Lessons Learned From Distributed Queries for the Clinical and Community Data Initiative (CODI)

July 2024

COLORADO HEALTH INSTITUTE
Informing Strategy. Advancing Health.

CODI
CLINICAL AND COMMUNITY DATA INITIATIVE

PHII PUBLIC HEALTH INFORMATICS INSTITUTE

THE TASK FORCE FOR GLOBAL HEALTH

The Clinical and Community Data Initiative (CODI) is a national effort spearheaded by the Centers for Disease Control and Prevention (CDC) to create individual-level longitudinal datasets that link people across organizations to understand chronic disease trajectories and interventions. CODI was first implemented in Colorado as part of the Colorado Health Observation Regional Data Service (CHORDS) Network in 2018.

The CODI implementation in Colorado, including two multi-step distributed queries and the execution of Privacy-Preserving Record Linkage, encountered several challenges. This document addresses sustainability, governance and technical challenges and summarizes lessons learned throughout the process of running complex multi-step distributed queries for CODI research.

Challenges and identified solutions associated with record linkage are described in a separate document developed by the CHORDS network, "Implementing a Quality Assurance Toolkit for the Privacy-Preserving Record Linkage Process in a Distributed Data Network of Health Care and Community Partners."

Acknowledgments

Support for the Clinical and Community Data Initiative (CODI) was provided in part by the Robert Wood Johnson Foundation (RWJF). The views expressed here do not necessarily reflect the views of the Foundation.

This report was also supported by Cooperative Agreement number 6–NU38OT000316, funded by the Centers for Disease Control and Prevention (CDC). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

This work was also supported in part by NIH/NCRR Colorado CTSI Grant Number UL1 RR025780. Contents are the authors' sole responsibility and do not necessarily represent official NIH views.

Executive Summary

CDC’s Clinical and Community Data Initiative (CODI) brings together people, processes and information technology in clinical and community organizations to build trust, create shared goals and link data. Linked local data can help answer questions about factors that affect individuals’ health.

CODI was first implemented in Colorado with three large health systems, two community-based organizations and an academic institution. The health systems and community-based organizations agreed to share data for the project and participate as data partners. CODI’s Colorado implementation used a distributed health data network to address questions related to the child obesity epidemic. The network model, which may be unfamiliar to public health professionals and health researchers, integrates health information technology with legal agreements and subject matter expertise to link and share data across organizations. Healthcare organizations are required to protect patients’ data and maintain high levels of data security under the Health Insurance Portability and Accountability Act (HIPAA). The model used in Colorado allowed organizations to share data for childhood obesity research and surveillance while ensuring patients’ privacy was protected.

CODI in Colorado took on two distinct research projects: one project explored changes in the prevalence of child obesity over time and a second project focused on whether the level of a child’s participation in a pediatric weight management intervention correlated with improvements in weight-related health outcomes on average. Each research project followed a multi-stage query process, where standardized queries were repeated several times with each data partner before analyzable datasets could be produced and shared with researchers. The query processes were designed as multi-stage queries to exchange the minimum data necessary.

Multiple challenges were encountered during the query execution process. In addition to technical challenges in getting the queries to run, there was turnover among the individuals involved in the process–including individuals who had been heavily involved in the query development–making it difficult to make even minor modifications after they left. The initial governance agreement to support CODI also had to be revisited.

Working through the challenges revealed several valuable lessons. First, sustainable public health informatics infrastructure is critical. Second, even in sophisticated organizations with experience managing distributed health data networks, keeping queries as simple as possible reduces longer-term maintenance needs. Third, data models need to be used somewhat regularly to ensure the technology and processes continue to run smoothly–they need exercise. These lessons may prove useful for other organizations involved in distributed health data networks or exploring data linkages between community and clinical data systems.

Background

Querying Relational Database Management Systems

According to Merriam-Webster's dictionary, to query is "to ask questions of especially with a desire for authoritative information." The term is somewhat anachronistic in modern discourse. However, among the community of software developers and other technical experts who use databases professionally, "query" carries a particular meaning. To a developer, a query is computer code that pulls data out of a database where data are stored in a useful format that can, for example, support a particular data analysis.

Queries are crucial in the context of relational database management systems (RDBMS), in which data are organized into multiple tables, each of which looks similar to a digital spreadsheet editor such as Microsoft Excel. As the "R" suggests, tables in an RDBMS are designed to relate to one another in particular ways. For example, two tables storing different types of information may contain an identifier for a specific person (e.g., a driver's license number) so that various types of data about that individual can be merged or joined together.

Software developers write queries for the RDBMS (i.e., "to ask questions of" the RDBMS) that produce query results in tabular formats similar to Microsoft Excel spreadsheets. Often, these queries are written using a particular computer coding language, such as a Structured Query Language (SQL). Several software products allow developers to store data in a RDBMS, write queries and view query results. Query code and query results are often displayed on the same computer screen. Popular SQL applications include Microsoft SQL Server Management Studio, MySQL and PostgreSQL, each of which has a slightly different SQL dialect. Some software programs designed for other purposes, such as complex statistical analysis and/or data visualization, can connect to and query a RDBMS. This allows data analysts to pull data out of a RDBMS and directly into a program they prefer, such as SAS or R.

Queries can be simple, consisting of two lines of computer code, or complex, with hundreds of lines of code or queries nested within queries nested within queries. Queries can access a single RDBMS table or many. When a query is written in a way that violates the logic and structure of an RDBMS, an error message is produced. Queries can also produce results that contain errors even though the RDBMS logic was not violated. For example, a query result may not contain all the individuals the developer wanted to include.

Electronic health record systems and distributed data networks

Electronic health record (EHR) systems are organized using complex RDBMS structures. A single healthcare organization's EHR system may have more than 18,000 tables. Even health systems using the same EHR vendor may have drastically different EHR systems due to differences in clinical workflows, billing processes and clinician preferences.

Distributed data networks, such as the networks comprising the National Patient-Centered Clinical Research Network (PCORnet), support clinical and public health research on a scale unimaginable in the past. Software developers with expertise in EHR systems write computer code to distill the complex structure of a specific EHR system into a simplified common format or a common data model.

When this common formatting or mapping process is repeated across multiple healthcare delivery systems, the common data model databases at different institutions can be queried with the same code. When the mapping is carried out consistently according to detailed data model guidelines, queries at different health systems produce data that can be directly compared or aggregated for analysis.

Local context

A national effort, CODI was first implemented in the Denver metropolitan area with three large health systems, two community-based organizations (CBOs), and the University of Colorado Anschutz Medical Campus serving as the Data Coordinating Center (DCC). The Data Coordinating Center managed the CODI project query development and query distribution. With the exception of the CBOs, these CODI partners had all participated in PCORnet networks prior to the CODI pilot and were familiar with distributed data networks designed for clinical research. Additionally, CODI partners participated in the CHORDS network, a separate (non-PCORnet) distributed data network designed for chronic disease surveillance at the local level. CHORDS leveraged the same data sharing infrastructure as PCORnet and had a very similar common data model. National leaders in childhood obesity treatment and surveillance were affiliated with CHORDS partners. Being relatively small and regionally-focused, the CHORDS network had flexibility in its governance processes to accommodate an initiative like CODI.

The history of the CHORDS network contributed to its selection for the CODI pilot and supported the successes of the Colorado implementation. The importance of the years-long relationships between the organizations and individuals participating in CODI cannot be overstated. These individuals had years of experience working together on the fundamental building blocks of CODI–including making shared decisions about cross-institution data sharing policies and procedures, mapping EHR data to a common data model, experimenting with record linkage techniques, developing queries for distribution, and administering the implementation of queries in diverse health systems.

Query development process

The CODI queries were informed by priority research questions developed by a group of leading experts in childhood obesity treatment and prevention. With their knowledge of the gaps in existing research literature and the potential capabilities of distributed data networks in addressing those gaps, the CODI research subgroup developed a list of potential priority questions. From that longer list, two questions were prioritized and selected for further development.

Computer scientists and informatics experts, who were contracted to support CODI's technical needs, worked with researchers and CODI partners to develop detailed use case specification documents for each of the two priority research questions:

1. What do CODI longitudinal data estimate the prevalence to be by weight category in children with various characteristics who are between the ages of 2 and 19 years in 2017, 2018, and 2019 (as of January 1 of each year) in Denver catchment and how do these estimates compare to other surveillance estimates if available?

2. Among children aged 2-19 years as of January 1, 2017, in the Denver area who participated in an intervention in CY2017, is intervention dose associated with health outcomes among children and did that association vary by patient characteristics or intervention type?

Based on the specifications documents, the same team of technical experts designed two multi-step query processes depicted in Figures 1 and 2 below. The multi-step queries were designed to limit the exchange of potentially identifiable information. Each of the queries incorporated relatively complex methods into the steps. In the case of the first research question, a separate computer program designed to improve data quality was incorporated in Step 5. In the case of the second research question, the query was designed to identify matched controls for comparison to intervention participants (in steps 3 and 4).

After the multi-step query processes were reviewed with CODI partners, a computer programmer who had not previously worked on the CODI project developed computer code in R and SQL to execute the multi-step queries. This development adhered to the query models, the use case specification documents, and the CODI data model.

Figure 1. Outline of Steps in Prevalence Use Case Query in CODI 1.0 in Colorado

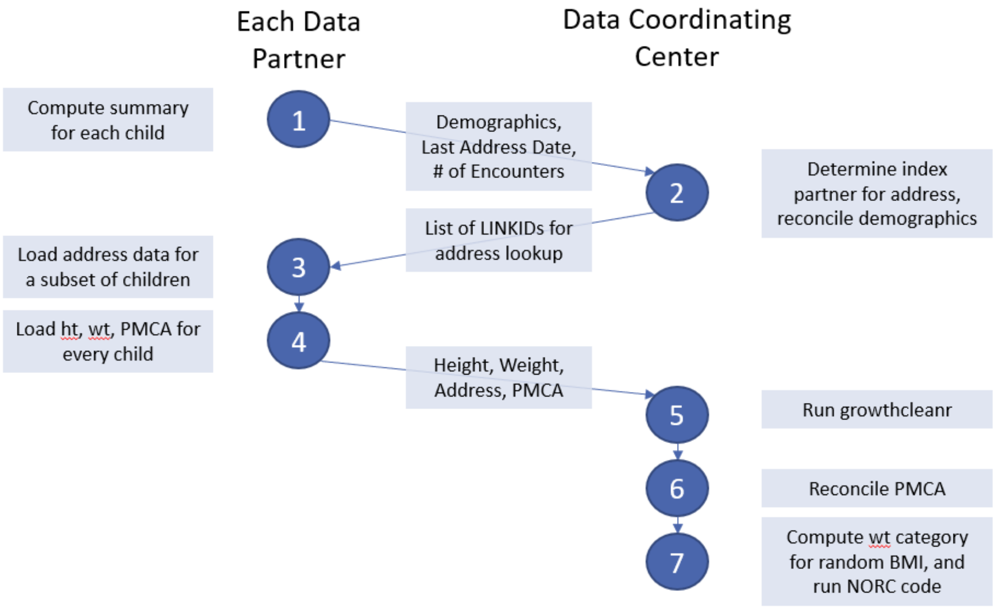
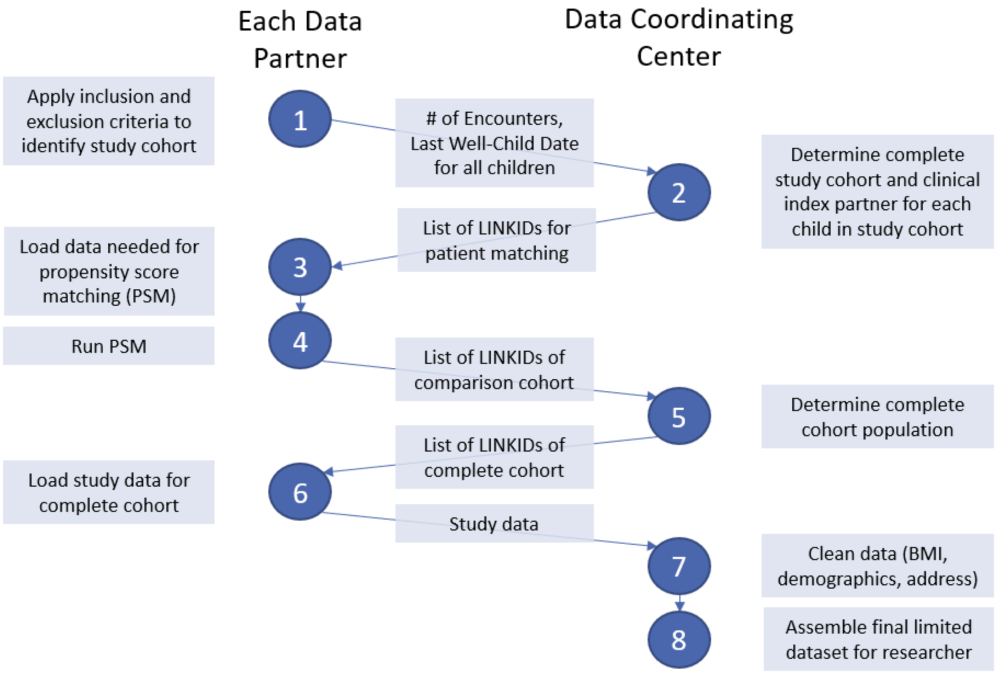

Each Data Partner
Data Coordinating Center
Compute summary for each child
1
Demographics, Last Address Date, # of Encounters
2
Determine index partner for address, reconcile demographics
List of LINKIDs for address lookup
Load address data for a subset of children
3
Load ht, wt, PMCA for every child
4
Height, Weight, Address, PMCA
5
Run growthcleanr
6
Reconcile PMCA
7
Compute wt category for random BMI, and run NORC code

Figure 2. Outline of Steps in Dose-Response Use Case Query in CODI 1.0 in Colorado


Each Data Partner
Data Coordinating Center
Apply inclusion and exclusion criteria to identify study cohort
1
# of Encounters, Last Well-Child Date for all children
2
Determine complete study cohort and clinical index partner for each child in study cohort
List of LINKIDs for patient matching
Load data needed for propensity score matching (PSM)
3
Run PSM
4
List of LINKIDs of comparison cohort
5
Determine complete cohort population
List of LINKIDs of complete cohort
Load study data for complete cohort
6
Study data
7
Clean data (BMI, demographics, address)
8
Assemble final limited dataset for researcher

Challenges encountered

Two rounds of queries

The code that was developed for CODI did not run as designed. The Data Coordinating Center (DCC) and individuals at CODI partner sites (together, the "CODI research team") developed modifications to the code and executed modified queries. However, due to errors in the linkage process that were discovered during execution of the queries, the resulting datasets could not be used for analysis.

Over a year passed between the initial execution of the queries and an attempt to execute the queries a second time after the linkage process had been redesigned and validated. During that year, the project experienced substantial turnover, both within the CODI research team and with contracted personnel providing technical assistance to CODI (e.g., designers of the query process). For example, two out of the three people involved in CODI at the DCC had left, and their positions were not filled. The loss of institutional knowledge and the passage of time proved to be challenges underpinning various other specific problems.

Governance revisited

An important outcome from the initial year of CODI was a master data sharing and use agreement that governed data sharing for any research projects within the CHORDS network that used the CODI infrastructure. This agreement, refined and executed by all CODI partners, gave the DCC the ability to approve data sharing for new research projects as they arose.

As the CODI research team embarked on the second round of queries, the team required additional capacity from researchers and fellows at the Centers for Disease Control and Prevention (CDC). The master agreement had not been reviewed by the CDC during its development. CDC colleagues identified necessary revisions to the agreement in order for the CDC to sign onto the agreement as a new partner with CODI in Colorado. However, the CHORDS network lacked the resources to revise the already-executed agreement. The agreement did not give the DCC the ability to make any edits to the agreement to evolve as projects may require.

Ultimately, individual data sharing agreements between each CODI partner and the CDC were required to support the second round of queries. The process to determine that new agreements were required and to develop the agreements required substantial time (nearly one year) that the master agreement had originally been designed to save.

Technical challenges

When the CODI research team attempted to rerun the queries a second time, none of the partner sites could get the code to run again. Common issues related to complex design and code–as well as updates to software and data–emerged after intensive investigation and troubleshooting.

Complex query design

The queries were complex and relied on dozens of files written in SQL and R to produce the results for each step of the query. The SQL and R code were interdependent; R code was used to execute SQL files and retrieve SQL query results. This back-and-forth interaction between R and SQL happened multiple times within a single step of a multi-step process. When one of the SQL files or R files produced an error, it took considerable time and investigation to uncover what was causing the error–whether it was a problem for one implementer or all implementers–and how to fix the error.

Sophisticated code

Some of the SQL and R files were relatively simple, but not all. The individuals who had developed or modified the queries had advanced programming skills that the team available for the second round of queries did not have. The code often referenced a data file generated several steps back in the back-and-forth interaction between SQL and R. Troubleshooting errors in sophisticated code proved difficult.

Software updates

The R programming language integrates packages that are computer programs developed for specific tasks by programmers worldwide who are not employed by the same company. Packages are stored in libraries that, similarly, are distributed geographically and organizationally. One such library, on which the R code written for CODI relied, was removed from use during the second round of queries. One CODI partner could not run query code with an alternative library and set of packages, requiring a site-specific tailored approach that caused several data quality problems later on in the query process.

Data updates

Unlike some research datasets, CODI datasets were connected to operational EHR systems that were changing while the linkage errors were being corrected and the new data sharing agreements were being developed. New EHR data were being collected. Modifications were being made to the EHR systems themselves. The extract, transform, and load (ETL) code designed to reformat EHR data into the CODI data model continued to operate while the linkage and governance challenges were being addressed. The passage of time meant that data were being collected for a longer period than was originally intended.

This extended time period created opportunities both to assess the relationship between the COVID-19 pandemic and longitudinal changes in children's body mass index (BMI), as well as to assess interventions' effects over a longer follow-up period. These opportunities required the development of new analytic plans for the CODI use cases. The original CODI code (both R and SQL) also needed to be modified. Conversely, maintenance of the site-specific ETL code varied across partner sites. Data quality issues encountered during the second round of queries may have resulted from low-to-no maintenance at some CODI partners.

Lessons learned

Sustainable infrastructure is critical.

Some of the challenges encountered due to turnover, governance and extract, transform, and load (ETL) maintenance would have been more easily addressed if the Colorado Health Observation Regional Data Service (CHORDS) network had more diversified and sustainable revenue streams. In parallel to CODI work, CHORDS partners were exploring how to sustain core services and functionality. Grant-based funding for topic-specific projects, including but not limited to CODI, did not provide the sort of institutional support that is required for a distributed data network to maintain core operations, such as a Data Coordinating Center (DCC) or well-functioning ETL code. Networks funded through National Patient-Centered Clinical Research Network (PCORnet) may not encounter these challenges as acutely as the CHORDS network did during the CODI implementation.

Social infrastructure matters, too. The CODI research team, which remained available to execute the second round of queries, was able to find creative solutions to continue the project. The relationships developed through the CHORDS network and CODI work have endured. Progress to date would not have been possible without these bonds.

Keep it simple.

There may have been reasons for developing a query structure and code that were as complex as the CODI queries. For example, relying on the query code to perform propensity score matching might have been seen as a way to reduce the burden on researchers or limit data being shared with researchers. The complexity of the queries and the code were too much for the CODI research team to manage easily. CODI partners were simply unable to produce research datasets with the code provided.

A simpler query process that relied only on SQL code, which the implementers were familiar with and which is less prone to software update bugs, would have been easier to implement and troubleshoot when problems arose. If complexity was required for computing performance, code and associated documentation could have been written with the expectation that the code might not work as originally intended and might need to be modified.

Distributed data networks need exercise.

Both the technical and social aspects of distributed data networks need to be used to identify issues, improve processes and keep things working optimally. The fact that the data sharing agreement was deemed unsatisfactory for the second round of CODI queries may have been attributable to the passage of time and loss of momentum. A lot of time had passed since lawyers and project managers had reviewed the agreement and so they may have held different perspectives upon revisiting it.

Also, data quality issues can be identified and corrected more easily if queries are being run periodically. Aside from the two rounds of queries, there were no queries being run against the CODI data model.

Conclusion

Thanks to generous funding from the Centers for Disease Control and Prevention (CDC) and the Robert Wood Johnson Foundation (RWJF), the CODI research team in Colorado has taken the time to navigate challenges related to sustainability, governance and technology to advance CODI research. Despite the various challenges throughout the last five years, as of January 2024, the CODI research team has developed a completed dataset for the first use case and a draft manuscript addressing the first research question. The query code for the second use case, which was more complex for several reasons, has been modified and a research dataset is close to being ready for analysis.

Final publications for CODI research will be made available on the CHORDS network website: https://www.coloradohealthinstitute.org/research/CHORDS