Achieving Data Liquidity: Lessons Learned from Analysis of 38 Clinical Registries (The Duke-Pew Data Interoperability Project)

Authors: James E. Tcheng, MD[1]; Joseph P. Drozda Jr., MD[2]; Davera Gabriel, RN[1]; Anne Heath[1]; Rebecca W. Wilgus, RN MSN[1]; Mary Williams[1]; Thomas A. Windle[3]; John R. Windle, MD[3].

[1]Duke Clinical Research Institute, Durham, North Carolina; [2]Sisters of Mercy Health System, St. Louis Missouri; [3]University of Nebraska Medical Center, Omaha, Nebraska

Abstract

Background: To assess the current state of clinical data interoperability, we evaluated the use of data standards across 38 large professional society registries.

Methods: The analysis included 4 primary components: 1) environmental scan, 2) abstraction and cross-tabulation of clinical concepts and corresponding data elements from registry case report forms, dictionaries, and / or data models, 3) cross-tabulation of same across national common data models, and 4) specifying data element metadata to achieve native data interoperability.

Results: The registry analysis identified approximately 50 core clinical concepts. None were captured using the same data representation across all registries, and there was little implementation of data standards. To improve technical implementation, we specified 13 key metadata for each concept to be used to achieve data consistency.

Conclusion: The registry community has not benefitted from and does not contribute to interoperability efforts. A common, authoritative process to specify and implement common data elements is greatly needed.

Introduction and Background:

Clinical data interoperability, or the capacity to exchange high-quality, clinically relevant information as data from one system to another, has largely failed. Proprietary processes and custom solutions are required at each point of data exchange to extract, transform, and load data from one system to the next. While healthcare ontologies have been developed to support interoperability, these ontologies are positioned as secondary encoding schemas and are not usable natively as clinical vocabularies. While electronic health record (EHR) systems are now ubiquitous, healthcare continues to largely capture clinical information as analog text requiring data abstraction for computational purposes. The transformation and movement of clinical data from one system to another is expensive, labor intensive, and prone to semantic inaccuracy.

The capture of clinical data for submission to registries is illustrative. Clinical registries collect baseline and longitudinal data to evaluate outcomes, assess care performance and processes, and facilitate process improvement in cohorts of patients defined by a disease state, condition or exposure to a medical product. Registries have been positioned as a critical source of real-world evidence to advance clinical and regulatory science.[1] A common source of data for clinical registries is the EHR. However, the submission of data to registries depends largely on the manual abstraction of data from the EHR and manual re-keying of data into registry data collection systems, termed "swivel chair interoperability'.[2,3]

To assess the current state of interoperability with respect to registries, we evaluated the adoption and use of data standards by 38 national clinical registries in a project termed "Improving Healthcare Data Interoperability" sponsored by the Pew Charitable Trusts. The hypothesis was that data liquidity had not been achieved in the registry domain, and that *native data interoperability* spanning clinical documentation systems and registry database systems would provide the best pathway to accomplishing data liquidity. Specifically, we evaluated the current status of the adoption of data standards by clinical registries and created technical (database programming) specifications for a set of more than 50 clinical concepts commonly shared across registries. We evaluated these clinical concepts in the context of existing work including the federal Common Clinical Dataset and the draft 2018 US Core Data for Interoperability, HL7 Fast Healthcare Interoperability Resources, the work of standards organizations such as LOINC, and common data models including OMOP, PCORnet, and SENTINEL. The presentation will share the results of this project and identify suggested next steps.

The project had 3 Aims:

1) Evaluate registry case report forms, registry dictionaries, and / or registry data model artifacts from clinical professional society registries representing a minimum of 20 clinical specialties and patient populations, identifying the core (shared) common clinical concepts collected across the majority of registries
2) Compare the data elements used by the registries to capture the core common clinical concepts in the context of the 'big 5' healthcare data standards (SNOMED-CT, LOINC, RxNorm, ICD-10, CPT), developing the minimum metadata needed by database developers to natively implement a consistent technical representation of the concepts
3) Develop a roadmap for the transformations needed to achieve native data interoperability of a common clinical dataset across EHR systems, registries, and national data models

Methods:

We solicited participation in the project through broadcast communications to registry members of two registry associations (National Quality Registry Network of the PCPI and the Council of Medical Specialty Societies) and personal communications to a small (<10) number of professional societies. Approximately 75 registries were solicited on our behalf by the two associations (we did not manage the mailing lists per se), with our personal appeals overlapping the association-managed solicitations. Participation was voluntary; all who agreed to participate were included in the analysis. All registries were solicited several times. There were no declinations – the only exclusion was for failure to respond to our invitation. Registries agreeing to participate submitted blank registry case report forms (CRFs), data dictionaries, and / or registry data model representations to the Informatics Group of the Duke Clinical Research Institute, Durham, NC. For national data models, artifacts were sourced from publicly

available content in May-June, 2018. All artifacts were anonymized and kept confidential for purposes of analysis and reporting.

The process we followed is illustrated in Figure 1. We first analyzed the registry artifacts. Clinical concepts were tabulated and the corresponding data elements abstracted from the registry artifacts. We pre-selected for concepts that might be expected to be found in multiple registries (e.g., patient demographics, physical exam findings, procedures, medications, major outcomes). We identified those found across multiple registries as the candidate concepts. Data elements unique to a specific disease or procedure were intentionally excluded from the analysis. From this work, we determined that the data elements of interest could be grouped into the following domains: identifiers / demographics, comorbidities, common physical exam findings, procedure information, medications, lab results and patient outcomes. Attributes assigned by each registry to each data element were recorded in a cross-tabulation. Within each domain, the similarities and differences in data elements across registries were evaluated and a 'match status' (degree of concordance) was assigned (identical, nearly identical, no match). Identical concordance was defined as an exact match of data element label and permissible values. Nearly identical concordance was defined as representation of the concepts (particularly permissible values) in identical semantics without an exact match of permissible values (e.g., true / false versus yes / no). Similar concordance reflected variability in the representation of the concept and / or permissible values such that complete translation was not possible (e.g., yes / no versus yes / no / not available). No match indicated the inability to completely translate in terms of semantic meaning from one representation to another. Identical and nearly identical were grouped as "concordant" (i.e., semantically interoperable), while any disagreement was classified as "discordant" (i.e., not concordant).

A similar approach was applied to data elements from the federal Common Clinical Dataset along with national data models including the Observational Medical Outcomes Partnership (OMOP), the FDA SENTINEL initiative, and Patient Centered Outcomes Research Network (PCORnet). Again, similarities and differences in data type, semantic meaning, permissible values, etc., were noted and a concordance status was assigned per the definitions above.

With identification of the set of core common clinical concepts, we next developed metadata (e.g., clinical definition, data type, permissible values, coded permissible values, and terminology bindings) for each of the concepts, focusing on only the metadata required to fully qualify each clinical concept for the purposes of building the concepts as interoperable data in databases. To the greatest extent possible, we relied and prioritized published clinical data standards along with the corresponding HL7 FHIR resource for specific metadata content.

We socialized both the Methods and Recommendations via two stakeholder webinars, an in-person meeting held at the Pew Charitable Trusts headquarters in Washington, DC, and an electronic survey (Qualtrics). Feedback from each was incorporated into the final project artifacts. Additional feedback received from direct communications between the project team and stakeholders was also incorporated. Proceedings from the webinars and in-person meeting were analyzed using NVivo (QSR International), summarized, and incorporated where possible. Recordings and transcripts, along with project work products, are available online at: https://dcri.org/registry-data-standards.

Results:

Of the registry owners solicited, 38 agreed to participate. There were no declinations (only non-responses). The artifacts (case report forms, data dictionaries, and / or data models) provided by the 38 registry owners were sufficient to perform the analysis across all 38 registries. Approximately 50 data elements were identified as candidates for the comparative evaluation and metadata development. The original intent was to identify clinical concepts common to 50% or more of the participating registries, however, a very limited number of clinical concepts (sex, date of birth, medications, laboratory results) met this criterion. We therefore elected to use a much lower threshold (20%), deriving the following concepts to be explicitly developed: Patient Name, Date of Birth, Sex, Race, Ethnicity, Procedures, Unique Device Identifier, Vital Signs (height, length, weight, blood pressure, pulse), Lab Results (via a separate model), Medications (via a separate model), Care Team (physician only), Smoking Status (via a separate model), Alcohol Use, Substance Use, and Vital Status (Table 1).

Surprisingly few clinical concepts were collected with precisely the same data element representation from any one registry to the next; there were no concepts captured exactly the same way (i.e., concordant) across all 38 registries. Similarly, common data elements in the national common data models were also surprisingly discordant, with the technical (database) representations appearing to be derived from convenience rather than conformance with a particular data standard. An illustrative example is the clinical concept of sex. Of the 21 registries capturing this information, all intended semantically to reflect the biological concept of birth sex (not the social construct of "gender"). Yet one third (n=7) of registries assigned "Gender" or "Patient Gender" as the data element label. Despite a reasonably well-established value set, there were 6 primary variations of the value set across the registries: Male | Female, 3 sub-variations of Male | Female | Unknown, Male | Female | Other | Unknown, Male | Female | Undifferentiated, Male | Female | Unknown/Missing, and unspecified text (Table 2). Another example, tobacco use, is illustrated in Table 3. For this concept, not one of the 11 registries was concordant with another, despite the federal quality measure and resulting value set regarding same, and only 1 of the registries referenced SNOMED.

Most clinical concepts could be captured in a unidimensional manner (i.e., as a label : value pair). For these clinical concepts, 13 key metadata were specified for each data element, with the priority given to metadata with specific relevance to either (or both) clinicians and database developers (Table 4). Metadata describing the context in which data are collected (i.e., past medical history, within specified timeframe of procedure) were not included in this framework as the implications are different for each use case. Priority values for recommended metadata was given to values from predicate work, particularly the ONC USCDI, the NIH National Library of Medicine (NLM) Value Set Authority Center (VSAC), and HL7 FHIR profiles and resources (especially content listed in the FHIR Detailed Description tabs and FHIR Implementation Guides).

Critical discoveries identified during the specification of metadata included that 1) there was no single source or set of sources where all metadata could be identified – developing the metadata for each concept required the use of multiple sources; and 2) often there was not an obvious, single 'correct' choice, particularly for reference ontology bindings. Additionally, many common data elements and corresponding allowed values lacked explicit, unambiguous clinical definitions; in fact, the metadata with greatest variability and the least consistency were clinical definitions and definitions of clinical allowed values.

Two clinical concepts required multiple data elements to capture the semantics required by registries: medications and laboratory results. Medication and laboratory data elements were collected by most registries; 78% collected medication data, while 100% captured laboratory results data. Medication data elements from registries were compared to the AHRQ Health Information Technology Standards Panel model and models from HL7, ICD, NCI, LOINC, SNOMED CT, RxNorm (including RxClass), NDF-RT, NDC and UNII. While over 20 classes of medications metadata were identified in the environmental scan, for the representation of medications in registries, only 4 contexts were identified – the need for a complete list of a patient's medications by name, medications by therapeutic domain, medications by pharmacologic class, and medications administered in the context of a given procedure. Capture of laboratory data also was multidimensional, typically as a series of observations with multiple prompts related to timing and units of measure. Standards contributing to the laboratory result models include LOINC, SNOMED-CT, and UCUM. Proceedings from two federal multi-agency sponsored public workshops were factored into the recommendations of the laboratory results recommendations.

Conclusions:

Collectively, the registry community is not aligned with and does not contribute to interoperability efforts. The inability to exchange data between EHR systems and registries and the inability to share standardized data across registries are barriers to interoperability and achieving the goal of generating real-world evidence. The project further noted that the specification of data elements in federal initiatives such as the USCDI are not oriented toward clinical workflow and utility. A prime example is the representation of tobacco use per SNOMED-CT codes, rather than reflecting how questions about tobacco use are actually prompted in the clinic, used for clinical purposes, or evaluated in the medical literature.

The technical output of the project was a recommended technical implementation of core common clinical data elements for database programmers of electronic health information (including registry) systems. Should all parties

conform to the implementation, data liquidity and *native data interoperability* for the project concepts will have been achieved. The artifacts of the project are publicly available at www.dcri.org/registry-data-standards. Socialization efforts are underway. Next steps include first the acceptance and implementation by the registry community, and second the execution of projects that demonstrate the interoperable exchange of data between health care providers (via their electronic health information systems) and registry owners.

The work products also include a white paper that lists details of the project, recommendations for federal partners, registry owners, healthcare organizations, and standards development organizations. The environmental scan found that the current landscape lacks a single, authoritative approach to advance interoperability within and across the registry community. Key recommendations include identification of a common authoritative process to identify, define and specify standards for common clinical data elements and an agreed upon process for governance thereof. Also needed is a common data element repository or common clinical data element library to support the technical adoption of standard common data elements. Similarly, common data element, model tooling, and terminology repositories for candidate data elements are needed.

Collectively, we can leverage many opportunities that will elevate our health care ecosystem to one where the capture of clinically relevant data at point of care also serves the needs of care delivery, outcomes evaluation, quality/performance measurement, and medical product evaluation/surveillance. Doing so increases the availability of real-world data to create real-world evidence, contribute to knowledge generation and translate knowledge into practice to improve public health.

References

1. Shuren J, Califf RM. Need for a National Evaluation System for Health Technology. *JAMA.* 2016;316(11):1153–1154. doi:10.1001/jama.2016.8708

2. Blumenthal S. The use of clinical registries in the United States: a landscape survey. EGEMS (Washinton DC). 2017 Dec 7;5(1):26. doi: 10.5334/egems.248

3. Conn J. 'Swivel chair' interoperability: FDA seeks solutions to mesh EHRs and drug research record systems. https://www.modernhealthcare.com/article/20150801/MAGAZINE/308019979/swivel-chair-interoperability-fda-seeks-solutions-to-mesh-ehrs-and-drug-research-record-systems

Acknowledgement

This work was supported by a grant from the Pew Charitable Trusts. The authors wish to acknowledge the many contributions of our Pew collaborators Ben Moscovitch, Josh Rising MD, and Anqi Lu.

Figure 1. *Improving Healthcare Data Interoperability* Project Process

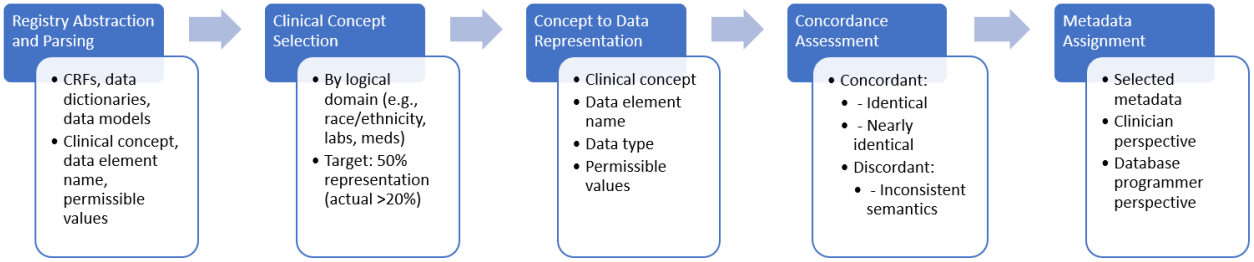

Table 1. Classes of Common Clinical Concepts Across Registries Modeled

1. Patient name
2. Date of birth
3. Sex
4. Race
5. Ethnicity
6. Procedures
7. Unique Device Identifier (UDI)
8. Vital signs (height, length, weight, blood pressure, pulse)
9. Laboratory results (via a separate model)
10. Medications (via a separate model)
11. Care team: physician
12. Smoking status (via a separate model)
13. Alcohol use
14. Substance use
15. Vital status (death)

Table 2. Example – Concordance Analysis of Clinical Concept: Sex

SEX		
Data Element Name	**Allowed (Permissible) Values**	**Number of Concordant Instances (1=unique)**
Sex	Male Female Other Unknown	1
Sex	Male Female Undifferentiated	1
Sex	1=Female 2=Male 3=Unknown	2
Sex	Male Female	3
Sex of patient	[unspecified text]	2
Patient's Sex at Birth	m= male f= female u= unknown	2
Sex (at birth)	Male Female	1
PATIENT_SEX	M F U	1
Patient Gender	Male Female	2
Gender	Male Female	5
Gender	1=Male 2=Female -1=Unknown/missing	1

Table 3. Example – Concordance Analysis of Clinical Concept: Tobacco Use

SMOKING STATUS		
Data Element Name	**Allowed (Permissible) Values**	**Number of Concordant Registries (1=unique)**
Tobacco Use	Never smoker Current every day smoker Current some day smoker Smoker, current status (frequency) unknown Former smoker Smoking status unknown	1
Tobacco Use	Never Former Current - Every Day Current - Some Days Current - Frequency Unknown	1
Tobacco Use	Never Current Quit within past 12 months Quit more than 12 months ago Screening not performed for medical reasons	1
Smoking status	(i) Current everyday smoker (449868002) (ii) Current some day smoker (428041000124106) (iii) Former smoker (8517006) (iv) Never smoker (266919005) (v) Smoker, current status unknown (77176002) (vi) Unknown if ever smoked (266927001) (vii) Heavy tobacco smoker (428071000124103) (viii) Light tobacco smoker (428061000124105)	1
Smoking	0 = Never 1 = Prior 2 = Current	1
Current/Recent smoker (< 1 year)	No Yes	1
Does the patient currently smoke?	No Yes	1
If Current or Quit within 12 months, Smoking cessation counseling provided?	No Yes	1
Tobacco Type	Cigarettes Cigars Pipe Smokeless	1
Types of Nicotine Use	Smoking Chewing E-cigarette Patch Gum	1
Tobacco w/in 1 year - Cigarette	No Yes	1

Table 4. Key Common Clinical Data Element Metadata Needed for Clinical and Database Implementation

1. Clinical concept label (human prompt – for case report form, data entry screen)
2. Clinical definition
3. Clinical allowed values (human prompt – for case report form, data entry screen)
4. Clinical allowed values definitions
5. Database field label
6. Database field data type / format (e.g., char, date, integer, values set)
7. Database field business rules (edit checks, range checks, etc.)
8. Database allowed values (as stored in a database)
9. Object identifier (OID)
10. Reference ontology concept binding
11. Reference ontology allowed values bindings
12. FHIR URL references (profiles, resources)
13. Sources, references, other notes