2020

Common Data Model Harmonization (CDMH) and Open Standards for Evidence Generation

Final Report

FDA U.S. FOOD & DRUG ADMINISTRATION

NIH National Institutes of Health

The Office of the National Coordinator for Health Information Technology

CDMH Team Members

Team Member Organizations	Team Members
HHS ASPE PCORTF Team	Scott Smith Susan Lumsden
Food and Drug Administration Team	Mitra Rocca Scott Gordon Wei Chen Rashedul Hasan Sara Meiselman Alec Petkoff Smita Hastak Charles Yaghmour Sean Khozin Vaishali Popat
Office of the National Coordinator for Health Information Technology (ONC)	Albert Taylor Nagesh Bashyam (Dragon)
National Center for Advancing Translational Sciences Team National Institutes of Health	Ken Gersing Raju Hemadri
National Cancer Institute Team National Institutes of Health	Denise Warzel Elad Sharon
National Library of Medicine	Lisa Lang Robin Taylor
Adeptia	Sandeep Naredla

Common Data Model Harmonization (CDMH) and Open Standards for Evidence Generation

Final Report

Table of Contents

1 EXECUTIVE SUMMARY

The use of real world data (RWD) [1] to generate real world evidence (RWE) has been encouraged through the 21st Century Cures Act [2] and by the U.S. Food and Drug Administration (U.S. FDA) for a number of purposes, including safety surveillance and to augment data from traditional randomized clinical trials. Ultimately, the vision is to accelerate the learning from health care data to improve patient care and to accelerate the development and effectiveness of new therapies by supporting regulatory decision making.

The Department of Health and Human Services (HHS) Office of the Assistant Secretary for Planning and Evaluation (ASPE) engaged FDA, several NIH Institutes including National Center for Advancing Translational Sciences (NCATS), National Cancer Institute (NCI) and National Library of Medicine (NLM), and the Office of the National Coordinator for Health Information Technology (ONC) to collaborate on the Harmonization of Common Data Models (CDMs) and Open Standards for Evidence Generation (Common Data Model Harmonization) project through the Patient-Centered Outcomes Research Trust Fund (PCORTF).

The CDMH project was intended to reuse data, methods, and other resources from existing research networks and to take advantage of available open, consensus-based standards. For example, the CDMH project was intended to use existing PCORTF investments and infrastructure at FDA and NIH to build reusable data mappings and transformation services.

This project was designed to harmonize specific CDMs such that important research information could be more easily shared and interpreted. Patient-Centered Outcomes Research Network (PCORnet), Informatics for Integrating Biology & the Bedside (i2b2/Accrual to Clinical Trials (ACT)), Observational Medical Outcomes Partnership (OMOP), and Sentinel CDMs were mapped to an intermediary model (the Biomedical Research Integrated Domain Group (BRIDG)) to auto-generate clinical queries.

The project was implemented in three phases: 1) Design and Development, 2) Implementation, and 3) Testing and Validation. The Design and Development Phase included requirements gathering to identify business and technical requirements to support query transformation across the four specific CDMs used in this project, and clinical query development to test the CDMH portal and to deliver results represented using clinical research standards. These requirements guided development of the architecture, tool design, data standards, mappings between CDMs, and the intermediary model to support the portal. The CDMH technical architecture and tools were implemented and tested to demonstrate their functionality, including the automation of the query development and execution process. In the Testing and Validation phase, a data validation process was developed to assess the results of a clinical query. Key deliverables included the Health Level International® (HL7®) Fast Healthcare Interoperability Resources® (FHIR®) CDMH implementation guide and data governance framework, methods and standard process for ongoing curation, validated mappings between the CDMs, a portal and clinical query software, and public access to data specifications and visualizations of the mappings.

In this report describing the development process and approach taken to harmonize the CDMs by leveraging standards and controlled terminologies, we include publicly available resources that can be leveraged by the research community.

2 BACKGROUND

In order to achieve a sustainable data network infrastructure, promote interoperability, and foster the creation of a Learning Health System as laid out in the Connecting Health and Care for the Nation a Shared Nationwide Interoperability Roadmap [3], there is a need to connect data across various CDMs leveraging data standards. By mapping various CDM data elements and leveraging PCORTF investments, one can reuse the data, methods and other resources from each network (e.g., Patient-Centered Outcomes Research Network (PCORNET), Sentinel, Observational Health Data Sciences and Informatics (OHDSI), Informatics for Integrating Biology & the Bedside (i2b2/Accrual to Clinical Trials [ACT]) thereby providing researchers with access to more extensive and more diverse observational data (with appropriate data partner permissions).

- *Multiple distributed networks and Common Data Models (CDMs) exist*
- *Harmonizing the CDMs of these networks can help create a true Learning Health System*

This project made use of existing PCORTF investments, including the Data Access Framework (DAF) [4] and the NIH Common Data Elements (CDE) Repository to build reusable data mappings and transformation services.

The audience for this project report includes scientists participating in current research networks and other stakeholders affiliated with NIH, FDA, Centers for Medicare & Medicaid Services (CMS), Centers for Disease Control and Prevention (CDC), and other organizations. By harmonizing various CDMs, the research community will have access to a larger and more representative information from EHRs, administrative claims, and registries, and additional demographic information (e.g., geriatrics). The combined information will be useful particularly for the study of rare events and for analyses from a global perspective.

In addition, tools developed by one network (e.g., Sentinel, PCORNET, i2b2/ACT and OHDSI) can be re-used by other networks. This solution would support a diverse range of queries from basic PCOR queries, to data mining to generate hypotheses for future research, to large-scale analysis, including randomized clinical trials.

2.1 AIMS AND OBJECTIVES

The goals of this project were to 1) facilitate the use of RWD sources, e.g., claims, electronic health records (EHRs), registries, electronic patient reported outcomes (ePRO), to support biomedical research and evidence generation, 2) to enhance regulatory decision making, and 3) to enhance existing health care and research data standards to support data integration and interoperability.

CDMH aimed to develop a method to harmonize four common data models of various networks, allowing researchers to ask research questions on much larger amounts of RWD than currently possible, and leveraging open standards and controlled terminologies to advance patient-centered outcomes research (PCOR).

Specifically, this project was motivated to create an infrastructure that allowed researchers to: 1) query four data networks (i.e., PCORnet, Sentinel, OHDSI, i2b2/ACT), 2) query across networks with access to larger and more diverse types of RWD, and 3) reuse the data, methods, analytic capabilities and knowledge across networks.

The objectives of the project were to:

- Develop a general framework (i.e., tools, processes, standards and governance) for the transformation of various CDMs
- Design a clinical use case to test the developed CDMH infrastructure
- Leverage data standards and controlled terminologies to advance Patient-Centered Outcomes Research
- Implement and test tools developed with several data partners
- Incorporate and reuse infrastructure developed by currently-funded PCORTF projects (i.e., Data Access Framework [4], and the NIH Common Data Element Repository [5]), as well as existing infrastructure (e.g., the Cancer Data Standards Registry and Repository (caDSR) [6])

2.2 PROBLEM STATEMENT

In order to enable research across healthcare networks, all of which use different formats and standards for their data, several large networks of RWD have made extensive investments to develop and deploy CDMs. CDMs are agreed-upon formats into which participating healthcare sites transform their data and which allow research across multiple health networks. However, each of these CDMs was created for a specific purpose and represent data in different formats with different rules. The data in different models cannot be easily shared among these networks.

2.3 DELIVERABLES

This project harmonized the CDMs developed by four data networks listed above to provide researchers in federal agencies and academia with access to data from a larger network of patients. With harmonization of the CDMs in these networks, the researchers will have access to not only EHR data, but to administrative claims data for conducting a diverse range of queries from basic PCOR queries, to large-scale sophisticated analyses, including clinical trials and post market-safety surveillance. Researchers now have at their disposal a unified access tool that allows researchers to multiple networks of observational data. Figure 1 illustrates the current and proposed common data architecture between CDMs and standards.

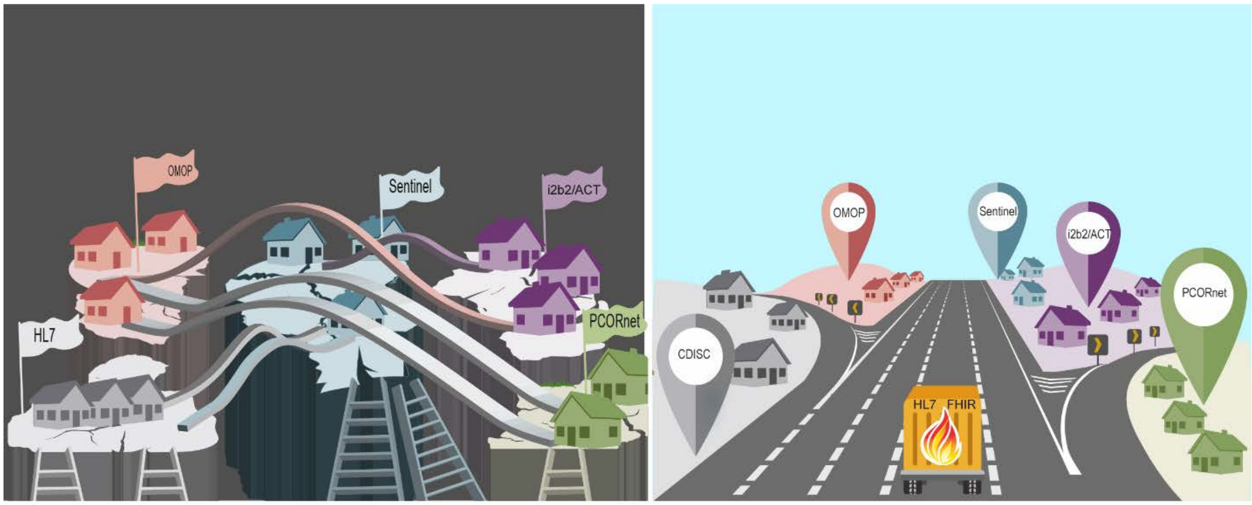


Figure 1. Current and envisioned Common Data Architecture between CDMs.

The key deliverables of this project were the following:

1. **Web-based portal and mapping tools:**

   - Validated mappings between four CDMs and the Biomedical Research Integrated Domain Group (BRIDG) model [1].
   - A web-based portal for access to Extract, Transfer and Load (ETL) and Mappings

2. **Standards-based technical architecture:**

   - Ability to query data using open standards and controlled terminologies
   - Query results represented in open standards and controlled terminologies

[1] BRIDG is a joint standard from HL7, CDISC and ISO. It is balloted periodically by all three SDOs.

3. **Implementation guides/data specifications:**

    - HL7 FHIR CDMH implementation guide, including the mappings from CDMs to BRIDG and CDMs to FHIR.
    - CDM data elements registered in NCI caDSR and NIH CDE Repository
    - Publicly available web site depicting the data element specifications

4. **Documentation:**

    - An environmental scan of existing CDMs mappings
    - Results of testing the mapped CDMs through an oncology drug safety use case
    - Methods and processes for ongoing curation and maintenance
    - Governance framework and policies for data use and processes
    - Web-based visualizations of the mappings between the CDM's data elements and data values based on BRIDG semantics

Deliverables have been placed in the public domain for interested organizations and PCOR researchers, who may want to apply this approach to collect data from these networks in their research.

3 METHODOLOGY

The purpose of the CDMH project is to enable the researcher to ask a question which will be translated and transmitted to multiple institutions (data partners) without having to worry about the data element mapping among multiple underlying data models and their database-specific capabilities, which will significantly expedite application development processes. Using this approach, a researcher can use a simple form-based web application (query builder) and use dropdowns and options to ask a question. Once the question is submitted, the CDMH tool translates (query translator) their selections into a database query specific to the institution's choice of common data model. The query is then transferred securely to data partners, who run this query against their data warehouses and send the result back. The results are then processed using an ETL mechanism into the CDMH database. Since the queries are sent to multiple institutions, there may be a delay in getting answers from everyone. However, the researcher can log back in and, using the CDMH portal's manage functionality (query manager), get status and view either partial or complete results (result viewer) as they are available.

4 MAJOR ACCOMPLISHMENTS

The accomplishments for the CDMH project have been grouped by objective.

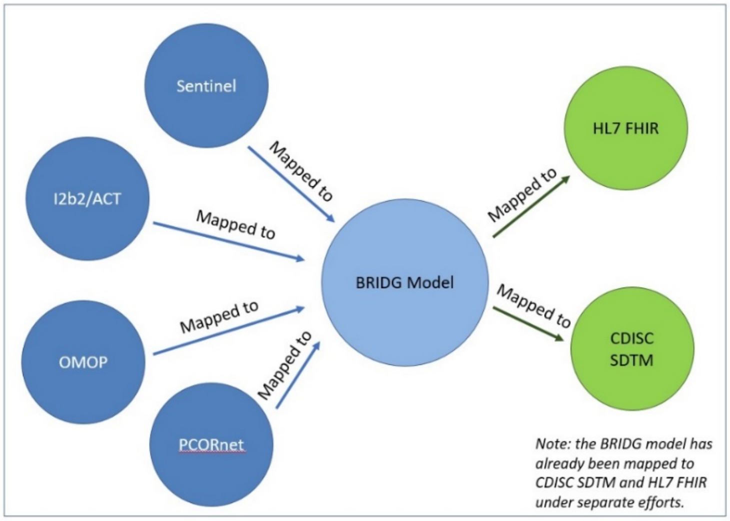
4.1 OBJECTIVE 1: DEVELOP THE CDM HARMONIZATION TOOL AND TRANSFORMATION PROCESS

This objective was to establish the core functions of the CDM Harmonization by creating the data logic which defines the relationship of data between CDMs and by selecting and developing the harmonization tool to power the transformation of data using the established data logic. In order to achieve this objective, the project team collaborated closely with key stakeholders from various CDMs, data partners, and Standards Development Organizations (SDOs).

The most appropriate intermediary model was identified to map data elements from each of the four CDMs to the concepts in the selected intermediary model. In addition, the team validated the mappings with the technical leads for each CDM. The validation of CDM mappings to the intermediary model was a critical step for ensuring its quality and performance.

4.1.1 Deliverable: Documentation of mappings from the four Common Data Models to the intermediary model (i.e. BRIDG)

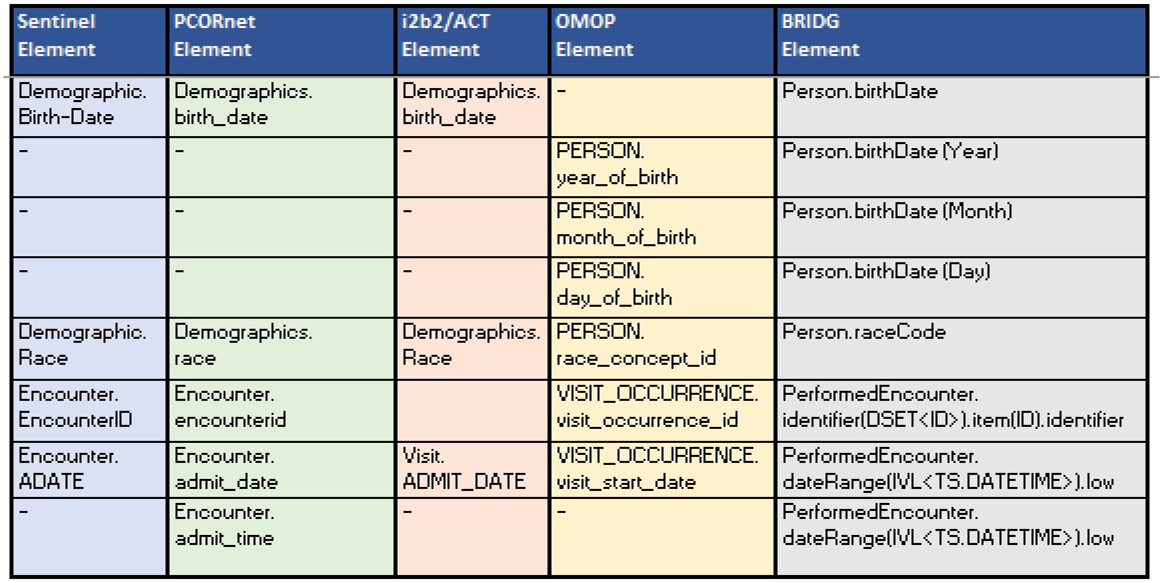
The project team collaborated closely with the technical leads from each of the four CDMs to validate the mappings from each CDM to the intermediary model (Figure 2). The BRIDG model was selected as the intermediary model, because CDISC SDTM and several HL7 v3 standards were already harmonized with this model. Appendix C provides detailed mapping process information. The data elements from the identified CDMs (i.e. Sentinel v6.0.2, PCORnet v3.1 and v4.0, OMOP v5.1 and I2b2/ACT v1.4) were mapped to the intermediary model. Mappings were validated and used to develop the backend query support for the portal to query across the CDMs and to retrieve RWD to answer research questions.



Note: the BRIDG model has already been mapped to CDISC SDTM and HL7 FHIR under separate efforts.

Figure 2. CDMs Mapping Process.

The mapping was documented in a single spreadsheet. Figure 3 shows the data elements from each of the CDMs and the corresponding BRIDG data elements.



Sentinel Element	PCORnet Element	i2b2/ACT Element	OMOP Element	BRIDG Element
Demographic. Birth-Date	Demographics. birth_date	Demographics. birth_date	-	Person.birthDate
-	-	-	PERSON. year_of_birth	Person.birthDate (Year)
-	-	-	PERSON. month_of_birth	Person.birthDate (Month)
-	-	-	PERSON. day_of_birth	Person.birthDate (Day)
Demographic. Race	Demographics. race	Demographics. Race	PERSON. race_concept_id	Person.raceCode
Encounter. EncounterID	Encounter. encounterid		VISIT_OCCURRENCE. visit_occurrence_id	PerformedEncounter. identifier(DSET<ID>).item(ID).identifier
Encounter. ADATE	Encounter. admit_date	Visit. ADMIT_DATE	VISIT_OCCURRENCE. visit_start_date	PerformedEncounter. dateRange(IVL<TS.DATETIME>).low
-	Encounter. admit_time	-	-	PerformedEncounter. dateRange(IVL<TS.DATETIME>).low

Figure 3. Sample of the mapping spreadsheet.

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry.

How it can be used: Researchers can use the mappings developed by this project as a tool to map their CDM to other models in order to conduct further analysis. The mappings developed in this project bridge the gap between these CDMs and may significantly result in a more comprehensive study.

Access to Resource: The mapping documents can be accessed at:
The CDMs to BRIDG mapping document: https://github.com/cdmhproject/cdmh
The BRIDG model with CDMH extensions is accessible: https://bridgmodel.nci.nih.gov/
The CDMH FHIR IG is accessible: http://www.hl7.org/fhir/us/cdmh/

4.1.2 Deliverable: Documentation of the Extract, Transform, and Load (ETL) selection and testing process

In order to automate the mapping process, the project team gathered business and technical requirements for an ETL tool, surveyed the market, and developed selection criteria for ETL tools. The team invited several ETL tool vendors to demonstrate their tool capabilities based on the CDMH project needs. The ETL selection process was also documented, in which a list of features for ETL tools and selection criteria were identified, and a weighted scoring spreadsheet was used to identify the most appropriate ETL tool vendor for this project.

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standard developers.

How it can be used: The ETL tool vendor selection criteria are available for researchers to use as a guideline or as a baseline to identify suitable ETL tools tailored to their studies.

Access to Resources:

1) A list of features for an ETL tool important to the CDMH project is accessible at: https://github.com/cdmhproject/cdmh/tree/master/Tools_and_Requirements.
2) A report capturing the ETL selection process based on the project's requirements is available at: https://github.com/cdmhproject/cdmh.

4.1.3 Deliverable: Implement data mapping processes using the selected ETL tool

The ETL tool was used to create the processes for data mapping (i.e., ETL-In (Transform Aggregate/Patient-Level data into CDMH) and ETL-Out (Export Patient Level Data to SDTM)) and data transformation (i.e., transform imported data into forms that are ready for export).

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standard developers.

How it can be used: Researchers can leverage the ETL tool to explore and obtain relevant information from raw data sources into standardized formats, so that additional data analysis can be carried out.

Access to Resource: The ETL tool at is available at NIH/NCATS https://github.com/cdmhproject/cdmh/tree/master/CDMH_Adeptia_Objects.

4.2 OBJECTIVE 2: TEST CDM HARMONIZATION TOOL IN A CLINICAL USE CASE

To assess the value of the developed CDM harmonization tool, the project team developed a protocol for the clinical use case, designed a query for the clinical use case, and executed the query in collaboration with several data partners.

4.2.1 Deliverable: the protocol for the clinical use case

A protocol in Appendix D was developed for the clinical use case selected for this project. The clinical use case focused on evaluating the safety of immunotherapy in cancer care in patients with auto-immune disorders and the identified clinical data elements were mapped to controlled terminologies (e.g., International Classification of Diseases, Clinical Modification, 9th and 10th Revision (ICD-9-CM, ICD-10-CM), RxNorm, National Drug Code (NDC), Systemized Nomenclature in Medicine - Clinical Terms (SNOMED CT)).

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standards developers, and epidemiologists.

How it can be used: Researchers can use the developed protocol, which includes controlled terminologies, as a use case for further development and testing of their own solutions.

Access to Resource: This protocol is available in Appendix D of this report.

4.2.2 Deliverable: Develop the queries for the clinical use case

The queries for each CDM were developed to focus on the clinical use case in collaboration with the technical lead for each of the 4 networks.

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standard developers, epidemiologists.

How it can be used: Researchers can employ the developed queries either as a template or as a starting point to be updated for the generation of queries tailored for their own data sources and domain applications.

Access to Resource: The queries are available on https://github.com/cdmhproject/cdmh.

4.2.3 Deliverable: Results of executed queries for the clinical use case

The project team collaborated with the following data partners, who executed the queries developed for each of the 4 CDMs (i.e. PCORnet, OMOP, Sentinel and i2b2/ACT) as the "Gold standard" and provided the results and lessons learned:

- Mayo Clinic/Yale University

  The team at the Mayo Clinic and Yale university executed the queries developed for PCORnet and OMOP and i2b2/ACT CDMs and assessed the technical and clinical validity of manually generated CDM queries.

- Elligo Health Research

  Data partners were identified by Elligo Health Research to test the Sentinel, OMOP and PCORnet use case queries and in one case, the original use case query specification. These data partners brought varied perspectives and tested different query types. They represented Hospital Corporation of America (Sentinel), Flatiron Health (proprietary methodology used to test original query specification), University of Chicago (OMOP and PCORnet), and IQVIA via ODHSI (OMOP). All data partners agreed to provide results in the aggregate, while University of Chicago also agreed to review line-level results.

- Sentinel Operations Center (SOC), Harvard Pilgrim Health Care Institute (HPHCI)

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standard developers, epidemiologists.

How it can be used: Lessons learned from the collaboration with multiple stakeholders (e.g... technical leads from each network and data partners) can be helpful to researchers.

4.3 OBJECTIVE 3: LEVERAGE OPEN STANDARDS AND CONTROLLED TERMINOLOGIES

One of the main goals of this project was to leverage standards and controlled terminologies in the CDM mapping and transformation process to support data integration and interoperability. In this work, BRIDG, CDISC SDTM, and HL7 FHIR standards and controlled terminologies were leveraged, including NDC, RxNorm, ICD-9-CM and ICD-10-CM.

4.3.1 Deliverable: Leverage HL7 FHIR standard in the development of the mapping tool

In this phase, the HL7 Common Data Model Harmonization FHIR Implementation Guide was developed to provide the instructions to use HL7 FHIR to implement the harmonized set of data elements created in this project. The Common Data Models Harmonization (CDMH) FHIR Implementation Guide (IG) focused on the mapping and translating observational data extracted for Patient Centered Outcomes Research (PCOR) purposes into FHIR format. Research data was extracted from the four targeted networks, each using a different data model to represent their data.

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standard developers and implementers.

How it can be used: Researchers can either leverage the CDMH FHIR implementation guide as a template to develop their own FHIR implementation guides or use the CDMH FHIR IG as a guide for developing their FHIR applications.

Access to Resources:

1) HL7 FHIR implementation guide: **http://www.hl7.org/fhir/us/cdmh/**
2) ONC CDMH resources: **https://www.healthit.gov/topic/scientific-initiatives/pcor/common-data-model-harmonization-cdm**

4.3.2 Deliverable: Leverage the BRIDG-to-CDISC SDTM mapping to export SDTM datasets

Additionally, the team developed a function to export, and then validate, query results in CDISC SDTM format out of the BRIDG-informed database. This made it possible to utilize results as part of submissions in the format required by FDA. The project team had to map controlled terminologies used by CDMs. For example, PCORnet, OMOP and i2b2/ACT CDMs use RxNorm terminology and Sentinel uses National Drug Code (NDC), and additional mapping from RxNorm to NDC was required.

The team was able to leverage the existing BRIDG-to-CDISC SDTM v3.2 mapping to implement a function in the mapping tool for exporting patient-level query results from the BRIDG-informed CDMH database in SDTM format. The SDTM format is a collection of SAS transport files, one per SDTM domain. Although CDMH does not contain study data, the sample SDTM dataset exported from CDMH contained simulated study data elements for completion. However, when study data is present in CDMH, the SDTM export function could be modified to include these data in the exported SDTM dataset.

For more information, please see Appendix F, which describes CDISC SDTM requirements for submitting drug information as one example and the required key terminologies.

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standard developers and implementers.

How it can be used: Researchers can select the results of a patient-level query leveraging the Common Data Model Harmonization and can export the results of the query to CDISC SDTM standard.

Access to Resource: http://www.hl7.org/fhir/us/cdmh/

4.3.3 Deliverable: Leverage NDC, RxNorm, ICD-9-CM, ICD-10-CM and other terminologies in the development of the mapping tool

Controlled terminologies (e.g., NDC, RxNorm, ICD-9-CM, ICD-10-CM) were leveraged in the CDM harmonization tool. The team used the FDA Global Substance Registration System (GSRS), a database that enables the efficient and accurate exchange of information on what substances are in regulated products. The GSRS database was used to identify the Dosage Form Code (e.g., C42916), Dosage Form Name (e.g., CAPSULE, EXTENDED RELEASE), Strength Number (e.g., "0.2"), Strength Numerator Unit (e.g., "g") and Strength Denominator Unit (e.g., "L"). The Query Builder User Interface (UI) used NDC and ICD-10-CM to represent Medications and Diagnosis codes in the UI. These values were translated to CDM-specific codes during the Query Translation phase. All these codes resided in CDMH BRIDG Code Value table. Various loader scripts were created to take the input files and load them to CDMH BRIDG.

Leveraging NLM RxNAV, a browser for several drug information sources, including RxNorm, was also explored to map NDC codes to RxNorm terms.

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standard developers.

How it can be used: This set of deliverables, including various codes and their mappings across multiple controlled terminologies (such as NDC, RxNorm, ICD-9-CM and ICD-10-CM),, the scripts loading input files to CDMH BRIDG, and the experience of leveraging NLM RxNAV for drug information sources can be used by researchers to identify suitable tools and mechanisms to harmonize terminologies.

Access to Resource:
https://github.com/cdmhproject/cdmh/tree/master/CDMH_Query_Management

4.4 OBJECTIVE 4: AUTOMATE THE QUERY DEVELOPMENT PROCESS

4.4.1 Deliverable: a web-based portal to create queries, access the ETL and mapping tools, and view results

To provide ease of use for the CDMH harmonization tool, a set of user-friendly tools were developed to 1) provide researchers with the ability to easily construct a question or query using the “CDMH Query Builder,” 2) monitor the status of queries using the “CDMH Query Manager,” 3) view the results of their query using the “CDMH Results Viewer,“ and 4) export the query results sets in a format of the users choosing (the “CDMH Export Function” tool).

The questions are entered by a researcher into the “CDMH Query Builder,” followed by translation of the query by the “CDMH Query Translator” into one or more of the CDM-specific query languages.

The status of the queries from various data partners that a researcher could monitor using “CDMH Query Manager” tool comes from a backend process that places the results received from the data partners in their specific CDMs in an Oracle database.

The “CDMH Results Viewer” allows the researcher to see the aggregated results from all the data partners and from all the different CDMs in a common format translated using an ETL process into BRIDG-based schema.

Finally, the researcher can export the aggregated result sets in various formats using the “CDMH Export Function.” For instance, a researcher at the FDA might want to create a CDISC-compliant SDTM file that can be imported using the FDA enterprise submission gateway, or a researcher working at the NIH might want to export the results into a CSV format for later analysis using a tool like SAS or R.

A high-level architecture diagram is shown below for reference (Figure 4). More detailed information on the technical solution and its various components are available in Appendix A.

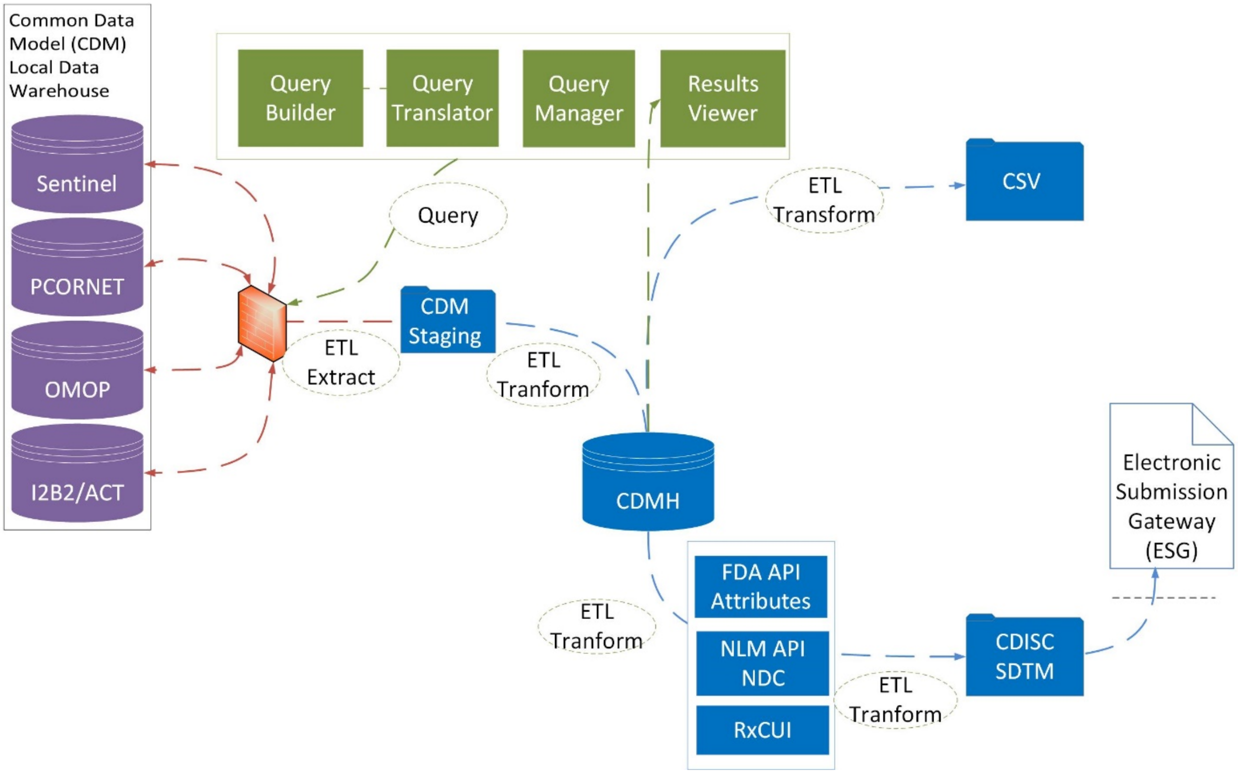


Figure 4. CDMH Technical solution - high level architecture diagram.

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standard developers.

How it can be used: Researchers interested in using the CDMH tools can contact the CDMH team for account creation. The CDMH team will create the account and will share the user guide with the researchers. Additionally, the GUI layout, variable selection, and the mechanism of generating complex queries from user selections via the web-based portal can function as a developers' guide when researchers intend to generate a similar web-based application.

Access to Resource: This is a web-based portal for access to the ETL and mapping tools available at: https://github.com/cdmhproject/cdmh.

4.4.2 Deliverable: Develop and execute the query for the clinical use case using the selected ETL tool

After the mapping and harmonization of the four CDMs, the project team automated the clinical query (Appendix D) development process for each model (Sentinel, OMOP, i2b2/ACT and PCORnet v3.1 and v4.0).

In order to validate the automatically generated queries made by the "CDMH Query Builder" and "CDMH Query Translator," queries were also developed manually for each of the CDMs, in close collaboration with the technical leads for each of the networks, to act as the control or gold standard. In parallel queries for OMOP CDM and two versions of the PCORnet, CDMs were auto-generated using the selected ETL “CDMH Query Builder” and Query Translator” tools. The results were compared from the manually generated query with the ETL-generated query. To develop the clinical query for the Sentinel CDM, the FDA collaborated closely with the Sentinel Operations Center (SOC) team at Harvard Pilgrim Health Care Institute. This query was written in SAS and was executed by the team. A tool to create an auto-generated query in a SAS format was not developed.

After mapping of the CDMs to the intermediary model, the auto-generated query was provided to data partners at Mayo Clinic/Yale University and to additional partners identified by Eligo Health Research to test the project’s approach and demonstrate its value. Data partners were identified by Elligo Health Research to test the Sentinel, OMOP, and PCORnet use case queries and in one case, the original use case query specification. These data partners (representing Hospital Corporation of America (Sentinel), Flatiron Health (proprietary methodology used to test original query specification), University of Chicago (OMOP and PCORnet) and IQVIA via ODHSI (OMOP)) brought varied perspectives and tested different query types. They all agreed to provide results in aggregate, and data partners at the University of Chicago also agreed to review line-level results.

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, and standards developers.

How it can be used: The query for the oncology use case can be used by various researchers, who can also log into the User Interface (UI) and download the results of the query.
Access to Resource: https://github.com/cdmhproject/cdmh

4.5 OBJECTIVE 5: DEVELOP A DATA VALIDATION PROCESS

1. Data partners stored and maintained data in different formats with varied metadata specifications. A data validation process was developed to ensure consistent processing of these potentially heterogenous datasets. The data validation process included obtainment of the data, its processing using the ETL tool, and the data's secure dissemination.

4.5.1 Deliverable: Develop the data validation process

The project teams stored the results received from the data partners on a secure server and the ETL tool loaded these results into the staging area. The staging area tables mirrored the format of the data partner environment.

Separate staging tables are created for Aggregate Results and Patient-Level Results for each partner.

- Aggregate Results: Aggregate consist of patient counts that meet the selection criteria specified by the query. The staging table for aggregate results reflects the CSV file format, plus metadata fields about the query and results
- Patient-Level Results: The are contained in a single .zip file containing multiple CSV files. Each CSV file contains a dump of all records from a database table that meet the selection criteria. The staging tables matches the table structure that exists at the data partner.

The ETL process is configured to reject the file from staging in case of any error during the load process.

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, and standards developers.

How it can be used: The specification of the implemented data validation process in CDMH is highly reusable and extremely viable to implement. Researchers can adopt this data validation process, including the generation of the staging tables, the setup of the mirrored data environments, and the separate evaluations of patient-level and aggregate results.
Access to Resource: The ETL process to stage the results reside in ETL tool at the NIH/NCATS environment and is accessible at: https://github.com/cdmhproject/cdmh

4.6 OBJECTIVE 6: INCORPORATE AND REUSE OF EXISTING NIH AND OS-PCORTF INFRASTRUCTURE

The project team incorporated and used previously funded OS-PCORTF resources, i.e., Data Access Framework – Research (http://hl7.org/fhir/us/daf-research/), and the NIH CDE Repository (https://cde.nlm.nih.gov/), as well as existing infrastructure, e.g., the Cancer Data Standards Registry and Repository (caDSR). The value of recording reusable, machine-readable definitions of models and data elements informs other initiatives such as the Cancer Moonshot Initiative, the *All of Us* Research Program. Deliverables from Section 4.1.1 helped NCI with creating caDSR - structured metadata and representing the mappings between CDMs and BRIDG. The NIH CDE Repository now hosts this new metadata content. This was possible because NCI and NIH used the same ISO standard as a basis for these repositories.

An intuitive visualization was created for end users to see how the models were related to each other and to the higher level conceptual BRIDG Model. This demonstrated one of the values of creating machine-readable metadata. This visualization could not have been created from traditional word processing and document creation software such as MS Word or PDF file reader/editor or Spreadsheet software. A visualization tool was developed to display not only the data mappings among various CDMs but also the relationships (e.g., the hierarchical or categorical relationships) of the data at different levels of granularity, from high-level concept categories in the BRIDG like Biologic Entity, to the lowest level of permissible values (PVs) for a given data element like Person Biologic Entity Ethnic Group Code.

The exchange of the caDSR CDE information with NIH CDE Repository demonstrated the value of creating structured CDE metadata based on a common metadata standard.

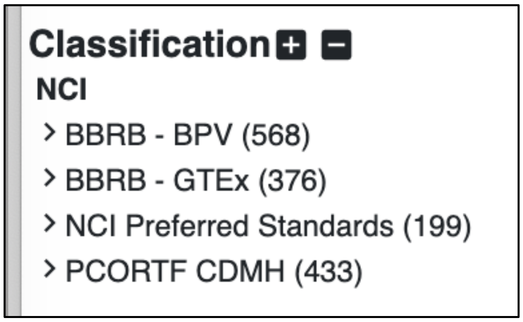
4.6.1 Deliverable: Demonstration illustrating the re-use of NCI caDSR

Common data elements (CDEs) for the four CDMs were created in NCI caDSR so that they are visible using the CDE Browser and in the NIH CDE Repository. The CDEs document the fields and their data values. A visualization tool was created to allow CDM owners and users to see how the data elements are semantically aligned and which data values are equivalent.

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standard developers, data managers

How it can be used: The researchers interested in reusing or combining data across the CDMs can use this human- and machine-readable documentation to map and transform the data. The researchers interested in collecting data that is directly compatible with any of the CDEs can use the metadata in NIH or NCI repository to design new data collection forms based on the CDM data models. The visualization can be used by researchers to intuitively explore the actual mappings among multiple CDMs in the CDMH use cases. The visualization solution can also be leveraged for the design and development of the researcher's own way of presenting relevant data either in CDMs or in other data forms.

Access to Resource: Visualization tool**:** https://vis-review-si.nci.nih.gov/,
CDE Browser: https://cdebrowser.nci.nih.gov under PCORTF CDM Program Area, NIH CDE Repository under NCI→PCORTF CDMs as shown in Figure 5.



Classification ⊞ ⊟

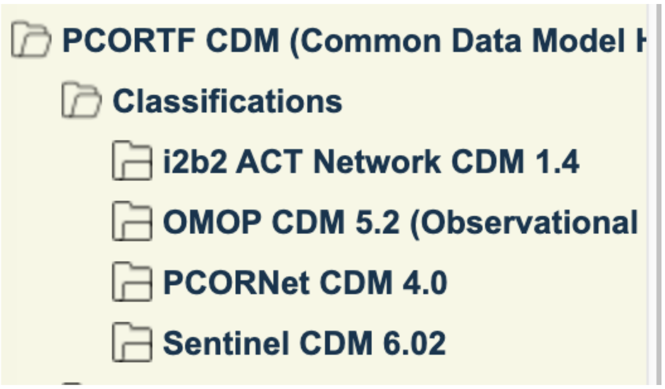
NCI

- BBRB - BPV (568)
- BBRB - GTEx (376)
- NCI Preferred Standards (199)
- PCORTF CDMH (433)

Figure 5. NIH CDE Repository entry for PCORTF CDM CDEs.

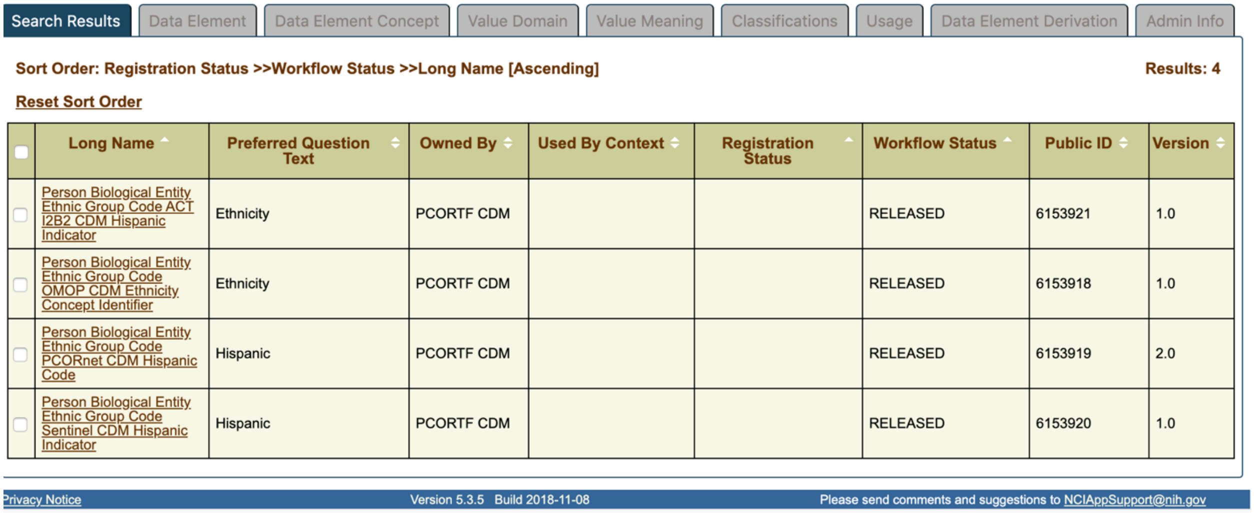
4.6.1.1 Overview of visualization

The metadata in the NCI caDSR was presented in the CDE Browser as lists of CDEs for each CDM as shown in Figure 6. All the CDEs in the 4 CDMs were aligned to the Biomedical Research Integrated Domain Group as a common conceptual model in caDSR. This was accomplished by linking the CDM details to NCI Thesaurus (NCIt) concepts. As an example, Figure 7 shows the CDEs for ethnicity in all 4 models. The details of each CDE can be viewed by clicking on the CDE and exploring the different tabs that display the standardized CDE name, the CDM name for the CDE, the concepts associated with the CDE, and the details across all the models for of any enumeration for the CDE. Figure 8 shows the permissible values for the PCORnet ethnicity CDE. The information can be easily downloaded in several formats, e.g., MS Excel and XML, to the support use of the information in various ways, such as to customize data collection tools or to produce custom reports.



PCORTF CDM (Common Data Model H
- Classifications
  - i2b2 ACT Network CDM 1.4
  - OMOP CDM 5.2 (Observational
  - PCORNet CDM 4.0
  - Sentinel CDM 6.02

Figure 6. CDEs are grouped by CDM in caDSR.



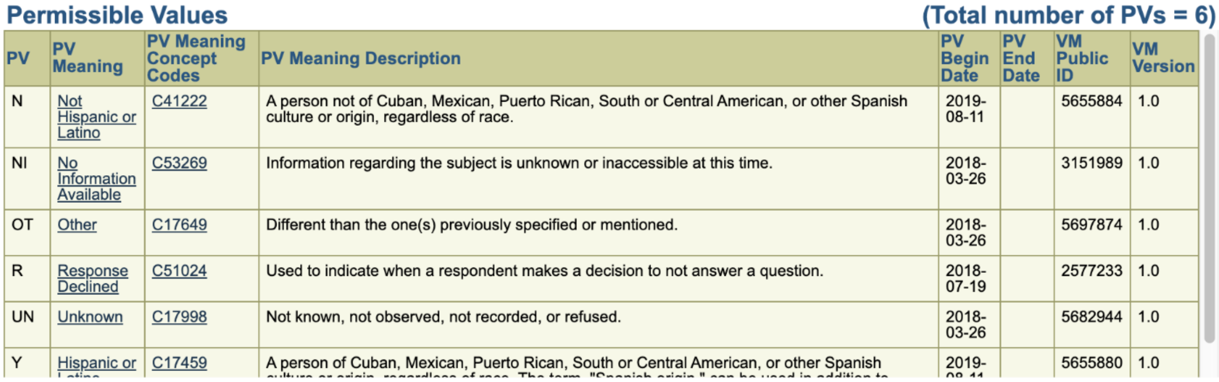
Sort Order: Registration Status >>Workflow Status >>Long Name [Ascending] **Results: 4**

Reset Sort Order

□	Long Name	Preferred Question Text	Owned By	Used By Context	Registration Status	Workflow Status	Public ID	Version
□	Person Biological Entity Ethnic Group Code ACT I2B2 CDM Hispanic Indicator	Ethnicity	PCORTF CDM			RELEASED	6153921	1.0
□	Person Biological Entity Ethnic Group Code OMOP CDM Ethnicity Concept Identifier	Ethnicity	PCORTF CDM			RELEASED	6153918	1.0
□	Person Biological Entity Ethnic Group Code PCORnet CDM Hispanic Code	Hispanic	PCORTF CDM			RELEASED	6153919	2.0
□	Person Biological Entity Ethnic Group Code Sentinel CDM Hispanic Indicator	Hispanic	PCORTF CDM			RELEASED	6153920	1.0

Privacy Notice Version 5.3.5 Build 2018-11-08 Please send comments and suggestions to NCIAppSupport@nih.gov

Figure 7. The CDEs for Ethnicity in all 4 CDMs.



Permissible Values (Total number of PVs = 6)

PV	PV Meaning	PV Meaning Concept Codes	PV Meaning Description	PV Begin Date	PV End Date	VM Public ID	VM Version
N	Not Hispanic or Latino	C41222	A person not of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin, regardless of race.	2019-08-11		5655884	1.0
NI	No Information Available	C53269	Information regarding the subject is unknown or inaccessible at this time.	2018-03-26		3151989	1.0
OT	Other	C17649	Different than the one(s) previously specified or mentioned.	2018-03-26		5697874	1.0
R	Response Declined	C51024	Used to indicate when a respondent makes a decision to not answer a question.	2018-07-19		2577233	1.0
UN	Unknown	C17998	Not known, not observed, not recorded, or refused.	2018-03-26		5682944	1.0
Y	Hispanic or Latino	C17459	A person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin, regardless of race. The term, "Spanish origin," can be used in addition to	2019-08-11		5655880	1.0

Figure 8. The permissible values for PCORnet CDE Person Biological Entity Ethnic Group Code.

The main disadvantage of this type of view arises from the inability of end users to see the relationships that are present in the metadata based on the concept information, including but not limited to the semantic hierarchical relationships in the BRIDG model, and the mapping of equivalent values across the CDMs.

To empower ease of visualizing the interrelationships between the elements, a web-based visualization was developed to specifically present the hierarchical relationships of metadata at different levels and the mapping of the data values of the CDMs. Figure 9 presents an example showing the hierarchical relationships of the metadata at different levels. The highest-level node, the root node, named "BRIDG-To-CDMs" is a notional entry point to the entire visualization. The next-level nodes are the child nodes of the root node, among which only Activity, Outcome and Performed Specimen Collection are expanded to show their child nodes. A user can keep expanding a node to show its child nodes until reaching the deepest level, level # 6, the leaf nodes. A leaf node contains the mapping relationships of a same/similar/related concepts in multiple CDMs, as shown in Figure 9. Using the same CDEs in the above figures, the mapping relationship in the figure is presented under, or indexed by, a PCORnet node, whose name is "Person Biological Entity Ethnic Group Code PCORnet CDM Hispanic Code – CDE – 6153919" (not shown). Figure 10 illustrates the use of standard terminology concept codes to link different values in each of the models. As an example, the value of "UN-Unknown-C17998" of PCORnet for Hispanic Code corresponds to the value of "U-Unknown-C17998" in Sentinel, "UNK-Unknown-C17998" in FHIR, and "UNKNOWN-Unknown-C17998" in CDISC CDASH, and there is no corresponding value in ACT or OMOP. The NCI Thesaurus concept code of C17998 means "Unknown."

The visualization included additional functionalities to allow a user to zoom in/out, pan to see information outside of the current screen, and expand and collapse nodes to show and hide child nodes. A "Search" functionality was implemented to allow a user to type search term for not only searching but also locating the node to the center of the screen for the user's convenience. The Search box also provided a list of suggested terms as the user types to enhance user experience.

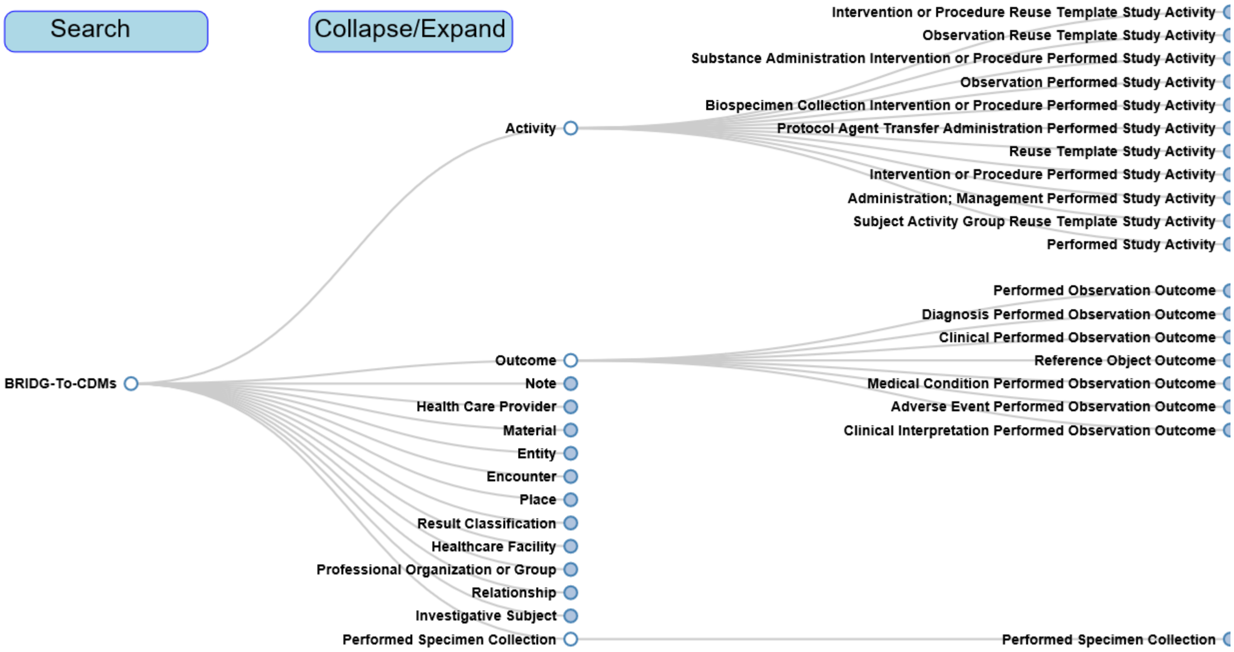

Search
Collapse/Expand
BRIDG-To-CDMs
Activity
Outcome
Note
Health Care Provider
Material
Entity
Encounter
Place
Result Classification
Healthcare Facility
Professional Organization or Group
Relationship
Investigative Subject
Performed Specimen Collection
Intervention or Procedure Reuse Template Study Activity
Observation Reuse Template Study Activity
Substance Administration Intervention or Procedure Performed Study Activity
Observation Performed Study Activity
Biospecimen Collection Intervention or Procedure Performed Study Activity
Protocol Agent Transfer Administration Performed Study Activity
Reuse Template Study Activity
Intervention or Procedure Performed Study Activity
Administration; Management Performed Study Activity
Subject Activity Group Reuse Template Study Activity
Performed Study Activity
Performed Observation Outcome
Diagnosis Performed Observation Outcome
Clinical Performed Observation Outcome
Reference Object Outcome
Medical Condition Performed Observation Outcome
Adverse Event Performed Observation Outcome
Clinical Interpretation Performed Observation Outcome
Performed Specimen Collection

Figure 9. A display of the hierarchical relationships of the data.

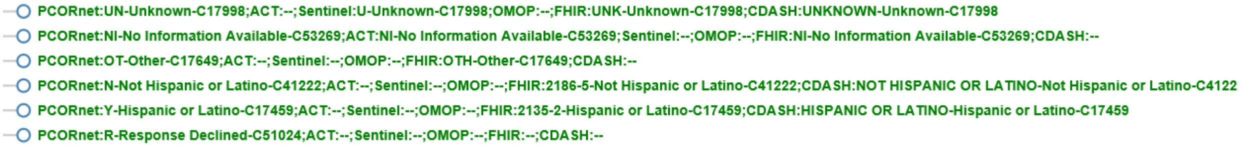

○ PCORnet:UN-Unknown-C17998;ACT:--;Sentinel:U-Unknown-C17998;OMOP:--;FHIR:UNK-Unknown-C17998;CDASH:UNKNOWN-Unknown-C17998
○ PCORnet:NI-No Information Available-C53269;ACT:NI-No Information Available-C53269;Sentinel:--;OMOP:--;FHIR:NI-No Information Available-C53269;CDASH:--
○ PCORnet:OT-Other-C17649;ACT:--;Sentinel:--;OMOP:--;FHIR:OTH-Other-C17649;CDASH:--
○ PCORnet:N-Not Hispanic or Latino-C41222;ACT:--;Sentinel:--;OMOP:--;FHIR:2186-5-Not Hispanic or Latino-C41222;CDASH:NOT HISPANIC OR LATINO-Not Hispanic or Latino-C4122
○ PCORnet:Y-Hispanic or Latino-C17459;ACT:--;Sentinel:--;OMOP:--;FHIR:2135-2-Hispanic or Latino-C17459;CDASH:HISPANIC OR LATINO-Hispanic or Latino-C17459
○ PCORnet:R-Response Declined-C51024;ACT:--;Sentinel:--;OMOP:--;FHIR:--;CDASH:--

Figure 10. A display of the data mapping relationships indexed by PCORnet among multiple CDMs.

4.6.2 Deliverable: Demonstration illustrating the re-use of NIH CDE Repository

The data elements were provided for the 4 CDMs to the NIH CDE Repository.

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standards developers

How it can be used: The information in the NIH CDE repository can be used to explore the CDEs in each of the 4 CDM models.

Access to Resource: NIH CDE At the Repository (https://cde.nlm.nih.gov/cde/search), click on NCI.

4.6.3 Deliverable: Demonstration illustrating the re-use of Data Access Framework (DAF) project

The DAF project used a phased approach to enable:

1. Local data access via queries within an organization (phase 1)
2. Targeted data access via queries between organizations (phase 2)
3. Data access for researchers (phase 3) to access multiple patients' data from multiple organizations

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standard developers.

How it can be used: Interested stakeholders and researchers can use the following three implementation guides (IGs) to enhance data access:

1. IHE DAF Document Metadata Based Access IG
2. Health Level Seven FHIR® US Core IG Release 1 (formerly known as DAF Core)
3. HL7 FHIR® DAF for Research IG Release 1

Access to Resources:

1. http://ihe.net/uploadedFiles/Documents/PCC/IHE_PCC_IG_DAF_National-Extension.pdf
2. http://hl7.org/fhir/us/core/
3. http://hl7.org/FHIR/us/daf/2016Sep/daf-research.html

4.7 OBJECTIVE 7: DEVELOP AND TEST EXPORT FORMATS

For this objective, the project team stored the query results received from participating data partners, exported the results to the CDISC SDTM standard (CDMH will incorporate clinical research data in the future), and validated the results. CDISC SDTM is the data standard used by the FDA to receive clinical research data.

4.7.1 Deliverable: Export to CDISC SDTM standard

The results were exported to CDISC SDTM.

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standard developers

How it can be used: For Patient Level Query from Results Viewer UI, researchers have the option to export the results in CDISC SDTM format. When researcher clicks on Export to SDTM, a background ETL process is triggered which translates the CDMH BRIDG values into SDTM values for that query and zip all the STDM files when they are ready. The researcher can save this zip locally.
Access to Resource: The Query Management UI to view Results Viewer is hosted in NIH/NCATS environment and is accessible at: https://github.com/cdmhproject/cdmh

4.8 OBJECTIVE 8: EDUCATION AND GOVERNANCE

The project team developed a governance framework document, educational material for use of the CDMH tools, and best practices for access to and use of the RWD.

The purpose of the governance framework was to outline governance policies and practices for access to and use of the RWD that are derived from data-sharing networks that connect CDMs, such as the Harmonization of Various Common Data Models for Evidence Generation (CDMH) [7] project. In the governance framework document, the project team described the foundations of data governance as they may apply to data-sharing networks, recommended a governance structure to establish and maintain policies and procedures, listed some documents that data-sharing networks should consider creating and/or adopting for their own governance and system access, and outlined other relevant issues that data-sharing networks may want to consider when establishing policies.

4.8.1 Deliverable: A public use environment for publication and access to mapping tools

A public environment was setup for researchers and provided them with access to the CDMH harmonization tools and materials.
Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standard developers
How it can be used: This material will describe the process, for using the CDMH solution, as well as the educational material for using the tools.
Access to Resource: A public environment accessible at: https://github.com/cdmhproject/cdmh was created.

4.8.2 Deliverable: Published project report, capturing the proposed governance framework, policies for data use and processes

This project report described the development process and approach to harmonize several CDMs, leveraging standards and controlled terminologies.
Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standard developers
How it can be used: The researchers can access publicly available resources developed as part of this project to guide their own development process for different domain applications.
Access to Resource: Public report accessible at: FDA and ASPE’s web sites and Appendix E.

4.9 OBJECTIVE 9: MAINTENANCE AND CURATION

The team developed a maintenance and curation plan and implemented a version control for software and data dissemination. The developed CDM mappings will require maintenance throughout the lifecycle of the project and beyond.

4.9.1 Deliverable: Develop and document methods and standard process for ongoing curation

Quality metadata is a powerful resource enabling consistent collection, analysis, aggregation, and discovery and reporting of data sets. The Standard Operating Procedure (SOP) for ongoing curation of CDM content leverages NCI's existing metadata governance plans and practices. A plan was developed to provide a framework outlining the decision rights and accountability of the CDM Stakeholders and Stewardship Organizations (SO), who store metadata content in the caDSR. The model owners fall into the category of an SO.

This management approach is aligned with the roles and responsibilities defined in International Organization for Standardization/ International Electrotechnical Commission (ISO/IEC) 11179, the international metadata registry standard, and implements the full set of recommended ISO practices. The governance process includes proactively reaching out to the content owners should any content be reused, and questions arise from the community. Once new versions of CDMs are published and the FDA updates the CDM mappings and sends them to NCI, NCI will curate new content or update existing and create reports for SO review. The report will contain details about each data element in the model, the semantic concept mappings for each data element and each valid value if the field is enumerated. The report will be reviewed and approved by the model owner before publishing the content to the NIH CDE Repository.

The report consists of the details about each CDM CDE including the CDM field name, the new standardized long name, curated question text suitable for eliciting the data, a text definition, the BRIDG class/attribute it has been mapped to and NCIt concept annotations.

The team developed a maintenance and curation plan and implemented a version control for software and data dissemination. The developed CDM mappings requires maintenance throughout the lifecycle of the project and beyond. A complete CDM Data Dictionary has been generated and made available to the public following the URL at the GitHub in the "Access to Resource" section below.

Target Audience: CDMs network collaborators, HHS researchers, academia, the biopharmaceutical industry, standard developers.

How it can be used: This SOP will be used to help model owners understand how to help keep the structured metadata updated and synchronized the NIH published version. The documentation will help the audience understand the mappings.

Access to Resource: https://wiki.nci.nih.gov/display/caDSR/caDSR+Metadata+Governance.
A complete CDM Data Dictionary:
https://github.com/cdmhproject/cdmh/tree/master/CDM%20Data%20Dictionary

4.9.2 Deliverable: Implement version control for software / data dissemination

The ETL tool was configured to use VisualSVN (Subversion) for version control. We used SVN to manage object modifications to prevent unintended overwrites. If any change was needed to any existing object they were checked out and updated. Similarly, we maintained Query Management UI code in SVN repository.
Target Audience: Standard ETL and Web developers.

How it can be used: The interested developers and researchers could save a working copy locally, make updates and push the changes once tested.

Access to Resource:
The version control system in the internal development stage is within NCATS: https://cdmhdev.ncats.io/!/#.
The version control system after the internal development stage including future maintenance is at GitHub: https://github.com/cdmhproject/cdmh.

5 DISCUSSION

5.1 LESSONS LEARNED AND CONSIDERATIONS FOR FUTURE WORK

This project confirmed the need for harmonization across CDMs to generate RWE and the need to leverage open, consensus-based standards. However, several challenges were encountered, including:

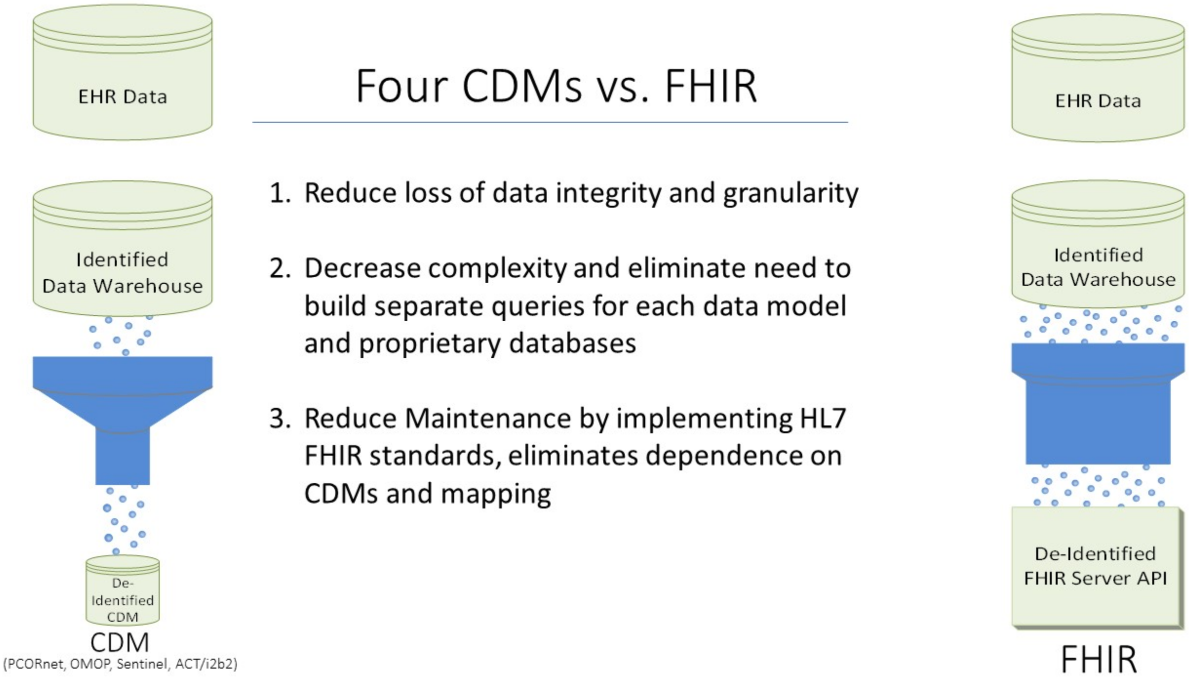
- Lack of data consistency and quality, and missing values at the point of data collection.
- Healthcare infrastructure heterogeneity. Various data partners used different flavors of relational databases. For example, a query written to run on an “Oracle” database system would not, without modification, run on a “Microsoft SQL Server” database system.
- Lack of data standardization at the point of care.
- Differences in data structure and domains between various CDMs.
- Differences in controlled terminologies in use across various CDMs.
- Maintenance of the CDMH infrastructure requires additional overhead.
- Data sharing challenges. There is a need for a universal data use agreement.

5.2 PROPOSAL FOR FUTURE WORK

The goal of Phase II of the CDMH project is to complete the implementation of Phase I as well as address the challenges of scaling query transformation and the loss of data inherent with all data models. To address these challenges FDA and NIH/NCATS are leading Phase II of the CDMH project. This new phase leverages the work that was done in this project and shifts the focus of the project to facilitating data transfer from data partners to HHS agencies using the HL7 FHIR standard. HL7 FHIR is a healthcare data exchange standard developed by HL7. Its use has seen widespread adoption in the last few years and has quickly become the common language of health care data exchange. This is, to a large extent, because of federal requirements to leverage the FHIR standard for use by health IT systems [8] [9]. Other federal agencies including NIH [10], FDA [11] and AHRQ [12] have encouraged or required its use, as has private industry, including technology giants such as Apple and Google. The reason for its high adoption is its ease of use. FHIR combines standard web-based tools and protocols (e.g., HTTPS, Restful APIs, JSON with FHIR resources (reusable units of information exchange and defined structured data)) to make movement of health care data easy to implement. In 2016, when Phase I of the CDMH proposal was written, HL7 FHIR was an evolving draft standard. However, since then FHIR Resources have matured. In addition to FHIR resources that focus on individual member, single patient and consumer specific exchange, a specification for the exchange of large volumes of data at a population level (HL7 FHIR: Bulk Data) has also been developed into a more mature state. Finally, though FHIR is not a data model, companies like Google and Microsoft have started offering commercial software FHIR repositories capable of storing FHIR results. FHIR’s maturation and ubiquitous adoption opens new possibilities to address some of the challenges encountered in this project.

5.3 EXTENDING CDMH CAPABILITY

The Phase II diagram (Figure 11) illustrates the data flow from either the 4 CDMs or the FHIR repository.



Four CDMs vs. FHIR

1. Reduce loss of data integrity and granularity
2. Decrease complexity and eliminate need to build separate queries for each data model and proprietary databases
3. Reduce Maintenance by implementing HL7 FHIR standards, eliminates dependence on CDMs and mapping

EHR Data
Identified Data Warehouse
De-Identified CDM
CDM
(PCORnet, OMOP, Sentinel, ACT/i2b2)
EHR Data
Identified Data Warehouse
De-Identified FHIR Server API
FHIR

Figure 11: Phase II Process Flow Diagram.

6 CONCLUSION

The CDMH project was is a multipart project that is conceptualized as a standards-based, HHS-wide framework that allows any HHS agency to interrogate RWD to improve patient health. The project focused on the harmonization of the four existing CDMs (i.e., Sentinel, i2b2/ACT, PCORnet, and OMOP) to facilitate access to a majority of the US population. The CDMH project team identified and tested tools, intermediary models, standards and controlled terminologies to harmonize several CDMs across multiple types of data sets (e.g., clinical and administrative claims) to support RWE. Through multiple activities, the CDMH team developed a set of tools and resources to enhance the retrieval and process of RWD. Each aspect of this project taught us lessons on the viability of using EHRs and administrative claims. These lessons were related to legal, ethical, and privacy concerns, as well more technical areas such as quality control, methodology, data standards and controlled terminologies and types of patient data (e.g. administrative claims vs. clinical). These lessons and the CDMH infrastructure developed in Phase I will be leveraged in Phase 2 of this project.

7 DISSEMINATION

In addition to the contracted deliverables for this project, the following activities were conducted to disseminate the work.

7.1 CONFERENCES

- Oral presentation:

  Bridging clinical research and clinical health care

  Washington, DC, February 28, 2018
- Oral presentation: Re-Imagine HHS

  Webinar, June 20, 2019
- Oral presentation: TransCelerate BioPharma Inc.

  Webinar, July 25, 2019
- Oral presentation: Clinical and Translational Science Awards (CTSA) Annual Meeting

  Crystal City, VA, September 26, 2019
- Oral presentation: HHS/ASPE Summer Webinar Series

  Webinar, July 23, 2019
- Oral presentation:
  AMIA 2018 Informatics Summit
  San Francisco, CA, March 12-15, 2018
- Oral presentation:
  Health Care Systems Research Network (HCSRN) Conference
  Minneapolis, MN, April 11-13, 2018
- Oral presentation:

  AMIA 2018

  San Francisco, CA, November 05, 2018
- Oral presentation:

  ONC Tech Forum

  Webinar, August 11, 2020
- Oral presentation:

  Learning Health Community eSource Symposium

  Webinar, August 20, 2020

8 ACKNOWLEDGEMENTS

The CDMH team would like to thank the members and collaborators for their work on this project as shown below in Table 1:

Table 1. Team Acknowledgements

Collaborating Partner	Participant
CDMH Team	• Mary Ann Slack • Mitra Ahadpour • ShaAvhrée Buckman-Garner • Bob Ball • Michael Nguyen • Esther Zhou • Jacqueline Puigbo • Jean Duteau • Mikhail Krivitskiy • Sarah Dutcher • Jason Gorsuch • Gregory Klebanov • Kathryn Matto • Richard Ballew • Vaishnavi Rao • Mike Flanigan • Elizabeth Ramsey
Mayo Clinic/Yale University CERSI	• Joe Ross • Nilay Shah • Wade Schulz • Paul Kingsbury • Guoqian Jiang • Patrick Young • Laura Ciaccio • Jessica Ritchie
Elligo Health Research	• Rebecca Kush • Michael Ibara Data Partner Representatives: o Sam Volchenboum (University of Chicago) o Brian Furner o Nicholas Brown o Christian Reich o Henry Morgan Stewart

Collaborating Partner	Participant
	o Amy Abernethy o Shrujal Baxi o Laura Koontz o David Volcano
Adeptia ETL Team	• Krishna Kumar
Common Data Models Technical Leads	• Jeff Brown • Christian Reich • Keith Marsolo • Jeff Klann • Michele Morris

REFERENCES

[1] FDA, "Real-World Evidence," [Online]. Available: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.

[2] FDA, "21st Century Cures Act," [Online]. Available: https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/21st-century-cures-act.

[3] "Roadmap 2024," [Online]. Available: https://www.healthit.gov/topic/interoperability.

[4] HealthIT.gov, "Data Access Framework (DAF)," healthit.gov, [Online]. Available: https://www.healthit.gov/topic/scientific-initiatives/pcor/data-access-framework-daf.

[5] "NIH Common Data Element Repository," NIH, [Online]. Available: https://cde.nlm.nih.gov.

[6] NIH NCI, "Cancer Data Standards Registry and Repository (caDSR)," [Online]. Available: https://wiki.nci.nih.gov/display/cadsr.

[7] HHS ASPE/FDA, "Harmonization of Various Common Data Models for Evidence Generation (CDMH)," [Online]. Available: https://aspe.hhs.gov/harmonization-various-common-data-models-and-open-standards-evidence-generation.

[8] "21st Century Cures Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program," Office of the National Coordinator for Health Information Technology (ONC), Department of Health and Human Services (HHS), 2020. [Online]. Available: https://www.federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification.

[9] "Medicare and Medicaid Programs; Patient Protection and Affordable Care Act; Interoperability and Patient Access for Medicare Advantage Organization and Medicaid Managed Care Plans, State Medicaid Agencies, CHIP Agencies and CHIP Managed Care Entities, Iss," Centers for Medicare & Medicaid Services (CMS), HHS.2020, [Online]. Available: https://www.federalregister.gov/documents/2020/05/01/2020-05050/medicare-and-medicaid-programs-patient-protection-and-affordable-care-act-interoperability-and.

[10] "NIH Fast Healthcare Interoperability Resources® Initiatives," NIH, [Online]. Available: https://datascience.nih.gov/fhir-initiatives.

[11] "CDER Data Standards Program 2019Annual Assessment," U.S. FDA CDER, March 2020. [Online]. Available: https://www.fda.gov/media/136436/download.

[12] "AHRQ's Digital Solutions to Support Care Transitions Challenge," AHRQ, [Online]. Available: https://www.ahrq.gov/mcctransitions-challenge/index.html.

[13] "PCORnet Front Door Policy," [Online]. Available: https://pcornet.org/wp-content/uploads/2016/05/PCORnet-Front-Door-Policy_Final_4.15.16.pdf.

[14] The U.S. Department of Health and Human Services Data Council, "2018 HHS Data Strategy: Enhancing the HHS Evidence-Based Portfolio," 2018. [Online]. Available: https://aspe.hhs.gov/system/files/pdf/261591/2018HHSDataStrategy.pdf.

[15] strategy.data.gov, "Federal Data Strategy," [Online]. Available: https://strategy.data.gov/overview.

[16] NIH, "Protecting Per sonal Health Information in Research: Understanding the HIPAA Privacy Rule, NIH Publication Number 03-5388," [Online]. Available: https://privacyruleandresearch.nih.gov/pdf/hipaa_privacy_rule_booklet.pdf.

[17] NIST, "data integrity," NIST, [Online]. Available: https://csrc.nist.gov/glossary/term/data-integrity.

[18] "Certification Companion Guide: Integrity," [Online]. Available: https://www.healthit.gov/test-method/integrity .

[19] Electronic Code of Federal Regulations, "Electronic Code of Federal Regulations, PART 11—ELECTRONIC RECORDS; ELECTRONIC SIGNATURES," [Online]. Available: https://www.ecfr.gov/cgi-bin/text-idx?SID=152af9e2fef388914fc41d9ba067b09c&mc=true&node=pt21.1.11&rgn=div5.

[20] HHS, "HHS - HIPAA Home - For Professionals - Special Topics - Research," [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/special-topics/research/index.html .

[21] Sentinel, "Principles and Policies: Conflict of Interest," [Online]. Available: https://www.sentinelinitiative.org/principles-and-policies-conflict-interest .

[22] "Subaward Agreement," 24 January 2013. [Online]. Available: http://sites.nationalacademies.org/PGA/cs/groups/pgasite/documents/webpage/pga_056056.doc .

[23] Health Care Systems Research Network (HCSRB), "HCSRN DUAT Toolkit," [Online]. Available: http://www.hcsrn.org/en/Tools%20&%20Materials/GrantsContracting/HCSRN_DUAToolkit.pdf .

[24] FDA, "Data Integrity and Compliance With Drug CGMP Questions and Answers Guidance for Industry," FDA, December 2018. [Online]. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-integrity-and-compliance-drug-cgmp-questions-and-answers-guidance-industry.

[25] HHS, "Breach Notification Rule," [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/breach-notification/index.html.

[26] "The (Re)usable Data Project," [Online]. Available: http://reusabledata.org .

[27] C. R. M. J. W. L. W. L. e. a. Carbon S, "An analysis and metric of reusable data licensing practices for biomedical resources," *PLOS ONE,* vol. 14, no. 3, 2019.

[28] A. J. a. M. C. Garfinkel S, "Understanding database reconstruction attacks on public data," *Commun. ACM,* vol. 62, no. 3, p. 46–53, 2019.

[29] OHDSI, "The Observational Health Data Sciences and Informatics," [Online]. Available: https://www.ohdsi.org.

[30] OHDSI, "OHDSI Forums," [Online]. Available: http://forums.ohdsi.org.

[31] "About PCORNet," [Online]. Available: https://pcornet.org/about.

[32] Partners HealthCare Systems, Inc, "i2b2 Design Document - Workplace Framework (WORK) Cell," [Online]. Available: https://www.i2b2.org/software/projects/workplace/Workplace_Design_15.pdf.

[33] i2b2, "i2b2 / ACT Network (SHRINE)," [Online]. Available: https://www.ctsicn.org/i2b2-shrine-act.

[34] Sentinel, "About Sentinel Initiative," [Online]. Available: https://www.sentinelinitiative.org/sentinel/about.

[35] Sentinel Initiative, "Active Risk Identification and Analysis (ARIA)," [Online]. Available: https://www.sentinelinitiative.org/active-risk-identification-and-analysis-aria.

[36] Sentinel Initiative, "FDA Catalyst," [Online]. Available: https://www.sentinelinitiative.org/fda-catalyst.

[37] Sentinel Initiative, "Sentinel Common Data Model," [Online]. Available: https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model.

[38] FDA, "Sentinel System Five-year Strategy 2019-2023," January 2019. [Online]. Available: https://www.fda.gov/media/120333/download.

[39] OHDSI, "OHDSI Data Standardization," [Online]. Available: https://www.ohdsi.org/data-standardization.

[40] J. Duke, G. Hripcsak and P. Ryan, "Open-Source Big Data Analytics in Healthcare," 2015. [Online]. Available: https://www.ohdsi.org/wp-content/uploads/2014/07/OHDSI-Tutorial-PreFinal-mod.pdf.

[41] "PCORnet Governance Policies," [Online]. Available: https://pcornet.org/wp-content/uploads/2018/02/PCORnet-Governance-Policy-v4_approved_31Jan2018.pdf.

[42] eHealth Initiative Foundation , "Developing a Governance and Operations Structure for the Sentinel Initiative," April 2009. [Online]. Available: https://www.pharmamedtechbi.com/~/media/Images/Publications/Archive/The%20Gray%20Sheet/35/021/01350210021_b/052509_sentinel_governance_report.pdf.

[43] PCORNet, "PCORnet Governance Policy v4," January 2018. [Online]. Available: https://pcornet.org/wp-content/uploads/2018/02/PCORnet-Governance-Policy-v4_approved_31Jan2018.pdf.

[44] PCORNET, "Research Committee Code of-Conduct," [Online]. Available: https://pcornetcommons.org/wp-content/uploads/2016/11/MS-PPRN-iConquerMS-Research-Committee-Code-of-Conduct-NOV2016.pdf .

[45] PCORI, [Online]. Available: https://pcornetcommons.org/resource_item/pcornet-governance-materials.

[46] PCORNet, "PCORNet Data Sharing Agreement DSA 2.0," [Online]. Available: https://pcornetcommons.org/resource_item/pcornet-data-sharing-agreement-dsa-2-0.

[47] "PCORnet Commons," [Online]. Available: https://pcornetcommons.org.

ABBREVIATIONS

ACT	Accrual to Clinical Trials
AMIA	American Medical Informatics Association
API	Application Programming Interface
AWS	Amazon Web Services
BRIDG	Biomedical Research Integrated Domain Group
caDSR	Cancer Data Standards Registry and Repository
CDASH	Clinical Data Acquisition Standards Harmonization
CDC	Centers for Disease Control
CDE	Common Data Elements
CDM	Common Data Model
CDISC	Clinical Data Interchange Standards Consortium
CDRN	Clinical Data Research Network
CFR	Code of Federal Regulations
CMS	Centers for Medicare and Medicaid Services
CNDS	Cross Network Directory Service
COI	Conflict of Interest
CPT	Current Procedural Terminology
CSV	Comma-separated Values
CTLA-4	Cytotoxic T-lymphocyte-associated protein 4
CTSA	Clinical and Translational Science Award
DAF	Data Access Framework
DDL	Data Definition Language
DSA	Data Sharing Agreement
DUA	Data Use Agreement
EHR	Electronic Health Record
ePRO	Electronic Patient-Reported Outcome
ESG	FDA Electronic Submission Gateway

ETL	Extract, Transfer, Load
EVS	Enterprise Vocabulary Services, founded and managed by NCI
FDA	US Food and Drug Administration
FHA	Federal Health Architecture
FHIM	Federal Health Information Model
G-SRS	Global Substance Registration System
HCSRN	Health Care Systems Research Network
HHS/ASPE	Department of Health and Human Services/ Office of the Assistant Secretary for Planning and Evaluation
HHS/ONC	Department of Health and Human Services/ Office of the National Coordinator for Health Information Technology
HIPAA	Health Insurance Portability and Accountability Act
HL7	Health Level Seven International
HL7 FHIR	HL7 Fast Healthcare Interoperability Resources
HTTPS	Hypertext Transfer Protocol Secure
i2b2	Informatics for Integrating Biology and Bedside
i2b2 ACT	i2B2 Accrual for Clinical Trials
ICD-9-CM; ICD-10-CM	International Classification of Diseases - Clinical Modification (Versions 9 and 10)
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
IEC	International Electrotechnical Commission
IG	Implementation Guide
IHE	Integrating the Healthcare Enterprise
IRAE	Immune-Related Adverse Events
IRB	Institutional Review Board
ISO	International Standards Organization
JSON	JavaScript Object Notation
LHS	Learning Health System

LOINC	Logical Observation Identifiers Names and Codes
MedDRA	Medical Dictionary for Regulatory Activities
NDC	National Drug Codes
NHLBI	National Heart, Lung and Blood Institute
NICHD	National Institute of Child Health and Human Development
NIDCR	National Institute of Dental and Craniofacial Research
NIH/ NCI	National Institute of Health / National Cancer Institute
NIH/NCATS	National Institute of Health / National Center for Advancing Translational Sciences
NLM	National Library of Medicine
NLM CDE Repository	NLM Common Data Elements Repository
NLM VSAC	NLM Value Set Authority Center
NLP	Natural Language Processing
OHDSI	Observational Health Data Sciences and Informatics
OMB	Office of Management and Budget
OMOP	Observational Medical Outcomes Partnership; CDM for OHDSI Network
PCORI	Patient-Centered Outcomes Research Institute
PCORnet	National Patient -Centered Clinical Research Network
PCORTF	Patient Centered Outcomes Research Trust Fund
PD-1	Programmed Cell Death Protein 1
PDL-1	Programmed Death-ligand 1

PedsNET	Pediatric learning health system and partner with PCORnet
PHI	Protected Health Information
PopMedNet	An open-source application used to facilitate multi-site health data networks.
PPRN	Patient-Powered Research Networks
PSCANNER	Patient-centered SCAlable National Network for Effectiveness Research
RDP	Re-usable Data Project
RWD	Real World Data
RWE	Real World Evidence
RxNorm	US-specific terminology in medicine that contains all medications available on the US market
SDC	Structured Data Capture
SDO	Standards Development Organization
SDTM	Study Data Tabulation Model
SFTP	Secure File Transfer Protocol
SNOMED CT	Systemized Nomenclature of Medicine Clinical Terms
SOP	Standard Operating Procedure
SPL	Structured Product Labelling
SVN	Subversion
UI	User Interface
USCDI	US Core Data for Interoperability
XML	Extensible Markup Language

APPENDICES

This section will contain technical details that should not be placed in the main sections.

APPENDIX A

TECHNICAL FRAMEWORK

A.1 CDMH Technical Architecture

The purpose of CDMH System as shown in Figure 12 was to enable researchers to create a query without having to worry about the underlying data models and their database specific requirements and capabilities. A researcher uses a simple form-based web application and standard code values to create their queries. The system automatically transforms the researcher's query into the data partner specific CDM and its database query. The system tracks the query and its status along each step to give the status of the query to the researcher. Finally, the system aggregates the results from multiple data partners and shows the results to the researcher. In addition, the system also utilizes the existing mappings to transform the results to CDISC SDTM for submission to the FDA.

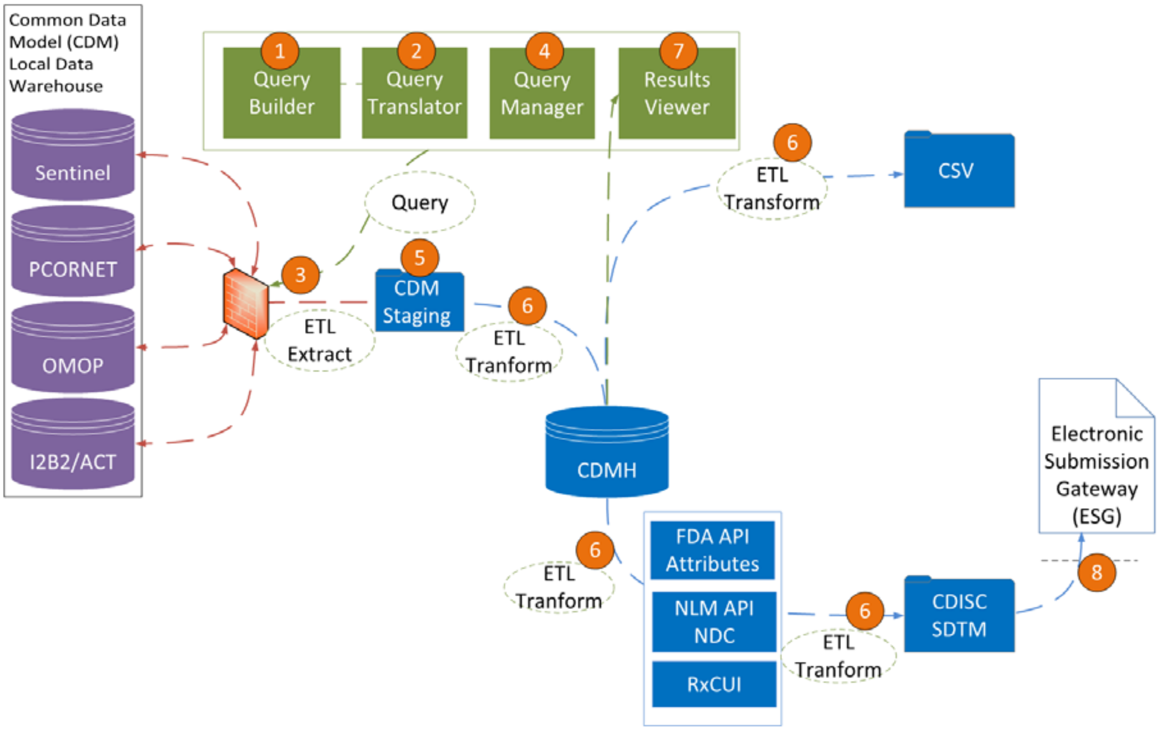
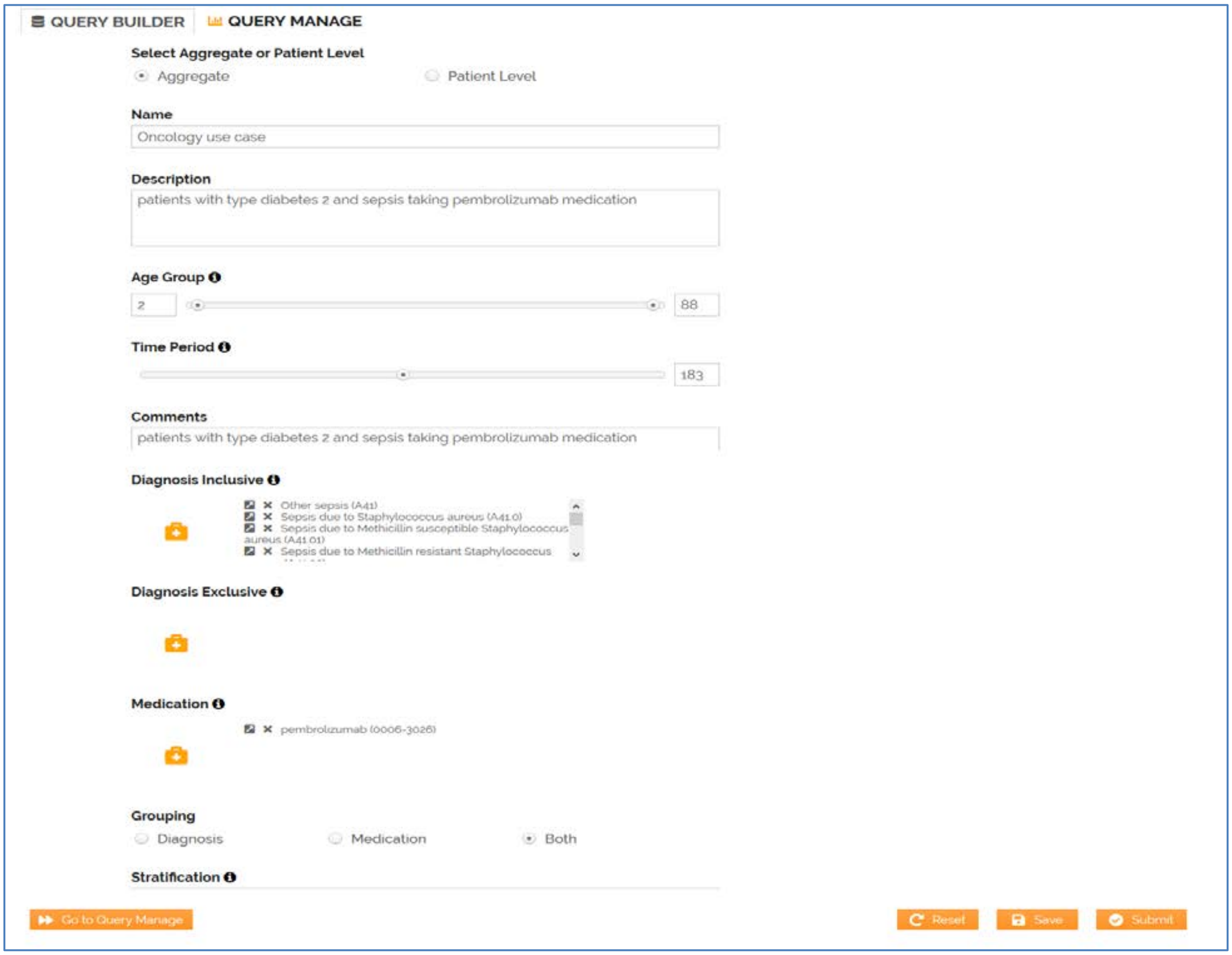


Figure 12. High-level architecture diagram with components labeled and explained in the following subsections.

A.1.1 Query Builder

The Query Builder tool enables the investigators to construct a clinical question using a web browser based graphical user interface (GUI), which is then transformed into a database query. The investigator doesn't need to know about the database tables, their joins or any efficient way to write the query. The investigator chooses a set of inclusion and exclusion criteria like patient diagnosis, medication, or time to onset from a set of filters or widget like radio buttons, checklists, etc. The software uses the user's choices to construct a set of SQL queries one for each registered. The query builder can save partial queries and update those queries anytime later. A user can utilize the previously submitted query and create a similar query using Query Manage tool.

A screenshot of the form-based Web user interface for user to build a SQL query is as shown in Figure 13.



Select Aggregate or Patient Level
● Aggregate ○ Patient Level

Name
Oncology use case

Description
patients with type diabetes 2 and sepsis taking pembrolizumab medication

Age Group
2 88

Time Period
183

Comments
patients with type diabetes 2 and sepsis taking pembrolizumab medication

Diagnosis Inclusive
- Other sepsis (A41)
- Sepsis due to Staphylococcus aureus (A41.0)
- Sepsis due to Methicillin susceptible Staphylococcus aureus (A41.01)
- Sepsis due to Methicillin resistant Staphylococcus

Diagnosis Exclusive

Medication
- pembrolizumab (0006-3026)

Grouping
○ Diagnosis ○ Medication ● Both

Stratification

Figure 13. A screenshot of the Query Builder Web UI.

The Query Builder component has the following capabilities:

1. **User Authentication**: Investigators are authenticated using Adeptia ETL connect user management capabilities.
2. **Graphical User Interface**: user can build a query based on specifying conditions for selected controlled vocabularies. Conditions will be limited to:
   a. Drugs (set of drugs from the use case*). Only NDC codes are displayed.
   b. Diagnosis (set of diagnosis from the use case**). Only ICD-10-CM codes are displayed
   c. Temporal relationship between the drugs and the diagnosis

      E.g., Give me all patients diagnosed with (**) after first receiving (*)

   d. All controlled vocabulary values are represented using CDMH's vocabulary set
   e. The user can specify the type of results required, i.e. aggregate results or patient-level results. For aggregate query type user can specify the group based on Diagnosis or Medication or both and stratification based on demographics (Gender/Race/Ethnicity).
   f. The User Interface (UI) is built using Sencha Ext JS framework and is hosted on Adeptia Connect.

Figure 14 shows the Query Builder component's workflow.

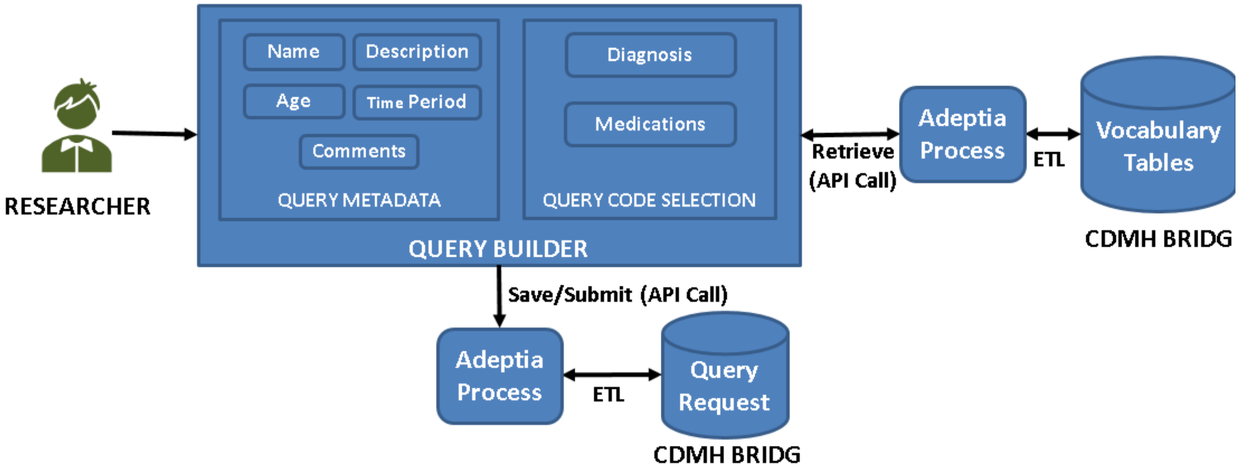


Figure 14. Query builder component workflow.

- When a researcher searches for a particular diagnosis or medication in UI using search string, the Query Builder invoke an Adeptia Process flow through a Rest API call, that searches the CDMH Vocabulary set and returns all diagnosis or medication containing the search string.
- When a researcher tries to submit a query, the Query builder validates if the query is complete, otherwise throws an error.

- When researcher saves or submits a new query, the researcher's selection is submitted to the CDMH database in JSON format. The query is then assigned with a unique id called QUERY_REQUEST_ID which is used to track that query during its full life cycle.

A.1.2 Query Transformation

The query transformation is the backend functionality that takes the researcher selection from the Query Builder tool and by using the CDM mapping tables, constructs or translates into 4 CDM specific constructs capable of querying each common data model.

The query transformation tool takes the generic SQL query constructed in the using the Query Builder Tool and by using the CDM mapping tables constructs or translates the CDMH generic query into 4 CDM specific constructs capable of query each unique common data model.

The Query Transformation functionality is achieved by the following modules.

- Code Translator
- Query Translator
- Query Submit to Partner

Code Translator

Maps the researcher selected controlled terminologies to CDM specific terminologies by utilizing the data-element and terminology mappings and stores a unique record in CDMH table for each data partner.

Query Translator

Constructs CDM specific query using the CDM specific codes mapped in Code Translator and stores the queries in the CDMH database

Query Submit to Data Partner

Sends the Data partner/CDM specific queries to the S3 bucket(s) to be picked up by each data partner(s) and sends email notifications to the data partner.

A.1.3 Secure Data Transfer

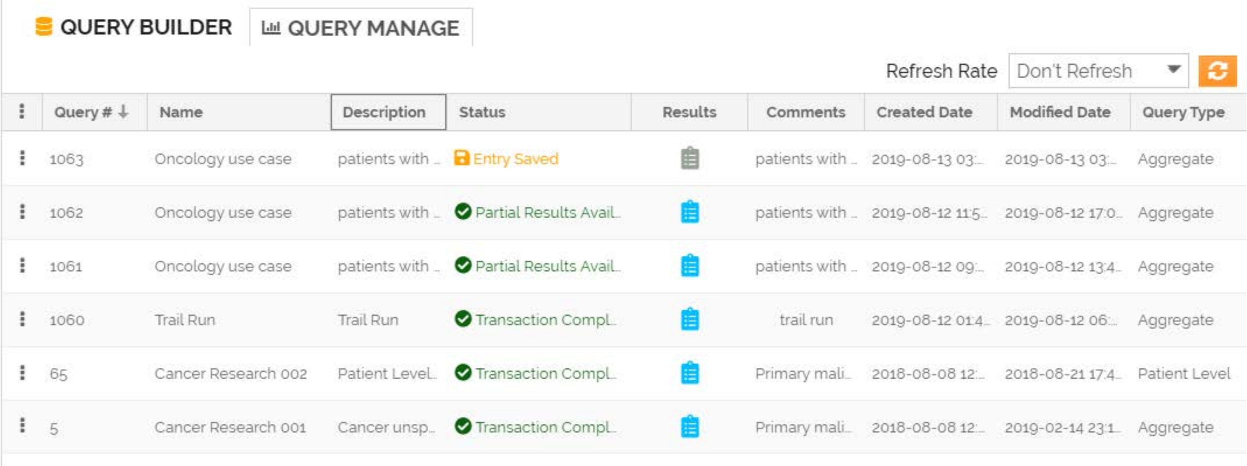
The transfer of information between data partner and HHS agencies is an important requirement that must be secure and trusted by all participants. In addition to security the ease of use of both inbound as well as outbound traffic is important to data partners participation. CDMH utilizes a Secure File transfer protocol (SFTP) site setup to secure the communication between the agency and the data partner. For each query from the researcher, the CDMH transfer process automatically places the file in the SFTP location for outbound traffic and monitors the folders for inbound results. These monitors communicate their status back to the Query management process which is one stop place to track the status of each query. The table below shows the SFTP server specifications in CDMH environment.

Table 2. CDMH SFTP Server Specifications

Purpose	**Serves as SFTP server to handle file transfers between ASPE environment and data partners**
OS	CentOS 7
Application	Sshd
Primary Disk (C Drive)	100 GB
Data Drive	250 GB
domain	ncats.io
Memory / CPU	T2.micro (1 vCPU,0.5 GB Memory)

A.1.4 Query Manager

As shown in Figure 15, the Query Manage tool allows the investigators to view, manage and track the queries within a browser-based graphical user interface. The Query Manage tools also enables the investigators to view open the Results Viewer for a query when the results are available from at least one data partner. Figure 15 shows the results of multiple queries and their result status and the links to view the results when they are available.



Refresh Rate: Don't Refresh

Query # ↓	Name	Description	Status	Results	Comments	Created Date	Modified Date	Query Type
1063	Oncology use case	patients with ...	Entry Saved		patients with ...	2019-08-13 03:...	2019-08-13 03:...	Aggregate
1062	Oncology use case	patients with ...	✓ Partial Results Avail...		patients with ...	2019-08-12 11:5...	2019-08-12 17:0...	Aggregate
1061	Oncology use case	patients with ...	✓ Partial Results Avail...		patients with ...	2019-08-12 09:...	2019-08-12 13:4...	Aggregate
1060	Trail Run	Trail Run	✓ Transaction Compl...		trail run	2019-08-12 01:4...	2019-08-12 06:...	Aggregate
65	Cancer Research 002	Patient Level...	✓ Transaction Compl...		Primary mali...	2018-08-08 12:...	2018-08-21 17:4...	Patient Level
5	Cancer Research 001	Cancer unsp...	✓ Transaction Compl...		Primary mali...	2018-08-08 12:...	2019-02-14 23:1...	Aggregate

Figure 15. Screenshot of Query Manager from the CDMH System.

The Query Manager component has the following capabilities:

- User authentication: Uses the same user authentication used by the Query Builder
- User interface: Allows to edit the saved queries or save an existing query as new query. Choosing any option takes the user to the Query Builder UI
- Provides metadata of each query
- Tracks the query status in its full life cycle
- Enables the option to view Results Viewer

A.1.5 Query Result Staging

Results Loader / Data Archiving and Data Management: The staging and management is best conceptualized as an archiving place for the raw results data. This function is to retain a permanent copy of the results before they are transformed into a harmonized intermediary model. Retaining an unaltered archived baseline copy of raw data is an essential requirement of data provenance. Data provenance is part of data integrity and ensures transparency of what, who and when data were altered. It is part of the data life cycle and must be available in case questions should arrive later.

A.1.6 ETL Transformations

CDMs-to-CDMH/BRIDG mappings: A copy of the results data is copied and transformed into a standard FHIR Resources values sets. The values sets are then placed or stored within the common data model called the BRIDG.

FHIR value sets are maintained and curated by multiple contributors. The curation of a values set is a major commitment of both time and resources, but it is essential for the long-term success of interoperability. By leverage large organizational commitment from SNOMED CT, ICD-9/10, RxNorm, NDC and many others we are assured of an updated, curated resources that can be counted on for both quality and longevity. Another advantage of using FHIR resources is the standardization of metadata about the various resources. HL7 metadata facilitates interoperability and quality by providing a machine-readable information that includes amongst other things the version and content of the value sets.

The details of the cross mapping of the CDM's to the BRIDG are described in detail in other sections of this report. Cross mapping is needed to combine semantically similar concepts that differ syntactically. For Instance, the FHIR Resource for gender is uses to OMB's administrative _gender value set. Mapping or Cross maps take a non-standard value and convert them into a more common value for instance if one of the common data models stored Male as value '1' but the standard FHIR resources uses the complete word 'Male' then all the value of '1' will be transformed into Male so the data can be combined with other CDM results.

All CDM data is stored in the BRIDG, the BRIDG data model was chosen for multiple reasons including it is a shared tool between both HL7 (ONC requirements) and CDISC, or FDA requirements. Thus, the BRIDG facilitates the transformation from EHR RWD information into the FDA required CDISC SDTM format. Finally, BRIDG was chosen because it is not one of the 4 CDMs and is therefore a neutral data model, that does not show favoritism to any of the existing CDM networks (Sentinel, OMOP, I2B2/ACT or PCORnet).

ETL-IN Processing data from the data partner into CDMH System

The ETL-IN component has the following capabilities:

a) Retrieve results dataset from the data partner(s) SFTP.

- Adeptia File events are defined to watch for incoming results and these events triggers ETL process to move results file SFTP to local when new files are available.
- Aggregate results are received in a CSV file, which will include the query ID, system, version of system, conditions met, and stratification
- Patient level results are received in a single .zip file containing multiple CSV files. Each CSV file contains a dump of all records from a database table that meet the selection criteria. Therefore, the number of CSV files will equal to the number of patient-level data tables in the database.
- The results file naming standard includes details like data partner name, source version and query id.
- The results files are stored as BLOB in CDMH and in archival folder.

b) Load the results into the staging tables

- There will be a single staging table for aggregate results that reflects the CSV file format, plus metadata fields about the query and results.
- The patient-level results staging tables will depict table structure of data partner tables.
- At any point staging tables store only single query related data.
- The results are stored as is.

c) Transform and load the results into the CDMH BRIDG-Informed DB

- The results from staging tables are loaded to CDMH BRIDG-Informed DB
- The data partner specific vocabulary values are translated to CDMH vocabulary while loading
- The results are tied using query id

d) *Utilize the data-element and vocabulary mappings*

e) Built using Adeptia

ETL-Out Exporting data from CDMH System into SDTM or download as Comma Separated Values (CSV)

The ETL-Out component will have the following capabilities:

a) Export patient-level results of a <u>specific query</u> to SDTM dataset for the user to store locally

b) Export patient-level results of a <u>specific query</u> to FHIR dataset for the user to store locally

c) ETL Out processes are triggered based on user selection from the Results Viewer user interface

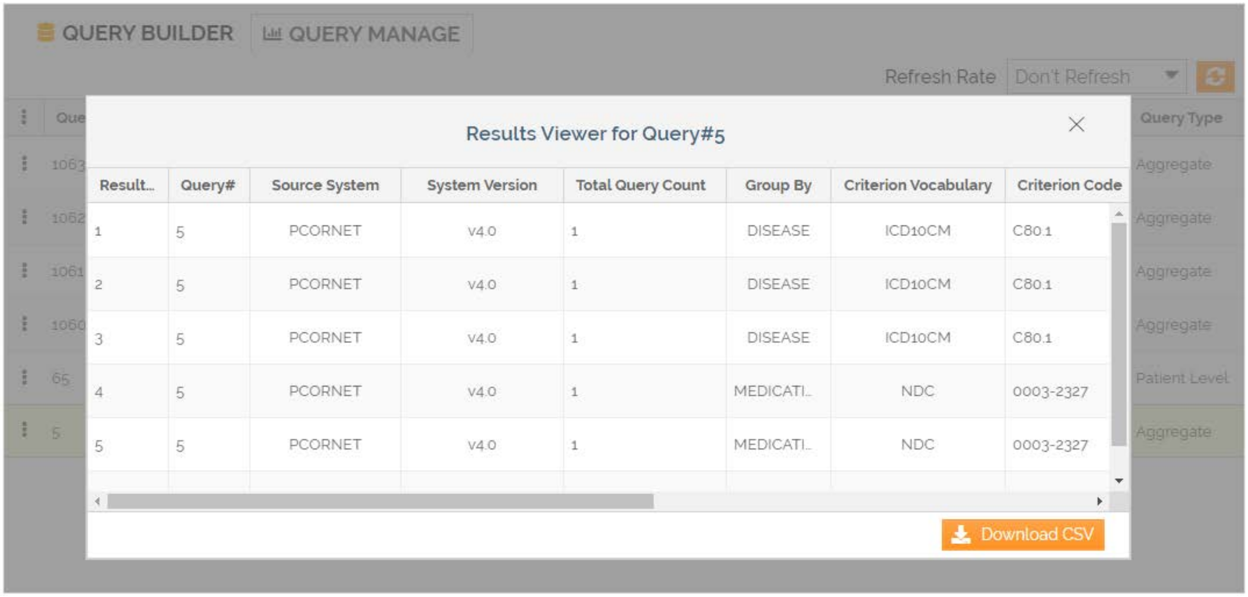
d) Built using Adeptia

A.1.7 Results Viewer and Analytics Export Function

The results viewer allows users to view the data in both a tabular and graphical format. An illustrative example might be the count of patients with heart disease, broken down by demographic characteristic like race, age, sex, etc. The results viewer is not meant to be a statistical tool like R, if an investigator wishes to do further analysis the viewer has an export function which creates multiple file formats including CSV. The CSV files can be easily imported into statistical products for analysis

The Report Viewer component has the following capabilities:

a) Use the same user authentication used by the Query Builder
b) Present the results in tabular format for aggregate level query

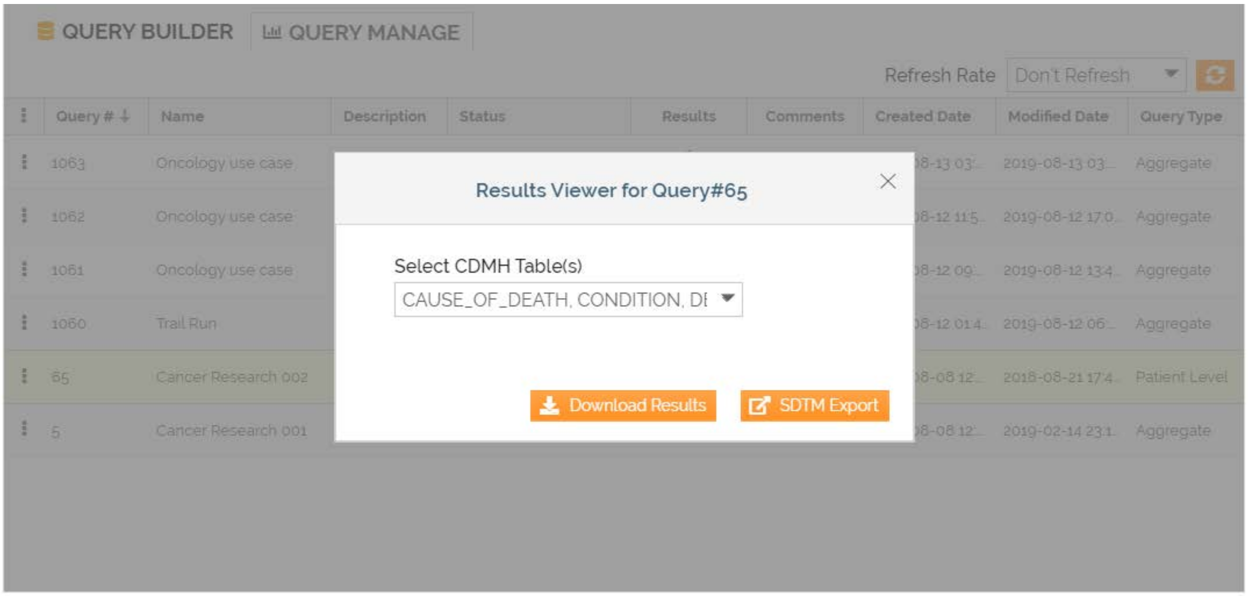


Results Viewer for Query#5

Result...	Query#	Source System	System Version	Total Query Count	Group By	Criterion Vocabulary	Criterion Code
1	5	PCORNET	v4.0	1	DISEASE	ICD10CM	C80.1
2	5	PCORNET	v4.0	1	DISEASE	ICD10CM	C80.1
3	5	PCORNET	v4.0	1	DISEASE	ICD10CM	C80.1
4	5	PCORNET	v4.0	1	MEDICATI...	NDC	0003-2327
5	5	PCORNET	v4.0	1	MEDICATI...	NDC	0003-2327

Figure 16. Screenshot of Result viewer interface in the CDMH System.

Export the results to a CSV file. Export the patient results to SDTM format.



Query #	Name	Description	Status	Results	Comments	Created Date	Modified Date	Query Type
1063	Oncology use case					8-13 03:...	2019-08-13 03...	Aggregate
1062	Oncology use case					8-12 11:5...	2019-08-12 17:0...	Aggregate
1061	Oncology use case					8-12 09:...	2019-08-12 13:4...	Aggregate
1060	Trail Run					8-12 01:4...	2019-08-12 06:...	Aggregate
65	Cancer Research 002					8-08 12:...	2018-08-21 17:4...	Patient Level
5	Cancer Research 001					8-08 12:...	2019-02-14 23:1...	Aggregate

Results Viewer for Query#65

Select CDMH Table(s): CAUSE_OF_DEATH, CONDITION, DI

Figure 17. Screenshot of export option in the CDMH System.

The user will have to log into the system to check whether the results were received, i.e. no automated notifications indicating that the query results are in place.

A.1.8 FDA Submission Process

One of the primary requirements of the data harmonization project was use submit RWD to the FDA. As of 2017 the FDA requires all data submission be in a file format defined by CDISC known as SDTM. The FDA requires the use of a standard semantics, and syntactical format with codified value sets to facilitate study results analysis. Without a standard format the entire FDA process would be impossible. New project submissions would come to a stop because staff would have to spend time finding the information and then defining equivalents of information o considering results.

Though SDTM facilitate data analytics the conversion to the CDSIC standard is an expensive and nontrivial undertaking. As previously discussed, the BRIDG data model facilitates the transformation of this process and is highly leveraged in the CDMH project. Absent the BRIDG resource, the CDMH project would be forced to build its extensive functionality, a multiyear multimillion-dollar endeavor.

a) **CDMH -to-SDTM mapping/ FHIR to SDTM Conversion**: CDMH as previously described stores the CDM data using FHIR resources within the BRIDG data model. This necessitates a transformation of the FHIR value set data into SDTM. The transformation coverts existing stored values into CDISC SDTM format which includes deriving the name value pairs or the unique code values referred to as “C-Codes”. C Codes are unique identifiers for values created by CDSC. CDISC much like other Standard Data Organizations maintains these value sets which ensures FDA receives up to date curated information.

   CDISC codes set are stored and maintained within NIH’s National Cancer Institute, NCI EVS and caDSR. By storing the CDMs to FHIR mappings within the NCI data infrastructure, CDMH is able to leverage the standards and resources of the federal government. All the cross mappings are centrally located within NCI caDSR with makes curation, maintenance. and updates possible.

b) **RxNorm to NDC mapping**: Because the PCORnet, OMOP and i2b2/ACT CDMs use RxNorm Codes, (Sentinel uses NDC Codes) additional mapping from RxNorm to NDC was required. NDC is needed in order to separate two concepts or variable (Route and Form) that are combined within the RxNorm Codes. For example, gas as defined by RxNorm includes both Nasal and Oral, and RxNorm for Injection could be intravenous, intramuscular or intraarticular. FDA understandably requires the both the Form and the Route as distinct concept, and both are required. The RxNorm to NDC mapping splits out Form and Route by using the RxNorm code and transforming it to a representative NDC code. An API to the FDA was developed which allow CDMH to use the representative NDC and derive a unique route and form c code identifier required for the SDTM file submission to the FDA

APPENDIX B

USER ACCEPTANCE TESTING (UAT TESTING)

User Acceptance Testing (UAT), is a process of verifying that a solution works for the user. It involves the testing (unit testing, functional, testing, integration testing, and system testing), the functionality of the software as well as documenting the process and results of the testing.   CDMH team from all the agencies completed UAT unit, functional, integration, and system testing. Test results and the documentation can be found in this appendix.

The tables below detail the list of functionalities to be tested for each component as a part of UAT and System Testing.

B.1 Query Builder

Task Name	Description	Method
UA Contents and Navigation	Verification that UA contains all the fields corresponding to Use Case and the user can navigate the screen	Review Query Build Screen
Mandatory fields functionality	Verification that Mandatory fields contain values	Review Query Build Screen
Query type functionality	Verifying capability of setting query type (Aggregate / Patient level) with corresponding fields display	Review Query Build Screen
CDMH Reference value sets access	Verifying that drop-down lists represent CDMH/FIHR vocabulary	Review Query Build Screen
Saving Query	Verifying capability of saving and editing query with assigning sequential Query ID	Review Query Manage Screen
Submitting Query	Verification of capability of Submitting query	Review Query Manage Screen

B.2 Query Translator

Task Name	Description	Method
Generating query with specified parameters	Verification that query captures parameters entered on the Query Build screen	Review Database
Translating query to CDM specific codes	Verification that query converts Codes and Vocabularies to CDM specific values	Review Database
SQL file verification	Verification that SQL file is created. The name should consist of Model name/version, partner id, query type and query id. The files should be placed in Modes specific folder	Local Adeptia Folder

Number of SQL files	Verification of number of files created for Aggregate (one) and Patient level (multiple files based on the CDM)	Local Adeptia Folder
Data Model Specific query validation	Verification that the query does not return any syntax errors	Running SQL file query in Model corresponding database schema

B.3 Data Loading

Task Name	Description	Method
Aggregate Level		
CSV file verification	Verification of file naming integrity and header section in CSV file	Opening CSV file
Loading to Staging area	Verification that the contents of CSV files data is loaded to Staging area aggregate table as is	Reviewing data in Model corresponding database schema staging tables
Loading data from Staging to CDMH	Verification that aggregate data is loaded to CDMH	Reviewing data in CDMH Database corresponding tables
Conversion of CDM vocabulary to CDMH	Verification that data is converted from Model specific codes and vocabulary to CDMH	Reviewing data in CDMH database corresponding tables
Conversion for RxNorm to NDC	Translation of RxNorm to NDC acquisition via API to FDA SPL database (G-SRS database)	Reviewing data in CDMH Database corresponding tables
Patient level		
File verification	Verification of file naming for zip and individual csv files. The received file naming should be consistent with the files sent.	Local Adeptia folder
Number of files verification	Verification that the number of files is correct for corresponding Data Model	Local Adeptia folder
Loading to Staging area	Verification that the contents of all CSV files data is loaded to Staging area tables as is	Reviewing data in CDMH Database corresponding tables
Loading data from Staging to CDMH	Verification that patient data is loaded to CDMH	Reviewing data in CDMH Database corresponding tables
Conversion of CDM vocabulary to CDMH	Verification that data is converted from Model specific codes and vocabulary to CDMH	Reviewing data in CDMH Database corresponding tables

Conversion for RxNorm to NDC	Translation of RxNorm to NDC acquisition via API to FDA SPL database (G-SRS database)	Reviewing data in CDMH Database corresponding tables

B.4 Results Viewer

Task Name	Description	Method
Reviewing Aggregate level data	Reviewing Aggregate level data per issued query in CMDH Vocabulary	Result Viewer page
Reviewing Patient level data	Reviewing Patient level data per issued query in CMDH Vocabulary	Result Viewer page
Generating CSV file output	Verifying capability of generating CSV format file for specific query	Result Viewer Page
Generating SDTM format output	Verifying capability of generating SDTM output and ensuring semantic accuracy	Result Viewer Page
Validating SDTM format	Reviewing data to insure SDTM format and structure integrity	SDTM

B.5 System Testing

Task Name	Description	Method
Aggregate Level		
Creating query	Creating query for Aggregate level	Review Query Build Screen
Query translation	Translating query to Data models specific vocabulary	Review Database
SFTP site capabilities	Verification of SFTP site for CDMH to CDM outbound files	Local Adeptia Folder
Loading data	Loading data to Staging and CDMH	Review Database
Reviewing query	Reviewing query results	Result Viewer page
Patient Level		
Creating query	Creating query for Patient level	Review Query Build Screen
Query translation	Translating query to Data models specific vocabulary	Review Database
SFTP site capabilities	Verification of SFTP site for CDMH to CDM outbound files	Local Adeptia Folder
Loading data	Loading data to Staging and CDMH	Review Database
Reviewing query	Reviewing query results	Result Viewer page
Reviewing SDTM output	Reviewing query results in SDTM format	SDTM

APPENDIX C

INTERMEDIARY MODEL SELECTION AND DETAILED MAPPING PROCESS

The initial phase of the project dealt with harmonizing all the data elements from each of the chosen CDMH models. The models chosen were Sentinel v6.0.2, PCORnet v3.1, i2b2-ACT v1.3, and OMOP v5.2. During the project, it was decided to harmonize PCORnet v4.0 as well as some of the data partners had already or would be implementing this new version.

To harmonize all the disparate models, the BRIDG model was chosen as the intermediary model. For each of the models, an initial BRIDG Mapping spreadsheet was created. Each model's data elements were imported as the Mapped Specification source. Then each data element was analyzed and mapped to an existing BRIDG element, if one existed. This analysis looked at the source model's definition of the element, any vocabulary that was provided for the field, as well as sample data to see exactly how the element was used. For each element that mapped to an existing BRIDG element, the BRIDG class, BRIDG element, and BRIDG mapping path were captured in the mapping spreadsheet. For CDM elements that did not map to an existing BRIDG element, a proposed element was entered. In some cases, these proposed elements were additions to existing BRIDG classes while in other cases a new BRIDG class was also proposed.

Once each of the models had been mapped to the BRIDG model, a Consolidated Mapping spreadsheet was created. This consolidated mapping spreadsheet combined all the mappings. There was a tab for each of the individual model mappings and one tab that listed the important information from each mapping. Review of the consolidated mappings was done to ensure that all mappings were consistent and that model elements that were mapped to the same BRIDG element were equivalent to each other.

Once the Consolidated Mapping spreadsheet had been finalized and reviewed and feedback incorporated back into the mappings, a CDMH Conceptual model was created in Enterprise Architect. This Conceptual model copied the mapped BRIDG classes and elements as well as any new BRIDG classes, associations, and elements. All the additions were marked with a CDMH stereotype.

The final Consolidated Mapping spreadsheet and the Conceptual model diagram were sent to the BRIDG team for review. For each of the additions, a rationale for the addition and a definition of the new element was presented. This review with the BRIDG team resulted in some changes to the proposed elements as well as some changes to the mappings as suggested by the BRIDG review team. Once the review was complete, the mappings and the Conceptual model were considered finalized and were submitted to the BRIDG team for incorporation into BRIDG v5.1. The consolidated CDMH semantics were harmonized with BRIDG R 5.0.1. At end the BRIDG Harmonization process, the CDMH project added 2 new classes and 17 new attributes to the BRIDG model. The BRIDG team released BRIDG 5.1 in March 2018 which incorporates the CDMH model semantics. BRIDG Release files can be downloaded from the BRIDG Website.

APPENDIX D

IMMUNOTHERAPY PROTOCOL

D.1 Background

Oncology as a field has recently been transformed by the emergence of immunotherapy in cancer care. Within immunotherapy several blocking agents targeting CTLA-4 and PD1/PDL1 ligands of cancer cells have recently been approved for a variety of indications including melanoma, non-small cell lung cancer, bladder cancer, head and neck cancer, renal cancer, and Hodgkin lymphoma. Many trials in additional indications are ongoing, indicating potential uses of these agents for a wide range of metastatic cancer conditions.

As immunotherapy represents a new modality of therapy which has not been in common use until now, post-market data is still scarce - confining our knowledge regarding safety and efficacy of these treatments solely on clinical trial data sources; whereas in the real world, patients may carry an array of comorbid conditions interfering with the safety and efficacy of such drugs which may otherwise not be identified in clinical trials. For example, drug approval trials for anticancer agents have historically excluded patients with HIV, autoimmune disease, and other severe organ dysfunction (e.g., end-stage renal disease), but recent literature provide evidence of higher number of 'immune-related adverse events (IRAE)' among patients treated with CTLA-4/PD-L1 inhibitors. Data accumulated from patients in real world settings would provide information related to important questions about what health care providers should avoid in terms of possible treatment combinations and which patients may safely consider these drugs as part of their anticancer treatment regimen. Therefore, there is practical use of RWD in order to create awareness among health care providers as well as to assist in further investigations in meeting the clinical needs. These databases can offer rich information related to the concomitant medications administered with immune checkpoint inhibitors, adding to our ability to monitor for possible emerging drug-drug interactions or the degree to which a therapy requires concomitant supportive care medications to treat adverse drug reactions. In addition, simply monitoring the duration of therapy with an anticancer agent in the metastatic setting could give a hint of the agent's efficacy when an approved indication is generalized to the real world setting (effectiveness), or provide hypothesis generating data on disease activity for non-approved indications for a given malignancy.

D.2 Objective

To demonstrate the approach for this project, four CDMs were harmonized to an intermediary model and tools have been developed as the result. The developed tools and queries were provided to networks of RWD which were asked to assess the safety of newly approved oncology drugs in combination with other immunotherapy agents.

The objective of this assessment was to facilitate the use of RWD (e.g., administrative claims, Electronic Health Records (EHRs), disease and product registries, electronic Patient Reported Outcome (ePRO)) to support evidence generation for regulatory and clinical decision making.

In this project, the team focused on the safety of newly approved immuno-oncology products. The primary aim of this assessment was safety. The potential for this type of analysis could also be relevant for determining the effect of the sequencing of various cancer treatments, or researchers could approach areas well beyond cancer.

D.3 Cohort Identification

A. INCLUSION CRITERIA

This assessment utilized data – beginning with 2011 or the first month in which one of the six agents of interest (Table 3) appeared on the FDA approved list of drugs.

Table 3. List of FDA approved drugs using Monoclonal antibodies (mAbs) against CTLA-4 and PD-1/PD-L1 ligands

Drug Name	mAb	Target	Approved indications	Year of approval
Yervoy	Ipilimumab	CTLA-4	Melanoma	2011
Keytruda	Pembrolizumab	PD-L1	Melanoma, NSCLC, HNSCC, cHL, Urothelial Carcinoma, MSI-H	2014
OPDIVO	Nivolumab	PD-1	Melanoma, NSCLC, Advanced Renal Cell Carcinoma, cHL, Squamous Cell Carcinoma, Urothelial Carcinoma, MSI-H, Colorectal Cancer	2014
Tecentriq	Atezolizumab	PD-1/PD-L1	Urothelial Carcinoma, NSCLC	2016
Imfinzi	Durvalumab	PD-L1	Urothelial Carcinoma	2017
Bavencio	Avelumab	PD-L1	MCC	2017

NSCLC: Non-Small Cell Lung Cancer, HNSCC: Head and Neck Squamous Cell Cancer, cHL: Classical Hodgkin Lymphoma, MSI-H: Microsatellite Instability-High Cancer, MCC: Merkel Cell Carcinoma

All patients who used the immuno-oncology drugs and were diagnosed with advanced cancer were included in this assessment. They must have been exposed to one or more of the CTLA-4 or PD-L1 check point inhibitors and have comorbid conditions of auto-immune disorders prior to the administration of the agents (Table 4).

Table 4. Inclusion criteria of cohort identification

#	Criterion
1.	All patients who have been prescribed the immuno-oncology drugs Received at least one dose of the agents listed in Table 3
2.	Diagnosed with advanced cancer (if available, it may be useful to examine those patients that do not have cancer but still receive one of these drugs, but this would be a small and exploratory analysis)

B. Exclusion Criteria

None

PARAMETERS OF INTEREST

B. Auto-immune Disorders (Priority #1) coded with ICD-9-CM, ICD-10-CM and SNOMED CT.

Duration of Therapy (priority #2) Duration is defined by the number of days between the beginning of the treatment and end of the treatment. The assessment will be interested in days between the first day of treatment and either -

- Last day of treatment (if the treatment ended prior to query) or
- Date of query (in case of ongoing treatments)
  In case of lapse of treatment, i.e. patients who stopped and restarted the therapy -
- Days between end of a therapy and resume of therapy
- Include the number of infusions of these intravenous medications
- Evaluate any gaps between infusions, which may be caused by immune-related adverse events and toxicity

APPENDIX E

DATA GOVERNANCE FRAMEWORK

E.1 Purpose of this Framework

One of the main goals of the Harmonization of Various Common Data Models for Evidence Generation (CDMH) [7] project was to leverage open, consensus-based standards to harmonize specific CDMs. Creating data-sharing networks of existing CDMs allows researchers to ask questions of much larger amounts of RWD than previously possible, and more rapidly advance patient-centered outcomes research (PCOR).

The sustainability and benefits of these data-sharing networks depend not only on technical infrastructure, but on an enduring governance framework. This governance framework document outlines some policies and practices for access to and use of RWD derived from data-sharing networks. It addresses elements of control, roles and responsibilities, policies, processes, and procedures, with the goal of creating trust and confidence in the data and the network.

Data-sharing networks that provide access to patient health data carry special obligations to institutional and regulatory authorities, data providers, and patients upon whose data the networks are built. A robust governance framework need not restrict reuse and research; it can enable access to data while promoting trust in the partners and confidence in the data. This encourages research collaboration, which benefits patients, while protecting their privacy and other interests. This framework aims to maximize value to researchers while maintaining appropriate controls.

Content in this framework is derived from publicly available sources. It is based on the experience of the National Library of Medicine (NLM) and its Federal partners on the CDMH project (FDA, NCI, NCATS, HHS/ONC), along with the four networks participating in the CDMH project: Informatics for Integrating Biology and the Bedside, Accrual to Clinical Trials (i2b2/ACT), Observational Health Data Sciences and Informatics (OHDSI), Sentinel, and the National Patient-Centered Clinical Research Institute (PCORI).

The policies and documentation were examined for these sources to identify how these CDM leaders approached problems related to governance. For example, PCORnet has a Front Door [13], an access point for stakeholders interested in PCORnet resources, through which they may request data, network collaborators, a study feasibility review, etc.

This framework describes the foundations of data governance as they may apply to data-sharing networks; recommends a governance structure to establish and maintain policies and procedures; and lists some documents that data-sharing networks should consider creating/adopting for their own governance and system access. Finally, it outlines some other relevant issues the Network may want to consider when establishing policies.

This framework is not intended to be exhaustive or prescriptive. Data-sharing networks should arrive at their governance structure, policies, and procedures through consensus with stakeholders.

E.2 Definition of Terms

The following definitions apply to terms used within this document. Capitalized terms adopt the specific definitions below.

Term	Definition
Aggregate Data	Counts or other statistical measures of De-Identified Data across individuals having certain attributes, e.g., groupings by diagnosis or age group.
Authorized User	An individual authorized to access the System to execute queries and obtain data.
Board	A governing body composed of representative members who are charged with sustaining the resources and vision of the Network.
Data Partner	Any organization participating in the overall network that supports a data model that can be queried by authorized researchers.
De-Identified Data	Data defined in accordance with the HIPAA Privacy Rule: 45 CFR Section 164.514 (a) with processes for de-identification set forth in 45 CFR Section 164.514 (b).
Individual Level Data	Data specific to individual patients, that may or may not be de-identified.
Investigator/Researcher	A user or authorized recipient of Network data.
Limited Data Set	A dataset defined in the context of the HIPAA Privacy Rule: 45 CFR Section 164.514 (e).
Minimum Necessary	Data defined in the context of the HIPAA Privacy Rule: 45 CFR Section 164.514(b).
Network	The entity formed by the contributing Data Partners, and its mission, activities, etc.
System	Network deliverables, including databases, query interfaces, and/or additional tooling.
Research Proposal	Outlines the subject of research for which the Network data are sought; supports an Application for System Access.

E.3 Foundations of Data Governance

Data governance is defined in *2018 HHS Data Strategy: Enhancing the Evidence-Based HHS Portfolio* [14] as “a set of processes that ensure that data assets are formally managed such that departmental needs are met.” *Federal Data Strategy: Leveraging Data as a Strategic Asset* [15], from the Office of Management and Budget (OMB), goes into more detail. It names ten high-level Principles, three of which it classifies under “Ethical Governance.” They are: 1. Uphold Ethics, 2. Exercise Responsibility, and 3. Promote Transparency. The practices that support these three principles are as follows:

11. Prioritize Data Governance: Ensure there are sufficient authorities, roles, organizational structures, policies, and resources in place to transparently support the management, maintenance, and use of strategic data assets.
12. Govern Data to Protect Confidentiality and Privacy: Ensure there are sufficient authorities, roles, organizational structures, policies, and resources in place to provide appropriate access to confidential data and to maintain public trust and safeguard privacy.

This data governance framework focuses on the principles of security and privacy, data integrity, and data quality.

E.3.1 Data Security

Security of data protects the privacy of individuals whose data are the subject of research. Physical, administrative, and technical safeguards are three elements of data security that are necessary to protect the privacy of individuals whose data are queried during research investigations. Accordingly, policies and guidelines for Investigators/Researchers are essential. Data confidentiality and the risk of a breach are primary factors in governance and consideration of request for access to data.

E.3.1.1 HIPAA Privacy and Security

The Health Insurance Portability and Accountability Act (HIPAA, 1996), codifies the “Privacy Rule” in Title 45 of the Code of Federal Regulations, Part 160, and Subparts A and E of Part 164 and the “Security Rule” in Title 45 of the Code of Federal Regulations, Part 160, and Subparts A and C of Part 164. The Privacy Rule was issued to protect the privacy of health information that identifies individuals who are living or deceased. The rule balances an individual’s interest in keeping his or her health information confidential with other social benefits, including health care research.

The Privacy Rule standards address the use and disclosure of individuals’ health information—“protected health information”—by organizations subject to the Privacy Rule—“covered entities”—as well as standards for individuals’ privacy rights to understand and control how their health information is used. The rule permits important uses of information, while protecting the privacy of people who seek care and healing.

Any exceptions to or claims of exemptions from such policies should be clearly stated as a matter of policy. For example, if a section of HIPAA Privacy or Security Rules should be found not to apply to the Network, this should be stated among the policies.

E.3.1.2 Institutional Policies

Institutional policies and interpretation of Federal regulation and guidance may also inform policies to be considered for governance. For example, NIH Publication Number 03-5388 [16] outlines considerations for researchers who handle protected health information (PHI). Policies from other institutions govern conduct of their Investigators/Researchers and may affect their participation. When these policies conflict with Network policies, the Network will need to reconcile or otherwise resolve the conflict.

E.3.1.3 System Scope and Data Granularity

The granularity of data that the System will provide—i.e., Aggregate Data or Individual Level Data—will have a significant impact on governance policies and procedures and may impact the technical architecture to provide appropriate control and data security.

The Network should make available details about the scope of data content in the System, such as data sources, numbers of records, database statistics, and applicable use cases, e.g., on a website or other resource dedicated to sharing information about the Network.

E.3.2 System Access

Users will be granted access following execution of an Access Agreement (see Section 4.4 for more on types of agreements) and will be authorized to perform actions within the scope of training and terms of access to the system. System access will be assigned only to individuals with individual accounts, and not through shared accounts.

The Network should decide if institutional affiliation will affect user authorization, or if authorization decisions will be made solely at the individual investigator level.

E.3.2.1 Institutional Review Boards

An applicant seeking access to the data must certify the approval of his/her Institutional Review Board (IRB) for the Research Proposal. Data may be disclosed only within the context of the approved Research Proposal. Networks must decide what evidence of IRB approval or exemption is sufficient.

Note: Approvals of system access may depend upon system controls to restrict access to specific types of data, such as Aggregate Data or Individual Level Data, which may have personally identifiable information (PII) or may be de-identified.

E.3.3 Data Integrity

In this framework, data integrity means the assurance that data have not been altered in an unauthorized manner during storage, during processing, or while in transit [17]. Data integrity is critical for research. Each system-to-system transfer and transformation introduces the risk of introducing an error or omitting data that are relevant to the clinical question. Accordingly, traceability from query results to the source system raw data, and the use of an audit trail support best practices. Relevant regulations and guidance can be found in:

- §170.315(d)(8) of ONC Health IT Certification Criteria (2015 Edition), *Integrity* [18].
- 21 CFR Part 11 for Electronic Records and Electronic Signatures, which applies to applies to drug makers, medical device manufacturers, biotech companies, biologics developers, CROs, and other FDA-regulated industries for clinical investigations [19].
- HHS HIPAA guidance for research professionals [20].

E.3.4 Data Quality

Data quality considerations fall into three main categories:

1. Accuracy—The data, definitions, and any translations (mapping to the CDM) preserve the original data.
2. Completeness—Through the mapping exercise, the project team determined completeness as it reflects the association of concepts from the individual data models to the Common Data Model. In practice, datasets are complete and do not contain omissions that could be interpreted as a null result.
3. Fitness for Purpose—Particularly in the secondary use (reuse) of data, an assessment of fitness for purpose in the context of specific use cases is required. Understanding both the power and limitations of a data set is essential.

Errors may be introduced at several stages, from the original data sources to delivery of query data to the Investigator/Researcher.

1. Errors in Source Data—Errors may be introduced at the point of capture in a patient record.
2. Inconsistent Use of Source Data Fields—Source data fields may be populated in an inconsistent manner, depending upon operator training or ambiguity or lack of precision in defining a concept, for example.
3. Errors in Mapping—Mapping errors should be reduced or eliminated through prior validation of the maps. Maps will require maintenance throughout the lifecycle of the Network.
4. Changes to Underlying Data Sources—System upgrades to underlying data sources may introduce errors and could impact query translation.

The care taken in mapping the individual data models to the CDM, and the verification of the maps, must be sustained over the lifecycles of all systems.

Maintenance of data quality is an ongoing exercise and should be evaluated periodically to assure that quality does not drift over time and that the data continue support the scope of identified use cases (fitness for purpose).

The Network should consider what its process will be when errors are found in data, how it will acknowledge the error and alert users of the data; if it will request that the appropriate Data Partner correct the error; etc.

E.4 Governance Structure

The Network will need a governance structure to establish and maintain rules, processes, and procedures like the ones described above. This could take different forms; e.g., the Network might establish a Board or administrative working group. For the purposes of this document we will refer to the governance structure as "the Board."

Whatever form it takes, the governing entity must:

1) Ensure that the Network's mission remains clear, appropriate, and relevant as times change
2) Determine that the Network's programs and activities support the organization's mission and achieve both their short-term goals and long-term purpose
3) Exercise fiduciary responsibility to obtain and appropriately use the resources required to sustain the Network's mission

E.4.1 Board Responsibilities

A Board provides for oversight and the establishment and maintenance of policies and procedures for the Network in the following areas:

- developing and approving the mission of the Network,
- developing, approving, and maintaining its governance structure and organization,
- reviewing applications for System access,
- establishment or remediation of policies, guidelines, and procedures necessary for compliance with applicable laws, regulatory guidelines, institutional policies, or contractual obligations,
- compliance with established policies and guidelines for use of the System,
- approving nominations for new Board members,
- providing guidance and oversight in fulfillment of the Network's mission, objectives, and requirements,
- fulfilling any funding entity, agency, or institutional reporting requirements,
- investigating suspected breaches of policy or system integrity,
- responding to any confirmed breaches of policy or system integrity

E.4.2 Board Membership

Individuals who are selected for Board membership should be personally committed to the mission and objectives of the Network. They should reflect and represent the diverse interests of stakeholders and may also represent or champion a particular perspective on data sharing, e.g., privacy and security.

E.4.2.1 Code of Conduct

Board members will strive to represent the interests of all stakeholders and to act in the best interest of the health research community, funding organizations, data partners, and patients. Board service should align with the stated mission and objectives of the Network, Board, individuals, institutions, and community served, and not seek to serve self-interest or personal advantage.

E.4.2.2 Confidentiality and Transparency

All Board members should maintain in a strict and confidential manner all protected or confidential information and should not disclose it to any other party without prior approval of the Board.

Conversely, the Network should make its operations and decision-making as transparent as possible. The Board should guide the Network toward a balance between confidentiality and transparency that respects stakeholder interests while protecting patients' health information.

E.4.2.3 Conflict of Interest

Conflicts of interest (COI) are defined in terms of the *risk* of undue influence, not necessarily actual bias or misconduct [21]. The Network should establish COI policies to prevent compromised decision-making. For some organizations, the 42 CFR Part 50. 604 requires that institutions conducting PHS-funded research "Maintain an up-to-date, written, enforced policy on financial conflicts of interest [22]."

In general, COI exist in any of the following situations:

1) Activities or relationships with other persons or organizations affect a participant's ability, or potential ability, to render impartial assistance or advice, or give the appearance of doing so
2) The participant's objectivity is or might be impaired
3) The participant has or might acquire an unfair competitive advantage

In such cases where a Board member or a family member, friend, or associate has an established or potential conflict of interest due to a financial or other benefit from a topic or decision under consideration, the Board member must declare the potential conflict of interest and recuse himself or herself from voting, and, possibly also from discussion or debate.

COI may arise not only from financial interests, but also from non-financial engagements with or commitments to other organizations and associations with interests related to the subject matter being addressed by specific organizations activities.

E.4.3 Charter

The Charter is the foundational document for the operation of the Network. It contains the essence of the Network, promotes a shared understanding, and acts as an agreement with respect to the roles and responsibilities of Board members and obligations to stakeholders. The

Charter should include all the standard sections that would apply to any organization: Definition (context, purpose, and mission of the Network), Governing Board (purpose, functions, membership, and organization), Committees and Offices, Board Responsibilities, and Board Operations (selection, leadership, operations).

In addition to the standard sections, the Charter for the Network may include sections specifically applicable to data-sharing. These are described below, in section 4.4.

E.4.4 System Access and Data Use/Sharing Agreements

In this section we describe some additional documents, particular to data-sharing networks, that the Network may want to create and/or disseminate. Another prerequisite for system access is an agreement that outlines the terms and policies applicable to participants in the Network and obtaining formal acknowledgement from all participants. This is critical for compliance of the Network with respect to its obligations to stakeholders.

E.4.4.1 Application for System Access

An Application for System Access provides identifying information for investigators and a justification for access to Network resources. It should include, at minimum, the Investigator's/Researcher's name, contact information, institutional affiliation, and protocol description (e.g., Research Proposal).

E.4.4.2 Data Use and Data Sharing Agreements

If an Application for System Access is not comprehensive enough, the Network should develop a Data Use Agreement/Data Sharing Agreement (DUA/DSA). Two examples from the CDMH project are the PCORnet DSA and the HCSRN DUA Toolkit.

The PCORnet DSA defines the standard terms that govern the sharing of data and its transfer from participating Network Data Affiliates to the PCORnet Coordinating Centers. It broadly defines and supports the following elements:

1) Data completeness/data characterization activities
2) Analytic queries requiring return of Aggregate Data
3) Analytic queries requiring return of a Limited Data Set
4) Analytic queries requiring return of Protected Health Information (PHI) other than what is permitted in a Limited Data Set

The Health Care Systems Research Network (HCSRN) provides a DUA Toolkit [23] as a "guide…to facilitate the establishment of Data Use Agreements (DUA) between HCSRN sites."

Additional topics for consideration include the following:

1) Limited Use of Data—Data are to be used for the expressed purpose of the Research Proposal and no other, without approval of the Investigator/Researcher's IRB and the Board.

2) Breach Disclosure—The participant agrees to notify the Network in the event of a definite or suspected breach of data security or unplanned/unapproved disclosure or PHI.
3) No-Re-Identification—The participant agrees not to attempt to re-identify and/or contact patients.
4) Data Security—The participant will certify that the data will be used and stored in accordance with specific technical requirements (e.g., firewall, encryption, non-use of portable computers, non-use of detachable data storage devices (with or without encryption),
5) Data Retention—The Network may want users to specify which data they plan to retain, and for how long, in order to fulfill requirements from their institution or a journal.
6) Data Destruction—The Network may want users to destroy or delete data after use; the timeframe in which this must be done should be stated explicitly.
7) Termination of Relationship
8) No Warranty—The participant acknowledges that no warranty is expressed or implied with respect to accuracy, completeness, or fitness for purpose of the data.
9) Liability and Limitation of Liability
10) Indemnification—The participant is required to hold harmless the Network in the event of the participant's negligence, misconduct, or other breach of agreement with the Network and to cover legal fees incurred by the Network because of the participant's actions.

E.4.5 Standard Operating Procedures for the System

The Board may establish Standard Operating Procedures (SOPs) for routine use of the System. Suggested SOPs include the following:

- Incorporation of an Additional Data Source/Provider—outlines the contractual and technical steps necessary to incorporate an additional data source.
- System Maintenance—ensures that the System is maintained appropriately and with minimal impact to users.
- User Management—ensures System integrity and access exclusively by Authorized Users.
- Monitoring and Auditing—addresses System performance relative to service level agreements or technical specifications/benchmarks and compliant use of the system within the scope of authorized/intended use. Periodic review of system audit trails, "secure, computer-generated, time-stamped electronic record that allows for reconstruction of the course of events relating to the creation, modification, or deletion of an electronic record," is necessary to comply with FDA guidance for electronic records and data integrity [24].
- Unexpected Events (e.g., breach of system integrity)— establishes clear guidance and procedures for investigating a known or suspected breach of unsecured protected health information and notifying affected patients [25]. The scope and detail will depend upon the nature of the data. It is recommended that the Board undertake a risk assessment to

determine the nature and potential impact of potential risks. HHS resources identify the existence of at least four factors to be used in determining the response to a potential breach of unsecured protected health information:

1. The nature and extent of the protected health information involved, including types of identifiers and the likelihood of re-identification;
2. The unauthorized person who used the protected health information or to whom the disclosure was made;
3. Whether the protected health information was actually acquired or viewed; and
4. The extent to which the risk to the protected health information has been mitigated [25].

E.5 Additional Concepts for Consideration

Data-sharing Networks may also want to consider the following concepts related to governance.

E.5.1 Publication

The analysis of Network data, as outlined in a Research Proposal, may result in an Investigator's/Researcher's publication of results in a scientific journal. The Network may want to determine, in advance, citation guidelines and standard language for Investigator/Researcher acknowledgement of Network resources.

In its own publications, e.g., annual reporting, the Network may want to identify Investigators/Researchers using the System, in which case it will need to obtain their permission.

E.5.2 Intellectual Property

The Network may want to determine, in advance, who would own the rights to new intellectual property derived from Network data, and what restrictions (if any) are placed on it. If possible, the Data Partners should arrive at a consensus on one agreement for all Network data.

E.5.3 Copyright and Licensing

The (Re)usable Data Project (RDP) [26] provides information concerning licensing others' data and categorizes a variety of models [27]. While licensing and copyright would offer protection for Network data owners, they may impose barriers to the effective use or reuse of data and software tools. The integration of data from multiple Data Partners introduces additional complexity, not just related to the restrictiveness or permissiveness of a Data Partner's terms, but simply because the terms are different.

One way to overcome this challenge is for the Network to establish, if possible, a single harmonized licensing agreement with the most permissive terms possible for the benefit of Investigators/Researchers.

Licensing of terminologies/vocabularies (e.g., SNOMED CT, ICD-10-CM, etc.) or products used in Network data may introduce additional restrictions, subject to the terms of their copyright owners.

E.5.4 Financial Matters

The Network should determine at the outset if there will be any cost for Investigators/Researchers or their institutions to participate*. The Network may also want to explicitly forbid the resale of data.*

E.5.5 Small Cohorts and Data Security

De-Identified Data in smaller cohorts is more easily re-identified, bringing greater risks to patient privacy. The Network should consider this risk carefully and identify safeguards to reduce the possibility. E.g., setting a minimum cell size to reduce the risk of re-identification of individuals, or further anonymizing Aggregate Data results below a specified threshold (e.g., differential privacy [28]*).*

E.5.6 Virtual Research Community

A common characteristic of the four Data Partners from the CDMH Project is their establishment of virtual research communities.

The OHDSI program [29] refers to itself as a "community" and maintains numerous resources for researchers, including a Community Forum [30] for engagement on numerous topics, from standards to software tool development to research. OHDSI facilitates data exchange only to the extent of voluntary participation, by its members helping other members.

PCORnet has implemented the PCORnet Commons, "a space for people involved, invested, and interested in health research to collaborate, share, and learn." The Commons is built around access to resources for data, research, and engagement.

New Networks may want to consider establishing a Virtual Research Community that supports users of the system and helps the users to engage with other researchers, perhaps using one of the above models as a framework. The community could also be a channel for announcing planned downtime, collecting feedback about System updates, etc.

E.6 Conclusion

This document outlines current practices and, where possible, highlights best practices, in data governance. It highlights potential challenges that face data-sharing Networks: easier access to better, more comprehensive, and harmonized data, while preserving the integrity, security, and confidentiality of the data sources. We hope that highlighting these practices, challenges, options, and opportunities will support data-sharing Networks in their startup phase.

E.7 Networks Participation in the CDMH Project

The wide deployment of health IT systems has created unique opportunities for providers, healthcare professionals, and researchers to access and use patient data that is already collected during clinical workflows. Below are overviews of the four CDMH collaborating networks, who take advantage of these opportunities in different ways, with the common goal of making health data widely available to researchers.

E.7.1 The National Patient-Centered Clinical Research Network (PCORnet)

PCORnet aims to advance the shift in clinical research from investigator-driven to patient-centered studies. A hallmark of PCORnet is that the patients, clinicians, and healthcare system leaders are all actively involved in the governance and use of the data, at a national level and locally within each participating network. Organizations and stakeholders seek to communicate clearly with and seek input from patients, clinicians, and all other stakeholders about how their data are used in clinical research.

The PCORnet CDM employs healthcare standard terminologies (e.g., ICD, SNOMED, CPT) to enable interoperability with, and responsiveness to, evolving data standards.

Some CDMs allow for the systematic analysis of disparate observational databases. The concept behind this approach is to transform data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes), and then perform systematic analyses using a library of standard analytic routines that have been written based on the common format [31].

PCORnet was launched with substantial financial investment from the Patient-Centered Outcomes Research Institute (PCORI).

E.7.2 Informatics for Integrating Biology & the Bedside (i2b2) / ACT Network

The i2b2 model [32] has products for data use that support modular open-source software programs for query, exploration, and analysis of clinical and translational genomics data.

The i2b2 / ACT Network is "a real-time platform allowing researchers to explore and validate feasibility for clinical studies across the NCATS Clinical and Translational Science Award (CTSA) consortium, from their desktops. The ACT Network helps researchers design and complete clinical studies, and is secure, HIPAA-compliant and IRB-approved. The ACT Network leverages the Shared Health Research Information Network (SHRINE) to support multi-site research projects by enabling study feasibility/cohort discovery at partnered institutions [33]."

E.7.3 Sentinel

The Sentinel Initiative was launched in response to the Food and Drug Administration Amendments Act of 2007 (FDAAA) and comprises several components including:

- Sentinel System, a national electronic system for medical product safety surveillance [34]
- Active Risk Identification and Analysis System [35]
- Biologics Effectiveness and Safety System (BEST)
- FDA Catalyst [36]

The first phase of this initiative was the Mini-Sentinel Pilot to inform the development of the Sentinel System. In September 2014, the FDA began transitioning from the Mini-Sentinel phase to the full Sentinel System, which officially launched in February 2016.

The Sentinel Operations Center leads development of the Sentinel CDM, which “allows Data Partners to quickly execute distributed programs against local data.” [37] The *Sentinel System: Five-Year Strategy 2019-2023* states that Sentinel might in the future “[harmonize] its SCDM with other established CDMs such as the Observational Medicinal Outcomes Partnership, PCORnet, Informatics for Integrating Biology at the Bedside (i2b2),” with interoperability as a long-term priority. [38]

Duke Margolis Center for Health Policy hosts the annual Sentinel initiative public workshops. These gather the Sentinel community and leading experts to share recent developments within the Sentinel Initiative, to provide training on the Sentinel System’s tools and data infrastructure, and to promote engagement and collaboration with patients, industry, academia, and consumers.

E.7.4 Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP)

OHDSI promotes engagement among health informatics researchers across numerous research domains and disciplines.

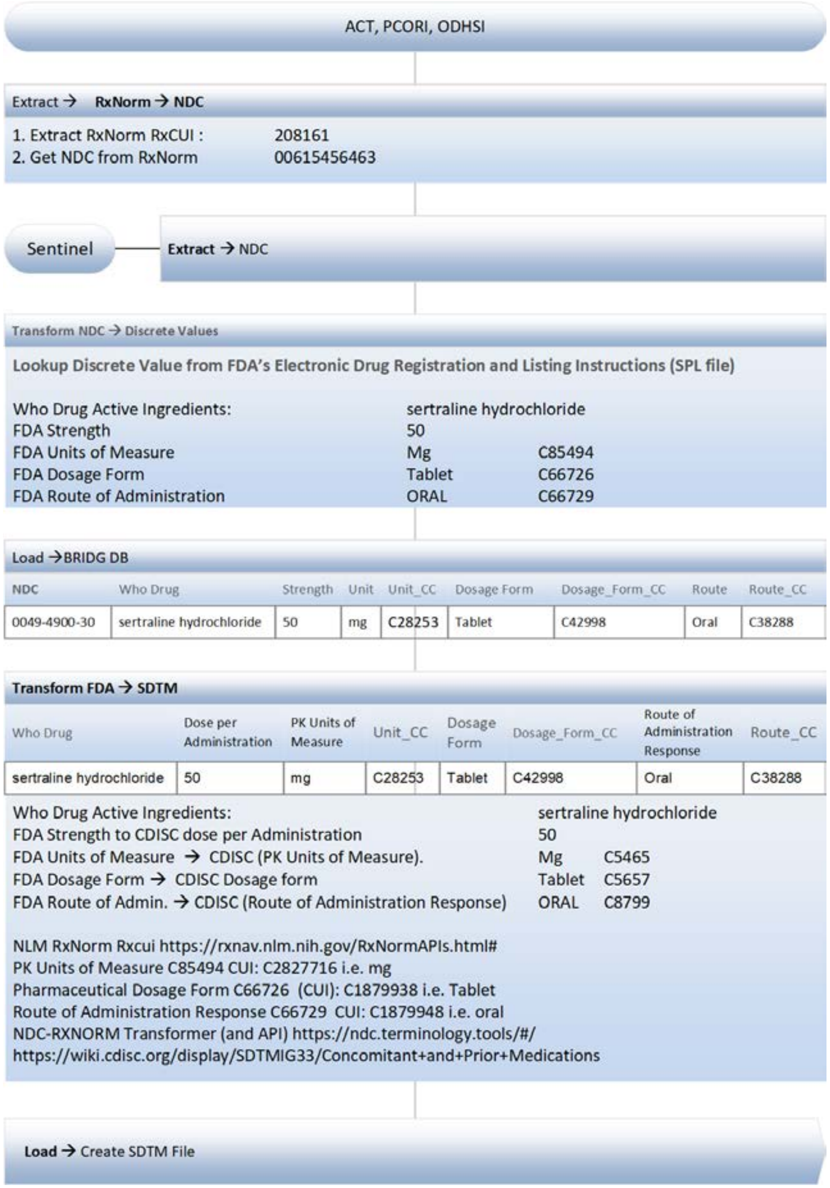
The OMOP CDM harmonizes disparate databases to a standardized vocabulary [39]. Guiding principles also outline architecture and design specifications which include project specific methodologies and hierarchical structures that represent the relationships between data.

To achieve the principle of inclusivity, OHDSI is an open collaborative. Anyone who can give time, data, or funding is welcome, and participation in the operation of OHDSI is expected [40].

APPENDIX F

NDC MAPPING FOR CDISC SDTM REQUIREMENT

The CDISC SDTM requirements for submitting drug information is to submit the following four data elements separately: drug code, drug form, route of administration, and dose unit. However, NDC codes encompass all these data elements under a single code. Therefore, and since the results set was received with NDC codes only, it was necessary to maintain mappings between NDC codes and the corresponding four data elements mentioned above. This mapping was then used when exporting CDISC SDTM datasets from the patient-level query results.



ACT, PCORI, ODHSI

Extract → RxNorm → NDC

1. Extract RxNorm RxCUI :	208161
2. Get NDC from RxNorm	00615456463

Sentinel

Extract → NDC

Transform NDC → Discrete Values

Lookup Discrete Value from FDA's Electronic Drug Registration and Listing Instructions (SPL file)

Who Drug Active Ingredients:	sertraline hydrochloride	
FDA Strength	50	
FDA Units of Measure	Mg	C85494
FDA Dosage Form	Tablet	C66726
FDA Route of Administration	ORAL	C66729

Load →BRIDG DB

NDC	Who Drug	Strength	Unit	Unit_CC	Dosage Form	Dosage_Form_CC	Route	Route_CC
0049-4900-30	sertraline hydrochloride	50	mg	C28253	Tablet	C42998	Oral	C38288

Transform FDA → SDTM

Who Drug	Dose per Administration	PK Units of Measure	Unit_CC	Dosage Form	Dosage_Form_CC	Route of Administration Response	Route_CC
sertraline hydrochloride	50	mg	C28253	Tablet	C42998	Oral	C38288

Who Drug Active Ingredients:	sertraline hydrochloride	
FDA Strength to CDISC dose per Administration	50	
FDA Units of Measure → CDISC (PK Units of Measure).	Mg	C5465
FDA Dosage Form → CDISC Dosage form	Tablet	C5657
FDA Route of Admin. → CDISC (Route of Administration Response)	ORAL	C8799

NLM RxNorm Rxcui https://rxnav.nlm.nih.gov/RxNormAPIs.html#
PK Units of Measure C85494 CUI: C2827716 i.e. mg
Pharmaceutical Dosage Form C66726 (CUI): C1879938 i.e. Tablet
Route of Administration Response C66729 CUI: C1879948 i.e. oral
NDC-RXNORM Transformer (and API) https://ndc.terminology.tools/#/
https://wiki.cdisc.org/display/SDTMIG33/Concomitant+and+Prior+Medications

Load → Create SDTM File

Figure 18. Transformation between CDMs.