Integrated Knowledge Management (IKM) Volume 2

Version 1 - Last Updated 11/21/2023

Table of Contents

I. IKM Background 1
1. IKM Volumes Background 3
1.1. Systemic Harmonization and Interoperability Enhancement for Laboratory Data (SHIELD) Overview 3
1.2. Real World Evidence (RWE) and Standards Overview 3
1.2.1. Data Interoperability Tooling Overview 4
1.2.2. Data Liquidity Overview 5
1.2.3. Metadata overview 6
1.2.4. Analytic Environments Overview 6
1.3. Interoperability Overview 7
1.4. Clinical Statement Overview 8
1.5. Terminology Overview 8
1.5.1. Industry Coding Standards 8
1.5.2. The Adoption of Three Key Terminology Standards: SNOMED CT®, LOINC® and RxNorm 9
1.6. Modularity and Versioning Overview 11
1.7. Safety Systems Thinking and High Reliability Organization (HRO) Overview 12
1.7.1. Motivation for Safety Systems Thinking and HRO 12
1.7.2. Safety Systems Thinking and HRO Background 13
1.7.3. Safety Systems Thinking and HRO Framework Considerations 14
1.7.4. Related Safety Systems Thinking and HRO Work 14
1.7.5. Safety Systems Thinking and HRO Conclusion 16
1.7.6. Safety Systems Use Case 16
1.8. References 17
2. Summary of Findings – Current State of RWE Analytics 19
2.1. Purpose 19
2.2. Introduction 19
2.2.1. Executive Summary 19
2.3. Real World Data (RWD) and Real World Evidence (RWE) Background 20
2.4. RWD Applications and Limitations 20
2.4.1. Current RWD Uses 21
2.4.2. National Covid Cohort Collaborative (N3C) 21
2.4.3. Patient-Centered Outcomes Research Institute (PCORI) and National Patient-Centered Clinical Research Network (PCORnet) 22
2.4.4. Observational Health Data Sciences and Informatics (OHDSI) 23
2.4.5. Electronic Medical Records and Genomics (eMERGE) 24
2.4.6. Sentinel 25
2.5. Learnings 26
2.6. References 28

List of Figures

1.1. Terminology and Structure Examples 8
1.2. Modules and Extensions 12
1.3. Common Knowledge Management Construct Challenges 14

Part I. IKM Background

Table of Contents

1. IKM Volumes Background ... 3
1.1. Systemic Harmonization and Interoperability Enhancement for Laboratory Data (SHIELD) Overview ... 3
1.2. Real World Evidence (RWE) and Standards Overview ... 3
1.2.1. Data Interoperability Tooling Overview ... 4
1.2.2. Data Liquidity Overview ... 5
1.2.3. Metadata overview ... 6
1.2.4. Analytic Environments Overview ... 6
1.3. Interoperability Overview ... 7
1.4. Clinical Statement Overview ... 8
1.5. Terminology Overview ... 8
1.5.1. Industry Coding Standards ... 8
1.5.2. The Adoption of Three Key Terminology Standards: SNOMED CT®, LOINC® and RxNorm ... 9
1.6. Modularity and Versioning Overview ... 11
1.7. Safety Systems Thinking and High Reliability Organization (HRO) Overview ... 12
1.7.1. Motivation for Safety Systems Thinking and HRO ... 12
1.7.2. Safety Systems Thinking and HRO Background ... 13
1.7.3. Safety Systems Thinking and HRO Framework Considerations ... 14
1.7.4. Related Safety Systems Thinking and HRO Work ... 14
1.7.5. Safety Systems Thinking and HRO Conclusion ... 16
1.7.6. Safety Systems Use Case ... 16
1.8. References ... 17
2. Summary of Findings – Current State of RWE Analytics ... 19
2.1. Purpose ... 19
2.2. Introduction ... 19
2.2.1. Executive Summary ... 19
2.3. Real World Data (RWD) and Real World Evidence (RWE) Background ... 20
2.4. RWD Applications and Limitations ... 20
2.4.1. Current RWD Uses ... 21
2.4.2. National Covid Cohort Collaborative (N3C) ... 21
2.4.3. Patient-Centered Outcomes Research Institute (PCORI) and National Patient-Centered Clinical Research Network (PCORnet) ... 22
2.4.4. Observational Health Data Sciences and Informatics (OHDSI) ... 23
2.4.5. Electronic Medical Records and Genomics (eMERGE) ... 24
2.4.6. Sentinel ... 25
2.5. Learnings ... 26
2.6. References ... 28

1. IKM Volumes Background

1.1. Systemic Harmonization and Interoperability Enhancement for Laboratory Data (SHIELD) Overview

SHIELD is a public-private collaboration that started in 2015 with a singular focus: improving the interoperability and utility of in vitro diagnostic (IVD) test data. To accomplish their mission of being able to “Describe the same test the same way anywhere in the Healthcare ecosystem”, SHIELD is working to develop a solution that focuses on:

- **Ecosystem Engagement** - Collaborating and sharing perspective across industry, agency, and discipline, as well as educating stakeholders about laboratory data interoperability.
- **Enhanced Analytic Data Storage** - Providing the community of secondary data users with high quality Real-World Data (RWD) in a central location.
- **Systems Thinking** - Re-engineering the laboratory data transfer process in a manner that prioritizes safety, integrity, and graceful evolution over time above all else.
- **Knowledge Management** - Promoting an integrated approach to identifying, capturing, evaluating, retrieving, and sharing laboratory data and ensuring that data is understandable, reproducible, and useful.

1.2. Real World Evidence (RWE) and Standards Overview

Real world evidence (RWE) using RWD collected as part of routine clinical care, has been heralded as an answer to many of the woes of our medical system. A vision has been made popular that suggests a wealth of untapped data – electronic health records (EHR), medical imaging, mobile apps, and more recently low-cost gene sequencing and wearable devices, unlocked using artificial intelligence and cloud computing – is a path to better diagnostics, personalized treatments, and early disease prevention for millions. Data interoperability of health data may help medicine realize the future in the same way in which other industries have advanced – such as banking, the “internet of things” and online shopping, thereby stressing the importance of establishing the Laboratory Interoperability Data Repository (LIDR) and related infrastructure. The vision of a national interoperable health information system has been elusive; however, because of clinical care data in data silos, incompatible health information technology (IT) systems, and proprietary software that make health information dif#cult to exchange, analyze, and interpret.

Given the number of distinct information systems and care delivery organizations in the ecosystem (e.g., primary care, specialists, laboratories, other care teams), multiple variations exist in how data is collected, manipulated, transmitted, and effectively used. The problem is not the systems per se, but rather the difficulty of data exchange between systems or data incompatibility. Medical data range in format from narrative, textual data to numerical measurements, recorded signals, photographs, drawings, and more. There is also the challenge of the volume of data recorded. Several different observations of a patient are often made concurrently, the observation of the same patient parameter made at several points in time, or both. [1] It is also important to keep a record of the circumstances under which data are obtained. For example, is a blood pressure reading taken on the arm or the leg, had the patient just exercised, what kind of device was used, etc.

Representing these data in various, disparate information systems introduces additional opportunities for inefficiencies, redundancies, and inconsistencies. Each system has a proprietary, standard, or ad-hoc information model and is typically configured to satisfy organization-specific needs, resulting in differences in data capture and storage between and within systems. This is the current reality, as electronic health data are represented in unpredictable and denormalized forms. This reduces the quality of data processing and ability to conduct safe and reliable analytics. Numerous scholarly and practical efforts illustrate the challenges of using observational data for safe patient care, comparative effectiveness research, and analysis including recording biases, workflow differences, and issues with variations in data collection, such as invalid, inconsistent, and missing data. [2]

Aggregated RWE and reliable analytics would benefit from implementable methodologies that address interoperability and data aggregation gaps to help establish a health data ecosystem capable for widespread use in regulatory efforts, public health surveillance, research, and care delivery. To achieve safe and effective information retrieval and reproducible search and query practices across various settings, the following constructs are needed to successfully build, maintain, and analyze data/resources across key biomedical concepts of interest.

1.2.1. Data Interoperability Tooling Overview

A generalized definition of interoperable health data that supports reliable RWE analytics asserts that data recorded and encoded at the time of creation should accurately reflect the meaning intended by the health care professional who created the data. In the current ecosystem, it can be a very manual, error-prone, and sometimes challenging process to determine if data received from entity A is equivalent to data from entity B and if it is interpreted across each setting as the data was originally intended at the point of origin. Data interoperability tooling, such as a Knowledge Management Platform, that can ingest disparate health data and knowledge sources and harmonize them into a common model and change management system could help in ensuring that the data are reliably computed and/or transmitted between health information systems without any change in meaning. The original meaning as intended by the first entity should be fully communicated electronically and understood upon receipt equally by the receiving entity. This type of knowledge platform could serve as data interoperability tooling to be able to determine equivalence between concepts and data sets from various organizations.

To facilitate such tooling, data must be represented by a normal form that can safely and reliably support data analysis that can be used to aggregate data creating using standard or non-standard input form or exchange mechanism. Examples of such data models include Health Level 7 (HL7) Clinical Information Modeling Initiative (CIMI) efforts, including the HL7 Analysis Normal Form (ANF) specification. The model to support a Knowledge Management Platform should meet the following evolutionary design criteria:

- Understandable: The data model for normalizing disparate health data can be processed by health IT systems and understood by most healthcare providers without reference to private or inaccessible information.
- Reproducible: Multiple users or systems apply the model and normal form to the same situations and source data with an equivalent result.
- Useful: The model is fit-for-purpose—it has practical value for data analysis, in support of clinical decision support, research, and population health that requires information aggregated across health IT systems.

Normalization of disparate health data and knowledge is defined as "the ability to identify every representational format that confers the same meaning as being equivalent (i.e., unambiguous representation)." [3] To be clear, the transformation/normalization would involve a data instance to data instance transformation. An example could be John Doe's Systolic Blood Pressure measurement taken on June 4, 2019 represented as a Fast Health Interoperability Resources (FHIR) Observation instance, which is then transformed to a common data model/normal form instance representing this same data. Transformation, in this case, is not a simple endeavor that one can hope to implement on clinical data. It will likely involve navigating disparate data structure trees and include variable representations to then generate a well-formed terminology expression. It is most likely possible to target sub-domains for consistent transformation, such as all quantitative laboratory results, but in some cases, it may be that each detailed clinical model needs its own unique transformation.

Currently, there are three basic categories of errors that might be associated with attempts at normalizing clinical data representation:

1. Errors associated with normalization of content of the terminology

2. Errors associated with normalization of the semantics of the terminology

3. Errors that result from ambiguous or misleading interaction between the structured clinical input and presentation of compound terminology to clinician end-users

A number of options exist for expressing transformation logic and for executing the transformation on specific instances of clinical data for normalization. These range from transformation languages to expensive middleware options commonly used in healthcare interfaces. The suitability of a chosen transformation language highly depends on the format of the source data, and the quality and accuracy of the transformation is left to the transformation author. Examples of transformation languages include extensible Stylesheet Language Transformations (XSLT), FHIR Mapping Language, and Query/View/Transformation (QVT). Often referred to as “data wrangling”, these transformation processes often require mapping data from one “raw” data form into another format with the intent of making it more valuable for downstream purposes such as analytics.

1.2.2. Data Liquidity Overview

To support RWE analytics and data aggregation, there is a desire to “reduce the amount of data wrangling” and one-off transformations across the ecosystem and instead enable “data liquidity”, or the ability of data to flow throughout the health data ecosystem easily and securely so that it is quickly available to clinicians, decision makers, laboratories, patients, and others when they need. A robust information architecture and scalable solution to enable this ideal state requires a common understanding of data and a method to support knowledge-representation and assertional and procedural rules based on common terminology and statement models. If a shareable architecture were able to be implemented across the ecosystem, health data and knowledge could potentially be represented in a way that is both applicable at the interoperability-enterprise and project-specific levels.

Another pragmatic methodology would be to engage vendors and incentivize them to implement an open, common data model and produce shareable output in interoperable, reproducible, and understandable formats that could then be ingested by other systems. Major U.S. EHR vendors including Epic and Cerner which represent > 80% of the U.S. healthcare EHR implementations represent a possible avenue to normalizing and aggregating EHR-based data in the U.S. on a broad scale. Individually, each vendor has independently attempted to create such clinical data resources and product offerings for their user communities. These efforts entail attempts to align, normalize and aggregate the EHR data held by each individual client into a large, de-identified data enclave for purposes of data analytics, quality improvement and research. Unfortunately, these independent efforts are proprietary and do not extend beyond the vendor-user community. Further, the data normalization and representation process remains a "black box" and cannot be independently verified for semantic correctness, data and information loss or gains, or appropriate application of metadata standards. In fact, these vendor-centric efforts face the same issues with vendor-centric data alignment and reuse on a smaller scale. [4-5] Despite these weaknesses, an opportunity exists to collaborate with EHR vendors to normalize and apply metadata standards safely and reliably to the EHR data created by their clients. To capitalize on this opportunity, vendors must be allowed to create unique value in their data offerings to the clientele but also be required to expose these data to the broader community such that data collected by numbers of vendors can be safely and reliably aggregated for research, public health, clinical quality improvement and regulatory uses as described by SHIELD.

1.2.3. Metadata overview

An important construct for enabling RWE analytics, knowledge management, data integration, and decision support, is adoption and reproducible encoding of health data using biomedical terminologies. Acceleration and development of EHR systems have precipitated the emergence of “standard terminologies” and their widespread adoption in the health data community. These standard terminologies include Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT®), the Logical Observation Identifiers, Names, and Codes (LOINC®) and RxNorm.

While these standards are widely implemented and adopted, aggregation and analytics can be difficult because there is no standard representation across these terminologies. Traversing the data models of these various terminology standards in an integrated way is non-trivial because SNOMED CT®, LOINC®, and RxNorm each use different formalisms and tools for their representation. Various terminologies have different semantics, models, release cycles, and versioning mechanisms. [6]

Ideally, to support RWE analytics and aggregation, enterprise terminology requires an integrated terminology using a common representation (metadata) and management. A common terminology model, such as HL7 Standardized Terminology Knowledge Base (also known as TermINology Knowledge Architecture [Tinkar]) would allow analytics on top of aggregated RWE data sets the following [7]:

- Ability to recognize equivalence between data from disparate health IT systems that use codes/terms from various standard terminologies (e.g., “Feels Feverish” in the Temperature Symptoms SNOMED CT® hierarchy versus “Feels Hot/Feverish” in the Observation and Sensation SNOMED CT® hierarchy. Both concepts are Findings in SNOMED CT® but there is no unifying way to identify equivalence).
- Ability to represent local concepts using codes and terms that are modeled extensions to standard terminologies (e.g., “COVID-19 negative test result” was needed in practical use before official SDO releases, or gaps like “mild anemia”, which was proposed, but not accepted, by both the international and U.S. SNOMED CT® release)
- Ability to identify flawed information or incorrect usage or representation of concepts from standard terminologies due to a lack of harmonization and multiple representations that exist across disparate standards (e.g., LOINC® and SNOMED CT® have overlapping concepts)
- Ability to safely change over time in a clear way so that changes are easily understood by implementers. (e.g., redundancy, major name)

1.2.4. Analytic Environments Overview

Integrated analytics environments assist in creating and automating data pipelines and workflows to deliver actionable information to a wide variety of end users. These environments extract data from various data sources, ideally into an integrated knowledge base, housed in a central ecosystem – thereby reducing data silos, enhancing security and governance, and eliminating system redundancies. For data analytics conducted using RWE, analytics environments will be dependent on a variety of knowledge bases and datasets that may have some common fields but also differing fields. A vast variety of biomedical and administrative knowledge present the basis for RWE analytics and knowledge management of these disparate datasets is a major element of consideration to facilitate RWE analytics. Knowledge management includes the creation, discovery, and application of knowledge assets to facilitate sharing and re-use across the ecosystem.

The ability to use interoperable syntax and semantics between data sources is a critical requirement to support effective knowledge management and reliable analytics of disparate RWE datasets. Use of industry standards for syntax and semantics helps facilitate integration; however, not all implementation considerations are planned for and adopted and additional efforts are needed to advance standards to improve data quality. The standards themselves update over time and there is a requirement for knowledge management in analytics environments to consider how standards and underlying code sets have changed over time and relaying these changes to data analysts and users who interpret data reports to best understand the information.

The following are best practices for the establishment of the LIDR taken from the SHIELD Community Roadmap:

1. Develop a freely accessible knowledge management architecture for laboratorians, clinicians, researchers, and regulators, which is needed to promote clinical interoperability, enabling the determination of equivalency between different test results to decide whether they can be safely used for trending, data aggregation, post-market efficacy studies, and research

2. Determine the usefulness of clinical interventions to improve patient care based on relevant laboratory knowledge and reporting data, such as public health reporting and clinical surveillance

3. Harmonize meaningful laboratory terminology standards, such as SNOMED CT® and LOINC®

4. Enhance the reproducibility of data exchange structures used to express laboratory procedures and outcomes, such as Clinical Data Interchange Standards Consortium (CDISC), FHIR, and Integrating the Healthcare Enterprise (IHE) Laboratory Analytical Workflow (LAW)

5. Promote the understandability of the laboratory test knowledge as interpreted and processed by supporting health IT systems such as Laboratory Information Systems (LIS), Laboratory Information Management Systems (LIMS), EHR

6. Establishment of this infrastructure will help facilitating reliable data governance and knowledge management that will assist understandable, reproducible, and useful RWE analytics. Widespread promotion across disparate health information systems will help reduce variation for large national analytic enclaves, such as the National COVID Cohort Collaborative (N3C) Data Enclave which aggregates and analyzes data from numerous, disparate electronic health record systems.

1.3. Interoperability Overview

Data interoperability is the ability to store, query, transform, and transmit data within and between systems while maintaining meaning and all associated and pertinent information. Interoperability primarily focuses on two aspects, structural and meaning interoperability.

- **Structural Interoperability** – While structure focuses on maintaining how systems format data, it cannot ensure that the receiving system will understand the structured data with the original meaning.

- **Meaning Interoperability –** Meaning interoperability works to ensure that the receiving system interprets data in an identical way as the original system.

Working to ensure that transmitted data maintains both structure and meaning is a critical component of our work towards an Integrated Knowledge Management (IKM) platform that can ensure lossless transmission of data. Currently, interoperability of data between and within both organizations and standards poses as a major patient safety issue in the healthcare system. A test with a given LIVD LOINC code can be performed within a lab and as it is transmitted to the LIS, the test can be assigned a different LIS LOINC code that does not capture all of the pertinent information. Even within the organization the interoperability of data is threatened, so when data is then transferred from the Lab’s LIS to a hospitals EHR that lack of data interoperability can compound as the LIS LOINC code is assigned another, potentially different EHR LOINC code.

Working to maintain both structure and meaning as data moves within and between systems will improve patient safety, product development, and health analytics and reduce regulatory and clinical burden on policy makers and healthcare providers.

1.4. Clinical Statement Overview

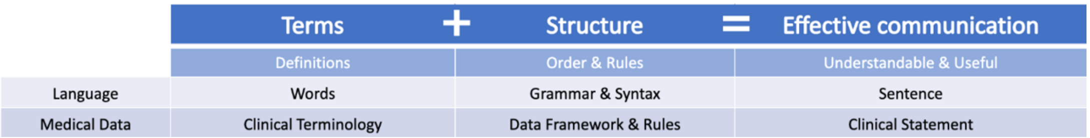
A clinical statement is a general informatics term. It is a definite and clear representation that a clinically significant fact or situation was observed to exist or happened, or that a particular procedure was requested. A clinical statement can be expressed as a narrative that provides a written account that can be naturally read by humans, as well as a normal form which is a machine-processable representation of the statement's data as a standardized and encoded fundamental form. [8]

Clinical Input Form is also a general informatics term. It describes the manner by which clinicians author clinical statements and enter them into their organizations' EHR. Clinical Input Forms (CIFs) have an impact as to how information is presented to the clinicians and how they enter the data. CIFs might be generated by natural language processing or may use models that constrain structured input to allow only certain values to be entered, such as a drop-down list or radio button, or breaking up large chunks of related information into smaller parts. [8]

Today, clinical statements are often represented in unpredictable and denormalized forms, which makes reliable and safe decision support challenging and reduces the quality of other types of data processing. ANF is an approach to clinical statements that ensures the statement representation is reproducible and scalable, with the adherence to principles of being simple, reproducible, and use case driven, with a clean separation between statement concerns and terminology concerns. [8]

In linguistics, words have their very own definitions, however, if it's not paired with proper sentence structure, the words convey very little meaning and are open to interpretation. When this framework is applied to the healthcare system, clinical data, and more specifically clinical statements, we need *terminology* and *structure* to convey a patient's condition or health history from one provider to another.

Figure 1.1. Terminology and Structure Examples



	Terms +	Structure =	Effective communication
	Definitions	Order & Rules	Understandable & Useful
Language	Words	Grammar & Syntax	Sentence
Medical Data	Clinical Terminology	Data Framework & Rules	Clinical Statement

ANF is a data structure that is designed to improve clinical decision support, data robustness, decrease data variation, and ultimately prevent terminology misrepresentations while following the principles of separation of concerns. There are several use cases that ANF can be applied to, such as questionnaire and questionnaire responses, decision-logic engines and frameworks, event-condition-action rules, natural language processing of consult notes, and many others.

1.5. Terminology Overview

1.5.1. Industry Coding Standards

Healthcare data exchange requires that healthcare institutions and laboratories reproducibly encode their test data using industry coding standards. For laboratory data, appropriate use of LOINC® and SNOMED

CT® is essential to ensure tests and results are accurately and reliably described within EHR, LIS, and public health reports. [9]

1.5.2. The Adoption of Three Key Terminology Standards: SNOMED CT®, LOINC® and RxNorm

Over the last several decades, SNOMED CT®, LOINC®, and RxNorm have been increasingly recognized as key resources for: 1. Knowledge Management, 2. Data Integration, 3. Decision Support, 4. High impact on clinical practice and 5. High impact on biomedical research. The following excerpt, written by Olivier Bodenreider, Ronald Cornet, and Daniel Vreeman discusses this further:

The recent acceleration in the deployment of EHR systems has precipitated the emergence of a few terminologies and their wide adoption in the clinical community. Two of them, the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT®) and the Logical Observation Identifiers, Names, and Codes (LOINC®), have become inter- national standards. The last one, RxNorm, is used mostly in the U.S., but similar national drug terminologies exist in other countries (e.g., the NHS Dictionary of medicines and devices (dm+d) [10] in the U.K., the Australian Medicines Terminology (AMT) [11] in Australia) and could have been substituted for RxNorm in this review. In addition to being designated standards mandated for use in U.S. governmental programs, such as the Meaningful Use incentive program [12], these three clinical terminologies have also been selected as the terminological backbone of the Observational Medical Outcomes Partnership (OMOP) common

data model (CDM) used for clinical data warehouses internationally by OHDSI, the Observational Health Data Sciences and Informatics collaborative. [13]

1.5.2.1. SNOMED CT® Overview

Regulations have had a significant impact on the adoption of SNOMED CT®: [13]

SNOMED International consist of member countries, of which the US is one of its inaugural members.

SNOMED CT® holds a yearly Expo as a forum for EHR vendors, health terminology specialists and the community of practice to exchange best practices and measure progress towards the implementation of SNOMOED CT® across the world.

In the US, EHRs are required to use SNOMED CT® for documenting the following: problem lists, procedures and some clinical findings, (i.e. smoking status).

In the U.K., many health information systems (HIT) must use SNOMEDT CT® as the clinical terminology standard within all electronic patient level recording and communications.

1.5.2.2. LOINC® Overview

Approximately 20 U.S. federal agencies have adopted LOINC® as a designated standard and mandated LOINC® for use in various programs [13], such as:

- The U.S. Meaningful Use Incentive Program now called Promoting Interoperability requires LOINC® in messages reporting laboratory test results, exchanging medical summaries, and sending data to cancer registries and public health agencies.
- The LIVD (LOINC® to IVD) Mapping Specification is required by the United States Department of Health and Human Services (HHS) for SARS-CoV-2 reporting and harmonizes how IVD test information is represented using LOINC®. The LIVD file format is currently led by IVD Industry Connectivity Consortium (IICC), and the JavaScript Object Notation (JSON) representation is a project at HL7.

- Federal, State, and Local Public Health Reporting Requirements use messaging standards such as HL7. Inside these messages, laboratories and clinical systems use LOINC® codes to identify which test is being reported.

- The Interoperability Standards Advisory of the Office of the National Coordinator for Health Information Technologies (ONC) lists LOINC® for many interoperability needs, including functional status, laboratory tests, imaging diagnostics, nursing observations, vital signs, and social determinants of health. *The Centers for Medicaid and Medicare Services (CMS)* adopted LOINC® for the patient assessment instruments required in post-acute care settings.

- Common Data models for Large-Scale Research Networks like the National Patient-Centered Clinical Research Network (PCORnet), Observational Health Data Sciences and Informatics (OHDSI), and the Observational Health Data Sciences and Informatics research group all use LOINC® in their common data models.

1.5.2.3. RxNorm Overview

Here are some examples of where RxNorm is used: [13]

- **Electronic Prescribing** -The National Council for prescription Drug Programs (NCPDP), a standards development organization, requires RxNorm as its standardized medical nomenclature for its SCRIPT standard for e-prescribing.

- **Information Exchange** - U.S. Department of Defense (DOD) and Department of Veterans Affairs (VA) rely on RxNorm to mediate drug information across their respective EHRs.

- **Formulary development** - CMS uses RxNorm in their Formulary Reference File – part of the medical drug benefits guidelines.

- **Reference Value Sets** - Electronic clinical quality measure drug value sets are defined in reference to RxNorm for the Meaningful use incentive program.

- **Analytics** - RxNorm is increasingly being used as a drug stand for clinical data warehouses. As an example, OHDSI research group uses RxNorm for representing drugs as part of its Observational Medical Outcomes Partnership (OMOP) common data model (CDM).

PCORnet also uses RxNorm in its common data model.

1.5.2.4. Interoperability Concerns

Despite widespread adoption, there are concerns regarding the interoperability and reliability of LOINC® between organizations. These concerns inspired several scholarly and practical efforts to champion and facilitate laboratory data exchange.

Examples of Scholarly and Practical Efforts to Assess SNOMED CT® from SNOMED CT® Interoperability Projects:

- There are licensing aspects to SNOMED CT® that complicates its use as a common format for LOINC®. This issue will be addressed by SHIELD by building on the SNOMED CT® foundation. Nonetheless, SNOMED CT's® commitment to develop and expand their use of description logics formalism as called out by the System of Logical Representation (SOLOR) as well as their alignment with the Desiderata as described by Cimino, et at, is a step in the right direction. [13]

Examples of Scholarly and Practical Efforts to Assess LOINC® from LOINC® Interoperability Projects:

- A study compared the use of LOINC® terms used in five medical centers and their alignment with vendor recommended LOINC® terms as published in the LIVD file. “We identified mismatches in how medical centers use LOINC® to encode laboratory tests compared to how test manufacturers encode the same laboratory tests. Of 331 tests available in the LIVD files, 136 (41%) were represented by a mismatched LOINC® code by the medical centers.” [9]

- A study auditing consistency of LOINC® encoding between three institutions stated, "There are variations in the way LOINC® is used for data exchange that result in some data not being truly interoperable across different enterprises." [15]
- A similar study described: "We also noted inconsistency across institutions regarding specificity of mappings. It appears that sometimes mappers link concepts to a more general LOINC® code, and at other times they link to a method specific LOINC® code. This causes inconsistency in mappings across institutions.” [12]
- A study assessing the accuracy of LOINC® terminology mappings for 10 commonly ordered tests found that “Of the 275 LOINC® codes reported, 54 (19.6%) were incorrect: 2 codes (5934-2 and 12345-1) (0.7%) did not exist in the LOINC® database and the highest error rates were observed in the property (27 of 275, 9.8%), system (27 of 275, 9.8%), and component (22 of 275, 8.0%) LOINC® axes.” [10]

Examples of Scholarly and Practical Efforts to Assess RxNorm from RxNorm Interoperability Projects:

- RxNorm lack of Description Logics formalism use limits its ability to use logic reasoners for inferencing. This limits RxNorms ability to find new patterns, relationships and identification of equivalence and interoperability. [13]
- While RxNorm has ‘defined ingredients’ [13], historically it has lacked the computational definition where concepts are defined by relationships, therefore determining equivalence and allowing inferencing. However, recent efforts by Bona et al.[15] have successfully created an OWL file to implement and define the classes necessary to model RxNorm and NDC concepts.

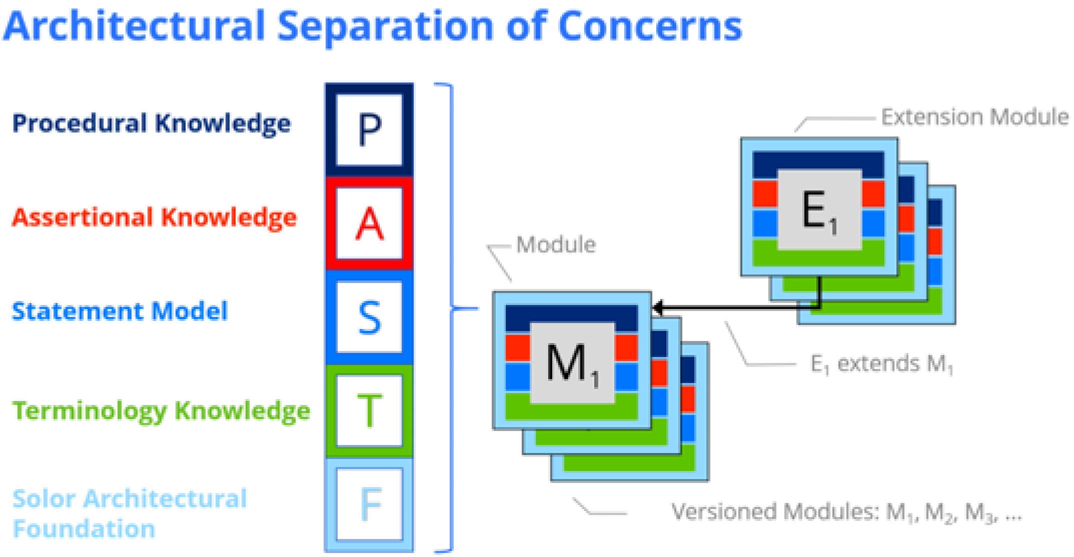
1.6. Modularity and Versioning Overview

When dealing with the complexities of the various architectural layers of the informatics architectural separation of concerns, one of the most important things to note is that any one of the architectural layers will be undergoing modifications at any given point in time, as various Standards Development Organizations go through each of their various drafting, balloting, and approval lifecycles. Therefore, it is important to establish a foundation for Solor as a versioning and modularity architecture that allows changes and subchanges to be referenced uniquely so that all parties can be on the same page as to a particular version.

For example, the following diagram shows how each module could be given a unique version number and contain all layers of the architectural stack. In the instance that a particular versioned module needs to be extended, an extension module could be added to that main versioned module without the need to go to a completely new full module version. This arrangement accounts for the constant change in the healthcare interoperability space while still allowing two organizations to baseline on the same version for testing or exchange purposes.

In software engineering, modularity refers to the extent to which software may be divided into smaller modules. Modularity emphasizes separating the functionality of a program into independent, interchangeable modules, such that each contains everything necessary to execute only one aspect of the desired functionality. A module interface expresses the elements that are provided and required by the module, and the elements defined in the interface are detectable by other modules. Modular programming is closely related to object-oriented programming, having the same goal of facilitating construction of large software programs and systems by decomposition into smaller pieces (i.e., 'polymorphism by encapsulation' or 'composition over inheritance'). With modular programming, concerns are separated such that modules perform logically discrete functions, interacting through well-defined interfaces. Often modules form a directed acyclic graph (DAG); in this case, a cyclic dependency between modules is seen as indicating that these should be a single module. In the case where modules do form a DAG they can be arranged as a hierarchy, where the lowest-level modules are independent, depending on no other modules, and higher-level modules depend on lower-level ones. A particular program or library is a top-level module of its own hierarchy, but can in turn be seen as a lower-level module of a higher-level program, library, or system.

Figure 1.2. Modules and Extensions



1.7. Safety Systems Thinking and High Reliability Organization (HRO) Overview

Today's current state for knowledge management constructs lack a standardized approach to ensure the quality of data as it is entered or transferred into the health care and laboratory systems. Without a clear and repeatable way to verify the accuracy of information within an EHR or laboratory system, there are opportunities for hazards and patient safety issues to emerge. Our team will design a system based on a Safety Systems Thinking approach and High Reliability Organization (HRO) principles that can improve the quality, reliability, and interoperability of data and address the concerns with current knowledge management constructs. We will use prior work and research from industry leaders and subject matter experts (SMEs) as foundational building blocks as we continue to research laboratory and health care systems, medical terminology standards, and best practices and lessons learned to support our development and design process.

1.7.1. Motivation for Safety Systems Thinking and HRO

The health care system in the United States has undergone a digital transformation over the past few decades as digital health information systems and EHRs have been widely adopted and implemented. Despite this expansive transformation, elements of interoperability, data quality management, and the associated patient safety risks are underrepresented as key subject matter experts and informatics stakeholders have not been optimally involved in the design and implementation process of health information systems.

The current state for health care knowledge management to ensure data quality and interoperability is to encode health data using medical terminology standards such as, LOINC®, SNOMED CT®, and RxNorm. These terminologies provide a standardized method to create and enter data elements into health care systems and, in an ideal world, support seamless interoperability between health care systems or health record systems. However, the application and use of these standards is typically not consistent during implementations between facilities within a health care system or between different systems. This non-standardized implementation leads to localized variants of terminologies and knowledge representations that limit interoperability and impacts opportunities to comprehensively understand data within a system.

Not only is the current knowledge management unstandardized, but relative uncertainty also exists regarding the quality and accuracy of the data representing knowledge. A standardized process to validate the vast amount health care data does not exist. The current situation further impacts the quality and accuracy of data, the ability to support research and clinical decisions, and patient safety. Existing data validation techniques need to be standardized and refined to support comprehensive and reproducible validation of data within systems and transferable data between systems.

1.7.2. Safety Systems Thinking and HRO Background

Improving existing knowledge management constructs requires an understanding of industry leading frameworks and principles. Safety Systems Thinking and HRO principles are two frameworks that will support the development of a standardized and repeatable knowledge management and data validation process with continual improvement.

1.7.2.1. Safety Systems Thinking Overview

A systems thinking approach to safety is a thorough and disciplined approach to identifying, analyzing, eliminating, and controlling hazards by analysis, design, and management procedures that operate throughout a system's life cycle. Hazards are created by unsafe conditions and processes that may lead to negative outcomes for stakeholders in the system, such as patients, providers, and researchers.

Traditional approaches to safety view hazards as a linear chain of events and focus on eliminating component failures. However, a system thinking approach to safety attempts to eliminate hazards by examining the various hierarchical control structures and feedback loops that impact the conditions of a system.

In systems thinking, analysis of the system begins in the early concept design stage and continually evaluates changes in the system or the environment that could indicate a shift to unsafe conditions. Reporting and information sharing channels are established, including hazard documentation, tracking, and resolution. These documentation channels are critical to support continual process improvement and are maintained throughout the lifespan of the system.

1.7.2.2. HRO Overview

A HRO is an organization that avoids accidents and hazards, even in complex environments where accidents are common. HROs are based on five key principles:

1. Preoccupation with Failure

   - The organization works to ensure stakeholders are aware of errors and understands how errors occur, the processes that allow errors to occur, and how to reduce the causes of errors.

2. Reluctance to Simplify

   - Stakeholders do not oversimplify the system and aim to fully understand all the processes and control structures that may impact that system.

3. Sensitivity to Operations

   - Organizations and stakeholders identify, and continually measure, key indicators of quality and safety that could alert to changes, even small ones, or indications of a shift towards unsafe conditions or hazardous environments.

4. Deference to Expertise

   - Organizations utilize the experience and knowledge of pertinent SMEs at every opportunity.

5. Commitment to Resilience

   - Organizations learn from errors through feedback loops and continuous monitoring, as they aim to continually improve.

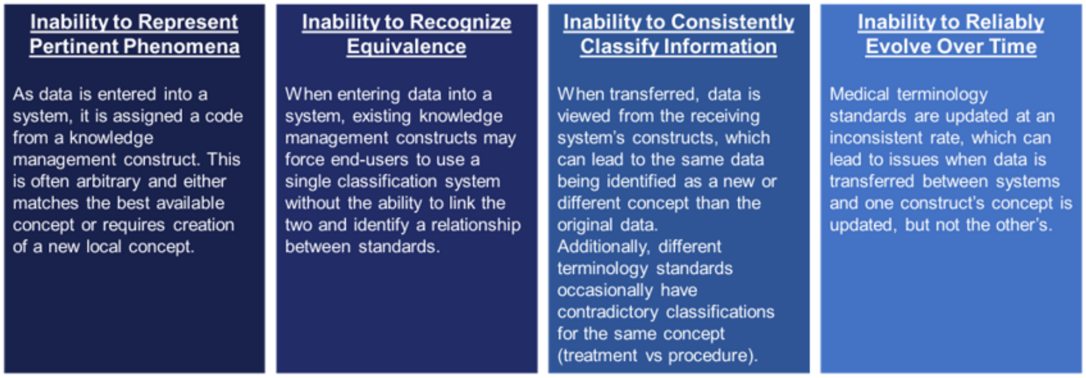
1.7.3. Safety Systems Thinking and HRO Framework Considerations

Following the rapid transition to EHRs, there is a desperate need to overhaul the current knowledge management constructs within the health care system to improve data quality and patient safety. While there will never be a perfectly safe system, an approach embedded in both Safety Systems Thinking and HRO principles can support major strides towards that goal. A combined approach will guide organizations through the successful implementation or redesign of systems to implement clear, concise, and standardized data validation techniques and support continuous process improvement through SME-led risk analysis and statistical, evidence-based review.

It is not uncommon for other high-risk industries to adopt HRO principles. Commercial aviation and nuclear submarines are industries that adopted HRO based principles to enable substantial improvements to quality and safety. Pairing these principles with a systems-based approach to safety and proven tools like Statistical Process Control (SPC), Systems Theoretic Process Analysis (STPA), and Causal Analysis based on System Theory (CAST) will further improve the safety, standardization, and quality of systems and knowledge management constructs.

As we use systems thinking and HRO principles to develop and design a system that addresses the existing flaws in the current knowledge management constructs, it is critical that we perform a comprehensive and proactive assessment of the pertinent laboratory and health care systems, their controls, best practices, and weaknesses. Figure 1.3, "Common Knowledge Management Construct Challenges" below outlines some common knowledge management construct issues that we will work to address as we design a system with proactive data assurance methods.

Figure 1.3. Common Knowledge Management Construct Challenges



Inability to Represent Pertinent Phenomena

As data is entered into a system, it is assigned a code from a knowledge management construct. This is often arbitrary and either matches the best available concept or requires creation of a new local concept.

Inability to Recognize Equivalence

When entering data into a system, existing knowledge management constructs may force end-users to use a single classification system without the ability to link the two and identify a relationship between standards.

Inability to Consistently Classify Information

When transferred, data is viewed from the receiving system's constructs, which can lead to the same data being identified as a new or different concept than the original data. Additionally, different terminology standards occasionally have contradictory classifications for the same concept (treatment vs procedure).

Inability to Reliably Evolve Over Time

Medical terminology standards are updated at an inconsistent rate, which can lead to issues when data is transferred between systems and one construct's concept is updated, but not the other's.

1.7.4. Related Safety Systems Thinking and HRO Work

Subject matter experts and other industry leaders have demonstrated use cases for the feasibility and applicability of a Systems Safety and HRO-based approach to improving data interoperability, reliability, and quality in the health care system. Our work will build upon the findings of these subject matter experts and will incorporate lessons learned to address known issues with current state knowledge management constructs.

1.7.4.1. Increasing the Value of Health Data by Engineering Safety, Quality, and High Reliability into the Data Life Cycle Overview

In this 2022 paper by Dr. Keith E. Campbell et al., the team outlines an Independent Validation and Verification (IV&V) approach that was designed with system safety, HRO principles, and statistical process control techniques. The generalized and iterative IV&V sought to improve the quality and interoperability of data through a three-phased approach, error discovery and assessment, statistical process control, and continuous process improvement.

By identifying and establishing feed forward pathways and baseline metrics for data and knowledge quality, the team provided a use-case for an IV&V approach's ability to statistically improve the quality and interoperability of prescription data. Dr. Campbell and team highlighted a variety of lessons learned that will need to be addressed to continue improvements to health care data, including the need to embed data and informatics SMEs into EHR modernization (EHRM) efforts and to expand HRO principles to Standards organizations and health care system data processes. [16]

1.7.4.2. A Risk-Based Methodology for the Quality Assurance of Healthcare Knowledge Artifacts Overview

Greg Rehwoldt outlines a real-world use case of developing and applying a risk-based methodology to improve the quality of current state knowledge artifacts in this 2018 dissertation. [17] This paper highlights a variety of issues with current state knowledge management, particularly the lack of effective quality assurance for existing knowledge artifacts. Current approaches to quality tend to focus on system specific events or stakeholder mistakes, but as the health care environment has evolved, quality assurance must validate data during creation and as it transitions between systems.

While improving patient safety and data quality are key focuses of this paper, financial theory is also applied to demonstrate how cost and time variables will impact quality assurance. It is critical that data validation and assurance methods are implemented optimally to reduce the risk of hazards or patient safety issues while accounting for organizational constraints and pressures. [17]

1.7.4.3. Capability Maturity Model Integration (CMMI) Background

The Capability Maturity Model Integration for Development (CMMI-DEV) is a risk management model developed at the Carnegie-Mellon University as a maturity assessment for an organization's processes and to support the identification of opportunities to improve management of risk and performance under stress. The model was designed as a process improvement initiative to support gap assessments and identify goals to support measured improvements to high value areas.

The CMMI-DEV examines 22 process areas and ranks an organization's maturity for each area and that area's process capabilities. By examining each area, this model identifies areas where improvements would provide the highest value and allows organizations to tailor their focus and efforts. [18]

1.7.4.4. Our National Rush to the EHR - An Analysis of Current Clinical Information Quality and Data Governance Practices Overview

The Enterprise Clinical Information Maturity (ECIM) framework outlined in this paper builds upon the previously mentioned CMMI-DEV model and supports the development of goals and baseline measurements to improve data quality and standardization. The ECIM framework also considers organization goals and the use of clinical data when evaluating current state processes. This comprehensive approach supports the identification of specific issues or opportunities for data quality and knowledge management improvements.

This paper highlights how current state electronic health information practices are not mature and identifies a variety of recurring issues related to knowledge management and data quality.

1. Electronic Medical Record (EMR) clinical documentation guidelines are poorly standardized and communicated to end-users.

2. The lack of internal data audits across the health care industry results in unknown patient record quality, accuracy, and completeness.

3. Key informatics stakeholders and SMEs are not adequately included in EHRM efforts.

4. Providers do not use current state knowledge management tools, like standardized terminology or classification systems, optimally.

5. Providers must learn to structure and enter clinical information with a standardized process to support data interoperability as natural language processing (NLP) abilities remain limited.

6. Systems lack a standardized approach to terminology and data that leads to many redundant, ambiguous, or improperly mapped terms in EHR data dictionaries. This can often lead to improperly labeled data as information is shared between systems.

7. Data issues that affect multiple records can occur as automated data flows apply terminology maps based on the received map, rather than the original source data. [19]

1.7.5. Safety Systems Thinking and HRO Conclusion

While this chapter has outlined key principles, frameworks, and research that will be used as we design an improved knowledge management system, there is still important research and work that needs to be done. Our team will need to continue researching existing laboratory and health care systems and terminology standards to build upon the years of research and SME experience.

Our system will be designed with Safety Systems Thinking and HRO principles and will support the application of a rigorous and improved data review process that verifies and validates data at each step of information exchange for accuracy, completeness, and reliability. We must determine how to optimally design a system that ensures data from multiple medical terminology standards and EHRs maintains all relevant information and meaning as it seamlessly moves within and between systems. The goal of our work is to optimally improve data accuracy and patient safety, while understanding the real-world pressures of the health care and laboratory systems

1.7.6. Safety Systems Use Case

The Synensys document, "System Safety within Laboratory Data Exchanges", [WILL INSERT LINK TO IKM.DEV WITH THIS PAPER] analyzes the complex challenges in laboratory data exchanges within healthcare, emphasizing the importance of accurate communication, standardized procedures, and contextual information in diagnostics. Various loss scenarios and potential risks are detailed, highlighting the consequences of miscommunication and inconsistency in laboratory data handling. The document is being cited as an example because it illustrates the application of system safety principles in healthcare. Its focus on engineering safety, quality, and reliability into laboratory data exchanges exemplifies the broader theme of implementing systems thinking and high reliability practices, providing a real-world perspective on these critical concepts. [20]

1.8. References

1. Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. 5th ed. Cham, Switzerland: Springer Nature; 2021.

2. Kahn MG, Brown JS, Chun AT, Davidson BN, Meeker D, Ryan PB, Schilling LM, Weiskopf NG, Williams AE, Zozus MN. Transparent reporting of data quality in distributed data networks. EGEMS (Wash DC). 2015 Mar 23;3(1):1052. doi: 10.13063/2327-9214.1052.

3. Elkin, P. Terminology and Terminological Systems. London: Springer; 2012.

4. Bernstam EV, Warner JL, Krauss JC, Ambinder E, Rubinstein WS, Komatsoulis G, Miller RS, Chen JL. Quantitating and assessing interoperability between electronic health records. J Am Med Inform Assoc. 2022 Jan 7:ocab289. doi: 10.1093/jamia/ocab289. Epub ahead of print. PMID: 35015861.

5. Wong A, Otles E, Donnelly JP, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med. 2021;181(8):1065–1070. doi:10.1001/jamainternmed.2021.2626

6. Bodenreider O, Cornet R, Vreeman DJ. Recent Developments in Clinical Terminologies - SNOMED CT, LOINC®, and RxNorm. Yearb Med Inform. 2018 Aug;27(1):129-139. doi: 10.1055/s-0038-1667077. Epub 2018 Aug 29. PMID: 30157516; PMCID: PMC6115234.

7. Logica Health, Health Level Seven International, Vocabulary Working Group. HL7 Standardized Terminology Knowledgebase, Release 1. [Internet]. Creative Commons Attribution 4.0 International (CC BY 4.0); 2021. Available from: HL7 Standardized Terminology Knowledgebase, Release

8. Analysis Normal Form Informative Ballot. HL7 CIMI Work Group. Sept 2019. http://www.hl7.org/documentcenter/public/ballots/2019SEP/downloads/HL7_CIMI_LM_ANF_R1_I1_2019SEP.pdf.

9. Cholan RA, Pappas G, Rehwoldt G, Sills AK, Korte ED, Appleton IK, Scott NM, Rubinstein WS, Brenner SA, Merrick R, Hadden WC, Campbell KE, Waters MS. Encoding laboratory testing data: case studies of the national implementation of HHS requirements and related standards in five laboratories. J Am Med Inform Assoc. 2022 May 25:ocac072. doi: 10.1093/jamia/ocac072. Epub ahead of print. PMID: 35639494.

10. Stram M, Seheult J, Sinard J, et al. ; Members of the Informatics Committee, College of American Pathologists. A survey of LOINC® code selection practices among participants of the College of American Pathologists Coagulation (CGL) and Cardiac Markers (CRT) proficiency testing programs. Arch Pathol Lab Med 2020; 144 (5): 586–96.

11. Stram M, Gigliotti T, Hartman D, et al. Logical observation identifiers names and codes for laboratorians: potential solutions and challenges for interoperability. Arch Pathol Lab Med 2020; 144 (2): 229–39.

12. Lin M, Vreeman D, McDonald C, et al. A characterization of local LOINC® mapping for laboratory tests in three large institutions. Methods Inf Med 2011; 50 (2): 105–14.

13. Bietenbeck A, Boeker M, Schulz S. NPU, loinc, and SNOMED CT: A comparison of terminologies for laboratory results reveals individual advantages and a lack of possibilities to encode interpretive comments [Internet]. De Gruyter. De Gruyter; 2018 [cited 2022Oct19]. Available from: https://www.degruyter.com/document/doi/10.1515/labmed-2018-0103/html?lang=en

14. Lin M, Vreeman D, McDonald C, et al. Auditing consistency and usefulness of LOINC® use among three large institutions—using version spaces for grouping LOINC® codes. J Biomed Inform 2012; 45 (4): 658–66.

15.Enhancing the drug ontology with semantically-rich representation of National Drug codes and RxNorm unique concept identifiers [Internet]. BMC Bioinformatics. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3192-8

16.Campbell KE, Montella D, Brown SH, Powell SM, Campbell SW, Cholan RA, et al. Increasing the Value of Health Data by Engineering Safety, Quality, and High Reliability into the Data Lifecycle 2022.

17.Rehwoldt G. A Risk-Based Methodology for the Quality Assurance of Healthcare Knowledge Artifacts (dissertation). 2018.

18.You C, Sherer T, Mabee D, Jacobs M, Koke J, Boer G, et al. Capability Maturity Model Integration (CMMI), background notes - azure boards [Internet]. Capability Maturity Model Integration (CMMI), background notes - Azure Boards | Microsoft Learn. [cited 2022Dec15]. Available from: https://learn.microsoft.com/en-us/azure/devops/boards/work-items/guidance/cmmi/guidance-background-to-cmmi?view=azure-devops#related-articles

19.Hawry PL, Clouse T, Spisla C, Konieck D. Our National Rush to the EHR - An Analysis of Current Clinical Information Quality and Data Governance Practices 2013.

20.Campbell SW, Case JT, Cholan R, England A, Geary C, Green A, et al. FDA System Safety within Laboratory Data Exchanges End of Base Year Report.

2. Summary of Findings – Current State of RWE Analytics

2.1. Purpose

In an effort to resolve the challenges posed by data interoperability and unlock the full potential of RWD and RWE across the health care system, the SHIELD collaborative was formed. The SHIELD collaborative is a multi-agency/stakeholder public-private initiative to develop and implement collaborative policies and business models to overcome lab interoperability barriers. Due to the complexity and vastness of this issue, SHIELD has embraced an ecosystem perspective that recognizes no single entity has the authority or influence to meaningfully impact the state of lab interoperability. SHIELD stakeholders include in vitro diagnostic (IVD) manufacturers, commercial and clinical center laboratories, professional organizations, EHR vendors, Pew Charitable Trusts, numerous federal agencies, standards development organizations, and patient advocates. SHIELD's overall objective is to ensure lab interoperability such that the same type of device is described the same way throughout EHRs.

The purpose of this Summary of Findings – current state of RWE analytics including but not limited to information retrieval and Knowledge Management and discovery – is to explore the current state of real world data and real world evidence as it pertains to lab data interoperability, FDA market approval and post-market surveillance, public health, clinical care, and related areas.

2.2. Introduction

2.2.1. Executive Summary

RWD and RWE have received significant attention in recent years from the health care community and federal health organizations. The potential value of RWD which can include information from electronic health records, laboratory testing, pharmacy records, environmental testing, and patient self-reports offers the possibility to harness the day-to-day aspects of health care delivery and outcomes as a source of information and evidence (i.e., RWE) that is far broader than typical, "closed loop" clinical trial data sets. The applications of RWD and RWE, both in isolation and in conjunction with traditional clinical trial methodologies, represent immense potential in advancing medical research, developing and evaluating new therapies, assessing post-market safety and efficacy, improving patient care, expanding public health surveillance, and enabling effective pandemic response measures.

Unfortunately, RWD and RWE face numerous obstacles that hinder their reliable widespread use. The realization of RWD's full potential is faced with a variety of challenges, ranging from data quality and bias, lack of standardization, data access challenges, and privacy concerns. One of the most significant challenges is that of data interoperability. Data interoperability does not simply mean the ability to electronically share health information between and across health systems, this is relatively trivial, and is done routinely. Data interoperability implies that health data created in one environment can be "picked up and moved" to another environment and be computationally processed safely and reliably to convey exactly the same meaning as the source institution intended and computed. The movement of a Portable Document Format (PDF) medical report from institution A to institution B is done routinely, but none of the data held within the report is computable, thus not interoperable. Discrete EHR data can be computable within its source system but cannot be electronically exchanged to a second institution and reliably and safely computed to convey the exact same meaning as initially intended because these data, even data between identical EHR systems, do not carry sufficient context to convey appropriate meaning to be safely and reliably reconstructed between system instances. Finally, even data encoded using standardized medical terminology, such is laboratory data, that is electronically exchanged and computable between systems does not safely guarantee the data will be considered to have the exact same meaning between systems. Numerous examples of this phenomenon have been discussed in literature. [1] To realize the potential of RWD which is suitable as RWE for regulatory, research, public health and clinical care purposes, data interoperability must be achieved.

2.3. Real World Data (RWD) and Real World Evidence (RWE) Background

RWD is defined as data relating to patient health status and/or the delivery of health care routine collected from a variety of sources as part of receiving care or daily living. It encompasses most sources of health data derived from common sources such as traditional clinical trial settings, including EHRs, claims and billing activities, disease registries, pharmacy records, as well as non-standard sources such as social media and wearable devices. [2]

Large quantities of clinical data are generated throughout the various stages of an individual's health care journey. When a patient visits a care provider, has their vitals recorded, reports their symptoms, undergoes laboratory tests, receives a prescription, and submits an insurance claim, valuable clinical data is created at each step. The data from each individual stage is stored in a variety of ways. Some data is discrete in nature, some is encoded, and still other data is recorded using natural language or text. In addition, data collected is contained within a siloed data ecosystem based on software system specific parameters that are different from other systems. Though information can be shared between entities to facilitate care from step to step, aggregation and standardization is difficult to reliably perform meaningful analytics without substantial data curation efforts.

Similar to RWD, RWE is defined as clinical evidence about the usage and potential benefits, or risks of medical treatments, technology, drug, or intervention derived from sources other than clinical trials. [3] It refers to analysis of data that has been collected outside the traditional randomized clinical trial (RCT) setting. RWE has immense value in advancing medical research, particularly when deriving the required information in a clinical trial would be unfeasible or unethical. [4]. By design, RWD and RWE are generated independent of a specific research question or study design. This can be a benefit in some cases and a barrier in others. Independent of the research question(s) of interest, aggregating RWD across multiple sites is necessary to establish sufficient representation of the study and control cohorts is difficult based due to data interoperability barriers.

2.4. RWD Applications and Limitations

Despite significant challenges, RWD and RWE present immense untapped potential within public health, research, safety and regulation, and post-market surveillance. A variety of health care entities have taken interest in RWD in recent years, including federal agencies, pharmaceutical companies, medical device manufacturers, and patients. One of the primary applications of RWE thus far has been in post-market surveillance monitoring drug safety and efficacy as well as researching health outcomes that are difficult to study with traditional methods, such as with lesser understood conditions, smaller patient populations, and long follow-up periods. [4] However, during the SARS-CoV-2 pandemic, RWD comprised the primary source of information available to the medical community to understand COVID, evaluate the safety and efficacy of IVD testing, antigen testing, and vaccination safety/efficacy. As an example, the State of Nebraska Department of Epidemiology and the Nebraska Public Health Laboratory operated large numbers of specimen collection centers for COVID testing. Each patient was required to report symptom status for loss of smell or taste prior to testing by antigen or polymerase chain reaction (PCR). Analysis showed that the presence of one or both symptoms had a positive predictive value (PPV) of 80% for COVID. These RWD provided the RWE to support pool testing of specimens to conserve testing reagents during a period of severe test supply shortages. [5]

The use of RWD that is of sufficient quality to be used as RWE for market approval, post-market surveillance of medicines, and in IVD presents a potentially rich environment and supplement to the current gold-standard controlled clinical trial methodologies. To this effect, Congress signed the 21st Century Cures Act (Cures Act) into law in 2016 to help accelerate the development of medical products and bring new innovations to patient populations. [2] The law intends to increase federal focus on evaluating RWD and RWE for regulatory decision making for new approvals, support post-approval study requirements, and evaluate additional indications for approved treatments. [2] Unfortunately, the broad readiness of the U.S. healthcare ecosystem was not, and is not, in place to broadly harness and safely and effectively use the data in the U.S. healthcare domain as captured in the EHR systems, public health agencies, administrative systems, and potentially patient reported information. Such a gap in surveillance, data capture, data exchange, data harmonization and subsequent analytics place the U.S. at risk to repeat the “data scramble” in the event of another emerging disease, as experienced during COVID. Despite existing issues with widescale use of RWD, there are multiple examples of where and how RWD is being used effectively from which future direction can be informed.

2.4.1. Current RWD Uses

There are several national and international efforts to aggregate and expose EHR/RWD for use by researchers and public health scientists. Such efforts include the National Patient-Centered Clinical Research Network (PCORnet), Observational Health Data Sciences and Informatics (OHDSI), National COVID Cohoert Collaborative (N3C) and its potential spin offs. Common to each of these efforts is the extraction of EHR data from participating entities, transforming these data to a common data model, annotating/managing the data with standardized medical terminologies and aggregating entities’ datasets in consolidated or federated mechanisms for analyses.

2.4.2. National Covid Cohort Collaborative (N3C)

N3C is one of the largest collections of COVID-19 clinical data in the United States. The cloud-based collection of secure and de-identified data is part of the N3C’s vision to use RWD for advancing knowledge to address COVID-19. N3C aims to create a data pipeline to harmonize EHR into a common data model, allowing the research and clinical communities better access to reliable COVID-19 data to generate novel approaches to address the impact of COVID-19 and mitigate future pandemics.

The National Center for Advancing Translational Sciences (NCATS) N3C Data Enclave contains RWD as collected in contributing sites’ EHR systems on COVID-19 positive patients and patients with symptoms consistent with COVID-19, Severe acute respiratory syndrome coronavirus 1 (SARS 1), Middle East Respiratory Syndrome (MERS) and H1N1 flu (a.k.a. Swine flu). De-identified data is contributed monthly and is augmented by public health laboratory data and whole genome sequence data of paired patient/specimen isolates.

NCATS N3C Data Enclave data sets are well described and include data levels ranging from Limited Data Sets, De-identified Data Sets, and Synthetic Data sets all with in depth descriptions of the data and eligibility and access requirements. NCATS guarantees the confidentiality and security of all data in the enclave, and oversight is achieved through user registration, federated login, data use agreements and requests. The cloud-based environment is Federal Risk and Authorization Management Program (FedRAMP) certified. To further protect the security of the enclave, downloading data is impermissible and all work must be done within the enclave.

N3C has paved the way as a successful model for other data enclaves. Data codification is well described, and the data harmonization team ingests limited data sets and transforms them according to a harmonized Observational Medical Outcomes Partnership (OMOP) analytics data set. Data and code provenance are tracked through the platform. Privacy Preserving Record Linkage (PPRL) is used to securely connect records while maintaining the anonymity of an individual across multiple records and may enhance

COVID-19 RWD research in N3C through de-duplication of patient records, linking patient records from different sources, and cohort discovery. Organizations submitting data must have an approved Data Transfer Agreement (DTA), and an optional Linkage Honest Broker Agreement (LHBA), through which organizations maintain control over their data. Data analyses must be performed within the platform and is supported through R, Python, and Slate. R packages include SciPy and scikit-learn and are pre-installed, and others can be added by contacting the data enclave support team. Overall, the enclave provides a secure and focused space for research collaboration and has proven its impact through hundreds of peer reviewed, scientific publications.

The successes of the N3C data enclave are not without limitations. Data sets are geographically limited to the U.S., and only patients with known COVID-19 positive tests are included. This may preclude large populations of patients with alternative methods of diagnosis from being included. Those whose results were not recorded in an EHR but were performed else where such as at a Public Health Lab (PHL), or who had viral sequencing performed are not included. Control cohort methods are not well described and thus there is lacking any RWD on non-covid patients. The strength of security does not come without some downsides and can limit the participation and quality of data contributed by participants. There is no method for testing for data equivalence between sites, and data is not in real-time or near real-time. The enclave would benefit from expanding its horizon beyond COVID-19 to other conditions and general disease surveillance. [6]

2.4.3. Patient-Centered Outcomes Research Institute (PCORI) and National Patient-Centered Clinical Research Network (PCORnet)

The Patient-Centered Outcomes Research Institute (PCORI) is an independent, nonprofit, and leading funder of patient-centered comparative clinical effectiveness research in the United States. The primary goal of PCORI is to increases useful information, speed the uptake and use of information, and influence funded research to be more patient centered. PCORI funded research includes comparative clinical effectiveness research (CER) which helps patients and other stakeholders to compare two or more medical treatments, and other research that focuses on patient centered outcomes. PCORI funds researchers to share their data sets and documentation for reanalysis and reuse. PCORI encourages use of data from the studies and provides funds to allow other researchers to verify and build on those findings to generate new evidence and make this information available to healthcare decision makers. PCORI funded more than 2,000 research studies and related projects, including project that support the methods and capacity for conducting that research. [7] PCORI approved $1 million to fund the development of automated ways to identify people more efficiently and accurately identify people with rare diseases via EHRs, enabling treatments to start sooner and helping researchers offer individuals opportunities to participate in studies. [7]

In 2013, PCORI funded development of the PCORnet, a national resource for health data and research expertise, with the goal of providing patient centered outcomes research in a timely manner. [8] PCORnet is a well established and fully integrated network with a research ecosystem uniquely networked to offer RWE essential to support major public health issues. [7] The Primary goal of PCORnet is to expand the role of patients and care providers in the research process and leverage EHR and registries to conduct research that is inexpensive, faster, and patient-centered in its design. [9]

Data is accessible via the PCORnet distributed network drawn from millions of EHRs with growing links to patient-reported and payer data to create a standardized data set that facilitates large-scale research. These data are accessible via Clinical Research Networks (CRNs) and facilitated by a coordinating center. [10] PCORnet CRNs are integrated delivery systems with one or more hospitals, outpatient clinics (primary and specialty care), and emergency departments. Across the 9 CRNs, there are 251 institutions that are organized into 61 data contributors. A data contributor manages the PCORnet data mart for one or more institutions. Overall, PCORnet Clinical Research Networks currently include 337 hospitals, 169,695 physicians, 3,564 primary care practices, 338 emergency departments, and 1,024 community clinics serving medically underserved populations. These healthcare institutions and their clinicians serve as a diverse set of clinical trial sites for pragmatic research conducted in everyday clinical care settings. [8]

In PCORnet, 66 million patients' health records are available for observational studies and provide vast scale to power research for conditions affecting even small numbers of people. Data quality must pass rigorous standards that meet conformance, completeness, plausibility, and persistence criteria. Every network partner operates under the shared Common Data Model, which standardizes data points from across the network into a common data format on which users can query. Data security in ensured by not aggregating it all into a single data pool but keeping it protected behind network partner's firewalls. Queries and responses are controlled through a secure Distributed Research Network Portal. Although the data is safe, a limitation of this approach is that queries are returned with answers, not data. Therefore, a researcher does not have access to the data.

The scope of PCORnet-accessible data, plus its embedded expertise and patient insights, makes the Network a powerful tool for researchers. Studies supported by PCORnet infrastructure answer clinical questions that impact patients' lives, including 2 recent studies that leverage PCORnet to study which populations are impacted by Long COVID and highlights PCORI-funded research results. [7, 11]

2.4.4. Observational Health Data Sciences and Informatics (OHDSI)

Founded in 2014, the OHDSI is an interdisciplinary collaborative that aims to maximize value in health data through large-scale analytics and a common data model. The stated mission of OHDSI is "to improve health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care". In terms of scale, the program consists of over 3,000 contributors (including universities, private companies, government agencies, and others) with data records that represent 810 million unique individuals. OHDSI works to the benefit of patients, providers, and researchers as well as larger organizations such as hospital systems and government agencies. The solutions and products that OHDSI develops are all open source. [12]

Successes

OHDSI has made significant progress in the 8+ years since its founding, emerging from the legacy OMOP program. Among its largest accomplishments, the collaborative has pieced together a federated database with patient records amounting to about 12 percent of the world's population. From a publicity standpoint, OHDSI has been referenced by the European Medicines Agency and other policy makers as well as journals, including Journal of the American Medical Association (JAMA).

OHDSI organizes into working groups and currently operates 27 of them, ranging in topic from focus regions (Asia-Pacific and Latin America regions) to specific care areas (oncology, psychiatry). In addition, they have working groups that focus on FHIR, the OMOP Common Data Model (CDM), and Natural Language Processing (NLP).

In April 2022, OHDSI and SNOMED International formalized a five-year collaborative agreement in which OHDSI and the associated user communities will receive "comprehensive ontologies on specific healthcare domains and content such as devices, social determinants of health, disease severity scores and modifiers of cancers, as well as better concept definitions and resolutions of composite concepts in large-scale observational research." While this is an enormous benefit to OHDSI, the benefit to SNOMED International manifests in the form of user data from interaction with their product. OHDSI also has an existing agreement with HL7 to integrate FHIR with the OMOP common data model in hopes of improving abilities to track and share data in the healthcare and research industries.

OHDSI has ventured into product development, including the Health Analytics Data-to-Evidence Suite (HADES), which is a set of open-source R packages specifically focusing on population level estimations, patient-level prediction, cohort construction, and evidence quality. They have also created ATLAS, a web-based application that supports design and execution of observational analyses to generate RWE from patient level data. [13]

Shortcomings

While OHDSI has demonstrated significant value, there are obstacles limiting its applications. OHDSI's broad reach and many collaborative partnerships add value to the completeness and generalizability of information, the fact that the data is largely international complicates its use in federal regulatory applications. Similarly, the structure of multiple disparate contributors restrictions data flexibility and transparency. The data is fully and irreversibly de-identified, which can limit analyses that require more individual or time stamped data, as well as complicating retrospective quality control. The lack of transparency is exacerbated by the fluidity of medical terminology, as the rigidity of black-box data wrangling limit the agility required to respond to standard terminology changes.

One large consideration with OHDSI data is that of correlation versus causation. The nature of OHDSI's data collection lends itself well towards evaluating trends, identifying associations, and investigating intersections of drugs, treatments, and diseases. While a number of variables for multiple associations can be examined, it is a significant challenge to assess potential causation from observed correlation. Data attributes that enable causation analysis, such as timestamping and longitudinality, are lacking with OHDSI. These limitations demonstrate common sacrifices that are often made when prioritizing data standardization and interoperability, especially when the privacy requirements of medical data require a base level of manipulation to obscure incoming data.

2.4.5. Electronic Medical Records and Genomics (eMERGE)

The Electronic Medical Records and Genomics (eMERGE) Network is a National Human Genome Research Institute (NHGRI) funded organization that combines genetic data with EHRs to support genomic medicine research. Genetic data is collected and stored within biorepositories at nine eMERGE sites, combined with EHRs, and then researched to determine how an individual's genetics increase their disease risk or make them better suited for specific medications. Since its initiation in 2007, eMERGE has grown to develop 777 publications, discover 68 phenotypes, and encompass over 136,000 network participants. [14]

eMERGE conducts genome wide association studies (GWAS) with EHRs to sequence patients' genes, evaluate disease risk, return actionable results to clinicians and patients, and contribute findings to patient EHRs. [15] One facet of the research is using broad GWAS results to calculate disease risk in values known as polygenic risk scores (PRS). Additional work is performed to compare International Classification of Diseases, Ninth Revision (ICD-9) codes across network sites to assess comparability of large populations taken from different institutions, evaluate privacy concerns surrounding patient consent and continuing genomic research, and identify optimal methods for the development and cross-site deployment of algorithms to identify case and control cohorts. [6] The overall goals of eMERGE, in addition to advancing research regarding genetic disease and treatment associations, are to develop analytical tools, identify and communicate genomic research best practices, and return genomic results to patients and practitioners to improve clinical care. [14]

Since its inception, eMERGE has provided valuable insights into both genomics research and methodologies to effectively analyze data across various institutions. Early into eMERGE's work, investigators found projects had better outcomes when performed across the entire network versus within individual sites. Phenotype algorithms were typically developed within one site and then deployed at secondary sites where observed issues could guide iterative revisions to make the algorithm more robust and generalizable. Moreover, incorporating multiple networks allowed for larger population sizes and added statistical power. For example, eMERGE's expansion to include pediatric sites in addition to the existing adult data was challenging but enabled researchers to examine heritable diseases that presented in childhood and continued into adulthood. This type of analysis would have been unfeasible for one site alone. [16] Data quality is of paramount importance when collaborating across multiple sites. eMERGE has demonstrated effective quality assurance procedures by ensuring common definitions and coding exists for phenotype variables being deposited into the database of Genotypes and Phenotypes (dpGaP) and facilitating Quality Assurance/Quality Control (QA/QC) of data as it comes in. [17]

Similarly, the size, scope, and diversity of eMERGE's data allows the network to develop and execute additional analyses as data evolves and medical interpretations change. Cultivating a diverse biobank and a rich EHR linkage allows for cost-effective longitudinal genomic studies. Moreover, the continued collection of clinical data enriches the depth and value of the network while incurring minimal added costs. [16] The ability to build upon existing data is particularly valuable in fields where scientific understanding is still evolving, as with genomics and novel diseases or treatments. Despite the challenges inherent with multi-site collaboration, eMERGE 12 has demonstrated that interoperability risk can be mitigated to allow for robust and powerful analytics.

Despite its many successes, the eMERGE network experienced several challenges. The personal nature of genetic data emphasizes the requirements for patient privacy and de-identification. When aggregating data across various timepoints and institutions, there may be additional timepoints where privacy can be compromised. As such, proper de-identification must be prioritized at each phase of data collection. Also inherent to genetic data is the issue of patient consent. Within eMERGE's ever-evolving network, ethical questions exist regarding whether an individual's original consent allows for the sharing of data with database of Genotypes and Phenotypes (dpGaP) or expanding research to include additional phenotypes. [14] As data in consolidated systems grows, applications for its use evolve, and metadata analyses become possible, focus must remain on how to ethically honor patient consent as scope changes.

One of the primary obstacles eMERGE experienced was that of data interoperability. While data was checked for quality and common data codes were established, limitations within the transferability of information persist. There is no guarantee of data completeness within EHRs, especially when care is received outside the primary provider's system. Additionally, while nomenclatures and ontologies such as SNOMED-CT® and LOINC® exist to uniformly represent medical concepts, there are no established processes to facilitate the secure, generalizable, and interoperable data exchange required to provide continued care between different health care systems. [16] This issue is exacerbated by the ever-changing knowledge and medical interpretations surrounding genomic data as any interoperability solutions must remain flexible to keep up as the field evolves. eMERGE developed various practices to address these challenges and communicated findings to the larger genomic research community. However, a consensus on data standards and data harmonization, while laborious, is required to perform multi-site analytics effectively. [14]

2.4.6. Sentinel

The FDA launched the Sentinel Initiative to help develop new ways of evaluating the safety of approved drugs, devices, and treatments in response to the FDA Amendments Act (FDAAA) of 2007 that mandated the use of EHR to perform safety assessments more effectively. [18]

The primary goal of the Sentinel system is to inform FDA decision-making using RWE generated from a distributed data network, a common data model, curated RWD, and flexible analytic tools. [19]

Sentinel collects EHRs that cover covers approximately 700 million person-years of longitudinal observation time, much of which comes from insurance claims and pharmacy records. [17] It pulls from a network of established data partners containing over 170 inpatient facilities and some of the nation's largest health insurance providers, including Aetna, Anthem, Humana, and Kaiser Permanente. [18] The Sentinel System is not a data repository, but rather a distributed network in which each data partner maintains ownership and operational control of their data. In addition to the collected partner data, which is delivered quarterly or annually, Sentinel can access select full-text medical records, inpatient data, including lab results from the nation’s largest hospital network that covers approximately five percent of inpatient care and over 10 billion pharmacy and medication encounters. [17]

Data is transformed into the Sentinel Common Data Model (SCDM) format for standardization, allowing for efficient analysis across data sources. Sentinel has developed a collection of flexible computer programs to run routine queries and perform analyses that inform FDA decision making, ranging from medication utilization patterns among patient cohorts to evaluation of drug and device safety.

The Sentinel System presents significant potential with respect to the research, approval, and maintenance of medical treatments. The combination of expansive standardized data and flexible analytics allows the FDA to generate RWE and associated analytics rapidly and reliably with relative ease, all while minimizing risk to patient privacy. The capabilities of the Sentinel System are varied, including the ability to access full-text medical records, support RCTs, and streamline the collection of patient data from mobile devices. [17] This system has enhanced the FDA’s central purpose by allowing it to design and execute post-market studies that would have previously been required of medical product sponsors. [19] Similarly, the FDA maintains authority to mandate companies perform post-approval studies when potential safety concerns arise, and Sentinel could be used to drive those decisions. Sentinel’s Active Risk Identification and Analysis (ARIA) System evaluates the safety of approved medical devices and drugs and has performed hundreds of such analyses to date. [22] The operations of ARIA are of particular significance when considering the known limitations of the FDA’s Adverse Event Reporting System (FAERS), which the FDA itself estimates to have received only one to ten percent of adverse events that occur. [23]

Although the Sentinel System has already provided measurable impact since its implementation, it has some desirable holes. The most notable one reveals the limitations inherent to the data sources Sentinel accesses. The majority of Sentinel’s data is derived from administrative claims data, which often lacks the precision and accuracy required to perform sound analytics on health outcomes. Claims data may diagnose and code conditions unreliably and often in an oversimplified manner, a complication that is further confounded by the prioritization of insurance coverage over medical precision. While a heart attack may be diagnosed and billed accurately, outcomes like psychiatric side effects, digestive problems, cognitive impairment, and countless others are often reflected unreliably or not at all. [17]

The data within the Sentinel System exhibits additional flaws. There is a strong potential for selection bias, leading to unknown generalizability to underrepresented populations, particularly those lacking health insurance. Even among included individuals, data can vary. For example, follow-up times within commercial health care plans are often shorter than those within Medicare and Medicaid, making it difficult to perform parallel longitudinal studies. Data is delivered to Sentinel quarterly or annually with a lag time of six to nine months, limiting its usefulness when addressing rapidly developing situations like COVID-19. Finally, data consistency and interoperability are additional concerns, as data can vary in content, accessibility, and completeness across the multiple data partners. [17]

There exist functional gaps that limit potential applications, but once improved, the Sentinel System can further improve as a powerful centralized tool for RWE generation, regulation, analytics, research, and patient care. As with all applications of RWD, strategies for how to best capture, maintain, and analyze RWD and RWE must continue to develop and expand.

2.5. Learnings

Electronic health records have been in use in the U.S. since the 1970’s. However, the rush to implement the EHR nationwide did not occur until the introduction of the Meaningful Use (MU) in 2009. [24] The EHR has provided a large pool of patient records and healthcare data from across the globe for secondary research. As of 2021, 85% of physicians in the US electronically recorded social determinants of health (SDOH) data and 97% electronically recorded broad determinants of health (BDOH) data. Rates of electronic recording of SDOH and BDOH data were slightly higher among primary care physicians, with 89%

of primary care physicians electronically recording SDOH data and 98% electronically recording BDOH data in 2021.

In the past, research relied upon billing data from the UB-04 billing form as the source of structured data for research purposes. The UB-04 medical billing form is the standard claim form that healthcare organization use for the billing of inpatient or outpatient medical and mental health claims. [25-26] Although developed by CMS, the form has become the standard form used by all insurance carriers. This form captures all data in a structured format and captures diagnoses, treatments and procedures, type of admission, medications, and demographic information. While the UB-04 and similar administrative data have been considered rich sources of data for research and can be readily de-identified, shared, and stored, much valuable information is lost. Coding systems used for billing purposes dilute data by use of non-specific coding systems which bundle similar conditions or procedures under the same code. Diagnosis data is captured using ICD10CM codes which will bundle similar diagnoses under the same code.

Limitations to using RWD cannot be overlooked, ignored, or dismissed. Aggregation of data from multiple sources may impact the availability of data in real-time. Time and effort to clean or harmonize data from multiple or disparate sources increases time between availability of RWD for researchers compared to when the data was collected. This delay may impact the accuracy or quality of research data, meaning researchers are analyzing data no longer true of a given scenario.

Patient reported data is information reported directly by patients to the healthcare providers or researchers. Patient-reported data provides a rich source of data for researchers yet is challenging to collect and share due to a lack of data standards. Data standardization is challenged by the various sources of patient reported data from paper or digital surveys, EHR portals, medical devices, or other sources.

Ideally RWD for RWE is extracted directly from clinical documentation. EHR vendor-agnostic data and interoperability standards such as Health Information Exchanges (HIEs), MU, HL7, and FHIR were created to encourage or mandate data interoperability across healthcare organizations. Yet healthcare organizations continue to consider their healthcare data as a proprietary asset. Adding complexity to RWD is inconsistent and/or proprietary implementation health information technologies to meet the needs and requests of the healthcare organization. Data analysts or data wranglers are charged with making data understandable. Adding complexity to this challenge is inconsistent and/or proprietary implementation of each EHR implementation to meet the needs and requests of the healthcare organization. This inconsistent and/or proprietary database design and EHR implementation impacts query methodology for secondary use of clinical data across healthcare systems. Each EHR vendor database structure is different and proprietary. EHR vendors did not provide standardized terminology implementation support, resulting in overreliance on third party terminology vendors to provide the linkages between “provider-friendly” terminology and standardized clinical terminologies. These ‘provider friendly terms’ are often mapped to a broader than or inaccurate standardized language concept. Insight to the quality of these mappings is opaque and error prone.

Clinical documentation by nursing or allied health professionals is rarely considered as a source of data for research. Yet, their documentation may yield a rich source for data. Additionally, their documentation is more likely to be captured in a structured manner in the form of structured (and often researched and validated) assessments and flowsheets. Ideally, RWD for RWE is extracted directly from clinical documentation.

To address this issue, research databases were created to allow researchers access clinical data without actual access to the EHR. EHR database managers utilize “extract, transform, load” (ETL) to move data from EHR database(s) or other sources to a unified repository—typically a data warehouse. However, the ETL process relies on structured and standardized data. Much of the rich, useful data within a patient’s EMR is found beyond the reach of ETL. Narrative text with progress notes, diagnostic imaging and surgical reports, scanned pdf documents and free text fields contain a rich source of data to support RWE.

Manual data curation and NLP are possible solutions to extract data from unstructured text. Manual data curation is a labor-intensive, time-consuming process involving human review of clinical documentation to read, interpret and extract information into a structured database. Well-defined heuristics and quality assurance processes are required to ensure accurate and consistent data extraction. Data available from data curation may not meet time-sensitive requirements due to the time and effort required for data extraction and any associated QA processes. NLP is a form of AI that can be used for computers to understand human generated language and process the data to understand the full meaning of the original intent. NLP combines algorithms with machine learning and models to correctly 16 extract and label data and assign meaning. Today, the most common models used to support NLP include convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Yet, even the most well-defined data curation heuristics and processes may find clinician prose a challenge. Clinical findings with a diagnostic imaging or pathology report may not be expressed using explicit language. The actual clinical finding must be inferred from the clinical prose used by the provider in their report. Data curators or NLP may be challenged to interpret the actual clinical finding being described. Similarly, the presence or absence of a clinical finding may be hidden in the clinical prose as well.

2.6. References

1. Stram M, Gigliotti T, Hartman D, Pitkus A, Huff SM, Riben M, et al. Logical Observation Identifiers Names and Codes for Laboratorians. Archives of Pathology & Laboratory Medicine. 2020 Feb;144(2):229–39.

2. Office of the Commissioner. Real-World Evidence [Internet]. U.S. Food and Drug Administration. 2019. Available from: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

3. U.S. Food and Drug Administration. 21st Century Cures Act [Internet]. U.S. Food and Drug Administration. 2016. Available from: https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/21st-century-cures-act

4. Beaulieu#Jones BK, Finlayson SG, Yuan W, Altman RB, Kohane IS, Prasad V, et al. Examining the Use of Real#World Evidence in the Regulatory Process. Clinical Pharmacology & Therapeutics. 2019 Nov 14;107(4):843–52.

5. Abdalhamid B, Bilder CR, McCutchen EL, Hinrichs SH, Koepsell SA, Iwen PC. Assessment of Specimen Pooling to Conserve SARS CoV-2 Testing Resources. American Journal of Clinical Pathology. 2020 Apr 18;153(6):715–8.

6. National Center for Advancing Translational Science. National COVID Cohort Collaborative [Internet]. U.S. Department of Health and Human Services. NCATS 2022. Available from: National COVID Cohort Collaborative (N3C) | National Center for Advancing Translational Sciences (nih.gov)

7. Patient-Centered Outcomes Research Institute. Highlights of PCORI-Funded Research Results [Internet]. PCORI 2022. Available from: Highlights of PCORI-Funded Research Results | PCORI

8. Forrest CB, McTigue KM, Hernandez AF, Cohen LW, Cruz H, Haynes K, et al. PCORnet® 2020: current state, accomplishments, and future directions. Journal of Clinical Epidemiology [Internet]. 2021 Jan [cited 2022 Dec 20];129:60–7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7521354/

9. Timbie JW, Rudin RS, Towe VL, Chen EK, Hunter LE, Case SR, et al. National Patient-Centered Clinical Research Network (PCORnet) Phase I: Final Evaluation Report [Internet]. RAND Corporation About this book; 2015 [cited 2022 Dec 20]. Available from: https://www.jstor.org/stable/10.7249/j.ctt19rmdhd.10

10. Data [Internet]. The National Patient-Centered Clinical Research Network. [cited 2022 Dec 20]. Available from: https://pcornet.org/data/

11.Home [Internet]. The National Patient-Centered Clinical Research Network. Available from: https://pcornet.org/

12.Who We Are – OHDSI [Internet]. [cited 2022 Dec 20]. Available from: https://ohdsi.org/who-we-are/

13.OHDSI – Observational Health Data Sciences and Informatics [Internet]. [cited 2022 Dec 20]. Available from: https://www.ohdsi.org/

14.Welcome to eMerge [Internet]. [cited 2022 Dec 20]. Available from: https://emerge-network.org/about-emerge/

15.Electronic Medical Records and Genomics (eMERGE) Network [Internet]. Genome.gov. [cited 2022 Dec 20]. Available from: https://www.genome.gov/Funded-Programs-Projects/Electronic-Medical-Records-and-Genomics-Network-eMERGE#history

16.Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genetics in Medicine. 2013 Jun 6;15(10):761–71.

17.McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Medical Genomics. 2011 Jan 26;4(1).

18.About the Food and Drug Administration (FDA) Sentinel Initiative | Sentinel Initiative [Internet]. www.sentinelinitiative.org. Available from: https://www.sentinelinitiative.org/about

19.Brown JS, Maro JC, Nguyen M, Ball R. Using and improving distributed data networks to generate actionable evidence: the case of real-world outcomes in the Food and Drug Administration's Sentinel system. Journal of the American Medical Informatics Association. 2020 Apr 11;27(5):793–7.

20.Adimadhyam S, Barreto EF, Cocoros NM, Toh S, Brown JS, Maro JC, et al. Leveraging the Capabilities of the FDA's Sentinel System To Improve Kidney Care. Journal of the American Society of Nephrology. 2020 Oct 19;31(11):2506–16.

21.Findlay S. The FDA's Sentinel Initiative [Internet]. Health Affairs. 2015 [cited 2022 Dec 20]. Available from: https://www.healthaffairs.org/do/10.1377/hpb20150604.936915/

22.Drug Studies | Sentinel Initiative [Internet]. www.sentinelinitiative.org. [cited 2022 Dec 20]. Available from: https://www.sentinelinitiative.org/studies/drugs

23.World of Drug Safety Module [Internet]. World of Drug Safety Unit List: Overview Of Drug Safety. [cited 2022Dec16]. Available from: https://www.accessdata.fda.gov/scripts/cderworld/index.cfm?action=drugsafety%3Amain&unit=1&lesson=1&topic=8&page=

24.The American Recovery and Reinvestment Act, enacted in February 2009, includes many measures to modernize our nation's infrastructure, one of which is the Health Information Technology for Economic and Clinical Health (HITECH) Act. The HITECH Act supports the concept of meaningful use (MU)

25.National Uniform Billing Committee Official Data Specifications Manual. National Uniform Billing Committee (UBC) [Internet]. www.nubc.org. [cited 2022 Dec 20]. Available from: https://www.nubc.org/subscription-information.

26.Professional paper claim form (CMS-1500) | CMS [Internet]. www.cms.gov. Available from: https://www.cms.gov/Medicare/Billing/ElectronicBillingEDITrans/1500