A Conceptual Framework of Data Readiness: The Contextual Intersection of Quality, Availability, Interoperability, and Provenance

Brian J. Douthit[1] Guilherme Del Fiol[2] Catherine J. Staes[2,3] Sharron L. Docherty[1,4] Rachel L. Richesson[5]

[1] School of Nursing, Duke University, Durham, North Carolina, United States
[2] Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, United States
[3] College of Nursing, University of Utah, Salt Lake City, Utah, United States
[4] School of Medicine, Duke University, Durham, North Carolina, United States
[5] Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, Michigan, United States

Address for correspondence Brian J. Douthit, PhD, RN-BC, School of Nursing, Duke University, 307 Trent Drive, Durham, NC 27710, United States (e-mail: brian.douthit@duke.edu).

Appl Clin Inform 2021;12:675–685.

Abstract

Background Data readiness is a concept often used when referring to health information technology applications in the informatics disciplines, but it is not clearly defined in the literature. To avoid misinterpretations in research and implementation, a formal definition should be developed.

Objectives The objective of this research is to provide a conceptual definition and framework for the term data readiness that can be used to guide research and development related to data-based applications in health care.

Methods PubMed, the National Institutes of Health RePORTER, Scopus, the Cochrane Library, and Duke University Library databases for business and information sciences were queried for formal mentions of the term "data readiness." Manuscripts found in the search were reviewed, and relevant information was extracted, evaluated, and assimilated into a framework for data readiness.

Results Of the 264 manuscripts found in the database searches, 20 were included in the final synthesis to define data readiness. In these 20 manuscripts, the term data readiness was revealed to encompass the constructs of data quality, data availability, interoperability, and data provenance.

Discussion Based upon our review of the literature, we define data readiness as the application-specific intersection of data quality, data availability, interoperability, and data provenance. While these concepts are not new, the combination of these factors in a novel data readiness model may help guide future informatics research and implementation science.

Conclusion This analysis provides a definition to guide research and development related to data-based applications in health care. Future work should be done to validate this definition, and to apply the components of data readiness to real-world applications so that specific metrics may be developed and disseminated.

Keywords
- data quality
- interoperability
- governance
- data management
- electronic health record

received
January 25, 2021
accepted after revision
June 9, 2021

DOI https://doi.org/10.1055/s-0041-1732423.
ISSN 1869-0327.

Background and Significance

The term data readiness has been growing in popularity in health technology and informatics circles, from its use in blogs of companies such as IBM,[1] to being offered as a service by various companies and agencies.[2,3] Yet, its meaning varies widely depending upon its use and application. A critical examination and conceptual exploration of the meaning of this term is necessary if data readiness is to be used to guide future research and practice. For example, in lay language, this term may convey the notion that "data" is in a state that is considered to be "ready" to use in a particular application. However, it is likely that there are many facets and interpretations that must be considered and assimilated to provide a common conceptual framework for discussion and use in health sciences. If left without critical examination, we may find ourselves using the term data readiness in conflicting circumstances that may undermine its use for describing informatics research and applications.

The act of analyzing terminology in informatics is not new; the exploration of "data quality" by Weiskopf and Weng[4] showed that a standard definition of a concept is important when attempting to operationalize measurement, having real-world implications for research and quality improvement. Their work revealed five dimensions of data quality, with each dimension representing unique concepts in the literature, many of which had overlapping and conflicting uses prior to this work. Their analysis demonstrated the confusion and possible misinterpretation of results (and ultimately consequent risk of misinformed patient care and interventions) that can result from an assumed but not explicit definition of a concept. The clear definition and expansion of the concept of data quality has since laid a critical foundation for the informatics community to advance methods and tools that can impact data quality.

Similarly, data readiness is frequently used in reference to specific clinical health information technology (HIT) tools but is often vague and left to individual interpretation. As such, the term might be used interchangeably for both *data* and *information* readiness which are distinctly different concepts, as supported by foundational informatics theories and literature.[5,6] To begin parsing between these two distinct concepts, data readiness should first be defined at face value to demonstrate how data should be conceptualized when considering its use in HIT applications.

To our knowledge, data readiness has not been formally defined nor operationalized in health care informatics, nor in any other field. A definition of data readiness could be applied to informatics applications, such as through the implementation of clinical practice guidelines as clinical decision support (CDS),[7] to provide a shared understanding of how to assess data for its ability to be used in pragmatic applications. Data readiness as a conceptual framework could help to guide the development of interoperable health solutions, synergistic technology development collaborations across different medical centers, and reuse of informatics solutions (such as apps and CDS tools) that could be integrated into heterogeneous electronic health record (EHR) systems and data repositories. This could provide a foundation for the development of metrics that can be used by organizations to prioritize applications most suited to their data and systems, and can further guide researchers and implementers to ensure that planned applications fit the data they intend to use.

Objectives

Using the extant literature, our objective is to provide a conceptual definition and framework for the term data readiness that can be used to guide research and development related to data-based applications in health care. To achieve this goal, we will (1) identify related and surrogate constructs associated with or determinants of data readiness, and (2) integrate these constructs into a parsimonious conceptual framework to define data readiness.

Methods

To define data readiness, we conducted a modified scoping review of the informatics-related literature of published manuscripts that use the term "data readiness." We searched the databases of PubMed, the National Institutes of Health (NIH) RePORTER, Scopus, the Cochrane Library, the American Medical Informatics Associated (AMIA) Knowledge Center, and the Duke University Library for all available business and information sciences databases. The initial search strategy included exact term matching for "data readiness," "readiness of data," and "readiness of the data." When applicable, "all fields" were selected for search results. No date restrictions were imposed; all results were included through July 2020.

Once the initial articles were retrieved, we confirmed that each article contained one of the search terms of data readiness. If one of the terms was found in the title, abstract, or body of the article, the article was included in the full review. Often, articles were picked up by the database searches due to a citation containing one of the search terms. If this was the case, the cited article was assessed for term matching and whether it was already identified in the initial search (i.e., a duplicate). In addition, while conducting this search, we sought synonyms for data readiness within the manuscripts, and included the synonym in the term matching strategy within our selected databases. Once the articles were identified and confirmed for term matching, two authors (B.J.D. and R.L.R.) conducted independent full-text reviews of each article to confirm if the term data readiness (and its associated phrases and synonyms) was used in a way that contributed to a conceptual definition. For example, we looked for the inclusion of definitions, frameworks, or figures explaining data readiness as a whole or in part, using the concept in a clearly defined use case, or operationalizing the concept for measurement. This iterative search process is outlined in ►Fig. 1.

With the final corpus of literature identified, we synthesized information from each article following a process similar to standard scoping literature review methodology.[8]

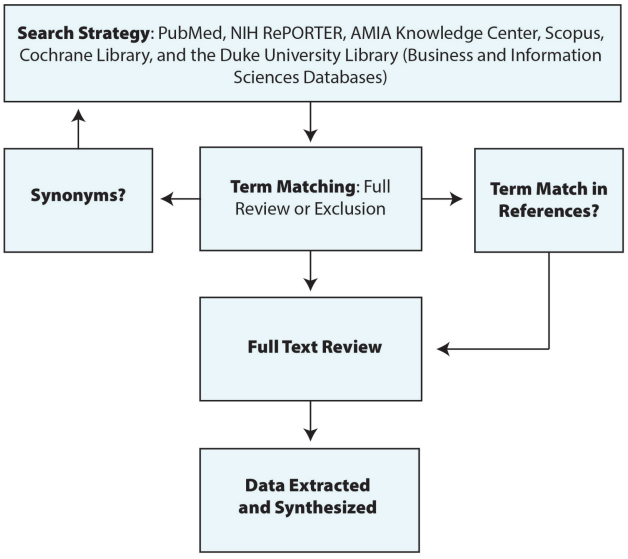

Fig. 1 Literature Search and Article Screening Strategy to Define the Concept of Data Readiness.

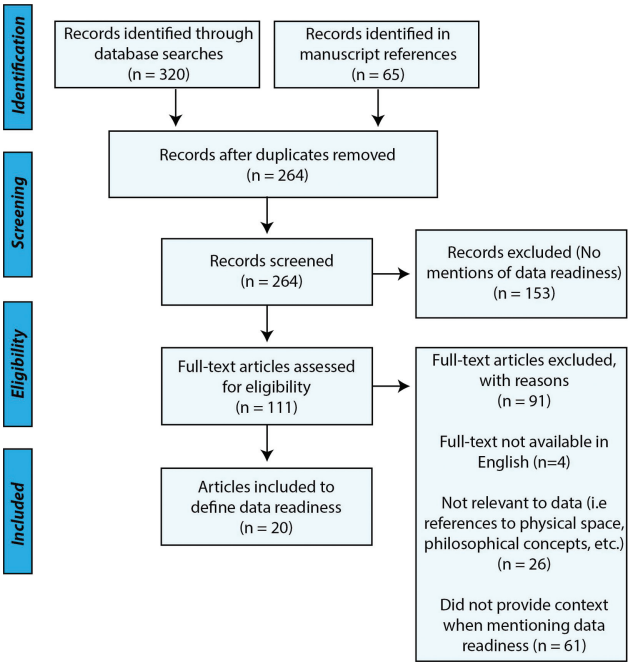

Fig. 2 Results of Literature Search for Publications to Define Data Readiness.

First, we extracted relevant information from each article, including the publication type, the objective of the article, and a summary of the information that the article provided which could contribute toward a definition for data readiness. Second, the authors reviewed and discussed the findings of each article and grouped the findings by common attributes. Third, the authors named each grouping, identifying these as key constructs for a definition of data readiness. Fourth, the constructs were applied to each article to assess for any findings that did not fit within the constructs. Finally, the authors added the key constructs that applied to each article in the summary table by group consensus.

Results

Search Strategy

The initial search produced 264 publications, after removing duplicates. Of note, only one grant was found through the NIH RePORTER. In total, 153 articles did not have any mention of data readiness and were excluded from the final reviews (often because "data" and "ready" were adjacent to each other but separated by a comma or period). Two potential synonyms were identified in the initial search: e-readiness and database readiness. E-readiness was not considered as a synonym as it already had a formal definition and did not pertain to health data, but rather to information technology infrastructure.[9] Four articles referring to "database readiness" were found, two of which met inclusion for exact term matching. However, neither yielded input to the definition of data readiness.

In all, 111 articles were included in the full-text review. Of the 111 articles, 20 met inclusion criteria to define the concept of data readiness. Four were excluded due to the full text not being available in English, 26 were excluded due to not being relevant to data (making references to concepts in physics and other concepts not relatable to health data), and 61 were excluded as they did not provide context to their use of data readiness to contribute toward a definition. Cohen's kappa between the raters was 0.68, or strong agreement.[10] Discrepancies were resolved through consensus between the raters. ►Fig. 2 outlines the search results in PRISMA format.[11]

The 20 publications were then reviewed and discussed to extract relevant constructs and contextual information to define data readiness following the methodology described in the Methods section. Following review by the authors, four distinct and salient constructs were identified: data quality, data availability, interoperability, and data provenance. ►Table 1 lists the included manuscripts, highlighting key information to define data readiness, the data readiness constructs in each manuscript, and a summary of the frequency of each construct.

Data Quality

Data quality is perhaps one of the more complex concepts included in data readiness, with multiple dimensions including accuracy, completeness, consistency, timeliness, uniqueness, and validity.[12,13] Many of these dimensions were represented in how the literature defined data quality in reference to data readiness. Regardless of the type of data, quality presented as the quantitative dimensions of data usefulness in the data readiness literature.

In tandem with quantitative measurement, data quality presented in one of two ways: as an antecedent to data readiness, or as a prospective benchmark. When data quality presents as an antecedent, the concept of "readiness" refers to the adequate quality of data to facilitate the task at hand. For instance, several publications[14–19] state that to use data, it must pass an initial test of data quality, otherwise it should not be used to make informed decisions. On the other hand, some literature referred to data quality as a prospective goal

Table 1 Summary of the literature's contributions to define data readiness

Author (year) and title	Publication type	Objective	Key contributions to define data readiness	Construct defined
Austin (2019)[26] A path to big data readiness	Opinion	To describe practical solutions to common problems experienced when integrating diverse datasets from disparate sources.	Characterizes readiness as the "intensity of data use" and whether it is collected. Also highlights the FAIR priniciples[46] for data management and data ownership.	Data availability; data provenance; interoperability
Campbell et al (2020)[34] Research data network ontologies for precision cancer medicine supporting I2b2 and OMOP	Panel abstract	To explore the use of ONC ontologies and reference terminologies to support research activities.	Furthers the importance of data standards to achieve interoperability.	Interoperability
Carr et al (2014)[14] Defining dimensions of research readiness: a conceptual model for primary care research networks	Literature review	To create a conceptual model of "research readiness" to be used to help identify key requirements for participation in research.	Readiness includes an assessment of how data are held (i.e., how it is distributed and centralized) and understanding its underlying quality.	Data quality; interoperability
Cheriath (2018)[37] Data Governance 2.0	Opinion	To provide an outline for using *data governance 2.0*, a solution to issues in data legacy infrastructure and data stewardship in business applications.	Data readiness is described as part of data management; conflicts somewhat with the other sources and delineates readiness from quality.	Data provenance
Chirkova et al (2018)[16] The data readiness problem for relational databases	Case report	To present and demonstrate a framework to address a "data readiness problem" (returning an accurate Boolean query) in a database	Defines readiness as the inclusion of data to fit the parameters of the intended query.	Data quality
de Lusignan et al (2011)[15] Key concepts to assess the readiness of data for international research: data quality, lineage and provenance, extraction and processing errors, traceability, and curation	Literature review	To define the key concepts which inform whether a system for collecting, aggregating, and processing routine clinical data for research is fit for purpose.	Covers many dimensions of data readiness; as noted in the title, includes data quality and provenance principles. Also describes data models and semantics in the body of the paper.	Data Quality; Data Provenance; Interoperability
Digital Curation Centre (n.d.)[27] 5 Steps to research data readiness – a guide for it managers	Brochure/opinion	To describe the five steps to research data readiness.	Describes readiness as the availability of data and addresses the importance of having data standards and an adequate data infrastructure.	Data availability; interoperability
Ellaway et al (2019)[38] Data, big and small: emerging challenges to medical education scholarship	Opinion	To provide an agenda to change data collection processes for use in medical education and scholarship.	Explores human-centered processes of data; quality and stewardship is essential to readiness.	Data quality; data provenance
Gibbs et al (2017)[17] IDS governance: setting up for ethical and effective use	Expert panel report	Panel report to describe an action plan to support integrated data systems for use in driving social policy.	Describes data readiness as: relevance and sufficiency, quality, collection frequency, granularity, history, privacy, and documentation (quality).	Data quality; data availability; data provenance

Table 1 (*Continued*)

Author (year) and title	Publication type	Objective	Key contributions to define data readiness	Construct defined
Ivers et al (2016)[47] Analysis of SME data readiness: a simulation perspective	Literature review	To investigate the data profile of manufacturing small and medium enterprises with specific emphasis on understanding the data readiness for discrete event simulation modeling.	Data readiness is described in terms of different data quality dimensions.	Data quality
Jennings et al (2018)[33] An instrument to identify computerised primary care research networks, genetic and disease registries prepared to conduct linked research: TRANSFoRm International Research Readiness (TIRRE) Survey	Cross-sectional study	To conduct a survey to assess for the ability of European databases to exchange data.	Data readiness is described as the ability to extract and synthesize data from different sources.	Interoperability
Klievink et al (2017)[28] Big data in the public sector: uncertainties and readiness	Exploratory qualitative study	To assess a framework for evaluating public organizations' big data readiness.	Data readiness is contextual to the organization. Stresses the availability of data for use in organizational goals.	Data availability
Lawrence (2017)[35] Data readiness levels	Opinion	To propose the use of data readiness levels to facilitate project management.	Data readiness includes three levels of assessment: ability to be exchanged, level of missing data, and the ability to show knowledge representation.	Interoperability; data quality
Lu et al (2014)[18] Data readiness level for unstructured data with a focus on unindexed text data	Cross-sectional study	To define the concept of "data readiness levels."	Metrics of quality are important to assess readiness. Data readiness is also dependent on the objective.	Data quality
Nanotechnology Signature Initiative (2013)[20] Nanotechnology Knowledge Infrastructure (NKI): enabling national leadership in sustainable design	Opinion	To discuss and define data readiness levels.	The quality of the data is key to assess its maturity, which in turn is a major factor in data readiness.	Data quality
Richesson (2016)[29] Quantifying system and data readiness for automated clinical decision support[a]	Research grant	To quantify the alignment of CDS data with EHR structures relating to data quality and provider preferences.	Data readiness explains the ability for data to be used in clinical practice guideline-based CDS; availability and quality are essential to successful guideline translation into CDS.	Data quality; data availability
The World Bank (2015)[30] Readiness assessment tool	Executive summary	To give instruction on conducting Open Data Readiness Assessments and associated guidance for leadership.	The assessment tool largely judges the availability of data for public health use.	Data availability

(*Continued*)

Table 1 (*Continued*)

Author (year) and title	Publication type	Objective	Key contributions to define data readiness	Construct defined
United Nations Office for Disaster Risk Reduction (2017)[31] Sendai Framework Data Readiness Review 2017–global summary report	Executive Summary	To report the findings of the Sendai Framework Data Readiness Review, and discuss each component.	To achieve data readiness, it must be available for use.	Data Availability
Vorhees Group, LLC (2007)[48] Institutional data readiness assessment tool	Institutional Data Readiness Assessment Tool	A tool to assess institutional data readiness; three part assessment with 5-level Likert scale	Describes data readiness as a product of human interaction with data; stresses importance of management.	Data provenance
Wen and Hwang (2019)[19] The associativity evaluation between open data and country characteristics	Cross-sectional study	To assess the levels of open government data among various countries.	Data readiness is described as an intersection between availability and quality; focuses on interoperability principles.	Data availability; data quality; interoperability
Frequency of data readiness constructs in the literature				
Construct name			Count	Frequency
Data quality			11	55%
Data availability			8	40%
Interoperability			8	40%
Data provenance			6	30%

Abbreviations: EHR, electronic health record; CDS, clinical decision support.
[a]This contribution is the only grant summary found in the NIH RePORTER; all other contributions were publications found through database searches.

of readiness, such as how the Nanotechnology Signature Initiative[20] posits the need to continually assess quality as a measure to predict its longevity. Regardless, the discovery of these temporal factors necessitates us to consider both short- and long-term strategies to assess data readiness.

In addition, we noted that different approaches to assess for data quality were recommended based on specific use-cases. As this is a well-studied field, several frameworks exist to help evaluate data quality in the context of HIT, including guidance from the Centers for Disease Control and Prevention,[21] the Canadian Institute for Health Information,[22] and many experts in the field.[23,24]

Data Availability

Data availability is defined as the accessibility of data in its desired format. Challenges in data availability may stem from a lack of interoperability, data capture issues, technical infrastructure, and privacy protection.[25] Some literature may use this term interchangeably with data quality, but data availability is distinct as it does not address the quantified usefulness of the data. Rather, it is defined as the ability to access the data in a way that is usable, acting as a bridge between the concepts of interoperability and quality in the definition of data readiness. This is supported by the fact that some publications used availability either separately or in addition to quality to define data readiness.[17,19,26–31] Data availability was also frequently used in reference to larger scale and public health applications. In these cases, there is a call for certain data to become available to aid in an international health collaboration.[19,30,31] For a potential metric, the proportion of data that are readily available may be calculated to inform effort expected or required in HIT implementation.

Interoperability

Interoperability is widely accepted as the ability of multiple systems to exchange and meaningfully use data.[32] This definition is reflected in the data readiness literature, as a focus of many article includes the ability to access and exchange data with a variety of sources. In applied instances, interoperability seemed to be a focus of national and international infrastructure[15,19,30] and research networks.[14,33,34] In addition, a significant body of literature noted interoperability as a fundamental construct to assess for the viability of HIT applications.[26,27,35] Interoperability is often used as an umbrella term for data exchange (reading and writing); in the literature we explored, it appears that the most frequent use of interoperability (if clearly defined) is the use of data standards to facilitate data exchange. We did not note any articles referencing specific measurements for interoperability. However, based on the content of the literature, one measure of interoperability may be the proportion of data elements represented using standard terminology and standard clinical models. Such assessments may be able to address both syntactic and semantic interoperability. In addition, dependent upon the application, the ability to read and write data should be considered as a tangible measure of interoperability.

Data Provenance

Data provenance refers to the ability to follow the lineage of individual data elements; from where it first appeared, to how it has been manipulated, to where it rests at the current moment.[36] This often accompanies questions regarding data ownership, management, and responsibility for keeping it up to date. The literature, like with interoperability, does not give specific measures to assess for provenance. However, provenance is stated to be an important prerequisite to data readiness, as confidence in the stability of the data relates directly to data quality, availability, and interoperability.[26,37,38] We also noted that provenance was referenced with regards to quality assurance and is an important factor in assuring longevity for data-based applications.[15,17,27] For a metric to evaluate provenance, a Likert scale or similar survey method could be used to judge confidence in the stability of data access and ability to track its input over time. Several frameworks regarding the assessment of data provenance could be used to develop a quantitative metric.[39–41]

Conceptual Framework of Data Readiness

In ►Fig. 3, we propose a guiding conceptual framework for data readiness, highlighting the complex interactions of the four essential constructs. Provenance, interoperability, availability, and quality all contribute to the feasibility of using data for any given health care application. Provenance is depicted first, as this is an overarching concept that has impact on the other three constructs. Without provenance, long-term interoperability and availability are uncertain, and data quality cannot be assured if the data are not able to be followed over time. Interoperability (defined as the ability of multiple systems to exchange data) should be assessed second, as without interoperability (especially in cases where data exchange or multiple sources of data are needed), the data become less available. The data would need to either be recollected (which in turn hampers feasibility[7]),

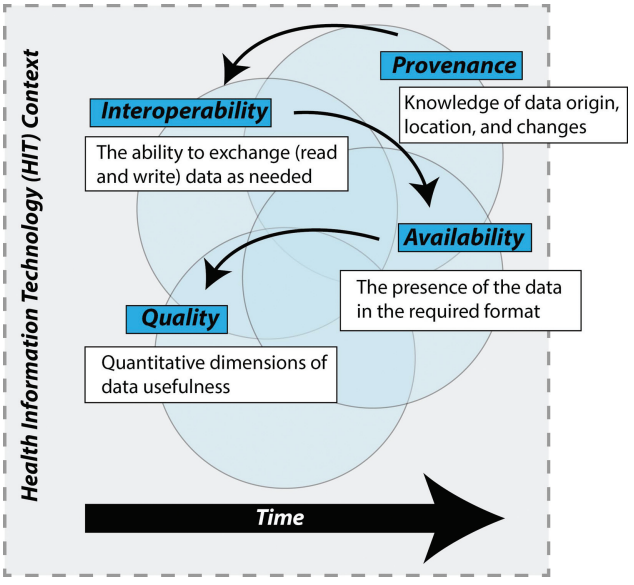

Fig. 3 Data Readiness Conceptual Framework for Data-Driven HIT Feasibility.

or would simply not be available, impacting feasibility. Availability precedes data quality, as without available data, quality cannot be assessed properly. Time is included at the bottom of the figure to represent the continual need to assess all four constructs over time, as a change in any one construct may affect feasibility. All of this is surrounded by the context of the HIT application, as any assessment of data readiness must be customized to the type of data that is required.

Discussion

Discussion of Results

Our concept analysis found that data readiness can be expressed as a hierarchical conceptual framework involving four constructs: data quality, data availability, interoperability, and data provenance. Data readiness is a context-dependent operationalization of these different constructs, and as noted from the body of literature identified in our search, each example focuses on different aspects of data readiness to fit its research question or goal. When using data readiness to inform research and HIT feasibility, all aspects should be considered and addressed *when possible or relevant*. With this caveat, we propose the definition of data readiness as the application-specific intersection of data quality, data availability, interoperability, and data provenance.

Data quality is a heavily studied area and is well recognized as a critical factor in the success of research and any data-based intervention.[23,42,43] It is also apparent that this is an important factor in data readiness, as 11 of the 20 publications reviewed determined data quality to be essential to assess. As data quality is a highly studied area, six major dimensions may be quantitatively analyzed: accuracy, completeness, consistency, timeliness, uniqueness, and validity.[12,13] As is noted in this literature, the approach in which to measure these constructs may vary depending on the type of data that is being used. Regardless, several frameworks exist to help guide these measurements.[13,22–24] Importantly, this review revealed the temporal nature of data quality, and thus the remaining concepts of the framework should also be evaluated at regular intervals or be assured that their assessments will last the lifetime of the HIT application.

Data availability appears to bridge data quality and interoperability together; if the data are not interoperable it cannot be accessed and used, while if data quality is not adequate the data cannot be used as intended. In reference to data readiness, the data must be available to use and in the correct format, or it is indeed not "ready" for health applications. To assess availability, we recommend both longitudinal and short-term approaches. For longitudinal assessment, if there are uncertainties of whether data will be available for the lifespan it is needed, the readiness of the data for HIT use is in question. As for short-term assessments, timeliness of data availability is also important to consider; some applications may require real-time data that may not be available at the time it is needed. In addition, even if the data are available, the format of the data must be considered, as it is possible that there is not a sufficient granularity to serve the needs of certain applications. As previously suggested, using the proportion of data that is readily available (considering accessibility, format, and other potential roadblocks such as security and privacy) can be a helpful metric to assess this dimension of data readiness.

In an era where concepts such as the learning health system, research networks, and international data repositories are becoming a reality, interoperability (the ability for one system to meaningfully use data from another system)[32] has become a topic of great importance in health care. It is not surprising that this construct is a key factor in data readiness, as the most current health care tools and applications demand interoperable systems. Implementing the most up-to-date health care practices and conducting cutting-edge research require access to many different sources of data, but there are many factors that make up the construct of interoperability. In the data readiness literature, the use of standards to facilitate data exchange is key to the construct of interoperability. However, this becomes complicated when considering issues of syntactic and semantic interoperability; standards alone do not predict interoperability.

Assessing interoperability is a unique challenge. One reason for this challenge is likely due to the dynamic nature of standard specifications (e.g., the Fast Healthcare Interoperability Resources [FHIR][44] standard) led by groups such as HL7 where national and international focus is driving the development of common data representation. Because of rapid change, it is difficult to develop metrics, especially when considering the lifespan of a HIT tool. However, one strategy might be to examine the proportion of required data elements that are covered by data and clinical model standards. Such an assessment would need to be re-evaluated at periodic intervals but would at least provide some insight into the feasibility of achieving interoperability.

In comparison to the other three constructs, data provenance asserts a human-centric factor of data readiness. The literature suggests that without a strong understanding of the origin, storage, and changes of the data, the long-term success of an HIT implementation is at risk. This is especially true when long-term and research-funded HIT solutions are considered. One example may be a data repository of patient data collected for research; it is crucial that "ownership" of the repository be clear, and audit of the data (both collection and changes) be traceable. If the data in the repository are recorded in dissimilar formats, the ability to use the data is in question. Although data provenance is important for data readiness, it is not necessarily a requirement for any of the other three constructs. Rather, it is important to consider overall, especially with long-term projects. While a lack of dependence may be true, interoperability can facilitate provenance. For example, use of FHIR can support data provenance through clear definition of data mapping and clinical models.[45] However, similar to the other three constructs, provenance is application-dependent. An individual querying data from a local EHR for a cross-sectional study would be less concerned about provenance than a team developing a global repository.

To summarize, the concept of data readiness is relevant to informatics applications—either explicitly or implicitly. All data-driven HIT applications use data, and hence require access to the data to operate, and require an assessment of readiness to determine the feasibility and plan work for implementation and integration of the tool. To assess data readiness, one must consider all four constructs identified in our definition. Depending on the context of the data's use, assessments for each construct should be customized as needed. As we found in the literature we reviewed, the constructs of data quality, data availability, interoperability, and provenance are not siloed; rather, it appears that complex interactions and sequential analysis are needed to fully assess the concept of data readiness. Our proposed framework for data readiness may guide future research and assessments related to data-driven applications in health care.

Limitations

This work represents a synthesis of literature that encompasses specific references to the term "data readiness." Although we sought to identify synonyms, none were found in this review. Therefore, this work is influenced solely by studies that used the data readiness term (and other similar terms defined in the Methods section) verbatim. Studies using different terms to convey data readiness would have been missed. Our interpretation of the resulting publications, although rigorous, is a subjective process, so we must accept the possibility of misinterpretation. Further, as data readiness is highly dependent on the particular application of the data, our suggestions for evaluation may not fit every HIT application. Further work should be conducted to apply the data readiness framework to real-world HIT applications to inform these metrics. Finally, the scope of this model addresses *data* readiness, and does not explore *information* readiness.

Conclusion

We propose that data readiness can be defined as the application-specific intersection of four constructs: data quality, data availability, interoperability, and data provenance. This work provides a foundation to expand upon and apply to real-world applications that require health data. In future work, this definition of data readiness should be validated, with alterations being suggested as applicable. Importantly, each construct of data readiness should be evaluated with different applications, so that specific metrics for each component may be developed and disseminated to aid in future research.

Clinical Relevance Statement

This work clarifies a frequently used terminology used across informatics disciplines and different specialties. By defining this jargon, we may more clearly and effectively use it as a framework in research and HIT implementation. Without a succinct definition and framework for its use, our discipline is prone to miscommunication and misinterpretation of research results, organizational goals, and HIT assessments.

Multiple Choice Questions

1. What four components were found to define data readiness in the literature?
   a. Data quality, data cleanliness, interoperability, and data provenance.
   b. Data quality, data availability, interoperability, and data provenance.
   c. Data variety, data velocity, data volume, and big data.
   d. Data warehousing, data cleaning, data standards, and data volume.

   Correct Answer: The correct answer is option b. In this review, it was found that data quality, data availability, interoperability, and data provenance were the four components that defined data readiness. Answer a includes data cleanliness, which could be thought of as a subconcept of quality. Answer c lists the components of big data, which may be related but are not the components of data readiness. Answer d lists possible considerations and manifestations of data readiness but are not the components themselves.

2. In what ways can data quality present when assessing for data readiness?
   a. Data quality is always required when assessing for data readiness.
   b. Data quality assessments should include assessments of volume, velocity, and variety.
   c. Data quality can present an antecedent and as a prospective benchmark.
   d. Data quality is a presentation of data availability.

   Correct Answer: The correct answer is option c. The articles that comprised this review usually referred to data quality as an antecedent (a requirement for data readiness prior to the use of data) and as a benchmark that should be assessed on a regular basis. Depending on the application, both may be a part of the application-specific definition. When possible, both presentations of data quality should be considered.

Protection of Human and Animal Subjects

This research does not involve human subjects.

Conflict of Interest

None declared.

Acknowledgments

Brian Douthit is supported by the Robert Wood Johnson Foundation (RWJF) as a Future of Nursing Scholar. The views presented here are solely the responsibility of the authors and do not necessarily represent the official views of the RWJF.

References

1 Jain A, Saha D, Patel H, et al. Data readiness for AI. 2019. Accessed June 30, 2021 at: https://researcher.watson.ibm.com/researcher/view_group.php?id=10391

2 General Services Administration. Data readiness services for artificial intelligence. 2020. Accessed June 30, 2021 at: https://beta.sam.gov/opp/340d2fbf60e441bcb13ec019f0548c07/view
3 Tableau. Data readiness. Accessed June 30, 2021 at: https://www.tableau.com/es-es/support/consulting/data-readiness
4 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013;20(01):144–151
5 Ackoff RL. From data to wisdom. J Appl Syst Anal 1989;16(01):3–9
6 Shannon CE. The mathematical theory of information. Bell Syst Tech J 1949;27(03):379–423
7 Richesson RL, Staes CJ, Douthit BJ, et al. Measuring implementation feasibility of clinical decision support alerts for clinical practice recommendations. J Am Med Inform Assoc 2020;27(04):514–521
8 Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid-Based Healthc 2015;13(03):141–146
9 Hung WH, Chang LM, Lin CP, Hsiao CH. E-readiness of website acceptance and implementation in SMEs. Comput Human Behav 2014;40:44–55
10 McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(03):276–282
11 Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372(71):n71
12 DAMA United Kingdom. The six primary dimensions for data quality assessment. 2013. Accessed April 26, 2021 at: https://silo.tips/download/the-six-primary-dimensions-for-data-quality-assessment
13 Centers for Disease Control and Prevention. The six dimensions of EHDI data quality assessment. Accessed April 26, 2021 at: https://www.cdc.gov/ncbddd/hearingloss/documents/dataqualityworksheet.pdf
14 Carr H, de Lusignan S, Liyanage H, Liaw ST, Terry A, Rafi I. Defining dimensions of research readiness: a conceptual model for primary care research networks. BMC Fam Pract 2014;15(01):169
15 de Lusignan S, Liaw ST, Krause P, et al; Contribution of the IMIA Primary Health Care Informatics Working Group. Key concepts to assess the readiness of data for international research: data quality, lineage and provenance, extraction and processing errors, traceability, and curation. Yearb Med Inform 2011;6:112–120
16 Chirkova R, Doyle J, Reutter JL. The data readiness problem for relational databases. Paper presented at: Cali, Colombia: CEUR Workshop Proceedings. May 21-25, 2018
17 Gibbs L, Nelson A, Dalton E, Cantor J, Shipp S, Jenkins D. IDS governance: setting up for ethical and effective use. 2017. Accessed June 30, 2021 at: https://www.aisp.upenn.edu/wp-content/uploads/2016/07/Governance.pdf
18 Lu Y, Fang X, Zhan J. Data readiness level for unstructured data with a focus on unindexed text data. Paper presented at: ACM International Conference Proceeding Series. Beijing, China; August 4–7, 2014
19 Wen Y-F, Hwang Y-T. The associativity evaluation between open data and country characteristics. Electron Libr 2019;37(02):337–364
20 National Nanotechnology Initiative. NSI: nanotechnology knowledge infrastructure (NKI) data readiness levels discussion draft. 2013. Accessed June 30, 2021 at: https://www.nano.gov/node/1015
21 German RR, Lee LM, Horan JM, Milstein RL, Pertowski CA, Waller MNGuidelines Working Group Centers for Disease Control and Prevention (CDC) Updated guidelines for evaluating public health surveillance systems: recommendations from the Guidelines Working Group. MMWR Recomm Rep 2001;50(RR-13):1–35
22 Canadian Institute for Health Information. The CIHI data quality framework. CIHI Ottawa; 2009. Accessed June 30, 2021 at: https://secure.cihi.ca/free_products/dq-data_quality_framework_2009_en.pdf
23 Weiskopf NG, Bakken S, Hripcsak G, Weng C. A data quality assessment guideline for electronic health record data reuse. EGEMS (Wash DC) 2017;5(01):14–14
24 Chen H, Yu P, Hailey D, Wang N. Methods for assessing the quality of data in public health information systems: a critical review. Stud Health Technol Inform 2014;204:13–18
25 University of Delaware. Managing data availability. 2020. Accessed April 26, 2021 at: https://www1.udel.edu/security/data/availability.html
26 Austin CC. A path to big data readiness. Paper presented at: Proceedings - 2018 IEEE International Conference on Big Data (Big Data 2018). Seattle, Washington, United States; December10–13, 2018
27 Digital Curation Centre. 5 steps to research data readiness - a guide for IT managers. Accessed June 30, 2021 at: https://www.dcc.ac.uk/sites/default/files/documents/resource/5%20Steps%20to%20Research%20Data%20Readiness.pdf
28 Klievink B, Romijn BJ, Cunningham S, de Bruijn H. Big data in the public sector: uncertainties and readiness. Inf Syst Front 2017;19(02):267–283
29 Richesson R. Quantifying system and data readiness for automated clinical decision support. National Library of Medicine; 2016. Accessed June 30, 2021 at: https://grantome.com/grant/NIH/R15-LM012335-01A1
30 The World Bank Group. Readiness assessment tool. 2015. Accessed June 30, 2021 at: http://opendatatoolkit.worldbank.org/en/odra.html
31 United Nations Office for Disaster Risk Reduction. Sendai Framework data readiness review 2017 - global summary report. 2017. Accessed June 30, 2021 at: https://www.undrr.org/publication/sendai-framework-data-readiness-review-2017-global-summary-report#:~:text=Sendai%20Framework%20data%20readiness%20review%202017%20%2D%20Global%20summary%20report,-Documents%20and%20publications&text=Effective%20monitoring%20of%20progress%20in,and%20applicability%20of%20multiple%20datasets
32 Health Information Management Systems Society. What is interoperability? 2019. Accessed June 30, 2021 at: https://www.himss.org/library/interoperability-standards/what-is-interoperability
33 Jennings E, De Lusignan S, Michalakidis G, et al. An instrument to identify computerised primary care research networks, genetic and disease registries prepared to conduct linked research: TRANSFoRm International Research Readiness (TIRRE) survey. J Innov Health Inform 2018;25(04):207–220
34 Campbell W, Campbell J, Reich C, Belenkaya R. Research data network ontologies for precision cancer medicine supporting i2b2 and OMOP. Paper presented at: 2020 AMIA Inform Summit. November 14–18, 2020
35 Lawrence ND. Data readiness levels. arXiv preprint arXiv:170502245.2017
36 Wang J, Crawl D, Purawat S, Nguyen M, Altintas I. Big data provenance: challenges, state of the art and opportunities. Paper presented at: Proc IEEE International Conference on Big Data. October 29–November 1, 2015:2509–2516
37 Cheriath K. Data governance 2.0. 2018 (Journal, Electronic). Accessed June 30, 2021 at: https://www.infoworld.com/article/3268054/data-governance-2-0.html
38 Ellaway RH, Topps D, Pusic M. Data, big and small: emerging challenges to medical education scholarship. Acad Med 2019;94(01):31–36
39 Cheah Y-W, Plale B. Provenance quality assessment methodology and framework. J Data Inform Qual 2014;5(03):9
40 Cheah Y, Plale B. Provenance analysis: Towards quality provenance. Paper presented at: 2012 IEEE 8th International Conference on E-Science, Chicago, Illinois, United States. ; October 8–12, 2012

41 Karvounarakis G, Ives ZG, Tannen V. Querying data provenance. Paper presented at: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, Indianapolis, Indiana, United States. ; June 6–11, 2010
42 Mulissa Z, Wendrad N, Bitewulign B, et al. Effect of data quality improvement intervention on health management information system data accuracy: an interrupted time series analysis. PLoS One 2020;15(08):e0237703–e0237703
43 Gass JD Jr, Misra A, Yadav MNS, et al. Implementation and results of an integrated data quality assurance protocol in a randomized controlled trial in Uttar Pradesh, India. Trials 2017;18(01):418–418
44 Health Level 7. FHIR overview. 2019. Accessed June 30, 2021 at: https://www.hl7.org/fhir/overview.html
45 Health Level 7. Resource provenance - content. 2019. Accessed June 30, 2021 at: https://www.hl7.org/fhir/provenance.html
46 Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018
47 Ivers AM, Byrne J, Byrne PJ. Analysis of SME data readiness: a simulation perspective. J Small Bus Enterprise Dev 2016;23(01):163–188
48 Vorhees Group. Institutional data readiness assessment tool. 2007. Accessed June 30, 2021 at: http://www.voorheesgroup.org/voorheesgroup-tools/Institutional%20Data%20Readiness%20Assessment%20Tool.pdf