eGEMs
Generating Evidence & Methods
to improve patient outcomes

A Framework for Classification of Electronic Health Data Extraction-Transformation-Loading Challenges in Data Network Participation

Toan Ong, PhD;[i] Rosina Pradhananga, MPH;[ii] Erin Holve, PhD;[iii] Michael G. Kahn, MD, PhD[i]

ABSTRACT

Background: Contributing health data to national, regional, and local networks or registries requires data stored in local systems with local structures and codes to be extracted, transformed, and loaded into a standard format called a Common Data Model (CDM). These processes called Extract, Transform, Load (ETL) require data partners or contributors to invest in costly technical resources with specialized skills in data models, terminologies, and programming. Given the wide range of tasks, skills, and technologies required to transform data into a CDM, a classification of ETL challenges can help identify needed resources, which in turn may encourage data partners with less-technical capabilities to participate in data-sharing networks.

Methods: We conducted key-informant interviews with data partner representatives to survey the ETL challenges faced in clinical data research networks (CDRNs) and registries. A list of ETL challenges, organized into six themes was vetted during a one-day workshop with a wide range of network stakeholders including data partners, researchers, and policy experts.

Results: We identified 24 technical ETL challenges related to the data sharing process. All of these ETL challenges were rated as "important" or "very important" by workshop participants using a five point Likert scale. Based on these findings, a framework for categorizing ETL challenges according to ETL phases, themes, and levels of data network participation was developed.

Conclusions: Overcoming ETL technical challenges require significant investments in a broad array of information technologies and human resources. Identifying these technical obstacles can inform optimal resource allocation to minimize the barriers and cost of entry for new data partners into extant networks, which in turn can expand data networks' inclusiveness and diversity. This paper offers pertinent information and guiding framework that are relevant for data partners in ascertaining challenges associated with contributing data in data networks.

[i]Department of Pediatrics University of Colorado Anschutz Medical Campus, [ii]AcademyHealth, [iii]Health Care Reform and Innovation Administration, Department of Health Care Finance, Washington D.C.

Introduction

Learning Health Systems (LHS), as envisioned by their original proponents and the National Academy of Medicine, are being realized by the emergence of numerous national research data networks such as PCORnet, Health Care Systems Research Network, and specialty registries.[1–6] Within these diverse research networks, health care data collected at delivery settings such as hospitals or clinics are extracted, transformed, and loaded (ETL) from their local formats to a common structure and semantic, which is often referred to as the Common Data Model (CDM).[7–9] The ETL processes to transform data into a CDM are usually resource-intensive and error-prone due to numerous technical challenges. Further, because these challenges are often underestimated by germane data institutions, there tends to be under-investment in ETL. The technical barriers and financial needs can limit the types of institutions that can participate in ETL efforts to those with significant dedicated informatics resources, such as large health systems, academic medical centers, and large administrative or billing data aggregators. Potentially left behind are smaller institutions or practices, many of which provide health care services to underserved or complex patient populations.

Promoting inclusive participation of institutions into large data networks first requires an understanding of the ETL barriers that data contributors encounter, so that solutions to alleviate the threshold requirements for network participation can be derived. In this research endeavor, we conducted interviews with key informants to compose a comprehensive set of ETL challenges faced in the United States based clinical research data networks and registries that require data transformation into a network-wide CDM. A larger group of stakeholders, including representatives from data networks, federal agencies, and ETL solution vendors further expanded upon the initial list of ETL challenges during an in-person workshop. We subsequently developed a framework for categorizing the identified ETL challenges and created a model that links them to alternative approaches used by data contributors to participate in research data networks. In this framework, based on their level of participation, data contributors can be categorized as data coordinators, data partners, or data sites. Understanding the ETL barriers associated with each participation level allows sites to make informed decisions regarding the resources required to contribute and engage successfully in data sharing networks.

Background

Current clinical data collection systems, such as electronic health records (EHRs), store data in different formats, even between systems from the same vendor. In order to address site specific differences, data sharing networks define a CDM that delineates single data structures and values that are allowed for each variable.[8,10–13] Data contributors are required to transform their local data into the CDM structures in accordance with the precise definitions provided by the CDM developers. In addition to organizational and regulatory requirements, there are numerous technical processes associated with creating a CDM from an existing clinical system. In brief, data elements required by the CDM must be identified in the local system; tools and software code for extracting data must be acquired, implemented, and validated against the CDM specifications; the technical infrastructure for storing the resulting transformed data must be implemented; and data quality issues at each step must be identified, resolved, or at least documented.

The processes of extracting data from source systems, transforming them into a different data structure, and storing them in a separate environment, e.g. data sets or database management systems, are dubbed ETL. Forming high quality data at the network level from diverse data contributors requires extensive knowledge of the source data and conduct of accurate ETL operations.[14] The term "source data" used in this paper refers to the original data system, such as an institutional electronic medical record or data warehouse, and the term "target data" or "target data model" refers to the CDM that is required to participate in a data-sharing network. The ETL challenges are the tasks which are difficult to accomplish during the ETL operations. ETL challenges can be classified into two major categories: non-technical and technical. Non-technical issues are related to data governance needs, including human subject study review protocols and institutional data sharing policies, as well as administrative and budget constraints.[15–18] Technical complications encompass information technology infrastructure, domain expertise, and implementation of the ETL processes.[19] While non-technical considerations are pertinent and are often the driving force in shaping data network engagement, this paper focuses exclusively on identifying and addressing technical challenges.

Literature focused on technical ETL challenges with healthcare data address, inter alia, technical skill requirement for data integration and heterogeneity of source EHR systems. Priest et al. compared the infrastructure and the data submission process to a national collaborative network and noted that the ETL processes are resource intensive and data quality assurance is an iterative process.[20] The findings emphasize that limited resources, namely human resource, processing time, and server memory and space pose significant challenges for data contributors to participate in research data networks. Ross et al. stated that the heterogeneity of source EHR systems and source types—clinical, administrative, and claims—usually makes ETL "labor-intensive" and "prone to errors" because of the complex semantics of each source coding system and due to the frequent changes in source data.[12] Sittig et al. compared the six different informatics platforms for comparative effectiveness research and concluded that achieving data completeness by integrating data from across different healthcare institutions is a major challenge.[13] Data integration across different systems requires sophisticated technical skills in: 1) data mapping via data harmonization; 2) data linking via record linkage; and 3) data quality assessment and data processing (e.g., natural language processing) techniques. The lack of the required skills and knowledge of the source data impacts the quality of data output from an ETL process (e.g., incomplete and inaccurate data) and leads to invalid or biased analysis of results.[21,22] Denney et al. stressed on the importance of performing data quality check (i.e., correctness) to ensure data as the output of an ETL process is fit for use.[23]

Data owners participate in research data networks for various reasons including quality improvement, system redesign, data exploration, and clinical research. Fully understanding the scope of the ETL challenges encountered by new data owners is an important step towards finding relevant solutions. Recognizing the importance of this initial phase, the Electronic Data Methods (EDM) Forum collaborative project titled "Infrastructure Tools to Increase Health Data Network Participation" was developed in 2015 in order to identify and characterize the technical ETL barriers experienced by data owners participating in research data networks.

Methods

This research constituted of four phases: key informant interviews; identification and categorization of ETL challenges; validation and extension of the initial findings during a one-day stakeholder workshop; and final synthesis of a multi-dimensional framework for classification of ETL challenges.

Phase 1: Key Informant Interviews

In order to identify major ETL challenges, we conducted semi-structured 60-minute individual telephone-based interviews with nine key informants. Eight of the interviewees represented those responsible for ETL operations in research data networks, including national data research networks and registries; one ETL solution vendor was also interviewed. Convenience and snowball sampling methods were used to identify the informants. Sample research data networks represented included national networks such as eMerge, PCORnet, VA Cardiovascular Assessment, Reporting, and Tracking (CART) program, and national registries led by the American College of Cardiology. The interviews, conducted by the authors who are biomedical informatics researchers with experiences operating local and national research data networks, spanned across three sections: extract (E), transform (T), and load (L). The transcripts and notes from the interviews were qualitatively reviewed to develop an initial list of ETL barriers stratified by the three phases.

Phase 2: Identification and Categorization of ETL Challenges

Similar recurring ETL challenges identified from the key informant interviews were grouped under common descriptive labels. Since the original structure of the interviews entailed separation of the discussion into E, T, and L categories, the ETL challenges obtained from the interviews were categorized by the three steps associated with the ETL process. This approach is referred to as the phase-based approach. However, the review of discussions with key informants revealed that a challenge could span multiple phases. Therefore, we also formed a theme-based approach which categorizes the ETL problems based on the content.

Phase 3: Workshop

Using the theme-based classification, we deployed a pre-workshop survey to a multi-stakeholder group (n=18) of attendees to rank the preliminary ETL challenges within the identified themes using 5-point Likert scales. The group of workshop attendees included a subset of the initial key informants (n=3), although all were invited to participate. Eight (44 percent) of the participants represented data networks—five data coordinating centers, one data partner, and two data sites. Further, six federal agency representatives (33 percent) from the National Institutes of Health, Office of the National Coordinator for Health Information Technology, Agency for Healthcare Research and Quality, and National Library of Medicine were involved. Four participants (22 percent) were ETL vendor representatives. Four members of the project team also engaged in and facilitated the discussion. The survey based prioritization of ETL challenges informed a more in-depth discussion during the workshop to reflect on the challenges, identify new ones, and conduct final prioritization of ETL barriers.

Phase 4: Final Synthesis: Development of a Participation-Based ETL Framework

The final set of ETL obstacles identified from the interviews and the meeting was synthesized and classified according to the two classification approaches mentioned in phase 2. In addition, since the levels of participation, namely data site, data partner and data coordinating center, dictate a set of activities which a member of the data network must perform, we developed a participation based approach to categorize the identified ETL challenges. The primary framework for classification of the ETL challenges is represented by an amalgamation of the phase-based, theme-based, and participation-based categorization systems.

Results

A Framework for ETL Challenge Classification

The framework for classifying ETL challenges (n=24) organizes them according to three perspectives as introduced in the methods section: phase-based, theme-based, and participation-based. The phases in the phase-based approach are extract, transform, and load. The six themes in the theme-based approach include source data, technical difficulties, knowledge management, code management and versioning, data quality, and ETL operations. The levels in participation-based approach include data site, data partner, and data coordinating center. Figure 1 demonstrates the mapping of each ETL challenge into the axis of the three-dimensional framework. For example, on the phase-based axis, the "lack of knowledge of source EHR data" has major impact on the data extraction and transformation phases. On the theme-based axis, the same challenge occurs with source data processing; on the participation-based axis, data sites and partners encounter this issue. Many of the barriers such as "lack of technical expertise or tools to perform ETL operations" and "lack of knowledge sharing" span across all three ETL phases.

Phase-based classification:

Phase 1: Data Extraction

Data extraction is the process of collecting source data used to populate the target data model. Source data can come from a single or multiple source systems. Therefore, data extraction activities need to address challenges that are related to understanding the structure and semantics of source data, source data accessibility, and heterogeneity of the semantics and structure across multiple data sources.

Phase 2: Data Transformation

Data transformation is the process of mapping source data in their original structure and codes into the structure and coding of the target data model. Mapping without information loss between the source and target data elements requires expertise in both the source and target data. If the source data are comprised of data from more than one system, each data source has to be harmonized into the target data model. The challenges of data transformation include maintaining the integrity of all data sources with regard to the requirement and changes in the target data model, data transformation tool, and the knowledge management process.

Phase 3: Data Loading

Data loading is the process of loading the output of the data transformation phase into the target data model. Data loading usually entails data insertion and appending into the target data model. The challenges related to this phase are associated with the efficiency of the loading process, the data quality assessment process, the complexity of incremental data loading, and ETL operation routines.

Theme-based classification:

With each challenge included in this classification, we include a description, the definition of the challenge, and an example based on practical experience of the authors, key informants, and workshop participants.

Theme 1: Source Data

The source data theme encompasses the ability to use source data in the ETL process which is supported by an understanding about source data as well as their availability.

Knowledge of Source Data

- Description: ETL participants do not have sufficient knowledge about the structure, semantics, physical location, and relationship among the data elements in the source data.
- Challenge: For each data element within the target data model, one or more data elements in the source data must be correctly identified to be used as source data. This challenge is amplified when the source system comprises of multiple sub-systems from various vendors. Failing to locate and understand the semantics and relational linkages of source data will either create missing or inaccurate data in the target model.
- Example: A typical commercial EHR system contains thousands of tables with thousands of columns. While the overall data model structure is similar across EHR installation, the contents (value sets) of these tables are usually highly customized to meet local needs. Extracting data from complex data models requires extensive general knowledge of source EHR data structures and unique local knowledge of how these tables are used (workflows) and therefore customized.

Source Data Accessibility

- Description: Source EHR data are not accessible because the ETL team does not have direct access to the source data.
- Challenge: ETL is an iterative and trial-and-error process. If the programmers who develop the ETL scripts do not have direct access to source data, perhaps due to institutional security restrictions or data governance policies, the software has to be tested by a different group of people who have data access. This two-phase process may be inefficient and time-consuming.
- Example: An organization has strict change-control processes for introducing new software into the operational EHR environment. Programmers are only allowed to develop tools or codes using a limited development environment. Without direct data access, the ETL developers will develop the data extract tools or codes without knowing the actual efficiency of these tools. A second team at the hospital with EHR data has to execute the data extraction routines. If there are errors, the team at the hospital will have to describe these errors to the third-party developers, thus allowing for misinterpretation.

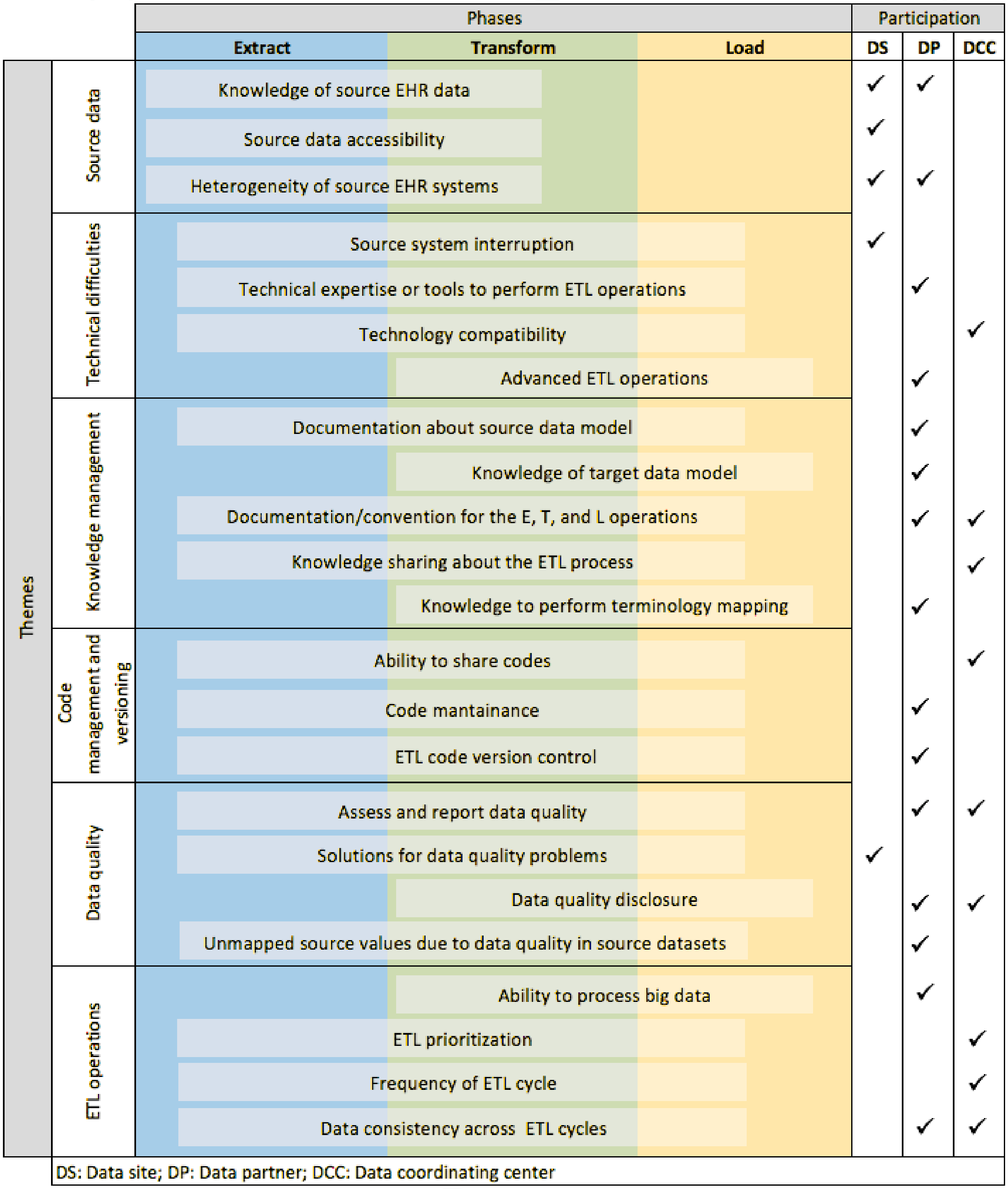
Heterogeneity of Source Systems

- Description: Data essential for the target model requires data elements that are distributed across multiple distinct source systems from one or more healthcare system vendors. Different source systems have different backend data models, terminologies, data access policies, and owners.
- Challenge: If a target data model requires source data from multiple source systems, these source data must be harmonized before populating the target model. The extra harmonization step requires extensive expertise about source systems and the target data model. Rarely does one individual possess the expertise above. Hence, a team of experts must work together to accomplish desirable harmonization results.
- Example: A table in the target data model is used to store all inpatient visits at a health care institution. In source systems, inpatient visit data span across three different departments—general surgery, emergency, and main hospital—with correspondent sub-systems. To populate the target table above, data from these sub-systems must be harmonized to agreed-upon structure and semantics for the single inpatient visit table in the target data model.

Theme 2: Technical Difficulties

Technical difficulties focus on complications associated with all technical components involved

Figure 1. A Framework for ETL Challenge Classification



		Phases			Participation		
		Extract	Transform	Load	DS	DP	DCC
Themes	Source data	Knowledge of source EHR data			✓	✓	
		Source data accessibility			✓		
		Heterogeneity of source EHR systems			✓	✓	
	Technical difficulties	Source system interruption			✓		
		Technical expertise or tools to perform ETL operations				✓	
		Technology compatibility					✓
			Advanced ETL operations			✓	
	Knowledge management	Documentation about source data model				✓	
			Knowledge of target data model			✓	
		Documentation/convention for the E, T, and L operations				✓	✓
		Knowledge sharing about the ETL process					✓
			Knowledge to perform terminology mapping			✓	
	Code management and versioning	Ability to share codes					✓
		Code mantainance				✓	
		ETL code version control				✓	
	Data quality	Assess and report data quality				✓	✓
		Solutions for data quality problems			✓		
			Data quality disclosure			✓	✓
		Unmapped source values due to data quality in source datasets				✓	
	ETL operations		Ability to process big data			✓	
		ETL prioritization					✓
		Frequency of ETL cycle					✓
		Data consistency across ETL cycles				✓	✓

DS: Data site; DP: Data partner; DCC: Data coordinating center in the implementation of an ETL process. These technical components include the systems which host the source and target data, software to support ETL operations, and information exchange method required among these technical components.

Source System Interruption

- Description: ETL operations interfere with the source EHR system operation. Even though data extraction is often a read-only task, the execution of big queries might impact the performance of the entire system.
- Challenge: Putting the operation of an EHR system on hold, especially at a large health care institution, is usually not feasible. On the other hand, extracting data directly from an operational EHR system might result in inconsistent data being extracted due to the real-time data updating or inserting actions.
- Example: If data extraction from the laboratory information system is performed at the same time new test results are being entered, the extracted data may be incomplete.

Technical Expertise or Tools to Perform ETL Operations

- Description: The required technical skills, such as query script writing tools, stored procedure programming knowledge, or software development environments for ETL are not available for ETL operations.
- Challenge: Despite the fact that ETL operations rely mostly on ETL specifications composed by domain experts who understand the source and target databases, these ETL specifications must be implemented using either a database scripting language or an ETL tool. Most health institutions do not have human capital with technical expertise in ETL programming tools.
- Example: A small cardiology clinic is interested in contributing EHR data to a regional stroke registry. However, other than an application supports technician who oversees the daily operation of the EHR system, there is no personnel to perform the required ETL operations.

Technology Compatibility

- Description: The backend technologies (e.g., database management systems, operating systems, and programming languages, etc.) are different between the source data model and the target data model or among the source systems.
- Challenge: Technology incompatibility requires additional steps to convert data from one format to another or hinders reusability of pre-composed scripts.
- Example: Data contributor A has Microsoft SQL Server as its backend database and data contributor B has Oracle as its backend database. A data transformation script composed in Oracle cannot be reused in SQL Server even if it is for the same data element.

Advanced ETL Operations

- Description: Some ETL operations such as data extraction from free text or data encryption for privacy preserving record linkage are not within capacities of a data contributor.
- Challenge: Advanced methods often require significant investment and resources which are difficult or expensive to acquire, especially when ETL for research is not a priority.
- Example: Extracting lab results from unstructured free-text data such as provider's notes requires natural language processing techniques which are only offered in highly-priced customized and advanced data mining tools.

Theme 3: Knowledge Management

Knowledge management challenges focus on knowledge generation, recording, maintenance, and sharing among different ETL teams in the same network or across different phases of an ETL process.

Documentation About Source Data Model

- Description: The source data models, source terminologies, and their relationships are not well documented.
- Challenge: The source EHR system is often highly complicated and locally customized. Knowledge about these source systems is often accumulated over time. It takes significant efforts to document all the details about the characteristics and changes in the source systems.
- Example: A new NULL flavor titled "missing" was added to a field in a table in the source data. However, the definition of that particular value was not provided, resulting in these "missing" values that were not appropriately transformed by the ETL process.

Knowledge of Target Data Model

- Description: Knowledge of the target data model was not fully captured by the ETL team due to lack of documentation about the target data model or the incomplete specifications of the target model.
- Challenge: Incomplete understanding of the intended semantics and conventions used by the target model usually leads to inaccurate or incomplete data being loaded.
- Example: A target data model has a permissible value in the "smoking status" but does not clearly define the specific measurement of smoking regularity.

Documentation/Convention for the E, T, and L Operations

- Description: Conventions and decisions made by different members of the ETL team during each step of the ETL process are not well documented. The examples of these decisions are terminology and schema mappings, data formatting conventions, or inclusion and exclusion criteria.
- Challenge: The lack of detailed documentation prevents the ETL process from being easily understood by others.
- Example: Data quality assessment process revealed that NULL values were used as a number of diagnosis codes in the target data model. Due to the lack of documentation, it was impossible to identify if the NULL value was intended to indicate missing data or unmapped codes in the source system or both.

Knowledge sharing about the ETL process

- Description: Common knowledge about ETL process cannot be shared between network data contributors due to the lack of knowledge sharing platform or agreement.
- Challenge: 1) the ETL team does not have the permission to share ETL work that includes information related to source EHR system due to contractual constraints imposed by the EHR vendor; and 2) the lack of acceptance of knowledge sharing technology.
- Example: Terminology mappings from codes used in the source data to standard open source codes have to be redone because the local codes are proprietary.

Knowledge to perform terminology mapping

- Description: The inability to perform accurate mapping between the terminologies used in the source and the target data models. In health care, terminology mappings require domain expertise about the structure and semantics of both source and target coding systems.
- Challenge: Terminology mapping is one of the most important and most difficult tasks in an ETL process. Incorrect terminology mapping leads to

inaccurate data being populated to the target data model.
- Example: A local pharmacist cannot map local drug codes to RxNorm because it is an unfamiliar terminology to the organization.

Theme 4: Code Management and Versioning

Code management and versioning challenges center on code utilization, management, and control to avoid duplicate programming efforts as well as to keep up with changes in the source and target data.

Ability to Share ETL Codes

- Description: Reusable ETL codes cannot be shared among data contributors with similar data infrastructure. Similar to the inability to share ETL knowledge, the ETL codes that include information related to proprietary products are often restricted by existing vendor contracts from being distributed, even between customers.
- Challenge: Inability to reuse the ETL codes developed by a collaborative data contributor might increase the cost of the ETL process due to duplicate efforts.
- Example: A data contributor has developed ETL code that extracts data from a vendor's imaging system but they are unable to share that code with other customers of the same system because they are precluded by intellectual property or trade secrets protection clauses in their vendor contract.

Code Maintenance

- Description: ETL code becomes outdated because changes occur in the source and target data models over time due to upgrades or new versions.
- Challenge: Modifications in the source and target data models occur frequently and independently. ETL code must be updated accordingly with respect to these changes. In addition, depending on the magnitude of the changes, each step of the ETL process has to be redesigned and revaluated.
- Example: The structure and semantic of smoking status data in the Vital table in the PCORnet CDM version 2.0 and version 3.0 are significantly different. The ETL process to populate smoking status data in PCORnet has to be revised due to this update.

ETL Code Version Control

- Description: Multiple versions of the ETL codes and terminologies are difficult to maintain and distinguish, especially when different technical experts are required to write ETL code for different systems. In addition, new versions are generated in response to initial data quality assessment findings, resulting in multiple versions of the ETL scripts that must be distinguished or merged.
- Challenge: Managing multiple versions of ETL code might lead to different versions being used by different ETL teams causing data discrepancies or even errors in target data.
- Example: An initial ETL script did not extract data from inpatients, resulting in significantly reduced data density. A revised script fixed this problem but created a new issue with Emergency Department (ED) patients. A third version fixed the issue with ED patients and was intended to be the final version. However, the team incorrectly submitted the second script, which still contained the issue with ED patients.

Theme 5: Data Quality

Data quality challenges entail assessing and reporting the quality of both source and target data of an ETL process. Data quality issues associated with source data must be disclosed and separated from data quality issued caused by the ETL operations.

Assess and Report Data Quality

- Description: There are insufficient guidelines or documentation to guide the development of data quality reports or the use of existing data quality tools.
- Challenge: Data contributors are "on their own" to determine how to assess data quality—which features to include and how to interpret results.
- Example: Sites are asked to assess data completeness, but there are no detailed specifications that define how to calculate this measure or which variables are most important to assess.

Solution for Data Quality Problems

- Description: Identified data quality problems are caused by issues in the source system that cannot be altered or corrected by the ETL process.
- Challenge: Data quality anomalies remain in the data set and continue to trigger data quality alerts.
- Example: Medication prescriptions contain dispensing information as free text rather than as discrete fields.

Data Quality Disclosure

- Description: Data quality measures include information that could reveal sensitive information about an organization's business practices.
- Challenge: Legal or perceived business threats prevent data quality findings from leaving the organization.
- Example: During a specific quarter, the laboratory system failed to transmit microbiology results into the EHR due to an unrecognized change in the HL7 interface.

Unmapped Source Values Due to Data Quality in Source Datasets

- Description: Coded fields do not conform to the allowed value set.
- Challenge: Incorrect codes fail to map, resulting in a large amount of missing data.
- Example: Coded diagnoses can be manually overwritten to be replaced with free text.

Theme 6: ETL Operations

ETL operations challenges cover barriers related to the entire ETL process such as ETL prioritizations or frequencies. The solutions for these challenges often have direct impact on workflow as well as resource allocation to every phase in an ETL process.

Ability to Process Big Data

- Description: A large volume of source data to be extracted and processed impacts the performance of the ETL process. Large volume data is a relative concept depending on the data processing capacity of the data contributor.
- Challenge: The ability to process large data extractions is associated with the availability of resources including technical expertise and information technology resources.
- Example: Physiological signal data or raw imaging data require very large data storage and backup capacity, network bandwidth for data transfers, and significant CPU and memory requirements. All of those resources require additional investment in data infrastructure which is not always available, especially for data contributors with limited resources.

ETL Prioritization

- Description: Prioritizing operational needs above ETL activities leads to delays in completing ETL activities. Because the resources for ETL are often limited and compete with other organizational needs, the ETL team has to set prioritization for each domain of the ETL process.
- Challenge: The prioritization process has to align with the utilization of data in the target model. Any

misalignment between data being transformed and data needed by the network will potentially delay the use of target data.

- Example: A research network with limited resources for ETL operations decides to focus on laboratory results in their initial ETL iteration while there is an urgent need for this network to respond to queries about medications.

Frequency of ETL Cycles

- Description: The frequency of ETL cycles results in time constraints for all activities of an ETL process. Shorter ETL cycles mean less time allowed for a particular step.
- Challenge: When the duration of an ETL cycle changes and other resources do not change, the efficiency of the entire process has to be improved. The magnitude of the required changes to the ETL operations also directly impact the frequency of ETL cycles.
- Example: Adding a new table to the target data model would not be a problem with a 6-month ETL cycle but would be a challenge for an ETL process that occurs on a weekly basis.

Data Consistency Across ETL Cycles

- Description: Inconsistent data being loaded into the target data model among the ETL cycles. Data inconsistency in the target model can be revealed by a data quality assessment tool.
- Challenge: Rapid changes in ETL operations such as terminology mapping and data de-identification might change the data populated in the target data model.
- Example: In the previous ETL cycle, 100 percent of concepts codes in the Diagnosis table were successfully mapped to standard concepts. However, in the current ETL cycle, the number of mapped concepts in the same table dropped by 50 percent even though no new concept code was added into that table.

Participation-based classification:

The key ETL challenges that were identified can be re-categorized based on level of participation of a member in a data research network. There are three different levels of participation: Data site (DS), data partner (DP), and data coordinating center (DCC) (Figure 2). Depending on the ETL phases and tasks performed due to their level of participation, members of a data network are subject to different challenges. (Right 3 columns of Figure 1) The participation-based classification allows participants in a research data network to identify their responsibilities in the ETL process. At each level, the framework demonstrates the tradeoff between the requirements of resources versus the control over the ETL process. The levels of participation are not mutually exclusive.

Data Sites: Data Contribution Only

DS are responsible only for facilitating the data extraction process; they are not responsible for actually creating or performing the technical ETL processes. This is the lowest level of participation which is appropriate for members with limited Information Technology (IT) resources. Data in their original structure and semantic are extracted and securely transferred to another location for processing. This other location has significant technical resources and knowledge to complete the ETL tasks on behalf of the site. However, depending on the requirements of the extraction process, DS may still need to provide basic IT resources such as memory and storage for this process.

DS often have extensive knowledge about the local EHR system and the source data. Therefore, the focus on a site is to ensure that required data elements are successfully extracted from the source systems. DS are also expected to provide adequate information about the source data to support the transformation and loading phases. The resources required of a DS are the lowest compared to other levels. Nonetheless, since DS are often care delivery-centric, they require technical support from DP and DCC. DS are not responsible for how the source data are transformed or standardized into the target data model.

Data Partners: Source Data Transformation and Loading into the Target Data Model

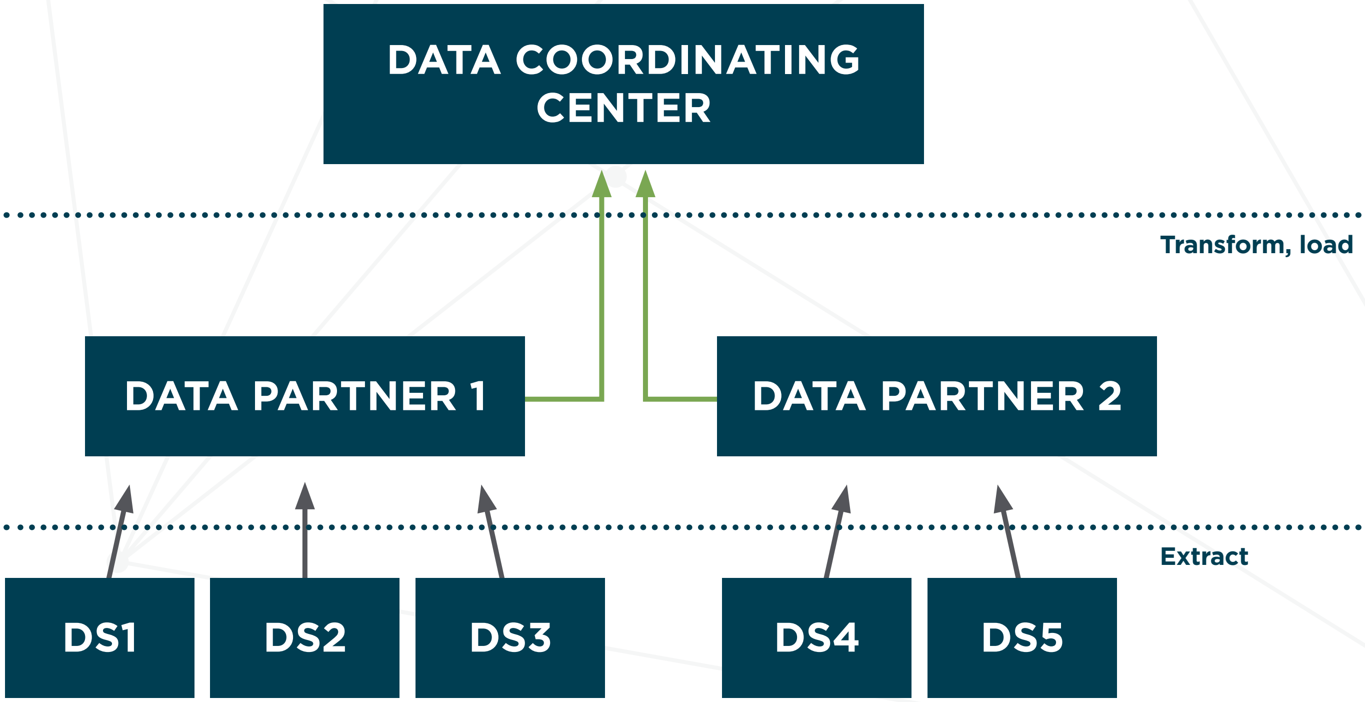
DPs are responsible for data transformation and loading into the target data model. They are expected to have more resources than DS. Partners work with sites to obtain knowledge about sites' datasets. A member can be both a DP and a DS. In this case, the member is responsible for all aspects of the ETL processes.

The responsibilities of a DP include collecting data from sites and transforming and loading data into a target model. Source data are transferred from DS to DP in the raw format (i.e., native data structure and terminologies). Therefore, detailed information about the source data are provided by sites to complete the transformation. DPs require significantly more resources and domain-specific and technical expertise than sites because they have to perform the most complicated tasks such as terminology mappings or natural language processing. DPs have full control over the transformation formula and the quality of data in the target model.

Data Coordination Centers: A Data Repository for Analysis and Dissemination

DCCs are responsible for managing the ETL operations across participants in the network. DCCs facilitate knowledge and code management

Figure 2. Participation Levels in a Research Data Network



and sharing as well as provide necessary technical support to DPs and DS. A DCC can also function as both a DP and a DS.

DCCs are often the receivers of the target datasets. DCCs are responsible for checking data quality in the target data model. They oversee the entire ETL process, manage knowledge sharing among data partners, and establish conventions and set the priorities for the entire data network. Expertise in the target data model and data quality measurements at the DCC level is required to ensure the quality and progress of the ETL process.

Discussion

The multi-dimensional framework of ETL challenges describes and organizes the types of barriers that need to be overcome in different phases of an ETL process by different entities. Depending on the amount of resources available and the priorities of the participating agent in the ETL process, some technical issues are more challenging than others. For example, terminology mapping was mentioned as a significant challenge for all the networks we interviewed. On the other hand, technical requirements such as powerful IT resources might not be an obstacle for some organizations with sufficient resources. From the perspective of all three phases of an ETL process, the challenges enumerated might be overwhelming and organizations that are responsible for all three phases must allocate resources sufficiently and efficiently. However, in practice, the ETL process is usually disjointed by the separation among the source data owner, the ETL team, and the target data owner. Based on the level of participation, a member in a research data network might need to be involved in one or two phases of the entire ETL process.

In addition to the challenges which were identified and categorized in the framework, key stakeholders engaged in this initiative raised other barriers in the ETL process, including lack of a uniformly consistent SQL syntax across widely used relational database systems, standard data identification methods, the transparency of the ETL process, and difficulties with staff training and retention. These complications were not incorporated into the main framework because they were either implicitly included in the existing challenges or they represented technical obstacles which the ETL team usually cannot address directly.

The methods for this research study include several limitations. First, we acknowledge that data governance-related challenges such as data access policies present significant barriers to the success of an ETL process. The data governance policies entail both non-technical and technical challenges, most of which were discussed at the workshop. However, these challenges are out of the scope of the paper, and therefore, were not systematically assessed. Second, although we included representatives from some of the major research data networks and registries, our sampling approach was non-random and may not have been representative of all types of data networks. Further, the sample size of key informants and multi-stakeholder group is small. Due to our sampling approach, along with the nature of the data, which were self-reported, some important ETL challenges may not be represented in this paper. All of our participants were engaged in research data networks. We did not include participants in other forms of data sharing, such as national clinical quality or patient safety collaborative, or non-research data efforts. It is unclear if these other use cases have different challenges although we suspect many of the issues identified in Figure 1 would apply to these other use cases. In order to overcome these potential shortcoming, a wiki of this project has been made available by the EDM Forum at (goo.gl/7DPmyv) to solicit additional recommendations and feedback about our findings.

Conclusion

ETL operations are at the core of technical activities to support research data integration. ETL operations in health care are usually resource-intensive and error-prone. Nevertheless, the challenges faced in these ETL operations were not recognized adequately in the literature. In this paper, we identify, describe, and categorize 24 technical ETL challenges associated with the participation in a research data network via a series of key informant interviews and workshop discussion with ETL experts from national and regional data networks. Proper understanding of the ETL challenges is an important first step towards the development of effective ETL solutions and best practices which lower the technical barriers for data owners to contribute data for research.

Acknowledgements

The EDM Forum Collaborative Methods Project was funded by the EDM Forum, which was supported by the Agency for Healthcare Research and Quality (AHRQ) through the American Recovery & Reinvestment Act of 2009, Grant U13 HS19564-01. We would like to thank all participants who participated in the interviews and the workshop for providing their valuable comments and recommendation for the materials presented in this paper.

References

1. Etheredge LM. A rapid-learning health system. Health Aff Proj Hope. 2007 Apr;26(2):w107-118.
2. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010 Nov 10;2(57):57cm29.
3. Institute of Medicine (US). Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary [Internet]. Grossmann C, Powers B, McGinnis JM, editors. Washington (DC): National Academies Press (US); 2011 [cited 2016 Aug 30]. (The National Academies Collection: Reports funded by National Institutes of Health). Available from: http://www.ncbi.nlm.nih.gov/books/NBK83569/
4. Margolis PA, Peterson LE, Seid M. Collaborative Chronic Care Networks (C3Ns) to transform chronic illness care. Pediatrics. 2013 Jun;131 Suppl 4:S219-223.
5. Roundtable on Value & Science-Driven Health Care, Institue of Medicine. Digital Data Improvement Priorities for Continuous Learning in Health and Health Care: Workshop Summary [Internet]. Washington (DC): National Academies Press (US); 2013 [cited 2016 Aug 30]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK207322/
6. Registry of Patient Registries (RoPR) - Abstract - Final | AHRQ Effective Health Care Program [Internet]. [cited 2016 Aug 30]. Available from: http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=690
7. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. Med Care. 2012 Jul;50 Suppl:S60-7.
8. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc JAMIA. 2012 Feb;19(1):54–60.
9. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System — A National Resource for Evidence Development. N Engl J Med. 2011 Feb 10;364(6):498–9.
10. Maro JC, Platt R, Holmes JH, Strom BL, Hennessy S, Lazarus R, et al. Design of a national distributed health data network. Ann Intern Med. 2009 Sep 1;151(5):341–4.
11. Brown JS, Lane K, Moore K, Platt R. Defining and evaluating possible database models to implement the FDA Sentinel initiative [Internet]. 2009 [cited 2014 Feb 15]. Available from: http://www.regulations.gov/contentStreamer?objectId=0900006480e6cd93&disposition=attachment&contentType=pdf
12. Ross TR, Ng D, Brown JS, Pardee R, Hornbrook MC, Hart G, et al. The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration. EGEMs Gener Evid Methods Improve Patient Outcomes [Internet]. 2014 Mar 24 [cited 2014 Apr 16];2(1). Available from: http://repository.academyhealth.org/egems/vol2/iss1/2
13. Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, et al. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogenous clinical data. Med Care. 2012 Jul;50 Suppl:S49-59.
14. Yoon D, Ahn EK, Park MY, Cho SY, Ryan P, Schuemie MJ, et al. Conversion and Data Quality Assessment of Electronic Health Record Data at a Korean Tertiary Teaching Hospital to a Common Data Model for Distributed Network Research. Healthc Inform Res. 2016 Jan;22(1):54–8.
15. Malin B, McGraw D. Data Governance. 2011 [cited 2015 Jun 17]; Available from: http://repository.academyhealth.org/cgi/viewcontent.cgi?article=1122&context=symposia
16. Holmes JH, Elliott TE, Brown JS, Raebel MA, Davidson A, Nelson AF, et al. Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature. J Am Med Inform Assoc. 2014 Jul 1;21(4):730–6.

17. Holmes JH. Privacy, Security, and Patient Engagement: The Changing Health Data Governance Landscape. eGEMs [Internet]. 2016 [cited 2016 Sep 5];4(2). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4827787/
18. Baker DB, Kaye J, Terry SF. Privacy, Fairness, and Respect for Individuals. EGEMs Gener Evid Methods Improve Patient Outcomes [Internet]. 2016 Mar 31 [cited 2016 Sep 5];4(2). Available from: http://repository.edm-forum.org/egems/vol4/iss2/7
19. Peek N, Holmes JH, Sun J. Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics. Yearb Med Inform. 2014 Aug 15;9:42–7.
20. Priest E, Klekar C, Cantu G, Berryman C, Garinger G, Hall L, et al. Developing Electronic Data Methods Infrastructure to Participate in Collaborative Research Networks. EGEMs Gener Evid Methods Improve Patient Outcomes [Internet]. 2014 Dec 2;2(1). Available from: http://repository.edm-forum.org/egems/vol2/iss1/18
21. Ware JH, Harrington D, Hunter DJ, D'agostino RB. Missing Data. N Engl J Med. 2012;367(14):1353–4.
22. Berti-Équille L. Data quality awareness: a case study for cost optimal association rule mining. Knowl Inf Syst. 2006 Mar 28;11(2):191.
23. Denney MJ, Long DM, Armistead MG, Anderson JL, Conway BN. Validating the extract, transform, load process used to populate a large clinical research database. Int J Med Inf. 2016 Oct;94:271–4.