Journal of the Society for Clinical Data Management

Satoshi Ueno, Yusuke Komiyama, Mariko Doi, Keika Hoshi. The Need for Data Standardization and Research Data Management Infrastructure to Promote the Use of Real-World Data. *Journal of the Society for Clinical Data Management*. 2024; 4(1): 6, pp. 1–9. DOI: https://doi.org/10.47912/jscdm.210

OPINION PIECE

The Need for Data Standardization and Research Data Management Infrastructure to Promote the Use of Real-World Data

Satoshi Ueno[*], Yusuke Komiyama[†], Mariko Doi[‡], Keika Hoshi[*]

Real-world data (RWD) is increasingly being used for regulatory decision-making and as a control group for new drug approval applications. RWD is also valuable for understanding risk factors (e.g., pre-existing medical conditions, personal protective equipment, travel, contacts, smoking, and exposure to animals) and vaccination status for the coronavirus disease 2019 (COVID-19). The methodology of utilizing RWD is inconsistent across healthcare institutions. However, the methodologies for utilizing RWD vary across healthcare institutions. Standardizing RWD for clinical use is crucial, and possible solutions include adopting Clinical Data Interchange Standards Consortium (CDISC) standards, tools, and concepts. This study examines the availability of CDISC and other international standards for the utilization of RWD with concrete examples and presents the potential platform for implementation.

We propose a temporary solution to convert clinical data warehouse (DWH) data into the Fast Healthcare Interoperability Resources (FHIR) format to comply with CDISC standards. This approach would allow for converting institution-level standards to national standards as an interim solution until FHIR is supported, mapping national standards to international standards. We believe that the ideal research environment is a data platform that adheres to both national and international regulations related to RWD applications. Within such a platform, users can share data freely, rather than rely on a specific facility or vendor. Data platform developments are progressing in Japan and globally. In Japan, initiatives to use research data on research data platforms are being conducted. We are experimenting with implementing tools and knowledge shared by CDISC.

Keywords: real-world data; electronic health records; reference standards; pragmatic clinical trials; research data management

Standardization and applications of data for various purposes

Current status of medical Information in healthcare institutions

Similar to RWD-related medical information, such as clinical data or test results, natural history data can also be used as a control group in new drug approval applications. RWD is increasingly utilized for new drug approval applications. However, electronic data management has not been centralized in Japanese healthcare institutions, where departments dealing with billing (health insurance claims), billing automation, and ordering have developed separate electronic systems. This indicates that the data collected and stored in electronic health records (EHRs) are not standardized even within individual institutions, preventing system-wide or cross-sectional searches and data extraction. Despite the publication of standards by the Ministry of Health, Labour and Welfare (MHLW) in Japan, the implementation and utilization of standards other than "ICD10 Based Standard Disease Code Master for Electronic Medical Records (HS005)" is less than 40% in medical institutions.[1] Thus, the adoption of national standards is not widespread. There is a pressing need to advance standardization so that medical information can be exchanged among medical institutions. This could facilitate patient-directed exchange of personal health records (**Figure 1**).

* Center for Health Informatics Policy, National Institute of Public Health, Wako, Saitama, JP

† Digital Content and Media Sciences Research Division, National Institute of Informatics, Chiyoda-Ward, Tokyo, JP

‡ Department of Epidemiology and Biostatistics, National Institute of Public Health, Wako, Saitama, JP

Corresponding author: Satoshi Ueno, PhD (ueno.s.aa@niph.go.jp)

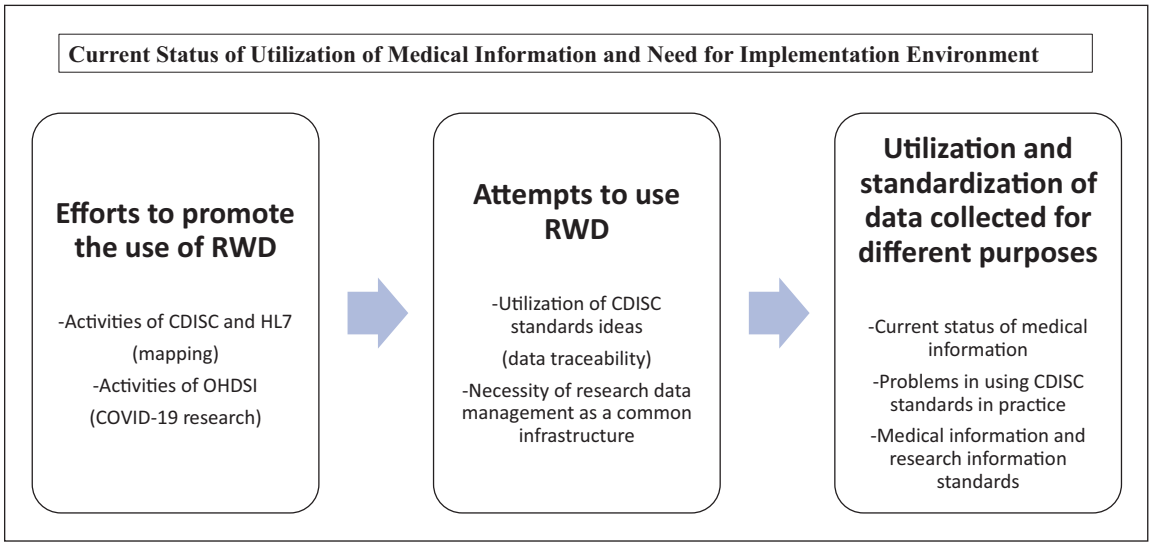

Figure 1: Current Status of Utilization of Medical Information and Need for Implementation Environment. The above diagram is a flowchart illustrating the structure of this paper. The current state of medical information makes it difficult to utilize data and implement standards effectively. Despite efforts by organizations to use RWD, a common infrastructure is lacking. To promote the use of RWD, a research data management infrastructure that implements publicly available standards and tools is required.

Challenges in implementing CDISC standards in clinical settings

The primary goal of medical information is to support and document clinical diagnosis, treatment, and monitoring. Another important goal is to claim medical service fees. As of 2017, 46.7% of general Japanese hospitals used EHRs. EHR adoption was larger (85.4%) in hospitals with ≥400 beds and 64.9% in hospitals with 200–399 beds.[2] However, the electronic chart systems are often customized for each hospital. No industry standards, such as variable names, were used, which impeded the integrated use of medical data stored across multiple systems.

The authors previously reported achieving downstream data standardization of EHR data using the Clinical Data Acquisition Standards Harmonization (CDASH) Implementation Guide (CDASHIG) v2.2, which is the standard for data collection. Fields necessary for research data were set according to CDASHIG, and data were mapped to clinical DWHs (**Table 1**). Of the 35 domains in CDASHIGv2.2 (2 Special-Purpose domains, 6 Interventions class domains, 6 Events class domains, 19 Findings class domains, and 2 Findings About event domains), only one domain, the laboratory Test Results (LB) of findings domain, was mappable to facility standard data elements and results. However, of the remaining 34 domains, 11 domains (i.e., Comments (Co), Demographics (DM), Prior and Concomitant Medications (CM), Procedures (PR), Substance Use (SU), Healthcare Encounters (HO), Medical History (MH), Physical Examination (PE), Questionnaires, Ratings, and Scales (QRS), Subject Characteristics (SC) and Vital Signs (VS)) were listed in the EHR, while the other 23 domains could not be confirmed. 7 of the 23 domains (ie, Exposure as Collected and EX – Exposure (EC), Adverse Events (AE), Clinical Events (CE), Disposition (DS), Protocol Deviations (DV), Inclusion/Exclusion Criteria Not Met (IE), and Pharmacokinetics Concentrations (Sampling) (PC)) were research-related concepts. They were not present in the EHR because there is no concept of protocol-defined targets or time periods. The HO domain contained information on the referring hospital. 10 domains (ie, CO, DM, CM, PR, SU, MH, PE, QRS, SC and VS data) were described as notes written in text, suggesting that text mining or other processing methods are required for their utilization, and that 4 domains (ie, CM, Meal Data (ML), Drug Accountability (DA), ECG Test Results (EG)) are data obtained from other departmental systems besides the EHR. Although the conditions are limited, as we only used a small number of cases from one hospital, only the LB domain, 1/35 (2.85%) of all 35 domains and 1/19 (5.26%) of the 19 Findings class domains, had data available, almost unchanged from the EHR.

The majority of EHRs are aimed simply at storing records and are written in article form. This format would require a colossal amount of text data processing (e.g., text mining or structured data formatting) to make it usable and sharable within and between institutions, except for laboratory test data, where items and values are input into separate fields. Standard features of EHRs only allow extraction of limited data, and systematic problems, such as lack of EHR standardization, and operational problems, such as medical staff failing to record information in the appropriate fields, are the causes. Furthermore, given that the current systems are historically based on systems intended for billing, there are challenges that must be overcome through cooperation with EHR vendors to promote RWD applications. FHIR is currently gaining attention as a standard for medical data as a way to solve some of these issues. Considering the efficacy of medical care and tests, it is important to enhance the reuse of electronic data because data accumulation for research purposes will lead to new findings.

Table 1: Assumed mapping of electronic medical record entries and data output from the DWH according to CDASHIG v2.2. A domain is a collection of data points with a common topic, such as adverse events or demographics. CDASHIG domains are aligned with SDTMIG domains for end-to-end data traceability (ref. 15). Domain names are presented according to CDASHIG v2.2 and divided into the following categories: Special-purpose, Interventions, Events, Findings, and Findings about events and interventions. Using EHRs found commercially in Japan, we checked whether the information collected in the CDASHIG domain exists in the EHRs and in the data output from the DWH.

Domain names	Electronic health record	Data output from DWH	Mappable data without processing	Data examples
1 SPECIAL-PURPOSE DOMAINS				
1.1 CO – COMMENTS	Y Article writing requiring text mining	Y Partially, difficult to identify	N	–
1.2 DM – DEMOGRAPHICS	Y Article writing requiring text mining	Y Partially, difficult to identify	N	Birthdate, Sex
2.1 INTERVENTIONS CLASS DOMAINS				
2.1.1 CM – Prior and Concomitant Medications	Y Article writing requiring text mining or for other departmental systems	Y Partially, difficult to identify, multiple information required	N	Concomitant Meds Start Date
2.1.2 EC – Exposure as Collected and EX – Exposure	NA for research concepts	NA	NA	–
2.1.3 PR – Procedures	Y Article writing requiring text mining	Y Test orders were available.	N	CT Procedure Start Date
2.1.4 SU – Substance Use	Y Article writing requiring text mining	Y Partially, difficult to identify	N	Smoking Alcohol Substance Use
2.1.5 ML – Meal Data	N for other departmental systems	N	N	–
2.1.6 AG – Procedure Agents	N	N	N	–
2.2 EVENTS CLASS DOMAIN				
2.2.1 AE – Adverse Events	NA for research concepts	NA	NA	–
2.2.2 CE – Clinical Events	NA for research concepts	NA	NA	–
2.2.3 DS – Disposition	NA for research concepts	NA	NA	–
2.2.4 DV – Protocol Deviations	NA for research concepts	NA	NA	–
2.2.5 HO – Healthcare Encounters	Y Referral information available	Y Partially, difficult to identify	N	Healthcare Encounter Occurrence
2.2.6 MH – Medical History	Y Article writing requiring text mining	Y Partially, difficult to identify	N	Medical History Term
2.3 FINDINGS CLASS DOMAINS				
2.3.1 DA – Drug Accountability	N for other departmental systems	N	N	–
2.3.2 DD – Death Details	N	N	N	–
2.3.3 EG – ECG Test Results	N for other departmental systems	N	N	–
2.3.4 IE – Inclusion/Exclusion Criteria Not Met	NA for research concepts	NA	NA	–

(Contd.)

Domain names	Electronic health record	Data output from DWH	Mappable data without processing	Data examples
2.3.5 LB – Laboratory Test Results	Y	Y	Y	Specimen Collection Date, Laboratory Test Name, Result, Unit, Normal Range Lower Limit, Normal Range Upper Limit
2.3.6 MB – Microbiology Specimen	N	N	N	–
2.3.7 MS – Microbiology Susceptibility	N	N	N	–
2.3.8 MI – Microscopic Findings	N	N	N	–
2.3.9 PC – Pharmacokinetics Concentrations (Sampling)	NA for research concepts	NA	NA	–
2.3.10 PE – Physical Examination	Y Article writing requiring text mining	Y Test orders were available.	N	Exam Date
2.3.11 QRS – Questionnaires, Ratings, and Scales	Y Article writing requiring text mining	Y Partially, difficult to identify	N	Date, test name
2.3.12 RP – Reproductive System Findings	N	N	N	–
2.3.13 RS – Disease Response and Clin Classification	N	N	N	–
2.3.14 SC – Subject Characteristics	Y Article writing requiring text mining	Y Partially, difficult to identify	N	Education Level, Employment Status, Marital Status
2.3.15 TU – Tumor/Lesion Identification	N	N	N	–
2.3.16 TR – Tumor/Lesion Results	N	N	N	–
2.3.17 VS – Vital Signs	Y Article writing requiring text mining	Y Partially, difficult to identify	N	Height, Weight
2.3.18 OE – Ophthalmic Examinations	N	N	N	–
2.3.19 RE – Respiratory System Findings	N	N	N	–
2.4 FINDINGS ABOUT EVENTS AND INTERVENTIONS DOMAIN				
2.4.1 FA – Findings About Events or Interventions	N	N	N	–
2.4.2 SR – Skin Response (Findings About Interventions)	N	N	N	–
2.5 ASSOCIATED PERSONS DOMAINS				
cf, DM Domain/CM Domain/MH Domain	–	–	–	Caregivers

Y; presence, N; absence, and NA; not applicable.

Standards of medical information and clinical research

The implementation and application statuses of the MHLW standards can be divided into several categories, as listed below.

First is the "HS005 ICD10-based Standard Disease Code Master for Electronic Medical Records", which is used for coding disease types. It is the most common system of codes and terminology, used by 92.0% (617/671) of hospitals and 71.6% (288/402) of clinics in Japan.[1] "Data formats" consist of "HS032 Standard Specification for Discharge Summary based on HL7 CDA Release 2", which standardizes sections of the discharge summary and enables the electronic exchange of these sections, used by <10% of both hospitals and clinics in Japan, and "HS011 Digital Imaging and Communications in Medicine (DICOM)", which is defined as a format for exchangeable medical images with the data and quality necessary for clinical use, used by 23% of hospitals in Japan.[1] "HS026 'SS-MIX2 Storage' Specification and Guidelines for Implementation", which specifies how Health Level 7 (HL7) 2.x messages from a medical information system should be archived to external storage, was the predominant guideline for "data exchange", which is followed by 37.3% of hospitals, while the other standardization guidelines were adopted by others, up to approximately 15%.[1] Given the above-mentioned information, although the MHLW Standards exist as a domestic standard of medical information management, their use far from widespread, given that the majority of healthcare institutions are collecting clinical data according to individual internal standards. This is also influenced by the fact that such databases are aimed at clinical use and billing, with seemingly no benefits in conforming to a standardized method. The MHLW has encouraged health information sharing networks and has listed the reduction of intersystem connection fees and testing periods, the continuity of data after system updates, interfacility intersystem data exchange, interfacility data exchange (regional health information sharing networks), and analysis of accumulated data as some of the advantages.[3] Using an international standard or a standard that can be directly linked to an international standard is the appropriate strategy. This is not always possible. Although this is a difficult situation, the method of linking intra-hospital standards to national standards and then to international standards should translate data into internationally accessible formats. Although data exchange and traceability are considered for intra-hospital systems, they are only implemented at individual facility levels. This is a long way from true standardization.

CDISC standards are used in clinical research, mainly for approval of regulatory applications. To date, EHR data that correspond to data in paper case report forms (pCRFs) and electronic case report forms (eCRFs) have been manually identified in medical records by study staff and recorded on or entered into the CRF or monitors to collect data. After this the data are source-data verified and are subject to audits and on-site investigations by regulatory authorities. The digitization of CRFs and application documents is now common practice, and interest in RWD is increasing. As such, there have been attempts to use data from EHRs and clinical DWHs. However, this involves extremely high costs as it requires custom computer programming at each institution. Furthermore, use of EHR data from only a limited number of facilities in a study could introduce bias. For RWD utilization, standards and tools that can bridge medical and research information are necessary for medical data to be used as research data. The CDISC RWD Connect project, which started in the fall of 2019, aims to identify the necessity of guidance and tools to facilitate the use of RWD.[4,5] The CDISC model may not be compatible for some clinical data and processes. CDISC's work to bridge the gap between healthcare and research should lead advances in automation and reuse of data from healthcare settings.

Efforts to promote the use of RWD

Disclosure of information on standards and tools

The FHIR standard formulated by HL7, has emerged as a set of standard specifications for healthcare data exchange in medical information.[6] The CDISC standards established by CDISC and the Observational Medical Outcomes Partnership (OMOP) common data model (CDM), established by the Observational Health Data Sciences and Informatics (OHDSI) are established for use in research.[7,8] The cooperation between CDISC, OHDSI, and HL7 has been advancing efforts to link HL7 FHIR with CDISC standards, and the development of OMOP CDM for efficient data collection is underway. Information related to implementation is published on the organizations' websites, and methods for using HL7 FHIR could be applied to generate research data from medical information for the utilization of RWD. In addition to the collaboration of various organizations, open-source programs and software have been published and are currently available, with new developmental plans continuously progressing.[9,11]

CDISC initiatives: mapping of CDISC standards to HL7 FHIR

In September 2021, CDISC published the "FHIR to CDISC Joint Mapping Implementation Guide v1.0".[9] This document defines the mapping between FHIR release 4.0 with the three CDISC standards: CDASHIG v2.1, SDTMIG v3.2, and LAB v1.0.1. It links EHR data to datasets that can be submitted to regulators according to CDISC standards.[9] Currently, the data model is limited to specific categories of data, such as laboratory tests (LB), vital signs (VS), adverse events (AE), medical history (MH), concomitant medications (CM), treatment and testing procedures (PR), and subject demographics (DM). The "FHIR to CDISC Joint Mapping Implementation Guide" also supports the creation of CRFs that align with data elements defined by FHIR resources and profiles embedded with CDISC variables and are useful in streamlining data collection time, eliminating redundant data entry, improving quality, and reducing costs. Future developments should enable the systematic collection of diverse data from healthcare sources.

OHDSI Initiatives: Case study from the COVID-19 pandemic

The OHDSI is an organization that seeks to standardize data from observational research. The open community of OHDSI has conducted numerous studies using its open-access tools. One notable example is their large-scale study on COVID-19, conducted over a short period. This study, involving 34,000 adults hospitalized due to COVID-19 in three countries, was published in October 2020.[10,11] As data can vary widely by organization, international standardization is necessary to present data in a common format that facilitates collaborative research, large-scale analyses, and the sharing of advanced tools and methodologies.[12] The OHDSI has published multiple software and tools, such as ATLAS and HADES (previously the OHDSI Methods Library) in the R package collection, which allows planning and performing patient-level observation data analysis; DATA QUALITY DASHBOARD, which is related to data quality; ACHILLES in the R package, which evaluates database characteristics and visualization; ATHENA, which assists in searching and reading terms; WHITERABBIT and RABBIT-IN-A-HAT, which are used in ETL design; and USAGI, which assists in creating code maps. This publicly available information can be used to conduct analysis using the same methodology, leading to more efficient and effective research reporting.[12]

Initiatives to encourage the use of RWD

Roles of CDISC standards: leveraging methodologies for each research stage

The CDISC standards ensure end-to-end data traceability. Some regulatory authorities, such as the US Food and Drug Administration (FDA), require the submission of data conforming to the Study Data Tabulation Model (SDTM) in tabular format for the Analysis Data Model (ADaM), which is aimed at analysis and reporting for the submission of Clinical Study Reports to regulatory authorities for approval applications.[13,14] CDASH is the CDISC data collection model that aligns with SDTM and supports standardized data collection, and the Protocol Representation Model (PRM) exists as a protocol standard. As such, CDISC standards, which consider end-to-end data traceability throughout the stages of the study protocol, data collection, data organization, and analysis, play an important role in data management. In terms of data collection, CDASH covers the problems related to data collection that are faced by healthcare institutions and is helpful in facilitating collaboration between processes, such as data management, that occur between data collection, handling, and reporting; programming; data analysis; and clinical evaluations.[15] Given that data traceability between CDASH and SDTM is ensured, using CDASH is key to the success of RWD utilization.

Given that end-to-end data traceability is guaranteed, the concepts of CDISC offer two perspectives on research data creation and validation. The CDISC standards are compatible with theories and best practices for research data management (RDM). For medical research, the CDISC PRM, CDASH, STDM, and ADaM standards have been published for planning, data collection, data tabulation, and analysis, respectively. Each standard and its associated tools and data standards can facilitate the use of healthcare data as research data. For example, they use PRM to define or organize items necessary for study start-up and protocol implementation, including information necessary for clinical trial registration.

The use of CDASH unifies data collection and data quality checking, and automates data acquisition from the healthcare system, thereby clarifying data storage at the time of collection. In July 2021, CDISC published the Therapeutic Area User Guide for COVID-19.[16] The widespread use of HL7 FHIR and mapping to CDISC standards facilitates RWD use.

Solutions and case studies

Considerations for implementing RDM platform

It is necessary to prepare an implementation environment at individual facilities to use publicly available information and tools. However, there are resource limitations. For example, building a custom environment for each study is not feasible with limited research funds. A shared environment should allow for study implementation with limited resources. Cloud platforms may offer further efficiency. Moreover, using the same methods and processes for data management across studies is ideal for ensuring data quality. Data standardization makes such platforms possible and facilitates the reuse of information systems and data.

Potentials of GakuNin RDM

In Japan, the Cabinet Office (CAO) is an agency of the Japanese Cabinet, officially headed by the Prime Minister. The CAO is responsible for handling the affairs of the Cabinet. The promotion of science and technology innovation policies established by the CAO requires universities to introduce systems that allow them to manage research data policies and metadata. The National Institute of Informatics (NII) developed GakuNin RDM as a national RDM service powered by the Open Science Framework (OSF). The service promotes research and prevents research misconduct nationwide at universities and research institutions.[17,18] GakuNin RDM is part of NII's research data platform, NII Research Data Cloud (NII RDC), specializing in handling data before it is published. Over 110 Japanese institutions have already adopted the service internally, which can be used as IT infrastructure for RWD collection.

The GakuNin RDM system is designed by building extensions to the Open Science Framework, an open-source software developed by the US Center for Open Science, to expand its services. GakuNin RDM is part of GakuNin, the Japanese authentication and authorization infrastructure (AAI), and can be accessed using federated authentication systems of various institutions allowing data sharing and management beyond institutional borders.[19] Furthermore, GakuNin RDM is connected to the Science Information NETwork (SINET), a high-speed network service for academic research. Institutional users of GakuNin RDM can allocate their preferred external cloud storage for saving their data.[20] Data saved while using the GakuNin RDM services are traced and managed by linking them to the time stamping server of the time

stamping authority. Although GakuNin RDM is a versatile RDM service, third-party tools such as plug-in software can be developed by researchers and institutions in specialized fields. GakuNin may therefore facilitate RWD use in Japan.

A universal system is an important criterion for building a national RWD platform using GakuNin RDM. Thus, it is necessary to develop a common RWD platform that is not limited to the functions needed by a specific project or organization while maintaining mutual compatibility between international standards and systems. As part of our research activities, we are working in collaboration with the team of the data center of the National Institute of Public Health (NIPH) in Japan to design 'FHIR to CDISC' as a plug-in software for GakuNin RDM. By providing the developed plug-in as a service to universities and research institutions nationwide through GakuNin RDM, we will make the service available to the entire healthcare field.

Example of GakuNin RDM usage at a national research institution in the public health domain

NIPH adopted the JAIRO Cloud, an academic repository service provided by NII and used by more than 800 institutions in Japan. JAIRO Cloud uses the open-source software repository software WEKO3. Because obtaining approval to use the GakuNin RDM organization-wide at the NIPH is difficult, the authors in this study applied to NII to use GakuNin RDM from NIPH, allowing a single division to use the RDM platform as a test operation. For the trial, we used Orthros, a service that allows users to use IdP on a trial basis, where the user's affiliated organization has not yet adopted the authentication system. As a next step, we need to improve the environment for using Orthros to participate in Academic Access Management Federation in Japan. In the future, we plan to investigate the operation secretariat system and cooperate with the institutional repository in parallel while introducing RDM services at the NIPH. Since there are limits to organizing this independently within the facility, maximizing the use of existing services is essential, thus providing an opportunity to revise its use.

Furthermore, researchers can ensure high levels of security when managing, storing, and sharing data and documents with co-researchers who have limited access. Storing research data used for the publication of the research as an organization should contribute to ensuring the reliability of that research and accumulating research data and methods. It is therefore necessary to ensure that organizational research platforms and management operations conform to the organization's rules.

Conclusion

Currently, two approaches are under investigation for RDM that consider the CDISC standards (**Figure 2**). The first approach is to apply the cases that have detailed the CDISC standards and an idea into RDM methodology. This would establish the methods of the RDM platform with reference to the CDISC methodology and best practice related to data handling advocated by CDISC. By keeping the utilization, application and integration of data in perspective, the use of data from other studies and fields would lead to new findings. The second approach is to implement the tools provided by CDISC and other standards development organizations in the RDM platform so that high-quality data can be used in

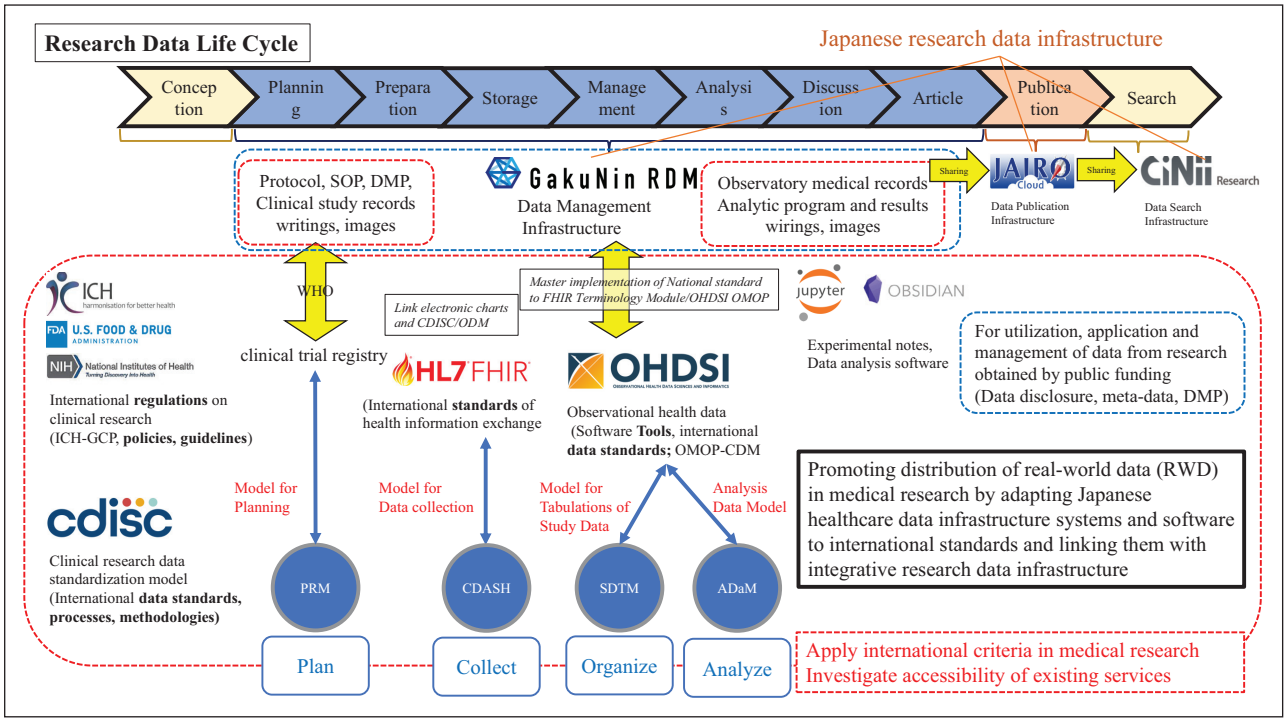

Figure 2: Research Data Life Cycle. The standards of environment development and implementation in Japan is incredibly complex. Although the environment is being developed, efficient implementation methods from planning to analysis are still in progress. One case study using existing standards and tools is presented.

analysis and reports by the same methodology. In the first approach, Table 1 suggests that it may be possible to map medical information for clinical laboratories only to the CDISC standards in its original form. Although orders were recorded for tests, diagnostic imaging settings and test results were stored in the system of the testing equipment. Text data was stored as text with reference to other results, which could not be easily identified, making it difficult to search for and extract data. In addition, it was difficult to locate the data due to the lack of a specific location for the description. Laboratory test results are considered the easiest data to standardize because the test item names, measurement results, units, etc., are stored in a fixed format and location. Moreover, based on the second approach of using publicly available mapping tools, any medical institution can generate data in compliance with CDISC standards using the same standards and tools. By applying HL7 FHIR to medical information and using the FHIR to CDISC Joint Mapping Implementation Guide, data can be collected in CDASH format. If the CDISC and OHDSI tools can be linked using HL7 FHIR as a hub, it will be a powerful tool in medical research. By using the CDISC standards, which have characteristics of both approaches, the flow advocated by CDISC can follow data collection, contributing to highly reliable data. Implementing the two above-mentioned approaches should improve data quality, ensure end-to-end data traceability, and assure established methodology. A substantial proportion of studies spend the majority of their funds on system construction. Shifting the allocation of resources from system development to actual research would contribute to better research developments.

To date, the results have mainly been presented as announced at the level of individual studies. However, conducting international research in a short time period is necessary to understand trends quickly, as we found in the case of COVID-19 research at OHDSI.[11,12] Moreover standardization, the use of published tools, and a jointly available implementation environment are also essential understanding global health trends. Although traditional research methods also have benefits, new potential methodologies must be sought, and the frameworks of research groups must be crossed to offer valuable research for the betterment of society. RDM is a necessary process for deriving correct results. As the research landscape changes over time, it becomes necessary to accept new ideas. A mindset that manages research data in line with the digital transformation is necessary for the publication of accurate results and protection of the researchers, since falsification and fabrication of data are common problems. Therefore, we would like to disseminate methodologies that are easy to incorporate and that promote research data utilization.

Glossary

CDISC: Clinical Data Interchange Standards Consortium; name of the organization that develops global industry standards for clinical trial data and provides standards for each research process.

RWD: Real-world data; data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.

HL7: Health Level 7; not-for-profit, ANSI-accredited standards developing organization dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services.

COVID-19: coronavirus disease 2019; infectious disease caused by the SARS-CoV-2 virus.

GakuNin RDM: Academic Access Management Federation in Japan (nickname: GakuNin) Research Data Management (RDM); designed by building extensions to the Open Science Framework, an open-source software developed by the US Center for Open Science, to expand its services.

JAIRO Cloud: Japanese Institutional Repositories Online (JAIRO) Cloud; an academic repository system provided by National Institute of Informatics (NII) in Japan.

CiNii: Citation Information by NII; a set of databases operated by the NII.

CDISC: Clinical Data Interchange Standards Consortium; name of the organization that develops global industry standards for clinical trial data and provides standards for each research process.

HL7 FHIR: Health Level 7 Fast Health Interoperability Resources; a standard for health care data exchange, published by HL7.

OHDSI: Observational Health Data Sciences and Informatics; a multi-stakeholder, interdisciplinary collaborative aimed at realizing the value of health data through large-scale analytics.

Acknowledgements

The research was supported by ROIS NII Open Collaborative Research 2022–(22S0102), JSPS KAKENHI Grant Number JP22K12905 and AMED under Grant Number JP22 mk0101191 (to S.U.) and JP22ek0109478 (to S.U.), Moonshot R&D Grant Number JPMJMS2021 (to Y.K.).

Competing Interests

The authors have no competing interests to declare.

Author Contributions

Satoshi Ueno and Yusuke Komiyama contributed equally to this work.

References

1. **Ministry of Health, Labour and Welfare.** Report on the actual conditions related to the standardization of medical information systems in Japan. Nippon niokeru iryojohoshisutemu no hyoujunka nikakawaru jittaichosakenkyuhoukokusho. Article in Japanese. Published March, 2020. Accessed April 27, 2022. https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryou/johoka/0000214316.html.

2. **Ministry of Health, Labour and Welfare.** Current status of informatization in the medical field. Iryobunya no johoka no genjo. Article in Japanese. Accessed April 27, 2022. https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryou/johoka/index.html.
3. **Ministry of Health, Labour and Welfare.** Medical information coordination network support navigation. Iryojohorenkei nettowaku shien-nabi. Hyojun. Article in Japanese. Accessed April 27, 2022. https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryou/johoka/renkei-support.html.
4. **Clinical Data Interchange Standards Consortium.** CDISC RWD Connect Report. Accessed April 27, 2022. https://www.cdisc.org/standards/real-world-data
5. **Facile R, Muhlbradt E, Gong M,** et al. Use of clinical data interchange standards consortium (CDISC) standards for real-world data: expert perspectives from a qualitative delphi survey. *JMIR Med Inform.* 2022; 10(1): e30363. DOI: https://doi.org/10.2196/30363
6. **Health Level 7 Fast Healthcare Interoperability Resources.** Accessed April 27, 2022. https://www.hl7.org/fhir/
7. **Clinical Data Interchange Standards Consortium.** Standards. Accessed April 27, 2022. https://www.cdisc.org/standards
8. **Observational Health Data Sciences and Informatics.** Data standardization. Accessed April 27, 2022. https://www.ohdsi.org/data-standardization/
9. **Clinical Data Interchange Standards Consortium.** FHIR to CDISC joint mapping implementation guide v1.0. Accessed April 27, 2022. https://www.cdisc.org/standards/real-world-data/fhir-cdisc-joint-mapping-implementation-guide-v1-0
10. **Observational Health Data Sciences and Informatics.** Hospitalized COVID-19 patients shown to be younger, healthier than influenza patients per recent global observational health study. Accessed April 27, 2022. https://www.ohdsi.org/covid19-characteristics-naturecomms-study/
11. **Observational Health Data Sciences and Informatics.** Software tools. Accessed April 27, 2022. https://www.ohdsi.org/software-tools/
12. **Burn, E, You, SC, Sena, AG,** et al. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. *Nat Commun* 11, 5009 (2020). Accessed April 27, 2022. https://doi.org/10.1038/s41467-020-18849-z
13. **US Food and Drug Administration.** Study data standards resources. FDA Data Standards Catalog v8.0. Published February 16, 2022. Accessed April 27, 2022. https://www.fda.gov/industry/fda-data-standards-advisory-board/study-data-standards-resources
14. **Pharmaceuticals and Medical Devices Agency.** New drug review with electronic data. Data Standards Catalog (2021-12-15). Accessed April 27, 2022. https://www.pmda.go.jp/english/review-services/reviews/0002.html
15. **Clinical Data Interchange Standards Consortium.** CDASHIG v2.2. (28 September 2021). Accessed April 27, 2022. https://www.cdisc.org/standards/foundational/cdash
16. **Clinical Data Interchange Standards Consortium.** COVID-19 therapeutic areas user guide v1.0. Accessed April 27, 2022. https://www.cdisc.org/standards/therapeutic-areas/covid-19
17. **Komiyama Y, Yamaji K.** Nationwide research data management service of Japan in the open science era. *IEEE*, 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI) 129–133; 2017. DOI: https://doi.org/10.1109/IIAI-AAI.2017.144
18. **Foster ED, Deardorff A.** Open science framework (OSF). *JMLA.* 2017; 105(2): 38. DOI: https://doi.org/10.5195/jmla.2017.88
19. **Yamaji K, Kataoka T, Nakamura M, Orawiwattanakul T, Sonehara N.** Attribute aggregating system for shibboleth based access management federation. 10th IEEE/IPSJ International Symposium on Applications and the Internet 281–284; 2010. DOI: https://doi.org/10.1109/SAINT.2010.14
20. **Kurimoto T,** et al. SINET5: A low-latency and high-bandwidth backbone network for SDN/NFV era. IEEE International Conference on Communications (ICC) 1–7; 2017. DOI: https://doi.org/10.1109/ICC.2017.7996843

How to cite this article: Satoshi Ueno, Yusuke Komiyama, Mariko Doi, Keika Hoshi. The Need for Data Standardization and Research Data Management Infrastructure to Promote the Use of Real-World Data. *Journal of the Society for Clinical Data Management*. 2024; 4(1): 6, pp. 1–9. DOI: https://doi.org/10.47912/jscdm.210

Submitted: 06 May 2022 **Accepted:** 29 November 2023 **Published:** 23 May 2024

Journal of the Society for Clinical Data Management is a peer-reviewed open access journal published by Society for Clinical Data Management.

OPEN ACCESS