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

Subcommittee Members: Lesley Curtis, PhD, MS (chair); Jeffrey Brown, PhD; John Laschinger, MD; Aaron Lottes, PhD; Keith Marsolo, PhD; Frederick Masoudi, MD, MSPH; Joseph Ross, MD, MHS; Art Sedrakyan, MD, PhD; Kara Southall, MS; James Tcheng, MD; Karen Ulisney, MS, CRNP; Charles Viviano, MD, PhD

Contents

Introduction

In 2012, the National Evaluation System for health Technology (NEST) was born 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]

In 2018, the Data Quality Subcommittee of the NEST Coordinating Center (NESTcc) was tasked with creating a Data Quality Framework for NESTcc Network Collaborators. The initial version of that framework, presented in this document, lays 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, this framework is grounded in the use of real-world data (RWD) gleaned from the clinical care setting instead of data collected specifically for research or evaluation purposes. This framework focuses on RWD from the electronic health record (EHR) rather than other clinically based data sources such as registries, which have been addressed elsewhere.[2]

This Data Quality Framework serves 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 capture and use of clinical information as high-quality data to support the generation of real-world evidence (RWE), which will ultimately, and most importantly, provide better care to patients.

This framework is composed of five sections that cover the topics most salient to achieving the highest data quality around medical devices:

1. 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, and results in trustworthy findings and conclusions.

2. Characteristics of Data: Choosing and using data appropriately first necessitates understanding and specifying the data needed, along with the context and limitations of potential sources of that data. Shortcomings of the data that potentially limit their application must also be identified.

3. Data Capture and Transformation: The use of EHR data for secondary analyses presents additional challenges in terms of data relevance and reliability. The processing and transformation of data into common data models provides a logical pathway for enabling analysis.

4. Data Curation: Curation turns raw data into information by organizing, assessing, and preparing the data for analysis. Data curation is an iterative process, with the goal to improve data quality over time.

5. NESTcc Data Quality Maturity Model: Maturity models are used by organizations to assess business capabilities, identify opportunities, and perform capacity planning. Maturity models also allow for benchmarking of relevant characteristics over time. The ability to capture data consistently and completely, to represent data via common data models, to validate the accuracy of data, and to then use the data through automated queries are examples of key processes that drive data quality. The five proposed stages of maturity reflect increasingly advanced and integrated levels of performance for health care systems to partner within the NESTcc ecosystem. The NESTcc Data Quality Maturity Model, by itself, does not ensure improvement but is rather an indicator of progress. The model can help researchers identify weaknesses, thereby enabling research teams to address them.

Governance

RWD are observational data that can be analyzed to produce RWE. RWD are defined by the U.S. Food and Drug Administration (FDA) as "data related to patient health status and/or the delivery of health care routinely collected from … electronic health records (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."[3] To support the generation of RWE

from RWD, core principles must be agreed on to establish governance, including policies and processes for organizational transparency and integrity; data access, management, linkage and aggregation, and use; and submission, management, review, and acceptance of analytic requests.[4,5]

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.[6] As no individual party is free from bias or conflict of interest, governance provides a basis to balance stakeholder influences and provide equal representation, thereby limiting the potential for bias or unethical behaviors and allowing trustworthy research. The Patient-Centered Outcomes Research Institute (PCORI) has identified stakeholders to include patients, clinicians, researchers, purchasers, payors, industry, hospitals and health systems, policy makers, and training institutions,[7] and these same stakeholders remain relevant to the RWD and RWE domains. Additionally, engagement of stakeholders is necessary throughout the lifecycle of evaluation, from study and analysis planning and conduct through dissemination of results.

NESTcc is fully committed to ensuring that the highest scientific and ethical standards are applied when using RWD to generate RWE. In doing so, evaluation activities (e.g., sharing patient data across various data sources) must incorporate patient protections such as patient privacy (e.g., HIPAA compliance) and comply with applicable local, state, and federal laws. Institutional review board review may be necessary. The best practices developed by the FDA Sentinel program offer a template for protecting patient privacy and institutional confidentiality when linking RWD across multiple health systems.[8,9]

The following are principles to guide health systems and other clinical organizations in forming policies and procedures for RWD/RWE:

Organizational Transparency and Integrity

- Leadership: Organization establishes executive leadership group for RWD/RWE
- Data Stewardship: Organization takes full responsibility for the organization's RWD
- Patient-centered: Patients are engaged in the RWD/RWE process and provide consent when applicable; organization adheres to ethical standards for responsible conduct of research
- Stakeholder Engagement: Key stakeholders, including patients, clinicians, and other health system and organization staff, are engaged in RWD/RWE project development and execution
- Transparency: All involved individuals from the organization are made clear to the public, potential conflicts of interest are publicly disclosed/reported, and organization's funding is publicly disclosed
- Oversight: Organization assembles independent advisory board with responsibility for organization's local data warehouse and research portfolio, which may include legal counsel to manage liability risk

Data Access, Management, Linkage and Aggregation, and Use

- Data Quality Assurance: Data are accurate and complete
- Data Storage: Data are securely stored, minimizing risk of secondary use or distribution without the appropriate permissions/agreements
- Data Permission: Appropriate agreements are in place for all data used for RWD/RWE, data are de-identified to the greatest extent possible, and patient protections are in place, while still allowing necessary analyses to be pursued; if identified data are used, analyses are conducted within secure network areas from which only aggregated or de-identified data can be removed
- Data Linkage: Linkage of RWD within and across sources is performed with appropriate oversight and processes in place, particularly patient privacy protection

Submission, Management, Review, and Acceptance of RWD/RWE Requests

- Clear Criteria: Criteria by which requests for RWD for RWE are considered are fair and publicly disclosed, including preclusion of access for non-scientific purposes, such as in pursuit of litigation, as well as qualifications for data security and storage
- Transparent Submission and Review Process: Requests for RWD for RWE are publicly disclosed and considered by an independent approval panel (and ethics review as needed), whose determinations are also publicly disclosed
- Commitment to Responsible Analysis: Requests for RWD for RWE include a description of collaborators (including affiliations and conflicts of interest) and proposed use of the RWD/RWE, including the research or evaluation question, data elements of interest, main outcome measures, and statistical analysis plan, which is publicly disclosed; considerations may be made for studies of as-yet-unapproved uses of medical products given commercial confidentiality
- Efficiency: Approved requests for RWD/RWE are managed expeditiously, within time frames that are as rapid as possible, from initiation to analysis to dissemination
- Data Use Agreements: Contractual requirements for data protection and privacy must be established for any approved RWD/RWE request in compliance with appropriate laws and regulations
- Commitment to Results Reporting: All analyses pursued as part of RWD/RWE projects are publicly reported (which could potentially include the project data dictionary and analytic code, as well as all results), including both lay and scientific summaries, regardless of plans to publish in peer-reviewed literature, and directly communicated to FDA when issues with medical product safety are identified; considerations may be made for studies of as-yet-unapproved uses of medical products given commercial confidentiality

Leveraging the use of RWD for RWE holds great promise for medical product evaluation. The principles described above should optimize the success of these efforts among health systems and other clinical

organizations, protect patient privacy, and guide the governance of policies and procedures for RWD/RWE.

Characteristics of Data

Generating evidence to inform and guide clinical and regulatory decisions requires data. For data to be useful, they must be both reliable (high quality) and relevant (fit to purpose) across a broad and representative population. A full understanding of the evaluation question(s) is a prerequisite for determining the assessments, outcomes, and endpoints needed for analysis, as well as the sources, settings, and methodologies needed for data accumulation or acquisition. Choice and use of data require understanding the limitations of the data source(s) and acknowledging that the shortcomings of the data may limit the questions that can be addressed. For example, retrospective observational data acquired from real-world sources, including EHR data, though typically more pragmatic and accurate for addressing real-world practice and outcomes of device use (i.e., questions of generalizability), lack the precision of data prospectively acquired in exploratory randomized clinical trials (i.e., questions of causality). However, rigorously designed prospective clinical trials that include assignment of therapy, randomization, and/or blinding can be embedded in existing RWD sources such as registries, supporting questions that address both causality and generalizability.[10]

The characteristics of satisfactory data are predicated upon a detailed understanding of the question which allows the investigator to prospectively define:

- The appropriate study population;
- The specific data elements required to measure device or medical product utilization;
- The specific data elements required to assess performance and outcomes (including adverse events and their timing) in the course of the disease or treatment;
- The appropriate settings and sources for data acquisition;
- The development or revision of standardized data sets (i.e., development of a common data dictionary for common data elements and key outcomes and endpoints);
- The experimental methods required (e.g., causal inference from prospective randomized controlled trial vs. informed decision making from available or collected observational data)

To generate information and evidence of sufficient quality for generating actionable insights and informing clinical or regulatory decisions, data must satisfy four characteristics:[3]

1. High quality;
2. Relevant to purpose and context;
3. Amendable to the application of appropriate analytic methods (i.e., convertible to evidence);
4. Interpretable using clinical and scientific judgment

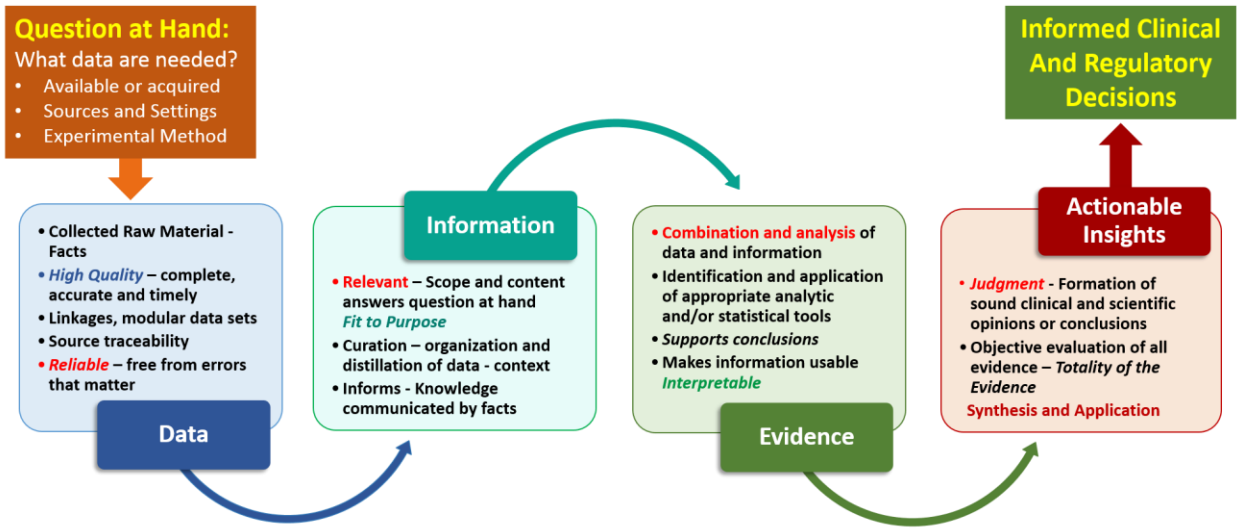
High-quality data are (to the greatest extent possible) complete, accurate, and free from errors that matter. The quality of the raw data increases when common definitional and temporal frameworks exist for disparate sources accessible for analysis. To obtain key endpoints or outcomes, adjudication, use of modular datasets with defined data elements, outcome verification from multiple sources, or other additional mechanisms might be needed to provide additional assurances of the accuracy of available data.

Data must be relevant or fit to purpose. This means the data are reliable and have the scope and content needed to answer the question(s) at hand. Pre-study planning and assessment of the various available data sources must be sufficient to determine whether existing data are contextually appropriate and complete, and whether additional data need to be acquired. Linkages of multiple high-quality data sets for either retrospective or prospective data generation may be used as needed to ensure all needed data are available. Accurate assessments of the totality of the data that will be available for the prespecified analyses are essential.

The combination and analysis of data and information is the final step in the production of evidence. Effective analysis requires the application of appropriate analytic and statistical tools. Prespecified statistical analysis plans are essential to minimize bias. Rigorous analysis makes information interpretable, transforming it into evidence. Objective evaluation of the totality of the evidence coupled with clinical and/or regulatory judgment leads to insights that can be used to inform clinical and regulatory decisions based on the question (see figure below).

Data and information should be viewed as a continuum, capable of developing evidence over the total product life cycle of a device or procedure.[1] Accessibility of the evidence as it evolves requires continuous data access coupled with seamless curation, analysis, and interpretation. Integrated data solutions that allow permanent linkages between previously isolated sources of data and development of open standards will foster a cooperative environment where duplication and costs are minimized and the value of evidence and the underlying infrastructure is maximized.[11,12]

Figure 1. Evidence generation and evaluation: Actionable insights for informed clinical and regulatory decisions (adapted from Califf RM, Sherman R, What we mean when we talk about data. MassDevice. December 11, 2015. https://www.massdevice.com/44947-2/)



Data Capture and Transformation

The use of EHR data for research purposes poses additional challenges to data relevance and reliability. Standardizing definitions for identifying patient cohorts and study endpoints or outcomes, increasing health care systems' interoperability to capture longitudinal patient data, and universally implementing the unique device identifier (UDI) capture will improve the relevance of EHR-based medical device research. Considerations for use of EHR data to conduct research also include understanding provenance, completeness, accuracy and consistency of the data, as well as awareness of what internal and external validation checks have been performed to evaluate the quality of data entry.[13]

Regardless of improvements in data collection systems to accommodate EHR-based research, researchers have little control over data recording and collection processes in clinical care facilities. Individuals using EHR data to derive RWE should understand how and why the data of interest were originally obtained as well as data provenance including subsequent data processing and other nuances that might affect reliability of the data. This understanding will help the researcher determine whether the EHR data are of suitable quality for a particular evaluation.

Improving data quality at the point of care and point of data entry should be the ultimate goal. Wherever possible, the key stakeholder communities should agree regarding the clinical concepts that need to be captured as data for use within the medical device evaluation ecosystem. This might be accomplished at the point of data entry (e.g., when tied to reimbursement or as required in clinical decision support). In these contexts, clinical concepts must be specified and defined as domain-specific common data elements (CDEs), which ideally use standardized definitions and are harmonized with common data models (CDMs) for optimal utility. Currently, most clinical information in an EHR is conveyed as free text

versus the ideal state where EHR data capture would be predominantly in discrete, structured fields. Clinical workflows and documentation systems will likely require modifications to ensure capture of structured data at the point of care.

Once data are captured as discrete elements, extraction, transformation, and loading (ETL) are more amenable to standardization. Discrete data elements, semantic interoperability, compatibility of data capture, and appropriate specification of CDM conventions allow for application of a CDM that subsequently permits execution of standardized analyses or queries by data partners.

Current capabilities may only allow a hybrid approach that combines auto-populating certain discrete structured data elements (e.g., demographics, numerical values, ICD codes) complemented by manual abstraction of other data into a CDM. Alternatively, or in combination, a process such as natural language processing could be used to obtain the data of interest from unstructured text. Of note, natural language processing cannot synthesize data elements or derive inferential conclusions. Moving toward greater agreement and use of computable phenotypes to assist with population identification and perhaps endpoint or outcome identification might address this issue.[14] Such an approach would need to be as comprehensive as possible and include input from all members of the health care ecosystem.

The capture of quality data is one component that determines the quality of the ETL process, which describes how data are extracted and transformed to conform to data standards and CDM specifications. These data are then loaded into a defined location and available for queries (e.g., via a distributed research network).[15] Additional consideration must also be given to applicable patient privacy requirements and agreements or contracts.

Designing the ETL process should follow established best practices, such as seeking input from CDM and data experts to design the ETL process, clinical experts to create coding maps for the process, technical experts to implement the process, and all stakeholders to design and implement quality control procedures.[16]

Data assurance and quality control are essential to the reliability of the RWD for RWE generation. Quality control processes should be integrated throughout, including a review of the ETL design documentation and verification and validation of each step of the ETL process.[16]

Consistency in data element definitions on the data capture side, along with the use of standards to support consistency in the ETL process, will allow researchers to have confidence in the quality of the data extracted from the EHR. Data aggregation is relatively straightforward when data are captured and transformed consistently and reproducibly.

Data Curation

Data curation is one of the steps used to turn raw data into information. Through the curation process, data are organized, assessed, and prepared for analysis. Many frameworks exist to guide this translation of RWD into fit-for-purpose data, but one approach is to consider a two-stage process. The first, foundational stage takes the raw data, applies a series of transformations and quality checks to make the dataset "research ready." The second, study-specific stage applies another series of transformations and quality checks to ensure that the dataset is "fit-for-purpose" for the specific question at hand (some networks/projects may combine both stages into a single process). Foundational data curation examines the data repository or datamart in the context of broad research concepts (e.g., are laboratory results mapped to an appropriate coding scheme?), whereas study-specific data curation also considers a specific-study context (e.g., are outcomes complete for the study population?). Surveys or metadata about data elements, the workflows that give rise to them, and source system provenance further inform the process of data curation and, when combined with information about data latency and extraction and transformation processes, help ensure that fitness-for-use can be assessed as needed. Examples of data curation processes developed by distributed research networks are shown in the table below.

Table 1. Data curation processes for specific distributed research networks.

	Collaborators		
Network	**Health systems**	**Payors**	**Approach to Data Characterization**
HCSRN	X	X	Detailed checks look at ranges, cross-field agreement, implausible data patterns, and cross-site comparisons. Partners execute data characterization package each time data are refreshed. Results are returned to the HCSRN Coordinating Center. Potential quality issues are flagged and mitigated at the partner level.[18]
Sentinel	X	X	Detailed checks look at ranges, cross-field agreement, implausible data patterns, and cross-site comparisons. Partners execute data characterization package each time data are refreshed. Results are returned to the Sentinel Coordinating Center. Potential quality issues are flagged and mitigated at the partner level.[19]
PCORnet	X	X	Includes *foundational data curation* process, which establishes a baseline level of research readiness for all network partners to support prep-to-research queries, and *study-specific data curation*, which includes assessments of outcomes/variables or other derived concepts for the cohort under study.[20]
OHDSI	X	X	Optional – each datamart can generate a standardized data profile that is viewable through a web-based tool (Achilles). Institutions can choose whether to share these profiles or retain them locally.[21]
ACT	X		Under development.

ACT = Accrual for Clinical Trials; HCSRN = Health Care Systems Research Network; OHDSI = Observational Health Data Sciences and Informatics; PCORnet = National Patient-Centered Clinical Research Network.

Key to the curation process are *data characterization routines*, which run against a collaborator's data repository or CDM and describe their performance against a series of *data quality checks* through descriptive statistics such as summaries of missing values, outliers, and frequency distributions. Many data checks rely on concepts analogous to conformance ("does the format of the data adhere to the underlying model?"), completeness ("are there values where we expect to see data populated?"), and plausibility ("do the values that appear make sense?"), as well as comparisons across collaborators.[17] As an example, the most recent PCORnet data characterization process consists of a set of SAS procedures that execute against the tables of the PCORnet CDM.[22] There are 31 unique data checks,[23] many of which apply to multiple fields or tables within the CDM (e.g., required fields are present, tables do not have orphan patient identifiers), for a total of 1,144 individual quality queries. These routines also generate additional tables of descriptive statistics, including the frequencies of specific data elements, crosstabs of data (e.g., procedure and procedure type), and counts of missing, non-missing, and distinct records. FDA Sentinel follows a similar process[24] and, as described below, NESTcc expects collaborators to utilize an approach that is suitable for the dataset and the question(s) being asked. While the data characterization routines are necessarily designed to assess quality within a collaborator's data repository, the summary results are aggregated and analyzed across a network to establish baseline trends and identify outliers or other anomalies.

Metadata About Data Provenance

The results of data characterization alone are not always enough to determine whether a given data set is fit to purpose. Information on provenance also plays a role, as there is widespread variability in how data are entered into EHRs or processed as claims, as well as how health systems and health plans extract those data to populate a given table within their repository or CDM. Knowledge about data collection practices and the decisions made to translate the source material into the target CDM can help provide additional context.[25] Many networks ask their collaborators to complete surveys that describe the provenance of their data sources, providing additional insight into the characteristics of their clinical workflows and/or source systems.[26,27] In some cases, provenance can also be derived automatically as part of the data capture or data transformation process (e.g., did the record originate from a billing system, or was it entered by a clinician?). This is important, because in studies on inpatient medication usage, for instance, one must know whether a datamart has included records only for medications that were administered to patients or all medications that were ordered, including prescriptions written prophylactically (or both), as they will generate markedly different characterization profiles.

Documentation of the Iterative Process of Data Curation

Data curation is an iterative process, with the expectation that characterization activities will help quality improve over time. Therefore, the operational definition of a given data check should stay consistent to allow comparisons over time. Networks may have data checks that are required or investigative. Given the variability in health system data, networks often limit required checks to those related to conformance. Investigative data checks may be remediable by a health system (e.g., >80% of laboratory results have a Logical Observation Identifiers Names and Codes [LOINC] code), or not be remediable due to source system limitations (e.g., <10% of medication orders include an end date). Investigative data

checks that are broadly remediable across the network are good candidates for having thresholds that are raised or lowered to reflect improvements in data quality (e.g., requiring that >50% of laboratory results be mapped to LOINC initially, gradually raising the minimum threshold to >80% as collaborators develop their mappings). Collaborators should track their efforts to address failed investigative data checks and networks should ensure that they perform purpose-specific curation for the population/question in these areas in order to determine whether the data support the study of interest. All of these steps should be documented and included as part of any analysis plan. As the base of RWE studies grows and the FDA releases more guidance, we expect to see best practices and standards emerge as to how to convey this information.

Data Curation Should Be Fit-for-purpose

The minimum requirements for data curation will vary depending on the dataset and the study, but there should be sufficient evidence that the data can answer the question of interest within the context of the intended use. For example, studies of overall utilization patterns for exploratory analyses will require a different level of certainty than a comparative study intended for policy or regulatory decision-making. Studies that use data from emerging domains (e.g., patient-generated data, information derived from natural language processing) may require a higher level of interrogation than a prep-to-research query using a well-known data source. Collaborators that participate in distributed research networks with formalized curation processes may be "pre-cleared" to support a range of activities if their data pass all relevant checks. Collaborators that are not part of any existing network will need to decide how much to invest in data curation. Ensuring that the resulting dataset can be used to answer operational questions that are of value to the health system/health plan is one way to justify the potential expense. Collaborators with data that have only been subjected to a cursory level of curation may still be able to participate but may find themselves restricted to high-level or preliminary exercises.

NESTcc Data Quality Maturity Model

Organizational maturity can be described as an expression of the capabilities of an organization in a specific domain, with the intent to foster continuous improvement across those capabilities. Maturity models organize levels of maturity into a framework, typically assessing culture, process, and/or technology.[28] Maturity models are typically self-administered by organizations to assess current state, model business capabilities, identify opportunities, and perform capacity planning. A key benefit is the benchmarking of relevant characteristics over time. In health care, the Healthcare Information and Management Systems Society (HIMSS) has published several maturity models, including a Health IT Usability Maturity Model (www.himss.org/himss-usability-maturity-model), EHR Adoption Model (www.himssanalytics.org/emram), and Adoption Model for Analytics Maturity (www.himssanalytics.org/amam). Specific to models developed for enterprise data governance, a detailed descriptive model from Stanford addresses the axes of people, policies, and capabilities across the dimensions of awareness, formalization, metadata, stewardship, data quality, and master data.[29]

To articulate a high level of expectations at different levels of organizational maturity with respect to RWD quality, we have developed the NESTcc Data Quality Maturity Model. The model is based on the

expectations of health care systems regarding source systems for RWD capture and management, principally via EHR and other clinical documentation systems.

We propose five stages of maturity of increasingly advanced and integrated levels of performance for health care systems to partner within the NESTcc ecosystem. The stages are at least partially aligned with previous maturity models, of which the HIMSS Usability Model is most informative:

Table 2. Comparability of Stages of NESTcc and Other Maturity Models

NESTcc Stage	HIMSS Usability Model	Capability Maturity Model Integration (CMMI) Model	Stanford Model
1. Conceptual	Unrecognized	Initial	Awareness
2. Reactive	Preliminary	Managed	Formalization
3. Structured	Implemented	Defined	Stewardship
4. Complete	Integrated	Quantitatively managed	Data quality
5. Advanced	Strategic	Optimizing	Master data

Stage 1 – Clinical processes capture information primarily in verbose, unstructured documents, not as discrete data; lack of organizational awareness of data utility, no effort to systematically manage health care data, lack of consistent or centralized governance, policies, and/or resources, data not organized centrally; data not available for organizational use and analysis; individual data units are project oriented or focused on immediate profits.

Stage 2 – Able to react to requests for analysis, respond to research requests – but mostly accomplished by manual chart review and abstraction; data management inefficient and expensive, with only sporadic recognition of data utility beyond immediate use; tacit support from leadership regarding need for centralized data governance and management, but only limited allocation of resources; data not available for organizational use and analysis beyond individual requests; individual data units are project-oriented or focused on immediate profits.

Stage 3 – Clinical systems manage transactional data types (e.g., orders, transactions, laboratory results, medication prescriptions) as discrete data; support from leadership (with resources provided) for centralized data governance and management of these data types at the enterprise level (e.g., support for ETL among internal systems); commitment to centralized enterprise data governance, management, and curation via managed processes, people, and technologies (e.g., enterprise data warehouse [EDW]); non-administrative queries (clinical questions, research) conducted mostly as one-offs via individual queries, still moderate-to-high cost to extract data for analysis; able to support a CDM but not done routinely and automatically; data transmission to registries still largely accomplished by manual chart review and abstraction.

Stage 4 – Granular and complete clinical data based on standardized clinical CDEs captured in the processes of care, integrated into those care processes; UDI captured in the processes of care and available in EHR and in the EDW; EDW routinely and systematically represents data externally via various

CDMs, including efficient queries, support for large number of research projects; leadership provides centralized data governance, management, and curation at the enterprise level, ensuring performance and data quality of local units and achieving financial sustainability.

Stage 5 – Data linkage and aggregation across systems enabled and open to external queries; interoperability of clinical data enabled; multiple sources of sustainable funding support for research; engagement of regulatory and industry enterprises with enterprise data; leadership responsible for centralized data governance, management, and curation at the enterprise level, business benefit well understood, with financial sustainability, and recognition and participation in initiatives external to the organization.

Key Data Process Domains that Drive Data Quality

Optimally, use of health care system RWD requires competency across several data process domains, including data consistency, completeness, and automation.[13] Building on those data process domains, the table below describes expectations at each NESTcc maturity stage. A foundational requirement is **consistent** clinical data based on standardized data dictionaries and/or applicable data standards. While data consistency can be most easily understood within the confines of an individual health care organization, ideally the data are semantically interoperable (i.e., have the same clinical and computational meaning) across organizations. Once standards have been implemented, the ability to capture complete data sets (including interpretation and accounting of the absence of data) characterizes the data **completeness** domain. The ability to represent data via **CDMs**, to validate the **accuracy** of data, and to then use the data through **automation** of queries are additional domains that describe business capabilities related to data quality.

Table 3. Organizational Operational Characteristics Typical of NESTcc Maturity Model Stages

	NESTcc Data Quality Domain				
	Consistency[a]	**Completeness[b]**	**CDM[c]**	**Accuracy[d]**	**Automation[e]**
1. Conceptual					
2. Reactive	+	+	+/-		
3. Structured	+	+	+	+/-	
4. Complete	+	+	+	+	+
5. Advanced	+	+	+	+	+

[a]Data Consistency: Relevant uniformity in data: Across all hospitals, providers, and outpatients (e.g., population/cohort identification, clinical documentation practices/policies between entities, workflow descriptions)

[b]Data Completeness: Presence of the necessary data elements for outcome assessment, CDEs used, all data are electronically available and either complete or with little missing data

[c]Data Models: CDMs include all data needed for decision making (e.g., clinical data elements, UDI)

[d]Data Accuracy: Validation: EHR data are validated systematically, with comparison to the source, independent measurement, upstream data source, and known standard or valid values (e.g., audits from charts)

[e]Data Automation: Queries able to be run automatically against CDMs

Conclusion

High-quality data are essential to support the post-market evaluation of medical devices and to inform regulatory decision-making. In this initial version of the NESTcc Data Quality Framework, we discuss 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 synthesize these topics in the NESTcc Data Quality Maturity Model, which enables collaborators to indicate their progress toward achieving the highest quality data. The next iteration of this framework will include the NESTcc Data Quality Self-Evaluation, a checklist that charts the specific actions organizations can take to move between stages of the maturity model. We welcome further discussion about how the framework can be operationalized by health systems, given the variability in maturity among individual clinics that compose a health system.

References

1. Shuren J, Califf RM. Need for a national evaluation system for health technology. *JAMA*. 2016;316(11):1153-4.
2. Agency for Healthcare Research and Quality. Registries for Evaluating Patient Outcomes: A User's Guide: 3rd Edition. Research Report. April 30, 2014. https://effectivehealthcare.ahrq.gov/topics/registries-guide-3rd-edition/research.
3. U.S. Food and Drug Administration. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices: Guidance for Industry and Food and Drug Administration Staff. August 31, 2017. https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm513027.pdf.
4. Cole A, Garrison L, Mestre-Ferrandiz J, Towse A. Data Governance Arrangements for Real-World Evidence. Office of Health Economics Consulting. November 2015. https://www.ohe.org/system/files/private/publications/420%20-%20Data%20Governance%20for%20RWE.pdf.
5. International Medical Device Regulators Forum, IMDRF Registry Working Group. Patient Registry: Essential Principles. October 2, 2015. http://www.imdrf.org/docs/imdrf/final/consultations/imdrf-cons-essential-principles-151124.pdf.
6. The Independent Commission on Good Governance in Public Services. The Good Governance Standard for Public Services. 2004. Office for Public Management, Chartered Institute of Public Finance and Accountancy, and Joseph Rowntree Foundation. https://www.jrf.org.uk/report/good-governance-standard-public-services.

7. Patient-Centered Outcomes Research Institute. The Value of Engagement. Posted: October 30, 2018. https://www.pcori.org/engagement/what-we-mean-engagement.
8. Sentinel. Background. https://www.sentinelinitiative.org/background.
9. Evans BJ. Panel 3: Appropriate Human-Subject Protections for Research Use of Sentinel System Data. Engelberg Center for Health Care Reform at Brookings. FDA Sentinel Initiative Meeting Series: Issue Brief. March 2010. https://www.brookings.edu/wp-content/uploads/2012/04/Panel-3-Issue-Brief.pdf.
10. Frobert O, Lagerqvist B, Olivecrona GK, et al. Thrombus aspiration during ST-segment elevation myocardial infarction. *N Engl J Med* 2013;369:1587-97.
11. Califf RM. Benefit-risk assessments at the US Food and Drug Administration: finding the balance. *JAMA.* 2017;317(7):693-4.
12. Faris O, Shuren J. An FDA viewpoint on unique considerations for medical-device clinical trials. *N Engl J Med*. 2017;376:1350-7.
13. Zozus MN, Hammond WE, Green BB, et al. Assessing Data Quality for Healthcare Systems Data Used in Clinical Research. Version: 1.0, last updated July 28, 2014. NIH Collaboratory. https://www.nihcollaboratory.org/Products/Assessing-data-quality_V1%200.pdf.
14. Richesson R, Smerek M, Rusincovitch S, et al. Electronic Health Records-Based Phenotyping. NIH Collaboratory Living Textbook of Pragmatic Clinical Trials. June 27, 2014. http://rethinkingclinicaltrials.org/resources/ehr-phenotyping/.
15. PCORnet: The National Patient-Centered Clinical Research Network. PCORnet Glossary. December 2015. https://pcornet.org/wp-content/uploads/2014/07/PCORnet_CDM_Glossary.pdf.
16. Observational Health Data Sciences and Informatics. ETL Creation Best Practices. Last modified June 28, 2017. http://www.ohdsi.org/web/wiki/doku.php?id=documentation:etl_best_practices.
17. Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)*. 2016;4(1):1244.
18. Health Care Systems Research Network. Data Resources. http://www.hcsrn.org/en/About/Data/. Accessed January 26, 2018.
19. Sentinel. Sentinel Operations Center. Sentinel Common Data Model - Data Quality Review and Characterization Process and Programs. Program Package version: 3.3.4. 2017. https://www.sentinelinitiative.org/sentinel/data-quality-review-and-characterization. Accessed January 26, 2018.
20. Qualls LG, Phillips TA, Hammill BG, et al. Evaluating foundational data quality in the National Patient-Centered Clinical Research Network (PCORnet®). *EGEMS (Wash DC).* 2018;6(1):3.
21. Observational Health Data Sciences and Informatics. ACHILLES for Data Characterization. https://www.ohdsi.org/analytic-tools/achilles-for-data-characterization/. Accessed January 26, 2018.
22. PCORnet Distributed Research Network Operations Center. PCORnet Data Curation Query Package. https://github.com/PCORnet-DRN-OC/PCORnet-Data-Curation. Accessed January 26, 2018.

23. PCORnet: The National Patient-Centered Clinical Research Network. PCORnet Data Checks v5. https://pcornet.org/wp-content/uploads/2018/05/PCORnet-Data-Checks-v5.pdf.
24. Sentinel. Sentinel Data Quality Assurance Practices. https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/sentinel-data-quality-assurance-practices.
25. Johnson KE, Kamineni A, Fuller S, Olmstead D, Wernli KJ. How the provenance of electronic health record data matters for research: a case example using system mapping. *EGEMS (Wash DC).* 2014;2(1):1058.
26. Curtis LH, Weiner MG, Boudreau DM, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf.* 2012;21(Suppl 1):23-31.
27. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care.* 2013;51(8 Suppl 3):S22-29.
28. Mettler T. Maturity assessment models: a design science research approach. *Int J Soc Syst Sci.* 2011;3(1/2):213–222.
29. Stanford University. Data Governance Maturity Model: Guiding Questions for Each Component-Dimension. http://web.stanford.edu/dept/pres-provost/cgi-bin/dg/wordpress/wp-content/uploads/2011/11/StanfordDataGovernanceMaturityModel.pdf.