims to prevent the leakage, damage, loss and tampering of personal information through technical and managerial methods.

Data safety processing requires the development of data encryption technology, procedures for risk assessment and emergency response covering all aspects of data curation, and implementation of inspections on effectiveness of security measures. These actions should be based on the type, quantity, nature and content of the various data involved in the study, especially for sensitive personal information.

4.3 Data extraction

Appropriate methods should be used to extract data, according to factors such as the storage format of the source data, whether they are electronic data, and whether they contain unstructured data, etc. The following principles should be followed in data extraction:

The method of data extraction should be verified to make sure that the extracted data meet the requirements of the study protocol. The extracted data and the source data should be conforming, and timestamp management of the extracted data and source data should be carried out.

Data extractors that are compatible or interoperating with the source data system should be used to reduce errors in data transcription, so as to

improve the accuracy and quality of data and efficiency of data collection in clinical studies.

4.4 Data cleaning

Data cleaning refers to the actions such as removal of duplicate or redundant data from extracted source data, edit checks of variable values, processing of abnormal values, as well as the processing of missing data. It should be noted that any changes to the data should be endorsed by the principal investigator or responsible personal of the source data. Otherwise, no data record should not be modified in order to protect the authenticity.

First, duplicate data and irrelevant data under the premise of ensuring data completeness should be removed. Duplicate data may be generated when merging data from different sources. Data unrelated to the research objective may be collected if the mapping relationship between the data source and the CDM is inaccurate. Removing these observations of no need from the dataset can reduce unnecessary work.

Then, logic checks and abnormal data processing should be carried out. Logic checks can correct errors in the source data or data extraction, for example: when the hospital discharge date is earlier than the admission date; when the birth date is inconsistent with the recorded age; when the results of laboratory tests are not biologically plausible; or when the qualitative assessment results are inconsistent with the criteria defined in the research protocol. Extreme care should be given in handling abnormal data to avoid potential generation of bias. The error findings and

abnormal data should be further verified before modifying the data and records of the changes should be kept.

Finally, missing data are to be addressed. The degree and reason for missing data, as well as the mechanisms by which missing data may occur can vary from case to case. If imputation of missing data is required, the correct imputation method should be adopted, with the consideration of the most reasonable assumptions of the missing mechanism.

4.5 Data conversion

After data cleaning, data conversion is the procedure that converts the original data into fitted RWD, by matching the data format standards, medical terms, coding system and derived variables with the requirements of the analysis datasets.

Reliable natural language processing algorithms can be used to convert text data, improving the efficiency of conversion while ensuring the accuracy and traceability of the data conversion.

When the derived variables are calculated, the source data variables and variable values for calculation, calculation methods and the definition of derived variables should be clearly defined, and timestamp management should be carried out to ensure the accuracy and traceability of data.

4.6 Data transmission and storage

The transmission and storage of RWD requires secure network environment, which controls the whole lifecycle of data from collection, processing, analysis to destruction. Encryption should be implemented during data transmission and storage. In addition, approval processes,

regulations of access control, role authority control and minimum authority access control policies should be established, the establishment of automated audit systems is encouraged，which monitor the processing and access of recorded data.

4.7 Data quality control

Data quality control is essential for ensuring the completeness, accuracy and transparency of study data. Data quality control requires the establishment of RWD quality management system and SOP under the following principles:

3.3.2 The accuracy and plausibility of source data

The accuracy and plausibility of source data should be ensured. For instance, EMR should formulate standards for quality control of medical records to meet analysis requirements. Outpatient disease description, diagnosis and medication information should be supported by related evidence. For any modification in the entry process, the responsible person must confirm and sign, and provide the reason for the modification, to ensure that a complete audit trail is kept.

3.3.2 Data completeness during data extraction

Data completeness should be fully considered in data extraction. Extraction fields should be evaluated and defined, and corresponding verification rules and database structures should be developed.

3.3.2 Data quality management plan

A complete data quality management plan should be developed. A systematic quality control and manual quality control plan should be

developed to ensure the accuracy and completeness of data. Complete checks and source document reviews should apply to key variable values, while other variable values may be subject to sampling-based checks according as needed and appropriate. For example, for demographic information, sampling of threshold values of numerical variables, coding mapping relationships, etc., can be conducted in a specified proportion to verify their accuracy and rationality.

4.8 Common data model

The CDM is a data model, based on a multi-disciplinary cooperation mode, for rapid, centralized and standardized processing of multi-source heterogenous data. It is primarily used to transform source data under different standards into a unified structure, format, and terminology for data consolidation across databases/datasets.

In view of the complexity of the structure and type of multi-source data, the differences in sample size and standards, it is necessary to extract, transform and load (ETL) the source data throughout the whole process of transformation into a CDM. It is necessary to ensure that the source data are syntactically and semantically consistent with the structure and terminology of the target analysis database (see Figure 2).

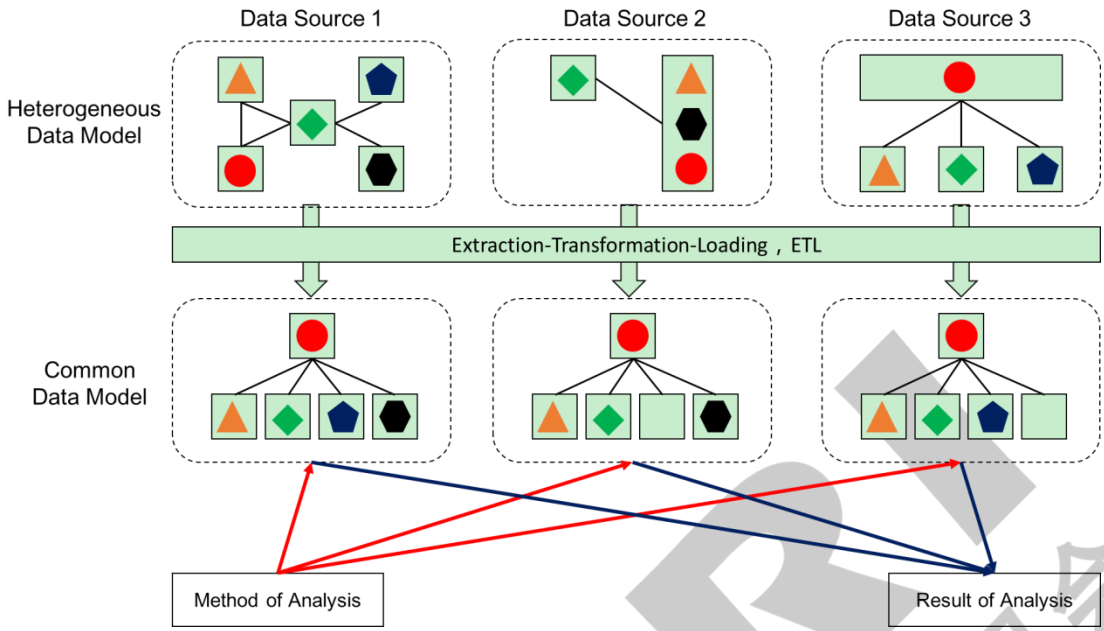

Figure 2: Transformation from heterogenous data model to a CDM

An ideal CDM should meet the following principles:

1) The CDM can be defined as a data curation mechanism, which can standardize source data and turn them into common structures, formats, and terminology, allowing data consolidation across multiple databases/datasets. The CDM should have access to source data and be a dynamically scalable and continuously improved data model with version control.

2) The definition, measurement, merging, recording and corresponding verification of CDM variables needs to be transparent, and clear and consistent rules need to be followed in data conversion across multiple databases.

3) The CDM should provide a common set of baseline concepts. Common variables or concepts related to safety and effectiveness should

be mapped to the CDM to be applicable to different clinical research questions, and can be compared with known study results.

4.9 RWD curation plan

The RWD curation plan should be developed in advance and be consistent with the research plan. If the curation plan needs to be revised during the study, it should be communicated with authority and submit the updated curation plan at the same time. The plan should document the objective of RWD use with respect to a regulatory decision, the study design that utilizes RWD, and record the source of the RWD, including but not limited to the following: types of RWD source data/source files, such as health information system (HIS) data, disease registry data, health insurance data, etc.; for RWD source data/source files, appropriate evaluation of previous applications and justification for utilization; curation of the RWD, i.e. the processes needed to transform the RWD data source to the analysis database; data models and data standards adopted; handling of missing data; actions taken to reduce or control the potential bias resulting from the use of RWD; quality control and quality assurance; and RWD fitness evaluation.

5. Compliance, safety and quality management system of RWD

5.1 Data compliance

RWD come from various sources, such as individual patient diagnosis and treatment, and patient privacy is a consideration in the collection, processing and use of the RWD. Any access to and use of RWD must be

approved by an ethics committee in order to safeguard the safety and rights of patients to the utmost extent. Participants of RWD curation need to strictly comply with the requirements of relevant laws and regulations, and the applicant must strictly implement and fulfill the obligations of identity protection and management.

5.2 Data security management

Data security management needs to be carried out pursuant to national laws and regulations and industry regulatory requirements, and necessary measures taken to maintain the safety of the information system and network facilities as well as the cloud platform that contains health and medical data. Protection of data security should cover the whole life cycle of data, including data collection, extraction, transmission, storage, exchange, and destruction. Encryption technology should be adopted to ensure the completeness, confidentiality and traceability of data during the processes of collection, extraction, transmission and storage. When a medium is used for transmission, control needs to be exercised over the medium. Different protection measures should be adopted for data forms of different media, and corresponding access control mechanism is established to audit, file, archive and audit access records.

Data audit and relevant SOPs provide records and the basis for data collection, extraction, transmission, maintenance, storage, sharing and use, and should include personnel, management and technical audits. Healthcare information system activity audit policies and appropriate SOPs should be developed and deployed. The scope of any audit should

include operations of any state of the data (including the act of logging in, creating, modifying, and deleting records). All operations should automatically generate audit records with timestamps (including but not limited to authorization information, operation date, operation reason, operation content, operator and signature, etc.) and be available for audit. Audit records should be securely stored, and access control policies should be established.

5.3 Quality management system

A comprehensive quality management system should be established to standardize RWD processing that is constantly optimized and improved in daily work. The basic quality elements should ensure the quality of RWD, and an operation procedure for the full RWD lifecycle management should be established. The functions of the computerized system should meet the requirements for RWD management and the relevant regulations on the computerized system; a complete personnel management system should be established, and personnel responsible for data collection, curation, analysis should receive appropriate training to meet the capability requirements of corresponding responsibilities, and authority standardized management should be in place; risk management processes from data collection to data submission should be established; standard information and document management specifications (paper and electronic versions) should be developed to ensure complete, accurate and transparent records of RWD processing and to protect data security and compliance.

6. Communication with regulatory authorities

In order to ensure that the quality of RWD meets regulatory requirements, the applicant is encouraged to communicate with the regulatory authorities in a timely manner. Prior communication is needed to identify whether the RWD support the generation of RWE according to the development strategy and specific study protocol. This includes RWD accessibility, whether the sample size is large enough, whether the data curation plan is reasonable and feasible, and whether the data quality can be guaranteed. During the study, if the data curation plan needs to be adjusted according to changes in the implementation of the study, this needs to be explained to the Health Authority and seek consent and endorsement. The updated study protocol and data curation plan should be recorded. The sponsor may consult the Health Authority on the application materials and database after the study is completed and before the submission of data.

Reference

[1] Cai Ting, Zhan SY. Develop the active surveillance system for vaccine safety in China [J]. Chinese Journal of Preventive Medicine, 2019,53(7): 664-667.

[2] National Health Commission, National Medical Products Administration. Guideline for Good Clinical Practice, 2020.07.01.

[3] The Center for Drug Evaluation of National Medical Products Administration. Technical Guideline on Data Management in Clinical Trial. 2016.07.27.

[4] National Medical Products Administration. Guidance on Using Real World Evidence to Support Drug Development and Evaluation (for Trial Implementation); 2020.01.07.

[5] Hou YF, Song HB, Liu HL, et al. Practice and Discussion on Active Surveillance by China Hospital Pharmacovigilance System [J]. Chinese Journal of Pharmacovigilance, 2019,16(4): 212-214.

[6] Zhou Li, OuYang WW, Li Geng, et al. Analysis of the current situation of registry studies in China[J]. Chinese Journal of Evidence-Based 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-procedural-guideline/reflection-paper-expectations-electronic-source-data-data-transcribed-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-model-europe-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: Glossary

Electronic Medical Record (EMR): Electronic records of individual patients' health-related information created, collected, managed and accessed by authorized clinical professionals within medical institutions.

Electronic Health Record (EHR): Electronic records of individual patients' health-related information created, managed, and consulted by authorized clinical professionals in multiple health care facilities, in compliance with nationally recognized interoperability standards for use.

Observational Study: Studies that explore the causal relationship between exposure/treatment and outcome with respect to specific research questions, without active intervention.

Patient-Reported Outcome (PRO): An indicator from the patient's own measurement and evaluation of disease outcome, including symptoms, physiology, psychology, functional capacity, satisfaction with medical services, etc. Both paper and electronic records are available, the latter being called ePRO.

Logic Check: The examination of the validity of clinical study data entered into a computer system. It primarily evaluates whether there are logical errors in the input data and its expected numerical logic, numerical range, or numerical attributes.

Data Standard: A set of rules on how to establish, define, format, or exchange specific types of data between computer systems. Data standard makes the information presented predictable and consistent, and in a form that can be used by information technology systems or scientific tools.

Data cleaning: The purpose of data cleaning is to identify and correct the noise in the data to minimize the impact of noise on the analysis results. Noises in data mainly include incomplete data, redundant data, conflicting data and wrong data.

Data Linkage: The merger, association and combination of data and information from multiple sources to form a unified data set.

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

Data Curation: Data curation refers to the curation of the source data for the purpose of statistical analysis of specific clinical research questions. Data curation includes the following aspects: data extraction (including multiple data sources), data safety processing, data cleaning (edit check and outliers processing, data completeness processing), data conversion (CDM, normalization, natural language processing, medical coding, derived variable calculation), data quality control, data transmission and storage, etc.

Common Data Model (CDM): A data system that realizes rapid, centralized and standardized processing of multi-source isomerous data based on the multidisciplinary cooperation model. It is primarily used to transform source data subject to different data standards into a unified structure, format, and terminology for data consolidation across databases/datasets.

Covariate: Variables that are expected by researchers or determined by

exploratory analysis to have significant impacts on the primary outcome variables. It is divided into baseline covariates and non-baseline covariates.

Source Data: Clinical symptoms and observations recorded in clinical studies, and all information on original records and certified copies used to reconstruct and evaluate other activities of the study. The source data are contained in the source file (including the original record or a valid copy thereof).

Real-World Data (RWD): A variety of data, collected through regular practice, that are related to an individual patient's health status and/or diagnosis, treatment and healthcare. Not all RWD are fit for use in RWE generation. Only RWD that meet the fitness requirements can generate RWE.

Real-World Research/Study (RWR/RWS): For specific clinical research questions, collection of data relating to the health status and/or diagnosis and care of the subject in a real-world setting (RWD) or summary data derived from such data. It is the process of obtaining clinical evidence (Real-World Evidence) of the use value and potential benefit-risk of the drug through analysis.

Real-World Evidence (RWE): Clinical evidence on drug use and potential benefit-risk from appropriate and adequate analysis of applicable RWD.

Appendix 2: Chinese-English Vocabulary

中文	English
预防接种不良事件	Adverse Events Following Immunization, AEFI
通用数据模型	Common Data Model, CDM
病例报告表	Case Report Form, CRF
数据治理	Data Curation
病例登记	Patient Registry
电子数据采集	Electronic Data Capture, EDC
电子病历	Electronic Medical Record, EMR
电子健康档案	Electronic Health Record, EHR
电子患者报告结局	electronic Patient-Reported Outcome, ePRO
观察性研究	Observational Study
患者报告结局	Patient Reported Outcome, PRO
结局变量	Outcome Variable
可追溯性	Traceability
逻辑核查	Edit Check
数据标准	Data Standard
数据清洗	Data Cleaning
数据元素	Data Element
数据治理	Data Curation
通用数据模型	Common Data Model, CDM
医院信息系统	Hospital Information System, HIS
衍生变量	Derived Variable
源数据	Source Data
真实世界数据	Real World Data, RWD
真实世界研究	Real World Research/Study, RWR/RWS
真实世界证据	Real World Evidence, RWE