Initial Quality Assessment of the Privacy-Preserving Record Linkage Process in a Distributed Data Network of Health Care and Community Partners

JANUARY 2022

Table of Contents

CHORDS Staff Contributing to This Report

- Emily Bacon, Bacon Analytics and Denver Health and Hospital Authority (lead author)
- Andy Gregorowicz, The MITRE Corporation
- Kenneth Scott, formerly Denver Health and Hospital Authority
- Alexandra Tillman, Denver Health and Hospital Authority
- Nina Bastian, Colorado Health Institute
- Paul Presken, Colorado Health Institute
- Sara Schmitt, Colorado Health Institute

This report was supported by Cooperative Agreement number 6- NU38OT000316, funded by the Centers for Disease Control and Prevention. 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.

Executive Summary

In 2018, the Centers for Disease Control and Prevention (CDC) funded the initial pilot of the Clinical and Community Data Initiative (CODI) in Colorado. CODI added capacity to perform privacy-preserving record linkage (PPRL) across health care and community partners, which allowed the pilot project to incorporate unique data sources of varying data quality without the exchange of personally identifiable information. This report summarizes the process and challenges for implementing CODI's PPRL solution for children ages 2-19 among three Colorado health care data partners and two community data partners, as well as findings and lessons learned from initial quality assurance (QA) efforts. Two primary lessons learned from running CODI's PPRL solution are that processing time is a limiting factor and if one data partner makes an error, all partners must rerun PPRL. Initial QA activities found low matching concordance for patient birthdates, suggesting either a data quality issue from a data partner and/or a problem in the PPRL matching process. Additional quality checks suggest that the PPRL process included a matching step, or project, that used a field with high missingness, potentially creating weak matches that drove birthdate discordance. The PPRL process is being tuned to eliminate the high missingness element and to remove redundant data to increase the accuracy of patient matches. A formal, proactive QA process for each step of CODI's PPRL solution is proposed, which could mitigate some of the challenges inherent to the PPRL process that purposefully de-identifies data.

Introduction

There is growing interest and technological capacity to link individuals across systems for a variety of uses, including the ability to link across social service[1] and health care organizations.[2,3] In 2018, the Centers for Disease Control and Prevention (CDC) funded the initial pilot of the Clinical and Community Data Initiative (CODI) to add capacity to perform record linkage across health care and community partners. In Colorado, CODI is leveraging a regional distributed health data network—the Colorado Health Observation Regional Data Service (CHORDS) Network. CHORDS, which began in 2011, is a network of 14 health care and behavioral health partners across the Metro Denver region that share federated electronic health record (EHR) data for public health surveillance and research[2]. CODI created a unique opportunity to longitudinally link patients between three CHORDS health care data partners and two community-based organizations. Whereas CHORDS had developed a process to perform record linkage in partnership with a health information exchange[4], the CODI approach to record linkage can eliminate the need to exchange personally identifiable information (PII) and incorporate non-health care data sources.

Traditional record linkage methods, also known as clear-text record linkage, use PII such as name, date of birth, gender, and address to identify the same individual across organizations. New and alternative record linkage methods, called privacy-preserving record linkage (PPRL), use a variety of techniques to obfuscate personal identifiers before data are shared externally for cross-organization record linkage. By obfuscating personal

identifiers, the PPRL process can protect individuals' privacy, while also enabling data to be integrated at the individual level across systems.

There are many PPRL methods that have been developed and made available through various software products. Both open-source and commercial software tools are available, though the CODI pilot prioritized open-source options to enhance future feasibility and scalability. The CODI team conducted an expansive search to select a PPRL software tool for the CODI pilot. The analysis identified 32 potential PPRL tools. After reviewing attributes of all of the tools, a performance evaluation was conducted on a synthetic data set between a commercial tool and two open-source PPRL implementations. Ultimately, anonlink[5] was selected because it was open source and performed similarly to the commercial tool in terms of linkage quality. Anonlink is python-based and uses Bloom filters as a method to efficiently and securely assign a unique network-wide identifier to each individual appearing in any participating data partner's records. For more information on CODI's PPRL process, an implementation guide is available.[6]

This report describes the process of implementing PPRL for the CODI pilot and initial quality assurance efforts. It is important to develop approaches for PPRL quality assurance to ensure that the process is as efficient and effective as possible. Efficient PPRL can facilitate more partners participating in record linkage projects, quicker turnaround to create new datasets, and reduced burden on data partners. Effective PPRL is essential to have confidence in results from linked analyses. Assessing PPRL for effectiveness is inherently difficult due to layered processes that preserve patient privacy. In other words, if PPRL preserves privacy, clear-text identifiers are not accessible for validation. In addition to a description of PPRL implementation and initial quality assessment, this report also provides next steps and recommendations for other groups interested in using the CODI PPRL process.

Initiating PPRL

Three CHORDS Network health care data partners and two community data partners piloted PPRL for CODI. The health care data partners represent large, diverse providers in the Denver Metro region, including a children's hospital, a safety-net health system, and a Health Maintenance Organization (HMO). The two community data partners included an organization that offers youth development programs that promote physical activity and wellbeing, as well as an organization that leads statewide efforts to connect families and individuals to food resources. The data coordinating center (DCC) based at the University of Colorado Anschutz Medical Campus served as the linkage agent. A linkage agent is an organization that performs linkage on behalf of data partners. The linkage agent receives de-identified PII and produces globally unique LINKIDs. Each LINKID represents an individual, and if an individual is present in multiple data partner systems, the LINKID will match across organizations. LINKIDs are provided to data owners and data partners. In turn, data owners and data partners provide the LINKIDs along with other patient information to construct longitudinal records. One of the health care data partners served

as the key escrow and technical partner for the community data partners. A key escrow is an organization responsible for generating an encryption secret, called a "salt," that is used in the de-identification process. The key escrow provides the salt value to data owners and data partners securely to ensure the security of the process. CODI's PPRL solution was evaluated by an external agency to assess the protection of patient privacy throughout the record linkage process. This evaluation, called expert determination, influenced how PPRL was executed, including how long sensitive data files could be stored. All organizational roles and responsibilities established for the CODI pilot, including a master data sharing agreement, were agreed upon through a formal CHORDS governance framework.

The CODI pilot developed two use cases to motivate the development of the CODI Research Data Model, Record Linkage Data Model, and the PPRL processes.[7] The first use case was designed to assess longitudinal changes in a number of pediatric patients' health outcomes in order to evaluate the real-world effectiveness of pediatric weight management interventions and other community programs. The second use case was designed to calculate the deduplicated prevalence of pediatric obesity across the Metro Denver region. The initial PPRL included children aged 2-19 during an interaction with one of the health care data partners from 2016-2019. This time period was selected because several partners' data were expected to be of higher quality after 2016 than they were prior to 2016 due to their electronic health record systems.

All CHORDS Network partners have their own virtual data warehouses (VDWs) that conform to the same common data model specific to CHORDS. Since 2018, the CHORDS data model included a LINKAGE table designed to store network-wide identifiers. This table provided a framework for an analogous table in the CODI Identity Management data model.[7] To implement PPRL, the CODI sites built a new table, IDENTIFIER, specified in the CODI Identity Management data model. The IDENTIFIER table stores PII for pediatric patients or program participants that can be used in the linkage process. Appendix A shows the fields included in the IDENTIFIER table. If a data partner could not populate one or more fields, it could leave those fields blank. Details on the table can be found in the PPRL implementation guide.[6] The IDENTIFIER table is stored separately from other tables in the VDW because it is not available to query for public health or research projects as it contains PII.

PPRL Implementation

The process to run PPRL involves nine steps outlined below. These are referenced in the results section of the report as they relate to the quality assurance process. For a more detailed description of PPRL implementation, please see the PPRL implementation guide.[6] CODI developed Data Owner Tools to support data partners and Linkage Agent Tools to support the linkage agent (i.e., the DCC).

1. Data partners populate the IDENTIFIER table with PII for all eligible individuals.

2. A secret value, also called a salt value or key, is generated by the key escrow.
3. The key is disseminated to data partners.
4. Data partners run their instances of the Data Owner Tools, which contain scripts to extract information from the IDENTIFIER table and supply it to anonlink, generating de-identified, hashed data with the key.
5. All partners delete key values to reduce the possibility of reidentification.
6. Data partners send de-identified data to the DCC.
7. The DCC runs its instance of anonlink to generate unique, network-wide identifiers (LINKIDs).
8. The DCC sends LINKIDs to data partners.
9. Data partners load LINKIDs into their virtual data warehouses (VDWs).

One component of PPRL that is particularly relevant to the quality assurance process is the way that anonlink was deployed for CODI to match individuals (Step 7). The DCC uses the software Linkage Agent Tools to run anonlink four separate times with different information to create linkages. Each time anonlink is run it uses a different project, or unique combination of individual characteristics, which informs the tool which data elements should be used. All projects include the same first three characteristics: name, sex, and birthdate, but the fourth characteristic is different. The four projects include:

- Given and family name, sex, birthdate, and address
- Given and family name, sex, birthdate, and zip code
- Given and family name, sex, birthdate, and phone number
- Given and family name, sex, birthdate, and parent's email address

Each project is run for all patients and when matches are identified, they are recorded in a database. Two records are considered a match if their hashed information, creating a data structure called a Bloom filter, are within a certain distance of one another. Distance between Bloom filters is measured using a Sørensen–Dice coefficient. Records are considered a match in one project if the coefficient is greater than or equal to 0.8 (i.e., 80% concordance). This threshold of 0.8 was selected after performing tuning on a synthetic dataset, standards of practice in the field, and an examination of data quality across the network. After all projects are run, final matches are selected. If matches are identified by any project and the resulting linkages do not conflict, then the linkage information is returned. If there is a case where linkage information conflicts, then matching is based on how many projects patients were successfully matched on.

For example, a patient at Organization A and one at Organization B successfully match on the [name, sex, birthdate, address] and the [name, sex, birthdate, phone number] projects but not the other two projects. The information that matched in the first two projects reached a high enough threshold of 0.8, so the Linkage Agent Tools would consider the patients to be the same individual at each organization. Enough PII matched to suggest that there's a very good chance the patients are the same person, despite matching on only two of four projects. Because clear text PII is often fraught with data

entry errors (e.g., misspelling names, incorrect birthdates, addresses, or emails), matching is based on a high probability that patient information is the same or very similar, rather than expecting each component of PII to be exactly the same. If all PII had to be exactly the same, the linkage process would miss many real links across organizations (e.g., produce many false negatives). After this iterative process of running each anonlink project, a final LINKID is generated to indicate patient matches and the LINKIDs are sent back to data partners to continue with Step 8 listed above.

Results

Running PPRL

The CHORDS data partners have collectively run PPRL twice since the beginning of the CODI pilot. This section describes each round of running PPRL and the challenges that arose in each round.

Round 1: There were two main challenges that arose during the first iteration of PPRL. The first challenge was that the processing time was extremely slow. The original version of Linkage Agent Tools used a query processing system called TinyDB to store and process linkage results obtained from anonlink. TinyDB was originally selected to minimize the dependency requirements when installing Linkage Agent Tools. Unfortunately, TinyDB was not able to maintain acceptable performance on the scale of the linkage information being generated in this setting. TinyDB was eventually replaced by a more powerful query processing system called MongoDB, which is designed to maintain and query larger data sets. After several improvements it took the DCC approximately 40 hours to complete on a single server with eight CPU cores.

The second challenge was a data quality issue. The degree of overlap between data partners seemed low compared to previous work using health information exchange identifiers. Upon further inspection of individuals who had PPRL identifiers assigned, one data partner discovered that they had implemented an extract, transform, and load (ETL) process that selected children who were aged 2-19 on the date of the ETL (which occurred weekly), rather than during the eligibility period. Thus, the resulting cohort from that data partner was incorrect. Had the security processes allowed the data partner to retain the key, they could have rerun Data Owner Tools and sent new, corrected, de-identified data to the DCC for linkage with no additional time required of other CODI data partners. But since data partners had deleted their key values (per Step 5 in the PPRL process), the team decided to initiate the PPRL process again. This experience confirmed that patients could not be reidentified once the key values were deleted. Thus, an upside to this error was that it inadvertently validated that the PPRL process as designed made patient reidentification impossible.

Two primary lessons learned from the first iteration of PPRL were 1) processing time is a limiting factor and 2) if one data partner makes an error, all partners must rerun PPRL.

Round 2: Another ETL error occurred on the second iteration of PPRL when a different data partner mistakenly switched column headers (GIVEN_NAME and FAMILY_NAME) in their IDENTIFIER table. Preliminary Round 2 results looked incorrect again compared to previous linkage work within the Network. The data partner with the error was able to regenerate the hashed PII file without requiring the entire group to rerun PPRL and thus PPRL results from the second run could be used.

Data Validation

After the second round of PPRL created usable LINKIDs, the next step was to validate the PPRL process. The CODI team was able to explore concordance in sex and birth date between linkage matches across data partners for a subset of matches. These demographic characteristics were chosen because they were available in the DEMOGRAPHICS table and were able to be queried from all partners. This was only possible after the initial CODI research use case queries (described earlier in the Initiating PPRL section) had been run, and the DCC had individual-level, linked data from multiple partners. This analysis was enabled through the CODI Master Data Sharing and Use Agreement and the Community Data Partner Memoranda of Understanding, which allow the DCC to conduct data quality analyses across sites.[8]

Table 1 shows the percent of pediatric records that matched and shows sex and birth date concordance across each combination of three health care and two community data partners. The percent of patients matching across health care and community data partners varied widely. Between health care data partners, the percent of matching patients ranged from 2-35%. Between community data partners, the percent of matching patients was only around 1%, suggesting little overlap in populations. Both community data partners matched a moderate to high percentage of patients to the health care data partners. Community Data Partner #1 matched between 13-33% of its clients to at least one health care data partner; Community Data Partner #2 matched between 17-60% of its clients to at least one health care data partner. Ranges could be driven by a variety of factors, including the size and service area of the organization and the types of patients or clients in an organization (e.g., inpatient and/or ambulatory services).

The concordance of sex for linked records was high across all the partners, ranging from 93.7-100%. For Community Data Partner #2, approximately 22% of clients were missing information on sex. In this case, clients with an unknown sex were not considered discordant, which drove up the percentage of concordant links for sex for this data partner. While record linkage creates the possibility to "fill in" a patient's missing value with data from a linked data partner, exploring this functionality was outside the scope of PPRL quality assurance analyses.

The concordance of birth date had a wider range from 47.1-91.4%. Lower concordance was driven by Health Care Data Partner #3, which returned as low as 56.3% concordance with Health Care Data Partner #1 and 47.1% concordance with Community Data Partner

#2. The ideal level of concordance between partners depends on multiple factors, including underlying data quality of data partners and the ways the data are going to be used. For example, a clinical use case that would be highly sensitive to false positives may set a higher concordance threshold of 90% or greater, whereas use cases where missing linkages could bias results (e.g., a diagnosis-based use case) may benefit from a lower threshold. CODI expected birth date concordance of at least 80% for its first two use cases. While birth date concordance met this threshold across all other health care and community data partners, the consistently low birthdate concordance for Health Care Data Partner #3 warranted further evaluation.

Table 1. Percent of matched pediatric patients and demographic concordance across three health care and two community data partners using privacy-preserving record linkage (PPRL)

% matched first DP / % matched second DP	% Sex concordance / % DOB concordance	Health Care Data Partner #1		Health Care Data Partner #2		Health Care Data Partner #3		Community Data Partner #1	
Health Care Data Partner #2		HCDP1: 35%	Sex: 99.8%						
		HCDP2: 11%	DOB: 85.8%						
Health Care Data Partner #3		HCDP1: 17%	Sex: 99.7%	HCDP2: 19%	Sex: 99.8%				
		HCDP3: 4%	DOB: 56.3%	HCDP3: 14%	DOB: 70.7%				
Community Data Partner #1		HCDP1: 1%	Sex: 100.0%	HCDP2: 1%	Sex: 99.9%	HCDP3: <1%	Sex: 99.8%		
		CDP1: 15%	DOB: 86.7%	CDP1: 33%	DOB: 92.5%	CDP1: 17%	DOB: 81.8%		
Community Data Partner #2		HCDP1: <1%	Sex: 96.1%*	HCDP2: <1%	Sex: 93.7%*	DP3: <1%	Sex: 98.8%*	CDP1: <1%	Sex: 100.0%*
		CDP2: 60%	DOB: 91.4%	CDP2: 45%	DOB: 85.9%	CDP2: 17%	DOB: 47.1%	CDP2: 1%	DOB: 66.7%

**Community Data Partner #2 had a significant number of individuals with unknown sex. These individuals were excluded from the denominator in the % concordant calculation.*

The CODI team took a multifaceted approach to evaluating low birth date concordance across several data partners. There are two potential drivers of low birthdate

concordance—a data quality issue from a data partner and/or a problem in the anonlink matching process. Because birth date concordance was low for all Health Care Data Partner #3 linkages, the CODI team began examining birthdate data quality.

Birth Date Data Quality Checks

The CODI team performed the following data quality checks with Health Care Data Partner #3 only:

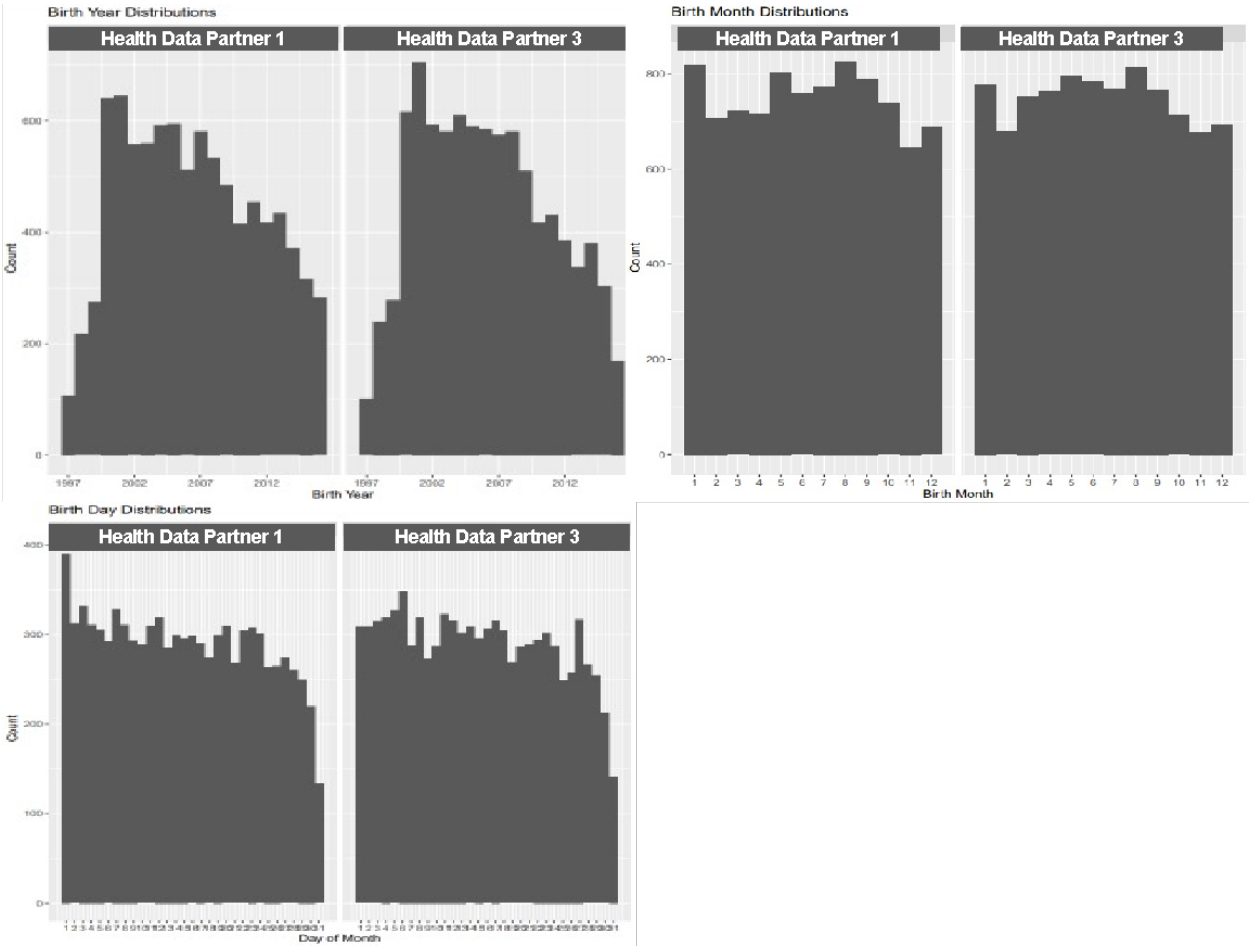
- Confirm birth dates in the patient DEMOGRAPHICS table (from which the IDENTIFIER table is created) match birth dates in the IDENTIFIER table
- Confirm the time zone for birth time and ensure Mountain Time like the other data partners
- Confirm dates are in ISO (YYYY-MM-DD) format in the IDENTIFIER table and are being exported as such in the pii.csv folder (part of the PPRL process)

Health Care Data Partner #3 confirmed that all criteria were met. Next, the CODI team examined birth date data quality across all data partners. The DCC performed the following data quality checks for all data partners:

- Distributions of birth month, day, and year for discordant birth date links; specifically looking for placeholder (e.g., January 1) or missing dates
- Number of discordant birth date matches that would be corrected by a month-day swap, indicating a common data entry issue
- Time between discordant birth dates across sites; specifically looking for clusters such as 1 day, month, or year differences that could indicate a common data entry issue or a low threshold for PPRL date matching

Figure 1 provides an example of the distributional comparisons for birth month, day, and year across discordant birthdates between Health Care Data Partner #3 and Health Care Data Partner #1. It shows expected distributions with relatively few outliers across organizations. Notably, there were more possible outliers in Health Care Data Partner #1 than Health Care Data Partner #3, but the latter had the most discordant birth dates.

Figure 1: Distribution of birth year, birth month, and birth day for linkages with discordant birthdates between two health care data partners.*



** Health care data partners are labeled as Health Data Partners in the figure.*

Table 2 shows the number of discordant birth date matches that would be corrected by a month-day swap for each data partner combination. Overall, very few birthdates would have become concordant if the month and day would have been swapped, indicating that this was not driving the high number of discordant birthdates.

Table 2: Number of discordant birth date matches that would be corrected by a month-day swap for each data partner combination.

Corrected by month-day swap/ Not corrected by month-day swap with sensical date	Health Care Data Partner #1	Health Care Data Partner #2	Health Care Data Partner #3	Community Data Partner #1
Health Care Data Partner #2	8 / 2,408			
Health Care Data Partner #3	14 / 3,779	29 / 8,591		
Community Data Partner #1	2 / 65	0 / 86	0 / 99	
Community Data Partner #2	0 / 18	0/18	0 / 22	0 / 1

Figures 2 and 3 show time between discordant birth dates for two sets of health care data partners. Figure 2 shows days between discordant birth date matches for Health Care Data Partners #1 and #3. Health Care Data Partner #3 was the partner with the highest rate of discordant birth date matches. There are three scales presented in each figure: discordant birth dates that are 0-31 days different, 0-365 days (1 year) different, and the entire range of difference (in days). For reference, Figure 3 shows the same analysis for Health Care Data Partners #1 and #2, which had high (85.8%) birth date concordance.

Among two partners with a high percentage of discordant birth dates, Figure 2 shows some clustering of discordant birth dates that differed by 1 day, 10 days, each month mark, and each year mark. However, there is a similar pattern in Figure 3 among two partners with a low percentage of discordant birthdates. One hypothesis explaining this pattern is that birthdates were not entered correctly by one or both data partners (e.g., 10/12/2009 vs. 10/13/2009). This would constitute an underlying data quality issue. Because the same pattern was also found among two partners that had a high percentage of birth date concordance, it is unlikely that this is driving the high birthdate discordance seen for Health Care Data Partner #3 and other sites.

A second hypothesis is that the anonymized strings produced in the hashing process make proximate birth dates (e.g., 10/12/2009 and 10/13/2009) match because they are very close together. This is by design, as the linking process attempts to find matches where typos have been introduced into the data. However, because the differences in days between discordant matches fall across a broad spectrum, it suggests that data entry

errors and matching thresholds alone are likely not responsible for birthdate discordance.

Figure 2: Days between Birthdate Discordant Matches for Health Care Data Partner #1 and Health Care Data Partner #3, shown at three scales from 0-31 day difference, 0-365 day difference, and the total range of difference.

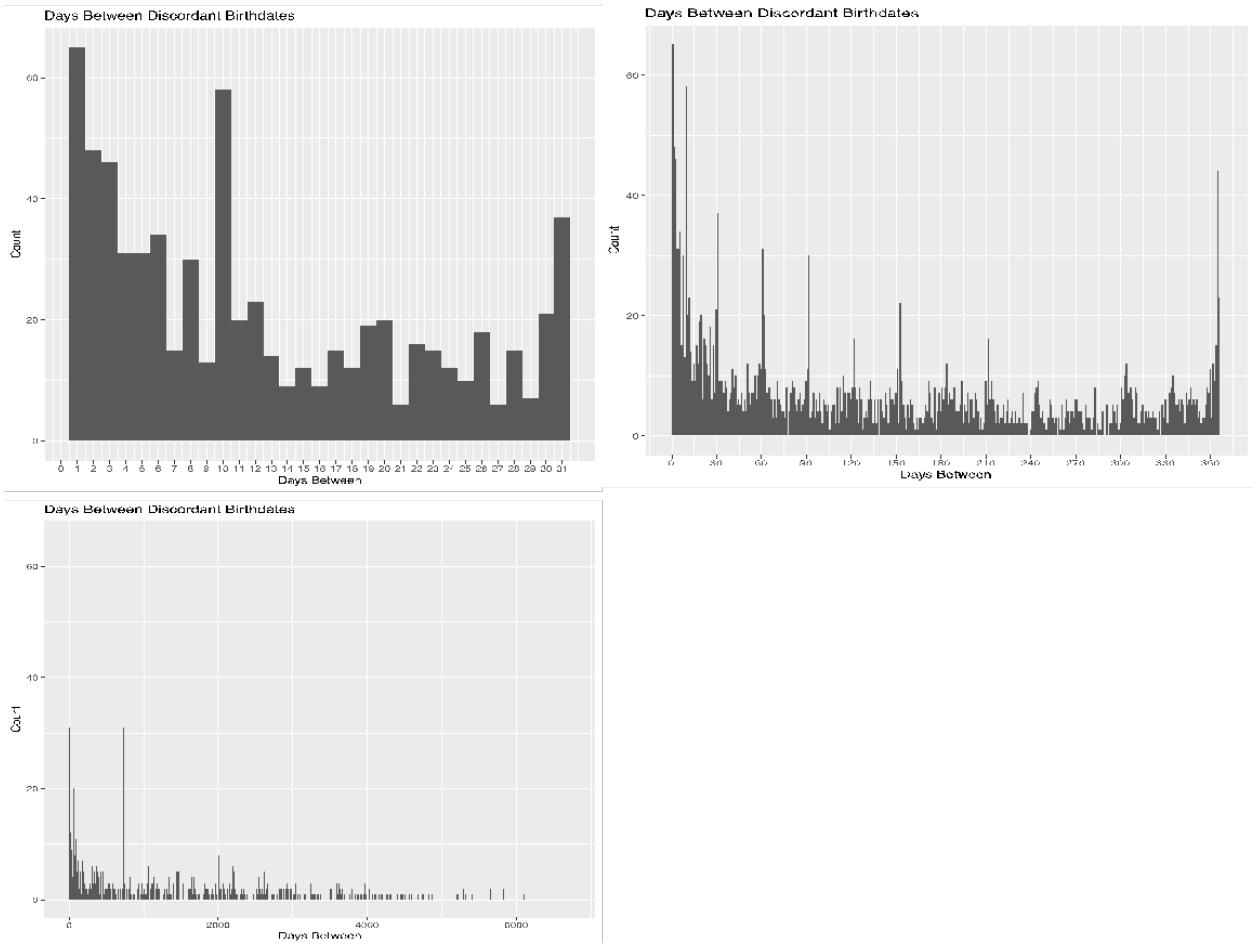
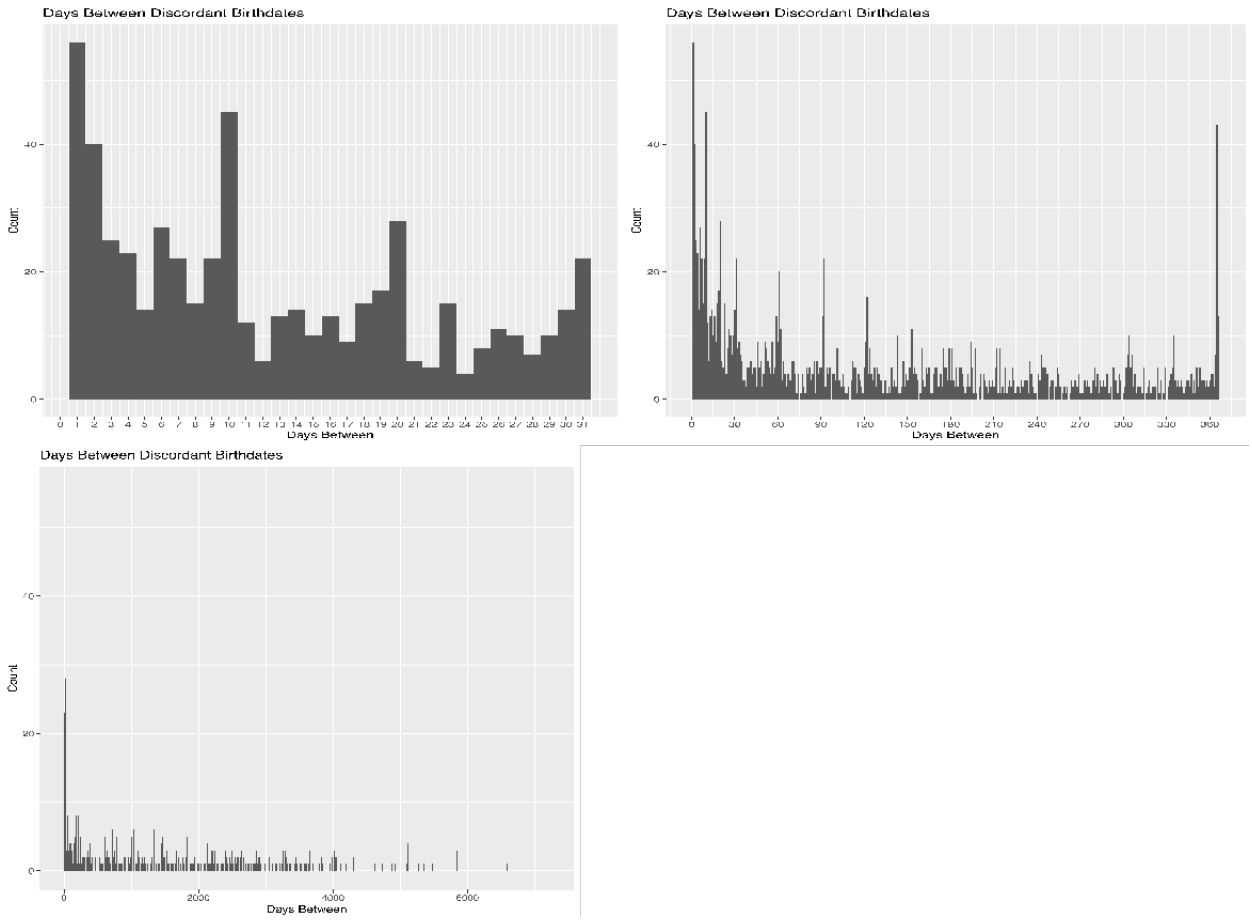


Figure 3: Days between Birthdate Discordant Matches for Health Care Data Partner #1 and Health Care Data Partner #2, shown at three scales from 0-31 day difference, 0-365 day difference, and the total range of difference.



Anonlink Matching Process

None of the examinations of birthdate data quality revealed issues large enough to create the extent of birth date discordance that was observed. The CODI team then began to look into the PPRL process as the second possible driver of birth date discordance. By design, PPRL has an extensive process of garbling individual identifiers so that individuals cannot be traced back to their PII in data partner systems. These security measures, while essential, pose challenges for PPRL quality assurance (QA) around PII used to link individuals. One way to identify whether an issue with the PPRL process was driving birth date discordance was to examine the MongoDB database created by Linkage Agent Tools that housed the output of the anonlink projects.

It is important to note that the database contains networks of record linkages. As an example, the database can contain a network where Person 1 at Organization A matches

to Person 5 at Organization B via the name, sex, birthdate, and address project. That network can also contain the same Person 5 at Organization B who matches to new Person 10 at Organization C via the name, sex, birthdate, and telephone number project. Additionally, the network could contain the same Person 10 at Organization C who matches to a new Person 2 at Organization A via the name, sex, birthdate, and zip code. This final piece of information introduces a conflict in the network, as the match network identifies two separate individuals at Organization A (person 1 and person 2) as a potential match. Linkage Agent Tools contains an algorithm to resolve these situations, favoring linkages which are established by a greater number of anonlink projects.

The following data quality checks were performed on the database of anonlink results:

- A query was developed examining the number of projects that matched for each link network. The lowest possible number of projects is 1 and the highest possible number of projects is roughly unbounded due to the fact that the link networks can contain a high number of conflicting links.
- A back tracing program was built to identify the specific projects used to match a single anonymized link ID across partners. Results included the number of different records from each data partner that were in the network for a given link ID and the specific anonlink project combinations for each potential record match. A subset of links with discordant and concordant birth dates were analyzed to see if patterns emerged.

The second data quality check, which looked at the way a specific link was formed from each data partner and anonlink project, revealed that many patients were only linking on the project that included name, sex, birth date, and parent email. These links were not being made on other projects, such as those including address or phone number. The CODI team discovered that if email data were missing the missing data could still be included in the matching process. The team asked the health care data partners to report the percentage of records in their IDENTIFIER table missing parent email address, and this revealed that 66-80% of parent email data was missing. This led to the hypothesis that the inclusion of a project using a field with high missingness may create a large number of weak or erroneous links, which could drive birth date discordance.

Next Steps

The CODI project team plans to continue investigating the PPRL process into 2022. To test the hypothesis of missing email data causing erroneous matches, the CODI team will create a script using the MongoDB database created by Linkage Agent Tools to examine what the results of the matching process would have been if only three anonlink projects were used. This analysis may show enough improvement to rerun the PPRL process with this change to obtain satisfactory results.

Additionally, the CODI team will re-examine the matching thresholds used by anonlink and potentially increase one or more matching thresholds. While this may reduce the number of false positive links, it could also increase the number of false negatives. To better understand the impact of raising the anonlink matching threshold, the team will record the number of linkages that have matching sex and birthdate that are lost when the threshold is raised.

If the two previous approaches do not yield satisfactory results, the CODI team may explore modifications to the algorithm used in Linkage Agent Tools to process the results from MongoDB and generate final linkages. Improvements could include assigning weights to particular anonlink projects or defining rules about which project or projects may be sufficient to create a linkage.

Finally, if the previous approaches are unsuccessful, the CODI team may consider modifying governance rules to compare clear text PII for data quality assurance processes or deploy an alternative PPRL solution to anonlink.

Conclusion and Lessons Learned

There were several challenges implementing PPRL that serve as learning opportunities for CODI, the CHORDS Network, and other groups interested in implementing a similar PPRL process. The initial run of PPRL was slow enough that it posed a timing and system bandwidth challenge for the DCC. Switching to a more powerful query processing system (MongoDB) improved processing time. Additionally, data model conformance issues in some partners caused PPRL to fail and demonstrated that a conformance issue for one partner requires all data partners to rerun PPRL. Once PPRL was successfully run, it was challenging to assess the quality of the matches. When a matching issue arose due to a low percentage of concordant birth dates, particularly for one health care data partner, identifying the source of the low concordance was complex. The important steps that PPRL takes to de-identify data also make it challenging to troubleshoot when problems occur.

CODI underwent an expert determination process to ensure that patient privacy was protected throughout each step of the record linkage process. The expert determination concluded that CODI should implement a synchronous approach, wherein salts and hashes are destroyed immediately after the LINKIDs are generated. An asynchronous PPRL approach would allow data partners and the DCC to retain salt and hashed values after the initial LINKIDs were generated. Retaining this data would facilitate QA review and facilitate or avoid re-running PPRL. Some sites considering PPRL could consider the balance between efficiency and privacy in finding a middle ground and eliminate the burden caused by the "weakest link" if PPRL fails for one partner. This may be particularly important when implementing PPRL across many organizations.

Other sites that wish to implement PPRL should consider developing formal QA review for each step of the PPRL process. Some of these QA elements were implemented by CODI but not always in a systematic way.

A formal QA process could include:

- Reporting from data partners on core areas of data quality for the IDENTIFIER table before beginning Step 1 of PPRL
    - Completeness (e.g., number/percent of missing records across each field)
    - Conformance (e.g., correct data types for each field)
    - Plausibility (e.g., number of placeholder birthdates, birthdates outside of allowable range)
    - Persistence (e.g., expected number of patients throughout the desired period)
    - Deduplication, or multiple records for a single patient (e.g., the same patient with multiple addresses)
    - Display a small sample of rows to a user for visual inspection
- A system for generating concordance of all identifiers used in linkage
    - Maintenance of hashed keys until after QA process is complete
    - Establishment of desired concordant threshold for each identifier (e.g., 80% sex concordance)
    - The DCC could run initial concordance analyses before research queries are generated and generate a matrix of concordance across partners
    - Ability to quickly distinguish partners or identifier characteristics that have higher than acceptable discordance across one or more areas
- A process for communicating with data partners about identification and resolution of data quality issues, particularly for community data partners and their technical partners
    - Includes expectations about timeliness of feedback and participation in diagnosing or fixing issues

Implementing PPRL for health care and community data partners has immense promise to provide longitudinal records showing the impact of physical, social, and behavioral interventions on an individual's overall health. While this analysis highlighted some of the initial data quality and implementation challenges in the process of PPRL using anonlink, there were a number of components about PPRL that were promising. Diverse data partners were all able to populate a standardized IDENTIFIER table to store PII for the PPRL process. Additionally, anonlink was user-friendly for data partners and its open-source nature makes it possible to share with other sites.

Overall, the number of pediatric patients that linked across most data partner organizations was high enough to create a sizeable sample for meaningful research use cases. This high sample size is still expected after tuning PPRL increases the quality of the matches. However, each use case for research, surveillance, or evaluation will have its

own criteria for matching quality and sample size that is important to evaluate throughout the PPRL process. Finally, the high concordance of sex across linked patients indicated that some demographic information was linking well. Future quality assurance processes can continue to refine the efficiency and effectiveness of the anonlink PPRL process for large distributed data networks.

References

1. Yousefi-Rizi, L., Baek, J.-D., Blumenfeld, N., & Stoskopf, C. J. (2021). Impact of Housing Instability and Social Risk Factors on Food Insecurity Among Vulnerable Residents in San Diego County. *Journal of Community Health*, 1-8.
2. Bacon, E., Budney, G., Bondy, J., Kahn, M. G., McCormick, E. V., Steiner, J. F., . . . Davidson, A. J. (2019). Developing a regional distributed data network for surveillance of chronic health conditions: the Colorado Health Observation Regional Data Service. Journal of Public Health Management and Practice, 25(5), 498.
3. Tumber, M. B., Bunzli, L., & Rosenberg, M. (2019). Addressing the social determinants of health: the Rhode Island state innovation model (RI SIM) experience. Rhode Island Medical Journal, 102(5), 22-25.
4. Scott, K. A., Davies, S. D., Zucker, R., Ong, T., Kraus, E. M., Kahn, M. G., . . . Bacon, E. (2021). A process to deduplicate individuals for regional chronic disease prevalence estimates using a distributed data network of electronic health records. Learning Health Systems, e10297.
5. CSIRO's Data61. (2017). Anonlink Private Record Linkage System. Retrieved from https://github.com/data61/anonlink
6. The MITRE Corporation. *CODI Privacy Preserving Record Linkage Implementation Guide*. 2020 [cited November 9, 2021]; Available from: https://github.com/mitre/codi/blob/main/CODI%20PPRL%20Implementation%20Guide.pdf.
7. The MITRE Corporation. *CODI Data Model Implementation Guide*. 2020 [cited November 9, 2021]; Available from: https://github.com/mitre/codi/blob/main/CODI%20Data%20Model%20Implementation%20Guide.pdf.
8. Colorado Health Observation Regional Data Service (CHORDS). CODI Master Data Sharing and Use Agreement. *CHORDS Governance Plan 3.0.* 2021. [cited November 9, 2021] https://www.CHORDSnetwork.org.

Appendix A:
IDENTIFIER Table Fields from the CHORDS 3.5 Data Model Manual

Name	Description	Definition	Data Element Provenance
IDENTIFIERID	An arbitrary, unique identifier.	int NOT NULL	CODI
PERSON_ID	A link back to the demographic table.	NVARCHAR(36) NOT NULL	CODI
BIRTH_DATE	Date of birth.	DATE NOT NULL Recommended format MMDDYYYY	CODI
SEX	Gender or sex of the person M = Male F = Female U = Unknown O = Other (Transsexual, Transgendered, or anything else that does not fit into one of the prior categories)	NVARCHAR(1) NOT NULL	CODI
GIVEN_NAME	A given name for the person. Often known as the person's first name.	NVARCHAR(50) NOT NULL	CODI
FAMILY_NAME	A family name for the person. Often known as the person's last name.	NVARCHAR(50) NOT NULL	CODI

MIDDLE_INITIAL	A middle initial for the person.	NVARCHAR(50) NULL	CODI
SSN	An SSN for the person.	NVARCHAR(9) NULL	Used in the HIE panel file.
INSURANCE_ NUMBER	An insurance number for the person as it appears on an insurance card.	NVARCHAR(50) NULL	CODI
MEMBER_ID	Member ID as assigned by the organization such as the medical record number.	NVARCHAR(50) NOT NULL	
HOUSEHOLD_ STREET_ADDRESS	An address for the person, including number/name/unit (i.e., the information sometimes referred to as street line 1 and street line 2).	NVARCHAR(50) NULL	CODI
CITY	City for the household address.	NVARCHAR(50) NULL	Used in the CORHIO panel file.
STATE	State for the household address.	NVARCHAR(50) NULL	Used in the CORHIO panel file.
HOUSEHOLD_ZIP	A ZIP code for the person.	NVARCHAR(50) NULL	CODI
HOUSEHOLD_ PHONE	A phone number for the person.	NVARCHAR(50) NULL	CODI
HOUSEHOLD_ EMAIL	An email address for the person.	NVARCHAR(50) NULL	CODI