NEST
NATIONAL EVALUATION SYSTEM FOR HEALTH TECHNOLOGY

NESTcc Data Quality Framework–*A Practical Guide to RWE for Medical Devices*

A Report of the Data Quality Subcommittee of the NEST Coordinating Center – An initiative of MDIC

National Evaluation System for health Technology Coordinating Center (NESTcc) Data Quality Framework

Subcommittee Members:

- *Adam Atherton, MS, PE, RAC, Independent Consultant*
- *Jeffrey Brown, PhD, TriNetX*
- *Jianxiong (George) Chu, PhD, Edwards Lifesciences*
- *Paul Coplan, ScD, MBA, MPH, Johnson & Johnson (co-chair)*
- *Adam Donat, JD, MS, Quest Diagnostics (co-chair)*
- *Nicolle Gatto, PhD, MPH, FISPE, Aetion*
- *Frederick Masoudi, MD, MSPH, Ascension*
- *Drew Nelson, BSN, CCRN, Medtronic*
- *Pamela Nesbitt, MS, Microsoft*
- *Steven Nichols, GE Healthcare*
- *Joseph S. Ross, MD, MHS, Yale University*
- *Art Sedrakyan, MD, PhD, Weill Cornell Medicine*
- *James Tcheng, MD, Duke University Health System*

Additional Contributors:

- *Sharon E. Davis, PhD, MS, Vanderbilt University Medical Center**
- *Doug Fridsma, PhD, Formerly Datavant*
- *Jordan Hirsch, MHA; Formerly MDIC/NESTcc***
- *Michael E. Matheny, MD, MS, MPH, Vanderbilt University Medical Center**
- *Panagiotis Mavros, PhD, CERobs Consulting LLC**
- *Mwanatumu Mbwana, PhD, RAC (US/EU), MDIC/NESTcc***
- *Sarah Merlino, MPH, MDIC/NESTcc***
- *Mingkai Peng, Johnson & Johnson*
- *Mary Beth Ritchey, PhD, FISPE, CERobs Consulting LLC**

** These contributors were compensated for their contributions and editing work on this publication. Their involvement reflects their professional expertise and was provided under a contractual agreement. Their contributions do not reflect the views or opinions of their respective institutions, organizations or employers.*

*** These contributors participated as MDIC/NEST employees.*

For Sub-committee members: Unless otherwise noted, sub-Committee members have voluntarily contributed their personal time to this publication. Their contributions reflect their independent views and experience, free from funding or other benefits that may influence the content of this document. Participation was in a personal capacity and does not reflect view of their respective institutions, organizations or employer.

Names of Committee Members who are federal employees have been temporarily excluded from this draft document. We are awaiting further guidance from the agency regarding the "Pause on Issuing Documents and Public Communication" issued by certain federal agencies.

Funding Disclosures

As part of its commitments outlined in the 2023 Medical Device User Fee Amendments (MDUFA V), the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) provides User Fee funds to the National Evaluation System for health Technology Coordinating Center (NESTcc) in the Medical Device Innovation Consortium (MDIC) to: i) support the development of RWD resources to facilitate appropriate access for research studies; ii) convene experts to develop best practices and, advance innovative methodology approaches with respect to RWE development and analysis. Funding for this work was made in part possible by FDA of HHS from industry user fees administered through a financial assistance award (FAIN# U01FD006292) to the MDIC. The contents are those of the author(s) and/or presenter(s) and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government.

Table of Contents

PREFACE

In 2012, FDA announced its vision for a medical device program to "quickly identify problematic devices, accurately and transparently characterize and disseminate information about device performance in clinical practice, and efficiently generate data to support premarket clearance or approval of new devices and new uses of currently marketed devices[1]." Soon thereafter, a multi-stakeholder effort began to establish the National Evaluation System for Health Technology (NEST) conducting much of its work acting as a Coordinating Center (NESTcc) bringing diverse stakeholders to the table.

NESTcc was established in 2016 with funding from the United States Food and Drug Administration (FDA) through a U01 Cooperative Agreement funded in part under the Medical Device User Fee and Modernization Act (MDUFA). Per MDUFA IV and MDUFA V commitments, NESTcc operates under the guidance of a Governance Committee to help ensure its work is in the best interest of the entire Medical Device Ecosystem, including health systems, patient groups, industry, clinicians, payers, and regulators. Its work is intended to help solve the unique challenges of using real-world data (RWD) to generate real world-evidence (RWE) in the study of medical devices.

NESTcc aims to support the sustainable generation and use of timely, reliable, and cost-effective RWE throughout the medical device total product lifecycle (TPLC), using high-quality RWD that are analyzed using robust methodological standards. Stakeholders across the medical device ecosystem, stand to benefit from improved use of RWD generated in the course of clinical care and everyday life to produce valid RWE. Opportunities include increased patient awareness of device safety issues, efficient and low-cost evidence generation for regulatory review and reimbursement purposes, and improved patient and provider ability to make care decisions based on robust evidence.

In 2018, NESTcc's Governing Committee commissioned two Subcommittees to develop Frameworks on Data Quality and Research Methods to support the development of high-quality RWE studies of medical devices. The Subcommittees included representatives from health systems, NESTcc Network Collaborators, medical device manufacturers, and the FDA. These original frameworks built upon existing work and utilized members' knowledge from similar initiatives like PCORnet, Sentinel, and MDEpiNet. They aimed to guide medical device ecosystem stakeholders in collaborating with NESTcc to ensure high-quality data and research methodology.

The first versions of both the Research Methods Framework and the Data Quality Framework were released in February 2020. The Data Quality Subcommittee reconvened in late 2020 to begin revisions to this Framework based on stakeholder feedback and lessons learned from 21 NESTcc RWE Test-Cases that were chosen through an Open Call Process. These test-cases explored the feasibility for medical device ecosystem stakeholders to work with RWD sources and NESTcc's initial set of Network Collaborators, and to identify areas where NESTcc could play a role in reducing transaction costs [e.g., contracting, Institutional Review Board (IRB) approvals, data sharing agreements, publication policies]. Descriptions of the 21 NESTcc Test-Cases are available on the NESTcc website[2]. Some test-cases progressed beyond feasibility leading to an FDA-approved label extension[3] [the first using solely

a comparative EHR database RWE study for a label extension approved by FDA's Center for Device and Radiological Health (CDRH)] and postmarketing safety surveillance studies[4].

The test-cases also revealed strengths and limitations of specific data sources and the challenges involved in creating datasets suitable for conducting regulatory-grade research[5-7]. NESTcc has now pivoted from the "test-case" environment to one of "implementation" using what was learned under MDUFA IV. One primary goal of the implementation case projects is the further development of the NEST Mark™ review approach to evaluation of RWD. The NEST Mark approach is designed to help de-risk the use of RWD for supporting regulatory filings using well-defined processes for evaluation of relevance and reliability of RWD specifically to generate RWE. As part of this process, NESTcc applies a systematic and consistent approach to evaluate essential data quality and study design elements from FDA's Guidance Documents and builds on the NEST Frameworks. This leads to a NEST Mark review report enhancing confidence that covered RWD are relevant and reliable to meet scientific and regulatory objectives for medical devices.

These next versions of the Methods and Data Quality Frameworks incorporate NESTcc's knowledge gained from test cases and early NEST Mark implementations, along with the subcommittee's extensive RWD experience. They include a broader range of data sources for quality assessment, consider recent RWD guidance documents, and offer additional RWE examples and best practices. They reflect the evolution in RWE innovation, aiming to provide a comprehensive resource for medical device ecosystem stakeholders.

On behalf of NESTcc, we would like to extend our heartfelt gratitude to each and every one of both our current and past subcommittee members for their incredible dedication and invaluable contributions. Their selfless commitment to creating these frameworks has provided an essential resource for professionals working with RWD and RWE in the medical device ecosystem. The expertise and hard work of the subcommittee will not only advance the field but will also pave the way for improved patient outcomes.

Jesse Berlin, ScD, School of Public Health and Center for Pharmacoepidemiology and Treatment Science, Rutgers University

Paul Coplan, ScD, MBA, FISPE, Johnson & Johnson MedTech Epidemiology & RWD Science

Adam Donat, JD, MS, Quest Diagnostics

Richard Smith, MBA, Senior Vice President, NEST

Jill Dreyfus, PhD, MPH, Senior Director of Evidence Generation, NEST

References

1. Gottlieb S, Shuren JE. Statement from FDA Commissioner Scott Gottlieb, M.D. and Jeff Shuren, M.D., Director of the Center for Devices and Radiological Health, on FDA's updates to Medical Device Safety Action Plan to enhance post-market safety. Food and Drug Administration. Accessed 12/22/2023, 2023. https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-and-jeff-shuren-md-director-center-devices-and-2
2. NEST Test Cases. https://nestcc.org/test-cases/
3. Dhruva SS, Zhang S, Chen J, et al. Safety and Effectiveness of a Catheter With Contact Force and 6-Hole Irrigation for Ablation of Persistent Atrial Fibrillation in Routine Clinical Practice. *JAMA Netw Open*. Aug 1 2022;5(8):e2227134. doi:10.1001/jamanetworkopen.2022.27134
4. Frankenberger EA, Resnic FS, Ssemaganda H, et al. Evaluation of intervertebral body implant performance using active surveillance of electronic health records. *BMJ Surg Interv Health Technol*. 2022;4(1):e000125. doi:10.1136/bmjsit-2021-000125
5. Blake HA, Sharples LD, Harron K, van der Meulen JH, Walker K. Probabilistic linkage without personal information successfully linked national clinical datasets. *J Clin Epidemiol*. Aug 2021;136:136-145. doi:10.1016/j.jclinepi.2021.04.015
6. Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. *Int J Environ Res Public Health*. May 14 2014;11(5):5170-207. doi:10.3390/ijerph110505170
7. Gottlieb S, Shuren JE. Statement from FDA Commissioner Scott Gottlieb, M.D. and Jeff Shuren, M.D., Director of the Center for Devices and Radiological Health, on FDA's updates to Medical Device Safety Action Plan to enhance post-market safety. Food and Drug Administration. Accessed 12/22/2023, 2023. https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-and-jeff-shuren-md-director-center-devices-and-2

1. INTRODUCTION

In 2018, NESTcc's Data Quality Subcommittee was tasked with creating a Data Quality Framework that could be used by all stakeholders across the NESTcc medical device ecosystem. The initial version of that Framework (posted on the NESTcc website in February 2020) laid out the foundation for the capture and use of high-quality data for post-market evaluation of medical devices. Aligned with NESTcc's pragmatic approach to device evaluation, the Framework was grounded in the use of RWD gleaned from the clinical care setting instead of data collected specifically for research or evaluation purposes. It focused on data quality issues for RWD from the EHR rather than other clinically-based data sources such as health insurance claims or registries, which have been addressed elsewhere[4].

This second iteration of the Data Quality Framework adds the detail necessary for it to be further operationalized. It is intended to serve as a guide to Network Collaborators and organizations that wish to collaborate with NESTcc, to ensure the quality of their data related to medical devices. The overarching goal of this Framework is to inform the retrospective capture and use of clinical information as high-quality data to support the generation of RWE, which will ultimately, and most importantly, provide better care to patients. The Framework provides principles, standards, and best practices that can be used for decision-making.

The Data Quality Framework is separated into two primary sections, and this is a significant reframing from the 2020 NEST Data Quality document. Now, the Framework better supports a documentation/reporting checklist and more closely reflects an operational process flow. The two primary sections are as follows:

1. First, Section 3 provides a summarization of the **elements to be considered for reporting** data quality in observational data with an accompanying checklist (see Appendix). It is important to document and characterize the data products that result from the extraction, transformation, load, linkage, and merger activities. These products, reports, and documentation become the foundational information for all downstream users of the data products and are considered fundamental to high-quality data sources. (See Section 3.)
2. Second, Sections 4-7 provide a **detailed description** of the **common processes** that occur during preparation of observational data for analysis. This material is intended to be used as a supporting reference to the primary data quality summary and checklist. The process categories include:
    a. **Database source**: Choosing and using data appropriately first necessitates defining a data source, along with understanding the quality of the data and its suitability for purpose that depends on where the original data were collected and sourced.
    b. **Extraction, transformation, load (ETL)**: ETL is a key process to pull data from RWD sources, applying well-defined rules for extraction, transformation, and loading into a formal, carefully documented data model that can support downstream analytics and applications.
    c. **Linkage of data sources**: Data linkage allows for connecting information about the same person or entity from different sources into a single analytical data set, offering more

information for clinical or research purposes. Line-level linkage of patient information can occur as part of both the transform and load steps of ETL (within an ETL) or between different data repositories that have already undergone ETL.

d. **Data governance**: Involving and engaging stakeholders is critical to good governance for RWD and RWE. Governance ensures stakeholder representation, limits the potential for bias or unethical behaviors, complies with patient data privacy laws, and results in trustworthy findings and conclusions.

Key operations within each category and relationships between categories are presented in ***Figure 1.***

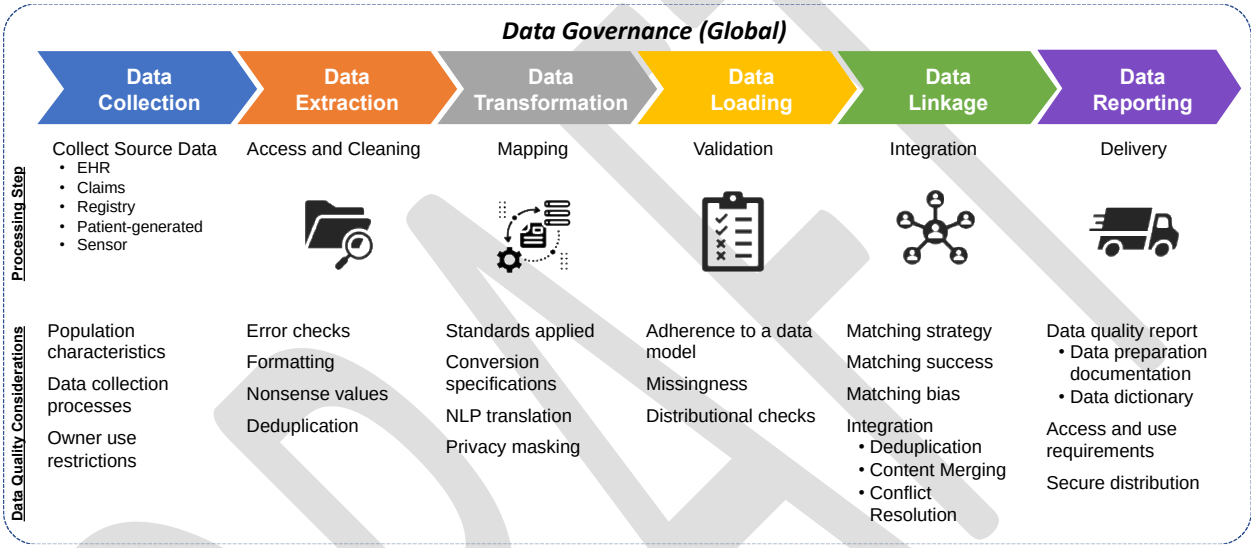

Figure 1: *Overview of the relationships between the sections of this document and the process workflow in which they are developed.*

Important Note: There is a natural overlap between the NEST Data Quality Framework and the NEST Research Methods Framework because specific details of the analysis plan require data transformation and data representation in a form amenable to direct analysis. This document focuses on the overall workflow of data preparation from original source data to characterization of transformed data products at a maturity level where they are ready for data transformation to a final analysis-ready data set as part of an observational study design and analytic plan (which are covered in the Research Methods Framework). In addition, a section on the NEST Data Quality Maturity Model was included in the first version of the NEST Data Quality Framework. No updates to the NEST Data Quality Maturity Model developed in the first Data Quality Framework were made in the second version; thus, the Data Quality Maturity Model is not included in this updated version. While the content of the NEST Data Quality Maturity Model will not be directly included here, many of the concepts are referenced throughout.

2. DEVELOPING THE DATA QUALITY

RWD are defined by the FDA as "data related to patient health status and/or the delivery of health care routinely collected from EHRs, claims and billing data, data from product and disease registries, patient-generated data including home-use settings, and data gathered from other sources that can inform on health status, such as mobile devices[8]." To evaluate the utility of any RWE study, an understanding of the underlying data is necessary. For this reason, data quality should be considered and evaluated throughout data curation– the process of creating, organizing, provisioning, and maintaining data for the use of others.

In general, datasets should be standardized, documented, and prepared for analytic use while avoiding or minimizing modifications to the source data where possible and attempting to support broad applicability and utility of the datasets. Promoting data completeness, quality, fidelity, and availability are necessary components of data curation. This is a complex responsibility with multiple dimensions and interrelated decisions.

Data Quality Dimensions. The field of data quality assessment for observational data is quite robust with a substantial body of work in the data science and informatics community. While there is variability in the terminology used to define the key aspects of data quality, there are several common, cross-cutting concepts that will be used throughout the remainder of the document[6,9] in alignment with the FDA's Data Standards Program Action Plan[10] and guidance on use of RWE[11]. In particular, the following seven key dimensions of data quality should be considered at each step along the pathway to data delivery (***Figure 2***):

1. Completeness: Data are recorded for all relevant events, items, or individuals, and data features are fully populated.
2. Validity: Data as recorded capture intended characteristics.
3. Accuracy: Data minimize errors in recording, coding and transformation.
4. Integrity: Data are designed to protect privacy, adhere to regulations, and minimize bias.
5. Reliability: Data are consistently collected, recorded, and processed.
6. Uniqueness: Data represent unique events, items, or individuals.
7. Timeliness: Data are collected, recorded, reviewed, and made available in a timely manner. Data represents information from critical periods of observation.

Figure 2. *Data quality dimensions to be considered throughout the process of data collection, processing, and delivery.*

These dimensions will inform the data quality assurance activities and resulting documentation products as data are prepared for delivery to analysis teams. Although some dimensions will be more relevant than others at each step, all seven should be considered throughout the process. A data quality report and data dictionary that addresses these dimensions across data collection processes, design decisions, summaries of features characteristics, and restrictions/considerations on use and reporting should accompany any distribution and release of prepared datasets to research teams.

Lastly, no RWD source will realistically meet all data quality metrics ideal for downstream use cases. The intent in providing data quality reporting is to support assessments of the applicability to a proposed use case and the robustness of the results of using the RWD. Sensitivity analyses or bias impact assessments may be conducted as part of study methodology to evaluate the impact of low data quality on some dimensions of the study results and conclusions, which are covered in the NESTcc Methodology Framework.

In this section, we recommend a data quality reporting structure that facilitates communication of RWD quality evaluations and limitations for analysis teams to consider when conducting RWE studies. This report should document the impact of all the decisions, actions, and requirements influencing the data curation process. Such data quality reports provide the basis for future users to know how the data were generated and prepared and to understand limitations, potential biases, and restrictions or ethical considerations of use. A thorough report also serves as the key point of communication with future potential users, e.g., in discussions between a medical device sponsor and the FDA on whether the data quality of a RWD source is adequate for regulatory decision-making.

The remainder of this section describes recommendations for each of five components of a data quality report for RWD aimed at communicating the seven key quality dimensions for future use of a RWD resource as summarized in ***Table 1***. Each section includes a list of specific information and questions to be answered in the report. **The Appendix provides a useful checklist to help ensure all recommended components for the data quality report are included.**

Table 1. *Categories of Data Operations to Prepare Observational Data for Reuse cross-references with Data Quality Dimensions.*

	Data quality dimensions						
Data quality report section	**Completeness**	**Integrity**	**Validity**	**Accuracy**	**Reliability**	**Uniqueness**	**Timeliness**
Overview		ü					
Sources	ü	ü	ü	ü	ü		ü
ETL	ü	ü	ü	ü	ü	ü	
Linkage	ü	ü	ü	ü			
Governance		ü					

We note that, within the integrity dimension, minimizing bias in the collection, preparation, and transformation of RWD is **critical** to usability and utility of future RWE studies. Bias in RWD may be introduced or avoided at any step in the data curation process. Bias in RWD may present across demographic groups, clinical status, geography, or time[12,13]. The data quality report should include a comprehensive effort to acknowledge potential sources of bias and document efforts to minimize bias

in the distributed data. We, therefore, have incorporated the consideration of bias within components of the data quality report rather than separate this integral aspect of data quality into a separate section.

2.1 The Dataset Overview Section

The data quality report should start by documenting the organization or individuals responsible for the datasets. This includes their role in developing and documenting the dataset, as well as key contact information for potential users to request access or submit questions. This section directly addresses the integrity dimension of data quality by ensuring transparency and communication.

Specific information to be included in the Data Overview section of the data quality report are described in Table 2. For further details on completing this section, see the Appendix.

*Table 2. Specific information to include in the Data Overview section of the Data Quality Report**

Name and contact information of the organization responsible for the data and information in the data quality report
Description of the role of the data distributor/owner in data collection and processing
Versioning and updating information

**For a more detailed listing, see the Data Quality Reporting Checklist in the Appendix.*

2.2 The Data Sources Section

The way in which data are originally collected can have a large impact on the data quality. The Data Sources section of a data quality report considers in what situation and for what purposes the data were originally collected, and the features of the data's validity, reliability, accuracy, and completeness. This section should document the organization(s)/institutions(s) at which data were collected, provide context surrounding data collection, give an overview of data collection practices, and provide an overview of the population represented in the published dataset. Portions of the data provided by patients' self-report should be distinguished from data generated from observations by providers, nurses, or other healthcare personnel.

Missingness associated data source provisioning should be clearly documented, including known reasons for lack of complete data collection and measurable impact on data completeness. Examples of data completeness limitations associated with data sources may be related to:

- collection practices (e.g., healthcare utilization and laboratory tests outside of a healthcare system's facilities are not captured)
- lack of full integration of the data system within healthcare institutions (e.g., surgical EHR tools and genomic data reporting)
- temporal system-level changes (e.g., staged implementation of survey instruments across institutions)
- incomplete data ascertainment (e.g., censoring at the time of data extraction)
- institution or clinic level training or workflow variations
- patients' willingness to divulge information.

Any temporal, geographic, or subpopulation variations in data completeness should be documented along with information on efforts to understand this variation. In the case of missing values associated with loss-to-follow-up, if possible, the report should characterize the proportion of patients without data covering the outcome ascertainment period and whether the data curation process attempted to correct for loss-to-follow up through linkage with other data sources. For example, some EHR-based datasets may be linked to claims data to capture hospital readmissions or emergency department visits in other healthcare systems. In such cases, any improvement in outcome coverage should be noted.

Any source-level practices and policies aimed at protecting patient privacy should be documented and compliance practices should reference relevant local, state, or federal regulations. The data quality report should note whether observations have been filtered, sampled, or restricted to exclude records for sensitive or protected populations. Any patient identifiers available in the dataset should be highlighted, as well as whether those identifiers can be used to integrate the dataset with additional data resources.

Table 3 provides a summary of information to document information regarding the data sources in the Data Quality Report. For further details on completing this section, see Appendix.

Table 3. Specific information to include for each data source in the Data Sources section of the Data Quality Report*

Summary information – *to give users an understanding of the data structure and volume* • Source data version or date of extraction • Type of data source (e.g., EHR, claims, survey, etc.) • Time frame of data collection • Location of data collection and organizational information • Level of observation (e.g., patient or encounter) • Number of unique observations and unique patients
Inclusion and exclusion criteria applied at the source level – *to give users a sense of the completeness of the datasets*
Basic demographic summary – *to give users a sense of population representativeness and generalizability*
Method of data collection – *to give users a sense of reliability and accuracy* • Context/Setting (e.g., routine care EHR entry or administrative claims, prospective registry, retrospective participation, or registry chart abstraction, abstraction) • Data provider (i.e., what data were provided by patient self-report vs provider collection) • Survey questionnaires used and accompanying source citations • Description of data collection/assurance practices
Completeness, consistency, & representativeness considerations • Proportion of represented population captured Note: A data source may only include patients seeking care at a given institution, and this may be influenced by various factors (e.g., geographic region; institutional, state, and national policies; perceived reputation for the care being delivered). Where possible, this section should document the proportion of all patients treated with the device that are expected to be represented in the dataset and note that users should consider the impacts of excluding those patients receiving the device at smaller institutions.

- Known patterns of data or study subject missingness not at random

Privacy considerations

- Documentation of any filtering, sampling, aggregation, or restrictions used to exclude records in the interest of protecting privacy
- Documentation of privacy protections required for use
- Explanation of informed consent policies or waivers obtained

**For a more detailed listing, see the Data Quality Reporting Checklist in the Appendix.*

2.3 The Data Extraction, Transformation, and Loading Section

Data processing workflows, as described in more detail in Section 5, involve almost every dimension of data quality and should be thoroughly documented. The documentation should include an overview of the data model that was implemented, along with key variable definitions, conventions, and standards that were used in populating the data. This information can be provided by a third party if a pre-existing data model is used or should be developed *de novo* if the data model is internally developed. The ETL process that describes the implementation of the data model specification should also be documented along with any nonconformance.

Key data quality assessments must be performed and documented at each stage of the ETL process by the team providing that service. This documentation include assurance that all source data are accounted for (added, changed, not included, etc.); results of consistency and reasonability checks; and impacts of any data cleaning steps for removal of infeasible values (e.g., number of observations impacted overall and across key subgroups).

Missingness may result from data processing decisions through removal of infeasible values (e.g., out of physiologic range), incomplete datasets [e.g., data that do not contain all required elements to conform to the selected common data model (CDM)], incorrect coding practices (e.g., non-standard codes that cannot be translated into standard terminologies), or privacy protection practices. Data processing details should describe sources of **missingness** and whether the data curation process injected missing values.The data quality report should also include the proportion of records with incomplete data, the degree of compliance with required features of the selected CDM, and a summary of missingness of key data elements such as outcome variables.

There are several widely used public CDMs[14], including the Observational Medical Outcomes Partnership (OMOP) CDM[15], the Sentinel CDM[16], the Patient-Centered Outcomes Research Network (PCORNet) CDM, and the Informatics for Integrating Biology and the Bedside (i2b2) CDM[17,18]. In addition to general documentation, we recommend that any data product using one of these CDMs also report on data quality findings and remediations from the community-provided data quality assessment tools[19,20].

Temporal complexities of data collection and processing should also be noted. For example, for data collected periodically over time, processes for aligning newly accrued outcome information with previously extracted exposure information should be described. In addition, for those cases in which **multiple data sources were integrated**, this section of the Data Quality Report may reference the data quality documents associated with the individual data sources. The methods of integration should be documented here or in the subsequent section on linkage.

The data quality report should also specify **privacy preserving transformations** made to the data to comply with local, state, and federal regulations as well as those enforced by the data owner. These adjustments may include collapsing of extreme values or aggregation of rare categories. The specificity of some values may also be restricted (e.g., reporting only year rather than specific dates). For datasets that include unstructured clinical narrative text, any masking of identifying information should be documented. Any additional effort to maintain realistic patterns in the data while ensuring de-identification should also be described. For example, algorithmic shifting of dates within an EHR can protect patient privacy while maintaining temporal relationships within clinical histories[21]. For each such privacy-preserving step in the data processing pipeline, the proportion of observations affected should be documented.

Summary documentation of the ETL process should also include a detailed data dictionary. For each variable, the dictionary should provide field name, definition, and format (i.e., integer, category, free text); specify mapping to standard terminology and alignment with a CDM attribute; define non-standard categories; document specific data cleaning and transformation steps impacting values; and report the proportion of missing values.

Table 4 provides a summary of information to document the ETL process in the Data Quality Report. For further details on completing this section, see the Appendix.

Table 4. Specific information to include in the Data Quality Report regarding ETL processes*

Measurement bias
Clearly defined data model • Documentation of the full specification, conventions, and definitions of the DM • Documentation of the process flow and transformation of data elements from the source into the DM
Documentation of the implementation of data model • Documentation of non-standard fields and elements • Documentation of any limitations or non-conformance of the DM
Extraction data quality checks • Assessment of data content (count, values, links between data elements) before and after extraction into data staging to ensure fidelity • Documentation of limitations in data extraction process
Transformation data quality checks • Documentation of which processes result in lost information, justify transformation decision • Documentation and assessment of generation of missing values if out of range, invalid, or unable to normalize • Documentation and characterization of imputation of missing data if performed • Documentation of data validity checks to evaluate data redundancy, duplication, mis-matches in table, and value connections • Account for all rows from source that are transformed (changed, removed, expanded, etc.) • Documentation of privacy preservation efforts – [e.g., data changes such as aggregation or removal of rare values or date shifting to remove patient protected health information(PHI)]
Load data quality checks

• Assessment of proportion of data that is mapped to controlled vocabularies • Documentation of final missingness of values • If merging data sources, assessment of data redundancy or duplications
ETL global - bias assessments • Frequency – document whether some measurements or tests were differentially ordered across patient subgroups • Precision and accuracy – document any measures that may have varying precision or accuracy in select patient subgroups (i.e., racial bias in pulse oximetry[22]) or over time (i.e., lab values recorded on upgraded equipment)

**For a more detailed listing, see the Data Quality Reporting Checklist in the Appendix.*

2.4 The Data Linkage Section

The process by which data were integrated across disparate sources (see Section 6) should be clearly described to provide insight into data completeness, validity, accuracy, and integrity. The data quality report should describe why line-level (i.e., matching records associated with the same individual), aggregate, or spatial linkage was undertaken and any known limitations of population overlap between data sources. The report should also detail the level of linkage (e.g., patient, institution, geographic area), the algorithm employed, and the identifiers used. If data were spatially linked, information on geocoding practices and success should be provided. The evaluation process for determining the accuracy of linkages should be described and metrics of success reported, including match rate and linkage accuracy. To provide insight into potential bias, these metrics should be reported overall and by key subgroups.

Privacy preserving practices associated with the linkage process should be noted, including whether data were de-identified after linkage and an evaluation of whether linkage created additional risk to patient confidentiality. For example, data distributed without patient identifiers that are linked to multiple geographic features may create opportunities for re-identification. If any post-linkage data masking or aggregation was conducted to address linkage-related privacy risks, these practices should be acknowledged in the report. This section should also clarify whether any identifiers remaining in the released dataset are permissible to be used for further linkage to additional datasets by analytic teams.

Prior work has provided guidance and checklists for reporting data linkage in pharmacoepidemiologic studies[23].

Table 5 provides a summary of information to document data linkage in the Data Quality Report. For further details on completing this section, see Appendix.

Table 5. Specific information to include in the Data Linkage section of the Data Quality Report*

Purpose of linking datasets and any patient consent required for such linkage
Detail the featured used for linkage and all steps of the linkage algorithm
Report match rate and linkage accuracy statistics, overall and by key subgroups
Provide an indication of linkage confidence within resulting linked dataset
Document overlap between populations represented in linked datasets and variation in linkage success rate and accuracy across patient subgroups

**For a more detailed listing, see the Data Quality Reporting Checklist in the Appendix.*

2.5. The Governance Considerations Section

The Governance section, focusing specifically on data integrity, captures key considerations described in detail in Section 7 and describes the context of allowable uses and restrictions governing the use of the data. Governance information includes information on dataset ownership and oversight of the data curation process; relevant shareholders of the available dataset and previously integrated data sources; applicable organizational and governmental policies/regulations; permitted use of the dataset for research; and instructions for requesting access. This section should also note expectations of users in terms of human subjects training and approval prior to access, as well as requirements surrounding data storage if access is granted.

Table 6 provides a summary of information to document data governance. For further details on completing this section, see the Appendix.

Table 6. Specific information to include in the Data Governance section of the Data Quality Report*

Citation guidance for future data users to include when reporting on RWE studies generated from the data
Document federal, state, local, and foreign laws and regulations applicable to use, formatting, and distribution of datasets, as well as organizational policies that may further limit or influence use
Identify and engage key shareholders, including patients, clinicians, and other health system and organization staff, and regulators (including payors) in RWD/RWE project development and execution
Specify requirements to ensure all shareholders understand and adhere to ethical standards for responsible conduct of research
Engage patients, clinical providers, and other relevant stakeholders in data definition and aggregation discussions to ensure categories, groupings, and labels align with stakeholder priorities and cultural norms
Establish a lead institution to oversee and coordinate preparation of RWD, including setting minimum expectations for contributing organizations' data infrastructure and technical maturity[5,24-26]
Define data stewardship standards to ensure organizations take responsibility for the management, storage, and use of the organization's RWD
Publicly disclose sources of funding, participating organizations, and both organizational and individual potential conflicts of interest
Store data securely, minimizing risk of further distribution and use without the appropriate permissions/agreements; data retention is described
Define expectations for information security infrastructure at all participating organizations, including establishing certification and continuing review requirements
Document clear criteria by which requests for RWD for RWE are considered, including mechanisms for determining appropriate disclosures, including preclusion of access for non-scientific purposes, such as in pursuit of litigation, as well as qualifications for data security and storage

**For a more detailed listing, see the Data Quality Reporting Checklist in the Appendix.*

3. UNDERSTANDING THE DATA SOURCES CATERGORIZATION PROCESS

Data sources considered for RWE studies intended to inform regulatory decisions about medical devices should be well characterized. This characterization includes a detailed description of the data source, information on the timeframe and latency of the data, technical and privacy aspects, and quality assurance-related information. For the purposes of this document, we have separated out discussions of data source collection and data quality assessment of source data from a discussion of data transformation processes, including ETL processes that are ubiquitous for observational data. This is somewhat arbitrary, because for any data beyond the point of initial generation, there are usually a variety of data transformation steps that take place to represent the data, and then that transformed data itself becomes the 'source' data for another process. However, this separation allows for clear discussion of different important elements of data quality.

3.1. Data Source Definition & Characterization

It is important to document as many characteristics and features of the source data as possible because the context of data collection can be lost during reuse in situations in which the data are neither reliable nor relevant (see Sidebar).

For the purposes of evaluating the quality of a data source, there are several categories of information that should be present; the term "metadata" is used to refer to high-level characteristics that describe a data source.

Metadata include information on data types, healthcare setting(s), purpose of data collection, how data are accessed, how data were obtained, any data transformations (including obfuscation), the full data dictionary, device information (types of identifiers and indication for use), the completeness of fields one would expect to be complete for all patients and needed for most downstream uses (e.g., age, sex, device identifier, production identifier), the timeframe and latency of the data, and key technical and privacy-related information. After initially summarized, these characteristics can be updated as the additional information accrues and underlying data are updated. High-quality data are (to the greatest extent possible) consistently collected and recorded, complete, and accurate.

Choice and use of data require understanding the limitations of the data source(s) and anticipating where shortcomings of the data may limit the questions that can be addressed. Preassessment of data reliability based on the characteristics allows researchers to quickly narrow available potential data sources to those most likely to meet the needs of the study and focus resources on detailed feasibility assessment to identify the best fit data source for the research question.

It is important to document and characterize the overall type and category of data being collected, as these overarching features impact all considerations of policy, governance, reuse, data transformation, data quality assessments, and the realm of applicable downstream uses. While a full discussion is out of scope, main categories of data used in observational healthcare data science will be noted here.

Reliability and Relevancy

Generating valid and trustworthy evidence from RWD requires data that are both reliable and relevant. *Reliability* refers to consistent practices of data collection, recording, and processing. Such consistency has implications for other dimensions of data quality, including validity, accuracy, completeness, and timeliness. Data reliability should be carefully characterized, assessed, and documented along with information on traceability and provenance[27]. Reliability frequently varies within different sub-domains of the source data.

Relevancy refers to the availability of key data needed to define exposures, outcomes and covariates, as well as representativeness and adequate sample size. Relevancy cannot be directly assessed when considering data quality for data sources without a specific study and a clear and specific research question. However, users of the data are frequently distant from data collection, particularly with regards to the sociotechnical and cultural details of the environment in which the data were generated. For this reason, it is critical to consider general categories of use when conducting data source quality assessment and to provide guidance (both recommendations and cautions) to data users with regards to which types of use cases may reasonably be performed on the data.

3.2 Categories & Setting of Observational Data

3.2.1 Publicly Available Data

These data vary the most widely of any data source category in this description, but the defining characteristic is the fact that the data are available to the public, do not contain any individual patient identifiable information, and in almost all cases, are not able to be linked to individual patient records in other data sets. Examples of these data sets are the FDA MAUDE medical device adverse event reporting data, which include reports of adverse events from institutions and providers. This data set collects a specific set of data elements and has a fully specified data dictionary but a high degree of variability in the *quality* of collected information. Another example of a publicly available data set is the FDA Unique Device Identification (GUDID) database, which has highly normalized data required from medical device manufacturers and is maintained by the National Library of Medicine and FDA[28]. Social media data are another large source of relevant information, where patients discuss their lives and circumstances online. Use of social media data sources in studies may result in significant biases because of how they are generated and due to substantial limitations in the accuracy of linkage to identifiable patient data.

3.2.2 Electronic Health Record (EHR)Data

EHRs are a commonly used source of observational healthcare data, particularly since the HITECH Act and the Office of the National Coordinator's initiatives resulted in wide adoption of EHRs over the last decade[29]. These data sources are very large and diverse and have highly variable data quality within data sub-domains depending on the needs of the organization, the size and technical sophistication of the organization, and the prioritization of resources within the healthcare system for data collection.

There are now a small number of widely used commercial vendors of EHRs in the United States, some providing access to large EHR databases, including aggregated EHR databases of many healthcare systems that cover 100 million+ individuals. Large US governmental EHR databases also exist (as do a few open-source vendors mostly used internationally and by the federal public health system). All EHR vendors in the United States have requirements for data collection, use of controlled terminologies, and key operational functions to be able to meet certification by ONCHIT[30]. These include requirements for UDI device data collection and tracking. For the most part, documentation of the data models used by EHR vendors are well-specified, documented, and have extensive use of controlled terminologies. Individual organizational implementations still vary significantly, and thus, documentation of data sources, implementation strategies and policies, and data quality assessments of source data collected, remain critical for downstream reuse.

A key limitation of many EHR databases is that if patients are treated at more than one healthcare system where each uses a different EHR database, the longitudinal patient journey is hard to ascertain and measure from any one of the EHR databases.

3.2.3 Administrative Claims Data

Administrative claims data are collected by healthcare organizations and normalized into structured data which are then submitted to third party payors specifically for the purpose of documentation of care and reimbursement. For this reason, these data are fairly robust in their reliability, completeness,

validity, uniqueness, and accuracy. They must adhere to payor requirements for collection, evaluation, and validation in order to be reimbursed. The requirements are associated with legal ramifications, audits, and reviews, including assessments for oversights, errors, and fraud. For this reason, these data sources have highly structured data and low levels of missingness. However, they only collect the specific data needed for reimbursement and do not have a large breadth of information.

Another benefit of these data is that when a payor is covering an individual for their health insurance, all claims are submitted to that payor which provides ascertainment coverage for healthcare received across healthcare systems during the covered time period. Also, administrative claims data frequently include a record of patients' membership in the population covered by the third party payors so that patient censoring due to insurance status ending can be measured.

3.2.4 Clinical Registry Data

Participation as a site in clinical registries (particularly those sponsored by professional associations) drives many of the performance improvement activities of a healthcare organization. Data in registries are captured in a very systematic way, resulting in high reliability. As analyses depend upon high-quality data, registry participation frequently requires local resources specifically assigned to assure data quality; document the degree of participation of personnel from each organization in developing the registry and ongoing involvement in collection and use for analyses. Compared with other sources of RWD, the quality of registry data is high, and missingness is proactively managed. The primary limitation, however, is that devices and data elements beyond those pre-specified by the registry program are not available.

3.2.5 Patient Reported Outcomes

The last category that deserves explicit discussion is the category of patient reported outcomes (PROs), which are usually collected through direct observation, surveys, interviews, healthcare portals, or mobile devices. This information is increasingly directly linked to healthcare delivery and thus has linkage between claims, registry, and/or EHR data. Important aspects of these data include the validity and reliability of the underlying survey, the frequency of survey administration, and longitudinal data completeness.

3.3 Data Collection, Evaluation, & Verification

A data source should include assessment, evaluation, and verification of data collection processes that generate the data. This requires direct access to the data collection process and thus may be inaccessible to users accessing data retrospectively for RWD/RWE studies (although they could request documentation of such processes and any past results). Error mitigation is critical at the time of data collection. Healthcare systems, registry managers, and payers, among others, have oversight responsibilities and auditing processes in place to support the primary reasons for data collection. This can involve any number of processes such as direct observation or data element adjudication and verification. Careful documentation of the data collection process informs present and future determinations of reliability, consistency, and relevancy for given use cases.

3.4 Data Availability Over Time

Use of observational data for medical device analyses frequently requires a nuanced understanding of data collection and data characteristics over time. Each category of source data may have differing levels of resolution, granularity, error, and noise over time. For example, EHR systems provide robust data correlating billable transactions of healthcare (e.g., office visits, diagnostic and therapeutic procedures, prescriptions, hospitalizations) with the time those transactions occur. Assignment of the time of clinical endpoints of interest may be less precise, requiring imputation (e.g., computable phenotypes) or even manual abstraction from other sources. The robustness, reliability, and consistency across data elements regarding time of events (typically expressed as dates and differences between dates) should be explicitly determined and documented with regards to:

- binding - to assure that all events of interest are assigned a date,
- assignment- reflecting whether the assigned date is the best representation of the occurrence of the event, and
- format consistency- to allow computational analysis.

As a standard consistency check, evaluation for timeframe face validity should therefore be included as a component of data curation.

Another aspect of time is how often and under what circumstances the source data elements are collected and under what circumstances they are updated and refreshed in a data warehouse or data environment. These characteristics are critical to understand how delayed data are from the present, and whether downstream use cases could reasonably use them for analyses and applications.

3.5 Dataset Completeness

The completeness of any data set is directly related to the purpose of the data collection, with varying degrees of rigor of collection and verification. In this setting, any data set is likely to have a variable amount of missingness, uncertainty, lack of documentation, infeasible values (out of physiologic/realistic range), and data that do not conform to the data dictionary as specified.

During initial data collection, acquisition, and documentation of source data, it is important to assess for missingness, erroneous values, mis-specification, and other forms of data corruption. As noted above, data quality assessments and remediation should be conducted as close to data collection as possible to optimize data quality. In general, incomplete and erroneous data should be documented and characterized and accompanied by recommendations for managing poor quality data. However, wherever possible, the data should be retained in original form, with data quality, assurance, and management processes performed during the ETL process. Defining the completeness of a data source may be debated, but at minimum should include extensive assessments and documentation of the characteristics of data elements as described elsewhere in this document.

3.6 Use of Data Standards

It is important to ascertain and document use of controlled vocabularies for data elements within each data source. This should include documentation of how the data collection is mapped to controlled vocabularies at the time of data collection or whether they are reviewed and assigned at a

later time. In addition, review of which vocabularies are used and what versions and sources the system collecting the data uses to generate observational data should be documented.

This documentation is an important step in characterizing source data for two primary reasons: 1) understanding of the data being collected is better when it is closer to data collection, and 2) consistency of semantics (i.e., meaning) is critical as data are transformed, linked, and used in downstream applications. These factors are key determinants of the quality and applicability of resulting analyses and insights. A description of each of the individual data elements of the source dataset (i.e., a "data dictionary") should be included, with notations as to how, where, why, and with what fidelity each data element was collected. The degree of adherence to data dictionary definitions for each data source element should be estimated and documented.

4. UNDERSTANDING THE EXTRACTION, TRANSFORMATION, & LOADING (ETL) PROCESS

The ETL process is required to extract data from real-world sources. In this process, the data are collected and aggregated, transformed, and loaded in a data warehouse or server environment where they are represented in a form amenable to downstream use cases and analytics.

It should be noted that for observational data, there are multiple possible approaches to ETL, and these are sometimes conflated during documentation and discussion. The main distinction for these different ETL frames is a tradeoff in how broad or narrow the applicability is for downstream consumption of the data. There are ETL processes that take source data- that is sometimes close to the original source and sometimes a product of a more sweeping ETL process- and transform the data into an analytic data set that is developed to be used for a specific use case. There is also an ETL process that collects and aggregates 'raw' source data together, transforms it, and makes it available within a data model with the goal of supporting a wide array of possible uses cases. *This latter process is the frame of ETL we are considering for this document.* The elements of the process are described here:

Extract: During extraction, raw data are copied or exported from source locations. This phase entails defining the data sources and implementing the necessary technical elements to transfer/copy the data from the point of origin to a location in which the data can be further used. Frequently, key identifiers and linkage between data sources are required to provide data that are adequately aligned and comprehensive.

Transform: After extraction, data are usually placed in a temporary storage which is also called a *staging area* or *data lake*. In this staging area, the raw data undergo processing, linking, normalization, and adaptation. Here, the data are transformed and consolidated for the scope of intended use as determined during initial development and specification of the target data model. This phase can involve some or all of the following tasks:

- Filtering, cleansing, de-duplicating, validating, and authenticating

- Performing calculations, translations, or summarizations based on the raw data. This can include changing row and column headers for consistency, converting units of measurement between metric and imperial units, editing text strings, and more
- Executing mappings or translations of controlled vocabularies to other vocabularies as necessary for standardization and interoperability
- Conducting audits to ensure data quality and compliance
- Removing, encrypting, or protecting data governed by industry or governmental regulators
- Formatting the data into tables or joined tables to match the schema of the target data model

Load: In this last step, the transformed data are moved from the staging area into a target data model. Typically, this involves an initial loading of all data, followed by periodic loading of incremental data changes or full refreshes to erase and replace data in the data storage environment. For most organizations that use ETL, the process is automated, well-defined, continuous, and batch driven. Typically, ETL takes place during off-hours when traffic on the source systems and the data warehouse is at its lowest. Sections 4.1-4.8 provide more detail on each aspect of the ETL process.

4.1. ETL: Extraction

4.1.1. Processes for Data Extraction

The exact processes for data extraction will vary depending on the source system(s). Regardless, the process by which data are extracted from each source system should be clearly documented, along with technical details of interfaces, systems, security, and data quality assurance provisions for that collection. Errors in source data extraction from multiple systems across an organization or over time can include significant missingness that can be challenging to detect without rigorous quality assurance processes. In many cases (e.g., publicly available data, patient reported outcomes, EHR data), application programming interfaces (APIs) that themselves conform to data transmission and representation standards are used to improve interoperability, standardization and the risk of errors[31]. [Note some examples of data transmission and representation standards include the Fast Healthcare Interoperability Resources ® (FHIR) standard or the Health Level 7® version 2 (HL7v2).] Procedures outlining the data extraction process should be maintained and followed so that those performing data extraction do so deliberately and consistently[32].

In addition, documented procedures allow for future users to understand the data extraction process.

4.1.2. Data Subject Matter Expertise

Integrated systems to move and process data from multiple reliable and relevant RWD sources into a single database/data warehouse (distributed or centralized) require involvement of a multidisciplinary team consisting of an **ETL leader**, **data architects**, **data engineers**, **data analysts/biostatisticians/data scientists**, **data source subject matter experts, clinical domain experts, database/warehouse developers** and **database administrators.** Following is a description of each of these key team members' roles**:**

- **ETL Lead:** Usually an experienced software engineer- oversees the whole team by developing and managing the infrastructure to cover the process.
- **Data Architect:** Projects the IT infrastructure for data engineers to develop

- **Data Engineer:** Develops the data pipeline (ETL tools) to build the ecosystem to gain access to the identified RWD sources via interfaces
- **Data Analyst/Biostatistician/Data Scientist/Epidemiologist Team:** Works closely with clinical experts to define all required data elements, data standards and models, outline the transformation process including the mapping from raw to derived variables, and develop and oversee the overall data quality and data assessments
- **Subject Matter Expert:** Provides important context on how the source data are collected, the workflows around the data, as well as conventions on how to link, merge, and align the data for downstream use in a data model
- **Clinical Domain Expert:** Usually a clinician with direct healthcare delivery experience in the specified clinical field (e.g., interventional cardiologist, orthopedic surgeons)- guides the whole ETL team to clearly define specific study objectives, identify the potential RWD sources, assess the reliability and relevance of the data sources, and define essential data variables to be analyzed
- **Database/Warehouse Developer**: Is responsible for the database/warehouse design, build and maintenance. Such person needs knowledge of relational database management systems (e.g., Oracle-Clinical, MediData-RAVE) with extensive work experience in structured query language (SQL)
- **Database Administrator**: Usually an IT specialist with extensive experience in database management- Manages the day-to-day operations of the data set including working with developers to troubleshoot any use issues

4.2. ETL: Transformation

4.2.1. Data Mapping and Normalization

Data mapping. Design of data mapping requires the understanding of both raw data source and target data source formats[15]. Documentation, workflow, and robust source data assessments should be conducted to understand raw data sources, including the tables, fields, and content. In many cases, comprehensiveness of mappings to controlled vocabularies are incomplete or require conversion to another set of vocabularies to conform to the target data model. Mapping from the fields and their contents of raw data source to data elements in the target data model should be clearly specified and documented. Mapping to *standard* vocabularies may be required. Provenance of source data identifiers and values may be stored in the target data model to ensure that data quality assessments extend back to the source, that additional data elements can be integrated more easily, and that information is not lost. Granular data quality assessment that includes evaluation of the scope of the desired data representation in the target data model from raw source data through to the final model can be helpful to identify transformation errors. Overall counts of data elements (patients, encounters, administrative codes, laboratory tests, etc.) in source and target representation should be assessed and verified.

Normalization. For data requiring normalization before loading, the normalization processes should be documented. An example of this could be mapping a proprietary medication controlled vocabulary to an open source, international vocabulary. Once conducted, the normalization process should be evaluated to ensure it performed as planned. It is frequently the case that the business owners of the source and destination vocabularies are changing their products over time, and the organization or

group that manages the crosswalk mappings also changes the elements over time. For this reason, documentation, updates to processes, and quality assurance tests over time are critical.

In addition, an evaluation should be performed to assess whether translation of data to vocabularies for the target data model results in consistent representation from all sources of data (i.e., normalization). For example, was a patient receiving dialysis identified by the SNOMED code for history of or procedure code for dialysis, and was this expressed consistently across contributing data sources included in the target data model? These data quality assessments should be documented so that someone later reviewing the process can see that the evaluation was performed and the result of the evaluation.

4.2.2. Information Transformation – AI/ML/NLP

Artificial intelligence (AI) tools, such as information extraction from unstructured clinical narratives using natural language processing (NLP), phenotyping using machine learning (ML), and image analysis using deep learning are important methods to transform raw information in data sources to useful variables.

Performance evaluation and bias assessment must be conducted on the target data source before AI tools are adopted in the information transformation process[33]. First, output from AI tools must align with the definition of variables in the study protocol and input required by AI tools must be available in the target data sources. Second, accuracy of output from AI tools on the target data source must be evaluated. Representativeness of training data for AI tools in terms of real-world diversity in race, geography, socioeconomic status, medical status, gender, and data collection process should be assessed for generalizability to the target data source[34]. A gold standard test dataset can be curated to evaluate the performance of AI tools through measures such as sensitivity, specificity, positive predictive value (PPV), accuracy and F-measures (for predictive performance). Manual review of AI tools' output allows for calculation of PPV. Subgroup analysis of accuracy helps to identify potential biases, and error analysis helps to find the cause of errors and bias. Third, workflow of AI adoption should be designed with involvement of data users, developers, system administrators and clinical staff. Standard operating procedures (SOP) for error mitigation should be developed, and output should be evaluated by qualified personnel. AI tools may also be used to assist or augment a data abstractor in an information transformation process. Fourth, a SOP of quality control should be developed to ensure consistent performance of AI tools over time and across different hospitals or healthcare systems.

4.2.3 De-identification

Steps should be taken to de-identify PHI. For example, consideration should be given to using an anonymous patient ID and whether any information in the data set would allow for identification (i.e., critical dates). Procedures for the assessment for PHI and subsequent de-identification should be documented.

4.2.4. Handling of Missing Data

Missing data are inevitable, and the percentage of missingness tends to be higher in RWD compared to traditional clinical trials. Missing data should be characterized during the ETL process for RWD as users will need to understand patterns of and reasons for missing values when evaluating a dataset

for appropriateness for RWE studies. However, no patient record should be deleted due to missingness. Instead, indicators for missing values and any known reasons for missingness should be documented. Missing values should not be imputed during the data curation process, leaving this step to individual downstream use cases. This allows each RWE study team to engage domain experts (e.g., clinicians and biostatisticians) for their advice on how to handle missing values in a way that is rigorous and appropriate for each given situation.

When documenting patterns and reasons for missingness, three types of missingness should be considered—missing completely at random, missing at random, and missing not at random[35]. (See Table 7) These categories characterize the underlying reasons for the missing values and whether there's a relationship between the missing values and the other available data.

Table 7. Three types of data missingness[35]

Missingness type	Description
Missing completely at random (MCAR)	In this scenario, the reason for the missing values is unrelated to the data in the dataset; thus, no relationship exists between whether a data point is missing and any values in the data set, missing or observed. In short, the probability for a data point to be missing is completely random and the missing data are just a random subset of the whole dataset.
Missing at random (MAR)	The reason for the missingness can be explained by the observed values on other variables; and thus, given some of the observed data, the missingness is conditionally at random. In another words, if the user can control for the conditional variable(s), they can get a random subset. For example, as compared to younger people, older people are more likely to skip a quality-of-life (QOL) questionnaire which is electronically administrated (online). Missingness in values in the QOL are related to the observed variables of age.
Missing not at random (MNAR, non-ignorable)	In this case, the missingness is not dependent on other variables in the dataset but dependent on the missing value itself. For example, in a longitudinal observational study, some patients who are doing well tend to drop out of the study, or some patients missed the follow-up because the sickness become so severe and thus unable to come back for the clinical visit.

Missingness can also be classified as **observable** and **non-observable**. For example, if some patients are missing a value for a variable for which a value was collectable at the time of care (e.g., missing demographic information), that is observable missingness. Some observable missingness may be intentional, such as systematic exclusion, masking, or filtering resulting from privacy protection policies of the original data source. Non-observable missingness would occur if a patient who did not come back for healthcare visits might have received healthcare somewhere else, in which case there is missing information about that person's health journey that is non-observable. In some situations, such non-observable missingness may be reducible through data linkage across sources.

4.3 ETL: Load

When ready, transformed data are then moved into a target data model. Often disparate data sources can be combined to create a more robust and useful dataset. In these cases, the data sources must first be standardized into a data model so that they are compatible with each other. Historically, many proprietary and variable data models were used to represent observational data, all with a goal of providing a defined set of data elements with a specification on normalization, harmonization, and integration, to support downstream use as pre-defined by that data model.

There are a number of mature CDMs for observational claims and EHR data (e.g., PCORNet, Sentinel CDM, OMOP CDM, i2b2 CDM)). Registries have their own specified data models, such as the American College of Cardiology National Cardiovascular Data Registries or the American Spine Registry. Regardless of the target data model used, information on the data sources, data elements, and transformation and normalization rules should be documented so that someone later reviewing can understand the model used for downstream applications.

In addition, when a CDM is used, validation and consistency checks should be performed to ensure data integrity is maintained in the resulting dataset (e.g. same number of patients, no missing values)[15]. Documentation of these validations and consistency checks should be maintained showing the checks performed and their result. Several observational data models have robust and extensive data quality assessments[36,37].

4.4 ETL End-to-end Testing

Validation and consistency checks should be conducted on key data elements in the analysis dataset. Depending on the dataset, complete validation of all data may not be feasible. In such cases, a risk-based approach, which prioritizes data curation efforts on the data elements most critical to future RWE study requirements, to quality assurance should be taken[32,38]. Plans for validation and consistency checks should be documented, including a description of any risk-based approaches.

Examples of methods useful for validation and consistency checks include:

- Ad hoc reviews
- A manual review of a sample
- A report/query-based review of the full dataset
- An automated report/query run against the full dataset with pre-defined rules and alerts

Procedures governing the ETL process should be planned out and documented before implementation. Pre-defined procedures help ensure consideration is made for maintaining the quality of data throughout the ETL process and that staff implementing ETL steps do so consistently. In addition, these defined procedures allow anyone working with or reviewing the resulting data to understand how it was handled and whether they can rely upon it for analysis and decision-making.

4.5 Process Change Management

Procedures and systems used for collecting, extracting, transforming, and loading data will sometimes undergo updates and change over time. For example, EHR updates may disrupt code mapping processes within an ETL pipeline. Accordingly, situational awareness of changes to sources data and ongoing validation of ETL processes is necessary. Version control documentation should be

maintained. This allows someone later reviewing data handling to understand exactly what procedures were in place at the time that the ETL processes they are reviewing occurred.

Further, when a process change is made, its impact on data quality should be assessed[39]. An example of such an assessment is to retain copies of the data pre- and post-change and perform an analysis of whether the post-change dataset reflects any unanticipated differences when compared to the pre-change dataset.

Instances may also arise where established procedures are not followed or cannot be followed (e.g., unanticipated challenges with the data, staff error)[40]. In cases where these deviations occur, the deviation should be assessed for its impact on data quality. The deviation, assessment, and any resulting actions should be documented to allow someone later reviewing to understand what happened and how the data were affected.

4.6 ETL Audit Trail

Steps that are performed for the ETL processes should be logged within an audit trail. This audit trail should allow someone later reviewing it to trace what actions occurred with the data[39]. Examples of actions or items to document could include what dataset the data of interest were extracted from, where they were extracted to, and results of transformation (the dataset before transformation, what transformation function was performed, and what dataset resulted).

In addition, metadata should be included as part of the audit trail to help fully document each step. Examples of important metadata to capture in an audit trail include date, time, and user[39]. These metadata also allow someone reviewing the audit trail to understand which version of the procedures were in use at the time a certain step took place, as procedures will sometimes be updated and change over time.

4.7 Data Aggregation

When data are aggregated across different sources, steps must be taken to ensure subjects are matched accurately. For example, in some cases a unique subject identifier can be used to link sources. Procedures for data aggregation should be pre-defined and documented. In addition, an assessment of the completeness of the linkage between sources should be conducted and documented.

4.8 Edit and Consistency Checks

A key function of curation is to assure the intrinsic “reasonableness” (face validity) of the data. While the processes for RWD source data capture *should* include edit checks (e.g., ranges compatible with the human experience for age, weight, laboratory values), business rules should be applied to derived analytical datasets to assess (and correct) infeasible data. Types of edit and consistency checks are dependent on the data element and can include (but are not limited to) out-of-range checks (e.g., maximum age of 150 years, minimum serum sodium of 100), data masks (applicable to formatted data such as dates), and consistency checks (e.g., rejection of an ejection fraction <30% with a diagnosis of heart failure with preserved ejection fraction). Evaluations for infeasible data should be listed along with resulting actions taken to modify, update, or otherwise “clean” data values in analytical datasets.

The goal is to ensure that the collection, cleansing, and transfer of the data was conducted reliably and that the data received is of the value the sending party intended. The intent is to ensure reasonableness of data.

4.9 Additional Considerations for the ETL Process

It may be useful to designate EHR datasets as (a) raw, (b) meeting the minimum requirements for the specified CDM, (c) incomplete datasets that require imputation and inference, and (d) cross validated datasets. In other words, a dataset's "completeness" depends on the intended use or desired characteristics. It is likely that expanded secondary use of EHR data will require an iterative process for evaluating and defining dataset representativeness.

Clear descriptions of the data as being in a "raw" state, as unchanged from the source as possible, or "pre-processed" state, where the raw data have been modified through cleaning, labeling, derivation, tokenization, binning, parsing, record removal, imputation, encoding, or otherwise is noteworthy. Retainment of raw data is ideal in most cases to facilitate inquiries and changes to desired processing protocols. Detailed documentation of the processing steps and their sequence, or making available the associated processing software or scripts, are minimal requirements to promote data transparency, while audit and monitoring log features, available as a service from most data pipelining and cloud service providers, offer a straightforward way to capture metadata on changes applied.

Where possible, use of data standards for interoperability that describe both definitions of concepts captured through the data elements and data exchange standards used (particularly if leveraged through ETL and aggregation processes) is encouraged. Potential impact of missingness of data in the analytic dataset (particularly mandatory and conditional data) should be described.

5. UNDERSTANDING THE LINKAGE OF DATA SOURCES PROCESS

Data sources may be linked to provide more information for research or clinical purposes. Data linkage is an important step to bring together information from different sources about the same person (i.e., line-level linkage) or entity (e.g., registries, EHRs, insurance claims, genomic data, patient reported outcomes, wearable data, mortality data). Patient data may also be linked spatially to environmental or sociodemographic data. Some examples of where linkage may be important in RWD/RWE studies include:

- Linking registry data to administrative claims to improve longitudinal follow up. For example, the Vascular Quality Initiative (VQI) registry data were linked with Medicare insurance claims data based on a two-step linkage process using procedure codes, date of surgery, patients' date of birth, gender, and three-digit zip code. The linked data were used to study the long-term patient outcomes after endovascular abdominal aortic aneurysm repair including reintervention and late rupture[41].
- Linking data in EHR databases to databases that have unique device identifying information, such as supply chain databases[42,43].

- Linking different claims databases to increase the sample size but de-duplicate members who may be in more than one insurance plan during the study period.
- Linking the original EHR order for routine clinical workflow and billing with medical images or diagnostic imaging reports by means of the Accession Number. Both the Digital Imaging and Communications in Medicine (DICOM) library and HL7 standards include several other unique identifiers that facilitate linkage for clinical and research purposes[44].

To assure the quality of linked data resources, the methods used for linking data across disparate data sources should be documented and the success of the linkage process should be evaluated. Below we describe aspects of the data linkage process and how they may impact data quality.

5.1 Identifying Information for Linking Records

Multiple potential data elements may be used to link records related to the same patient across distinct data sources. Various combinations of identifying features common to each data source to be linked may be considered for matching. The specificity and number of such data elements will increase the accuracy of linkage. When linking multiple data sources in sequence, a diagram highlighting the common features and their use across data sources may be a useful means to illustrate the linkage approach.

In those cases where patient identifying information (PII), such as name or social security number, is contained in each data source of interest, these features may be directly used to match records. However, the use of PII to link patient information requires patient consent, unless the data use approach and data protections are adequate to obtain a waiver of consent by an IRB. Common key(s) from PII used for data linkage include patient ID, encounter information, device type, device unique identifier, or medications, and must be explicitly specified.

Tokenization. When PII is unavailable or not permissible for direct use in data linkage, privacy preserving record linkage technology (PPRL) may be used to connect patient records across data sources while preventing disclosure of identifying information. PPRL removes PII and converts the information into a series of hashes or tokens that can be used to link records across different datasets. Often, multiple tokens are generated for a record that can be used for different linking strategies. For example, if a record contains the patient name, birthdate and SSN, a token can be created for that patient. Other tokens may be generated from different combinations of PII such as phone number, gender, and first/last name. Multiple available tokens increase the accuracy of the match.

5.2 Linkage Algorithms

Data linkage can be performed using a variety of methods and with varying degrees of confidence. The main categories are deterministic and probabilistic algorithms; the former may or may not allow for approximate matches and the latter may use a variety of underlying modeling methods including machine learning[45]. Both types of algorithms may be used with single or multistep matching. Whether deterministic or probabilistic, the algorithm should also take into account that appropriate record linkages could be one-to-one, one-to-many, or many-to-one.

Deterministic matching. The deterministic approach links records through exact match on a unique key or set of keys that are common across datasets. Keys may include patient/entity identifiers (e.g., one or more personal health identifiers or unique device identifiers) or a set of indirect identifiers (e.g.,

patients' date of birth, providers' identifier, postal code, date of surgery, procedure codes, surgery sites)[46]. The chosen identifier(s) and their formats should be explicitly specified.

Linking attempts result in one or more matches or a non-match. Links may fail due to data type differences, typographical errors, capitalization, character set differences, white space, representation (i.e., ideographic versus phonetic) or prepended/appended characters. Normalization or coercion may be required prior to attempted linkage to ensure reliable matching across systems.

Deterministic matching can be conducted using a single or multiple step process. For the latter, records are matched in a series of progressively less restrictive steps that are composed of fewer key variables or incorporate approximate matching. Approximate matching, also known as fuzzy matching, relies on patterns within the unique keys to match records. This overcomes some of the challenges of deterministic matching by not requiring exact matches on all keys. For example, dates may only be required to match within ±2 days or partial string matches may be permitted. A predetermined threshold, or confidence level, must be attained for a match to be considered valid. This type of matching is subject to false positive and false negative matches due to issues such as typographical errors, system limitations, improper thresholding, or narrow criteria. When multistep approaches are defined and/or approximate matching steps are included, the linked datasets should include a new field indicating the hierarchy of confidence in returned matches, allowing users to conduct sensitivity analyses or establish thresholds of acceptable matching for specific use cases.

Probabilistic matching. Probabilistic linking uses multiple keys, in combination, to identify and evaluate links, and is used when a unique key is not available or is of insufficient quality. In contrast to deterministic matching, the probabilistic approach uses a matching score that represents the likelihood of records belonging to the same individual or entity. The score may be matching weight, predicted probability from a ML model, or the edit distance between two strings (e.g., patients' names)[5]. The process of calculating the matching score should be clearly described. The selection of threshold of matching score to determine the matching status should be justified and validated. Probabilistic matches are subject to false positive and false negative matches. Thus, as with deterministic matching, a new field indicating the confidence level of each match should be reported in the linked dataset to allow users to conduct sensitivity analyses. If multiple matches to the same patient record are above the probabilistic threshold for accepting a match, and if only one-to-one matches are permitted, then the linkage with the highest matching score should be used (in most cases).

5.3 Accuracy Assessment

Data linkage is subject to type 1 and type 2 errors[47]. Type 1 error is the false linkage rate (or false positive rate) and can be calculated as the proportion of different individuals that are erroneously linked among all the linked individuals. The false linkage occurs when the same patient was falsely matched with different patients or different visits. False linkages may be detected based on contradictory or conflicting pieces of information (e.g., date conflicts, patient identifier conflicts).

Type 2 error is the missed linkage rate and reflects the proportion of the records belonging to the same individual that fail to be linked among the linkable individuals. For deterministic linkage, missed linkage could be due to missing values or changes of identifiers over time or across hospitals. For probabilistic linkage, the threshold of matching score would determine the type 1 or type 2 error of

the data linkage process. When a probabilistic approach is used, the analysis plan should include testing the impact of false linkage or missed linkage on the robustness of findings.

Two common quality metrics for data linking include the *match rate* and *link accuracy*. The match rate is the proportion of matches linked (i.e., true positive links divided by the total matches conducted). Link accuracy is the proportion of correct links (i.e., number of true positive links divided by the total links conducted). Some checks of accurate linkage may include whether patients or entities were matched at the same granularity (e.g., visit, time) or falsely matched with different patients/entities. Redundancy or duplication checks are also a good way to identify linkage issues. Identifying contradictory or conflicting pieces of information from a sample of the analytical data set is another way to check linkage quality.

5.4 Post-Linkage Characterization

Following completion of linkage, a set of linked data characterizations should include an evaluation of linkage feasibility and bias, both of which influence the completeness and reliability of any resulting linked datasets. The degree of overlap between patient populations in disparate data sources directly influences the match rate, even when all identifiers are perfectly reported. For example, when linking one database to a claims database, only some of the individuals included in the first database may have healthcare insurance through the claims database provider, so only a subset of the first database would be potentially linkable. An assessment of expected population overlap provides necessary context to the accuracy metrics noted in the prior section. If the populations represented in data sources have substantial overlap, then users may be concerned about completeness and representativeness of the linked data if match rates are low. However, if source populations have few individuals included in both datasets, then a lower match rate may be expected and accepted.

The purpose of the linkage will also influence the interpretation of linkage results. If the goal is to augment existing data with new kinds of data (like registry or EHR data linked to administrative claims data to reduce missing follow up patient records over time), it is important to maximize the proportion of patients or study subjects linked to avoid selection biases due to who is linked and to increase sample size for adequate statistical power. However, when creating a cohort of patients from two different datasets expected to have different patient populations (e.g., two distinct hospital databases), a lower proportion of linked patients is preferable because it represents the patients' experience being captured in more than one of the aggregated databases. Each matched patient may represent a duplicate in the dataset, and so matched records may be used to deduplicate the cohort.

For patients not linked between data sources, reasons and bias in linkage failure should be explored. Records may not be linkable due to differences in data completeness across data sources or changes in keys across sources or over time (e.g., name and address changes). Successful data linkage can be subject to information and selection bias. Given the potential for heterogeneity in population characteristics, clinical practices, and coding across data sources, some subgroups of individuals with a range of characteristics including gender, age, insurance types, ethnicity, deprivation, and health status could be over- or under-represented among records affected by linkage failure or error[47]. For example, rates of patient consent for research participation across subgroups, or patients with certain types of insurance (e.g., Medicaid) may be less linkable and result in information bias. Bias can occur if differential data linkage changes the association between the exposure and outcome of interest.

Comparison of patients' characteristics among linked and unlinked data should be conducted to identify potential sources of bias.

6. UNDERSTANDING THE DATA GOVERNANCE PROCESS

To support the generation of representative, robust, and reliable RWE from RWD that results in trustworthy findings and conclusions, a thorough consideration of governance policies and practices should be undertaken from evaluation of source datasets through data preparation and dissemination for research analyses. In addition, these policies and practices should be transparent to the patients and healthcare systems contributing data as well as to all users and consumers of the data as part of any use, consideration, and interpretation of the RWD/RWE. This section describes the following six aspects of governance that inform data quality expectations and assessment:

- Policy environment
- Stakeholder expectations/engagement
- Organizational transparency and integrity
- Privacy and security considerations
- Patient consent for use of routinely collected data
- Use and access requirements

6.1 Policy Environment

NESTcc is committed to ensuring that the highest scientific and ethical standards are applied when using RWD to generate RWE. Such standards are maintained regardless of whether the RWE is intended for purposes of a sponsor pursuing a first clinical indication approval or clearance for a novel medical product, a supplemental new clinical indication for an already authorized medical product, or effectiveness or safety-focused analyses to satisfy post-approval studies or post-market surveillance requirements. All RWE evaluation activities (e.g., sharing patient data across various data sources) must incorporate patient protections such as ensuring privacy of PHI and complying with applicable local, state, federal, and foreign laws and regulations.

In the United States, the Health Insurance Portability and Accountability Act (HIPAA) is particularly relevant to data quality assessments of RWD sourced from health care organizations[48]. This includes both EHR data and healthcare claims data. Adherence to HIPAA requirements may require limiting the precision of select data elements (e.g., geographic identifiers), masking select values (e.g., categorizing ages over 89 as '90 and older'), and removal of select identifiers. These practices influence data quality through data loss and accuracy of potential linkage to coordinating data sources.

Often, safety evaluations with RWD may necessitate access to identifiers, particularly device serial numbers or patient identifiers. Specific exceptions to the HIPAA Privacy Rule are available for post-marketing surveillance and research through waivers of informed

consent. IRB review may be necessary to evaluate and approve such exceptions[49]. There are a number of best practices developed by observational consortia, including Observational Health Data Sciences and Informatics (OHDSI)[50], FDA Sentinel[51], i2b2[52], and PCORNet[53] that have procedures for adhering to federal policies protecting patient privacy and institutional confidentiality when linking RWD across multiple health systems or other data sources (such as linking registry and claims data)[54].

Governance and policy considerations regarding *patient*-generated data are less well-defined and more context-specific. These data may be collected and maintained by HIPAA-covered entities or by organizations that are *not* subject to HIPAA regulations. In the latter case, the Federal Trade Commission (FTC) may have authority and expectations regarding data sharing, and integration may be highly variable[55].

The use and distribution of RWD is governed by a complex set of local, state, federal, and foreign regulations, in addition to policies data owners themselves may institute. When policies and regulations disagree, statute may determine which policies take precedence, or data managers may choose to adhere to the most restrictive policies by default. Preparation of data for research purposes will need to reconcile these competing demands to promote data quality in terms of both integrity and completeness.

6.2 Stakeholder Expectations and Engagement

Stakeholder involvement and engagement is a critical component of good governance for RWD/RWE. The "Good Governance Standard for Public Services" has described stakeholder engagement as a core value of good governance[56]. Key stakeholders include patients, clinicians, researchers, purchasers, payors, industry, hospitals and health systems, policy makers, and training institutions, as well as governmental agencies including regulators. A critical feature of generating RWE from RWD is that key stakeholders are not only users of this information, but potential aggregators and/or owners of the data being used for this work. As examples, EHR data are observations of patients' experience and clinical situation and care and largely represent the clinical and administrative care teams' reflection of their understanding of the patients' context and care. However, hospitals and health systems (EHR data), payors (claims data), and professional societies and industry (registry data) may all be sources for RWD used for RWE and each might reflect different stakeholder engagement viewpoints on the adequate and effective use of the data. Industry, regulators and other policy makers, and other stakeholders are all critical end users that leverage RWE for decision-making. Stakeholder engagement is necessary throughout the life cycle of evaluation, from study and analysis planning and conduct through dissemination of results.

Indeed, stakeholders have a strong influence on data quality and expectations. For example, professional societies define what and how features are collected in registries, and they influence completeness by determining which sites participate in registries. Health systems make decisions about data capture and formatting strategies with implications across all seven data quality dimensions. Organizations managing common data models, including OHDSI[15], PCORNet[53], i2b2[52], and FDA Sentinel[57], publish explicit policies, best practices,

standards, transformations, and processes for ensuring the quality of data from contributing member organizations.

The following principles can guide all stakeholders (data-owner, stewards, and end users), but particularly those that are sources of RWD such as hospitals, health systems, payors, and other organizations, in forming policies and procedures for RWD/RWE, engaged with NESTcc for the purpose of using RWD to generate RWE.

- Identify and engage key stakeholders, including patients, clinicians, and other health system and organization staff, and regulators (including payors) in RWD/RWE project development and execution
- Ensure that all stakeholders understand and adhere to ethical standards for responsible conduct of research
- Engage patients in the RWD/RWE process and collect consent when applicable
- Engage patients, clinical providers, and other relevant stakeholders in data definition and aggregation discussions to ensure categories, groupings, and labels align with stakeholder priorities, nuances of the various stakeholder experiences, and cultural norms
- Evaluate and document potential sources of bias related to stakeholder involvement from data collection through data delivery for research application

6.3 Organizational Transparency and Integrity

Establishing and assuring high data quality standards requires transparency and integrity from organizations engaged in RWD/RWE, but particularly those that are sources of RWD. All organizations that are collecting, storing, processing, managing, and analyzing RWD should establish clear policies and procedures that promote transparency, foster integrity, control potential conflicts of interest, and minimize and disclose sources of bias. Organizational integrity can be evaluated against maturity models that consider organizational readiness for participation in RWD/RWE studies, evaluating elements of technical infrastructure, workforce readiness, and internal governance practices. To ensure that RWE efforts are robust, reliable, and trustworthy, coordinating centers and all participating organizations should disclose such evaluations prior to participation and regularly thereafter to assess continued compliance and promote transparency.

Specific organizational activities that promote data quality and use of data for RWD/RWE studies are:

- Establish a lead institution to oversee and coordinate preparation of RWD, including setting minimum expectations for contributing organizations' data infrastructure and technical maturity [5,24-26]
- Establish executive leadership groups across participating organizations to establish and publish common RWD/RWE policies and procedures
- Obtain appropriate permissions for data use (e.g., patient consent, where required) and publication (e.g., IRB review and approval)
- Assemble an independent advisory board with responsibility for the organization's local data warehouse and research portfolio, which may include legal counsel to manage liability risk

- Define data stewardship standards to ensure organizations take responsibility for the management, storage, and use of the organization's RWD
- Publicly disclose sources of funding, participating organizations, and potential conflicts of interest at both the organizational and individual levels

6.4 Privacy and Security Considerations

Maintaining data quality of RWD requires ensuring data security and protecting privacy throughout data collection, processing, and sharing. Federal, state, and local regulations, as well as organizational policies may govern and define expectations and limitations of RWD to meet privacy and security requirements. For example, organizations managing CDMs may also specify data privacy infrastructure and protection expectations for participants[50,51,58].

Secure data storage, transmission systems, and analysis environments will require proper access privileges, encryption/decryption, security protocols, and risk mitigation standards. Specific data security policies may be guided by compliance with security requirements for FDA data systems, the Federal Information Security Modernization Act (FISMA), and National Institute of Standards and Technology (NIST)-800 series documents. FISMA compliance requires systems to establish and maintain security programs, conduct annual reviews, and meet NIST standards[59]. The emergence of cloud computing poses additional security considerations for RWD/RWE programs, and these are addressed in part by the NIST-800 documents.

Privacy considerations, including de-identification, privacy-preserving data transformations, and data sharing, may be guided by a combination of regulations (e.g., HIPAA) and participating organizations' policies. Inclusion of PHI and PII in RWD may be restricted by governing privacy policies. Necessary linkages across datasets may be impacted by privacy policies, including PPRL technologies (see Section 6). Evaluating the influence of privacy policies on data quality will need to consider impacts of privacy-preserving data granularity, specificity, and completeness. The potential for introducing bias must also be assessed, particularly when small subgroups are suppressed to protect privacy.

There are examples of computing systems and architecture that allow for secure collaboration and follow the relevant federal and/or international policies. One such system is the MDEpiNet's High-Performance Integrated Virtual Environment (HIVE) which has been approved by the FDA for use within their federal computing infrastructure[60]. This system ensures data security through the combination of encryption and data splitting, while preserving data quality through redundancy and metadata. Evaluation of security infrastructure by third parties may be required to ensure compliance and promote transparency.

6.5 Patient Consent for Use of Routinely Collected Data

Patient consent for the sharing of health data may be implied or explicit depending on jurisdictional or local privacy consent policies, and whether or not the data are being shared for treatment or research purposes. The *opt-out model* is more common in data sharing for treatment purposes, in which choosing to obtain care in a particular setting implies sharing of health data. There are typically controls that allow a patient not to participate in this type of sharing. The *explicit, opt-in model* requires that the patient first acknowledge they agree to share their health data before it is shared. This form of consent is most common in research and obtained through informed consent. Among

other things, informed consent must consider language, literacy, as well as mental or physical limitations that could have an impact on the patients' ability to provide consent. Decisions made based on those limitations could introduce bias into the dataset.

The informed consent process should be described in the study report, and any risks related to limitations or bias introduced through the consent process should be evaluated and described in the study conclusions[61]. The consent policy, process and records should be maintained and available upon request. Depending on the circumstances, records that cannot be traced to informed consent may need to be excluded from the dataset. These should be explicit in the governance of the data. However, routinely collected data are allowed to be used under the opt-out model more frequently, and with waivers of informed consent for their use as described in the updated Common Rule[49].

6.6 Use and Access Requirements

Protecting access to systems that collect, store, and manipulate RWD is an important step to help ensure the quality of a data source. Access to these systems should be limited to only authorized individuals accessing the minimum data necessary for defined research objectives[61]. An organization should consider what systems data will interact with throughout the ETL process and which individuals or roles should have access. In addition, consideration should be given to whether access is needed only at certain stages of the ETL process.

Appropriately protecting access to systems and the data they contain not only helps maintain data integrity but also the privacy and confidentiality of patient information. Systems access procedures should take into consideration and comply with any statutory or regulatory requirements that are relevant to the data source, including HIPAA[62] and the 21st Century Cures Act[63]. The HIPAA Privacy Rule explicitly defines requirements for data sharing and access involving covered entities. When RWE analyses require sharing of PHI, sources of RWD and study partners may enter into a Business Associate Agreement (BAA)[64]. BAAs outline required data privacy and security safeguards, as well as liability policies. In general, shared data should be de-identified to the greatest extent possible given the information needed to address research questions, link datasets, and comply with relevant regulations. Removal of identifying information has direct implications for subsequent data linkage, which may impact completeness and uniqueness. However, in cases where a limited dataset stripped of HIPAA-required identifiers would suffice, a Data Use Agreement (DUA) may be pursued[65]. DUAs outline acceptable use of the shared data and limit further disclosures, allowing data access and sharing while providing a higher degree of privacy protection.

Data access policies of participating organizations, BAAs, and DUAs directly influence the quality of data available for RWE analyses and how such analyses can be conducted. When sharing of sufficient data elements is permitted, analyses on combined and linked multi-institution data on a common secure server may be possible. Alternatively, distributed learning algorithms may be necessary when privacy policies at participating organizations do not permit sharing of data at necessary levels of detail. Federated learning algorithms can support analyses by iteratively sharing data summaries rather than patient-level data[66].

Documentation should be maintained demonstrating that data access policies minimize risks to confidentiality and privacy as described above and as required by specific data sharing agreements.

The level of documentation may vary based on factors such as the data source and specifics of the ETL process used. Examples of documentation that could serve this purpose include a report documenting analysis of access needs, procedures for limiting systems access, and a log of individuals showing which systems they have access to. Regardless of the format, the documentation should allow a reader to assess the completeness of access protections.

In summary, all organizations engaged in RWD/RWE, but particularly those that are sources of RWD, should establish clear policies and procedures to promote transparency, facilitate fair access and responsible use to ensure that RWE efforts are robust, reliable, and trustworthy. Specific organizational activities that promote data quality and use of data for RWD/RWE studies include:

- Documenting clear criteria by which requests for RWD for RWE are considered. Criteria should cover mechanisms for determining appropriate disclosures, including preclusion of access for non-scientific purposes (e.g., in pursuit of litigation), as well as qualifications for data security and storage
- For public databases, disclosing requests for RWD for RWE and evaluating requests by an independent approval panel (and ethics review as needed) whose determinations are also publicly disclosed
- Ensuring appropriate agreements (BAA, DUA) are in place for all datasets, data are de-identified to the greatest extent possible, and patient protections are in place, while still allowing necessary analyses to be pursued

7. CONCLUSION

High-quality RWD are essential for the use of regulatory grade RWE to inform regulatory decision-making, such as labeling extension requests or post-market evaluations of medical devices. In the initial version of the NESTcc Data Quality Framework, we discussed the most salient topics associated with achieving high-quality data, including data governance, characteristics of data, approaches to data capture and transformation, and best practices in data curation. We then synthesized these topics in the NESTcc Data Quality Maturity Model, which enables collaborators to indicate their progress toward achieving the highest quality data.

In Version 2 of the Framework, we have provided detailed practical principles, standards, and best practices that can be used for decision-making related to the quality of RWD sources. We describe seven key dimensions of data quality—completeness, validity, accuracy, integrity, reliability, uniqueness, and timeliness. Section 2 of this Framework serves as a reporting tool for evaluating these dimensions across the data curation process encompassing the collection, transformation, distribution, and protection of RWD to support RWE studies. This reporting tool is further operationalized in an accompanying checklist (see Appendix). The checklist is intended to be used as an operational framework by users of a data source to allow them to robustly consider and document usability, utility, and applicability to their intended use case.

The document also provides an overview and more detailed explanatory sections for the documentation and characterization of the source data, transformation processes, loading into a data model for re-use along with discussions of governance and linkage between data sources.

The Data Quality Framework is aligned with the NEST Research Methods Framework, which is meant to generate information and evidence contextually suitable to support regulatory decisions and to generate actionable clinical insights. To do this, the investigator must consider the question at hand within the context of the target data source. A well-articulated research question should include enough information to allow the investigator to prospectively define the population (indication and setting), exposure(s), outcome(s) and subgroups of interest; potential confounding variables, and the study period and follow-up duration.

In conclusion, the overarching goal of this updated version of the NESTcc's Data Quality Framework is to create standards and criteria by which the quality of data sources can be evaluated and determined. Together with NESTcc's Research Methodology Framework, these documents provide guidance that can help enhance the quality of data and RWE to better evaluate medical device safety, effectiveness, and quality, which will ultimately, and most importantly, bring better care to patients.

RERFERENCES

1. NEST Test Cases. https://nestcc.org/test-cases/
2. Dhruva SS, Zhang S, Chen J, et al. Safety and Effectiveness of a Catheter With Contact Force and 6-Hole Irrigation for Ablation of Persistent Atrial Fibrillation in Routine Clinical Practice. *JAMA Netw Open*. Aug 1 2022;5(8):e2227134. doi:10.1001/jamanetworkopen.2022.27134
3. Frankenberger EA, Resnic FS, Ssemaganda H, et al. Evaluation of intervertebral body implant performance using active surveillance of electronic health records. *BMJ Surg Interv Health Technol*. 2022;4(1):e000125. doi:10.1136/bmjsit-2021-000125
4. Gatto NM, Reynolds RF, Campbell UB. A Structured Preapproval and Postapproval Comparative Study Design Framework to Generate Valid and Transparent Real-World Evidence for Regulatory Decisions. *Clin Pharmacol Ther*. Jul 2019;106(1):103-115. doi:10.1002/cpt.1480
5. Blake HA, Sharples LD, Harron K, van der Meulen JH, Walker K. Probabilistic linkage without personal information successfully linked national clinical datasets. *J Clin Epidemiol*. Aug 2021;136:136-145. doi:10.1016/j.jclinepi.2021.04.015
6. Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. *Int J Environ Res Public Health*. May 14 2014;11(5):5170-207. doi:10.3390/ijerph110505170
7. Gottlieb S, Shuren JE. Statement from FDA Commissioner Scott Gottlieb, M.D. and Jeff Shuren, M.D., Director of the Center for Devices and Radiological Health, on FDA's updates to Medical Device Safety Action Plan to enhance post-market safety. Food and Drug Administration. Accessed 12/22/2023, 2023. https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-and-jeff-shuren-md-director-center-devices-and-2
8. United States Food and Drug Administration (FDA). Real world evidence. US Food and Drug Administration. Accessed 12/22/2023, 2023. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

9. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. *AMIA Annu Symp Proc*. 2013;2013:1472-7.
10. Data Standards Program Action Plan (US Food and Drug Administration) (2022).
11. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices (2023).
12. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. *J Med Internet Res*. May 29 2018;20(5):e185. doi:10.2196/jmir.9134
13. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. *J Multidiscip Healthc*. 2016;9:211-7. doi:10.2147/jmdh.S104807
14. Xu Y, Zhou X, Suehs BT, et al. A Comparative Assessment of Observational Medical Outcomes Partnership and Mini-Sentinel Common Data Models and Analytics: Implications for Active Drug Safety Surveillance. *Drug Saf*. Aug 2015;38(8):749-65. doi:10.1007/s40264-015-0297-5
15. OHDSI. The Book of OHDSI. https://ohdsi.github.io/TheBookOfOhdsi/
16. Raebel MA, Haynes K, Woodworth TS, et al. Electronic clinical laboratory test results data tables: lessons from Mini-Sentinel. *Pharmacoepidemiol Drug Saf*. Jun 2014;23(6):609-18. doi:10.1002/pds.3580
17. Weeks J, Pardee R. Learning to Share Health Care Data: A Brief Timeline of Influential Common Data Models and Distributed Health Data Networks in U.S. Health Care Research. *EGEMS (Wash DC)*. Mar 25 2019;7(1):4. doi:10.5334/egems.279
18. Deshmukh VG, Meystre SM, Mitchell JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. *BMC Med Res Methodol*. Oct 28 2009;9:70. doi:10.1186/1471-2288-9-70
19. Majeed RW, Röhrig R. Automated realtime data import for the i2b2 clinical data warehouse: introducing the HL7 ETL cell. *Stud Health Technol Inform*. 2012;180:270-4.
20. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. *PloS one*. 2019;14(2):e0212463. doi:10.1371/journal.pone.0212463
21. Danciu I, Cowan JD, Basford M, et al. Secondary use of clinical data: the Vanderbilt approach. *J Biomed Inform*. Dec 2014;52:28-35. doi:10.1016/j.jbi.2014.02.003
22. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial Bias in Pulse Oximetry Measurement. *N Engl J Med*. Dec 17 2020;383(25):2477-2478. doi:10.1056/NEJMc2029240
23. Pratt NL, Mack CD, Meyer AM, et al. Data linkage in pharmacoepidemiology: A call for rigorous evaluation and reporting. *Pharmacoepidemiol Drug Saf*. Jan 2020;29(1):9-17. doi:10.1002/pds.4924
24. Association AH. *Maturity Framework: Data-driven Health Care Organizations*. 2020. https://www.aha.org/system/files/media/file/2021/01/MI_Leveraging_Data_Maturity_Framework.pdf
25. Knosp BM, Dorr DA, Campion TR. Maturity in enterprise data warehouses for research operations: Analysis of a pilot study. *Journal of Clinical and Translational Science*. 2023;7(1):e70. e70. doi:10.1017/cts.2023.23
26. Shaygan A, Daim T. Technology management maturity assessment model in healthcare research centers. *Technovation*. 2023/02/01/ 2023;120:102444. doi:https://doi.org/10.1016/j.technovation.2021.102444
27. United States Food and Drug Administration (FDA). Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision Making for Drug and Biological Products. Draft Guidance for Industr. 09/01/2021 2021;

28. Unique device identification system. Final rule. *Fed Regist*. Sep 24 2013;78(185):58785-828.
29. Everson J, Rubin JC, Friedman CP. Reconsidering hospital EHR adoption at the dawn of HITECH: implications of the reported 9% adoption of a "basic" EHR. *Journal of the American Medical Informatics Association : JAMIA*. Aug 1 2020;27(8):1198-1205. doi:10.1093/jamia/ocaa090
30. 2015 Edition Health Information Technology (Health IT) Certification Criteria, 2015 Edition Base Electronic Health Record (EHR) Definition, and ONC Health IT Certification Program Modifications. Final rule. *Fed Regist*. Oct 16 2015;80(200):62601-759.
31. McGuinness JE, Zhang TM, Cooper K, et al. Extraction of Electronic Health Record Data using Fast Healthcare Interoperability Resources for Automated Breast Cancer Risk Assessment. *AMIA Annu Symp Proc*. 2021;2021:843-852.
32. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices: Guidance for Industry and Food and Drug Administration Staff (2017).
33. Locke T, Parker V, Thoumi A, Goldstein B, Silcox C. Preventing Bias and Inequities in AI-Enabled Health Tools. *Duke Margolis Center for Health Policy*. 2022;
34. Reddy S, Rogers W, Makinen VP, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health Care Inform*. Oct 2021;28(1)doi:10.1136/bmjhci-2021-100444
35. Lee KJ, Tilling KM, Cornish RP, et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. *J Clin Epidemiol*. Jun 2021;134:79-88. doi:10.1016/j.jclinepi.2021.01.008
36. Qualls LG, Phillips TA, Hammill BG, et al. Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®). *EGEMS (Wash DC)*. Apr 13 2018;6(1):3. doi:10.5334/egems.199
37. Lynch KE, Deppen SA, DuVall SL, et al. Incrementally Transforming Electronic Medical Records into the Observational Medical Outcomes Partnership Common Data Model: A Multidimensional Quality Assurance Approach. *Appl Clin Inform*. Oct 2019;10(5):794-803. doi:10.1055/s-0039-1697598
38. Services USDoHaH. *Oversight of Clinical Investigations — A Risk-Based Approach to Monitoring*. 2013.
39. United States Food and Drug Administration (FDA). Guidance for Industry: Computerized Systems Used in Clinical Investigations. *United States Food and Drug Administration (FDA)*. 2007;
40. SACHRP Recommendations: Attachment C: Recommendation on Protocol Deviations (2012).
41. Goodney P, Mao J, Columbo J, et al. Use of linked registry claims data for long term surveillance of devices after endovascular abdominal aortic aneurysm repair: observational surveillance study. *BMJ*. 2022;379:e071452. doi:10.1136/bmj-2022-071452
42. Dhruva SS, Ridgeway JL, Ross JS, Drozda JP, Jr., Wilson NA. Exploring unique device identifier implementation and use for real-world evidence: a mixed-methods study with NESTcc health system network collaborators. *BMJ Surg Interv Health Technol*. 2023;5(1):e000167. doi:10.1136/bmjsit-2022-000167
43. Jiang G, Dhruva SS, Chen J, et al. Feasibility of capturing real-world data from health information technology systems at multiple centers to assess cardiac ablation device outcomes: A fit-for-purpose informatics analysis report. *Journal of the American Medical Informatics Association : JAMIA*. Sep 18 2021;28(10):2241-2250. doi:10.1093/jamia/ocab117

44. Nind T, Sutherland J, McAllister G, et al. An extensible big data software architecture managing a research resource of real-world clinical radiology data linked to other health data from the whole Scottish population. *Gigascience*. Sep 29 2020;9(10)doi:10.1093/gigascience/giaa095
45. Carreras G, Simonetti M, Cricelli C, Lapi F. Deterministic and Probabilistic Record Linkage: an Application to Primary Care Data. *J Med Syst*. Mar 22 2018;42(5):82. doi:10.1007/s10916-018-0944-3
46. Mao J, Etkin CD, Lewallen DG, Sedrakyan A. Creation and Validation of Linkage Between Orthopedic Registry and Administrative Data Using Indirect Identifiers. *J Arthroplasty*. Jun 2019;34(6):1076-1081.e0. doi:10.1016/j.arth.2019.01.063
47. Harron KL, Doidge JC, Knight HE, et al. A guide to evaluating linkage quality for the analysis of linked data. *Int J Epidemiol*. Oct 1 2017;46(5):1699-1710. doi:10.1093/ije/dyx177
48. demekong PF, Annamaraju P, Haydel MJ. Health Insurance Portability and Accountability Act. *StatPearls*. StatPearls Publishing Copyright © 2023, StatPearls Publishing LLC.; 2023.
49. Menikoff J, Kaneshiro J, Pritchard I. The Common Rule, Updated. *N Engl J Med*. Feb 16 2017;376(7):613-615. doi:10.1056/NEJMp1700736
50. Kostka K. Preserving Privacy in an OMOP CDM Implementation. https://ohdsi.github.io/CommonDataModel/cdmPrivacy.html
51. Administration USFaD, Center SO. Sentinel System Principles and Policies. https://www.sentinelinitiative.org/about/principles-policies
52. Weber G, Klann K, Mendis M, Murphy S, Potenzone R, Rice P. i2b2 Common Data Model Documentation. https://community.i2b2.org/wiki/display/BUN/i2b2+Common+Data+Model+Documentation
53. Institute P-COR. *Policy for Data Management and Data Sharing*. 2018. https://www.pcori.org/sites/default/files/PCORI-Policy-for-Data-Management-and-Data-Sharing.pdf
54. Rosati K, Jorgensen N, Soliz M, Evans B. *Sentinel Initiative Principles and Policies, HIPAA and Common Rule Compliance in the Sentinel Initiative*. 2018. https://www.sentinelinitiative.org/about/principles-policies
55. Services USDoHaH. *Examining Oversight of the Privacy & Security of Health Data Collected by Entities Not Regulated by HIPAA*. 2016. https://www.healthit.gov/sites/default/files/non-covered_entities_report_june_17_2016.pdf
56. The Good Governance Standard for Public Services (Hackney Press, Ltd.) (2004).
57. Center SO. Data Quality Review and Characterization Programs. https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model/data-quality-review-and-characterization-programs
58. Institute P-COR. *Statement on Protecting Patient Privacy*. 2021. https://pcornet.org/wp-content/uploads/2022/01/PCORnet-Statement-on-Protecting-Patient-Privacy-2021-10-06-.pdf
59. Services USCfMM. Federal Information Security Management Act (FISMA). https://security.cms.gov/learn/federal-information-security-management-act-fisma#
60. MDEpiNet. High-Performance Integrated Virtual Environment (HIVE). https://www.mdepinet.net/hive
61. 21 CFR Part 11.10, FDA Guidance to Industry: Part 11, Electronic Records; Electronic Signatures — Scope and Application (2003).
62. Department of Health and Human Services. Security 101 for Covered Entities: HIPAA Security Series. 2007;

63. Office of the National Coordinator for Health Information Technology (ONC). Information Blocking. Accessed 10/30/2023, https://www.healthit.gov/topic/information-blocking
64. Amundson EP, Cole J. DAKOTACARE update: what is a "business associate" agreement? HIPAA OMNIBUS rule--privacy and security changes. *S D Med*. Oct 2013;66(10):432-3.
65. Bönisch C, Hanß S, Spicher N, Sax U, Krefting D. Reusing Biomedical Data as Agreed - Towards Structured Metadata for Data Use Agreements. *Stud Health Technol Inform*. Sep 12 2023;307:31-38. doi:10.3233/shti230690
66. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated Learning for Healthcare Informatics. *J Healthc Inform Res*. 2021;5(1):1-19. doi:10.1007/s41666-020-00082-4

NEST

NATIONAL EVALUATION SYSTEM FOR HEALTH TECHNOLOGY

CONTACT INFORMATION

For more information, please contact NESTcc at nestcc@mdic.org

NEST

NATIONAL EVALUATION SYSTEM FOR HEALTH TECHNOLOGY

Learn more: www.nestcc.org

Phone: (202) 559-2938

Email: nestcc@mdic.org