Lessons Learned: Optimizing Record Linkage Across Clinical and Community Organizations

Andy Gregorowicz, MS[a]; Dylan Hall[a]; Keith J. Miller, PhD, MA[a]; Emily Bacon, PhD[bc]; Alexandra R Tillman, MS[c]; Bryant Doyle[d]; Kaitlin Porter, MPH[a]; Raymond J. King, PhD, MSc[e]

[a] The MITRE Corporation; [b] Bacon Analytics; [c] Public Health Institute at Denver Health and Hospital Authority; [d] University of Colorado, School of Medicine; [e] CDC

CLINICAL AND COMMUNITY DATA INITIATIVE

Background

- Public health research can benefit from access to individual-level, linked longitudinal records that include clinical interventions and outcomes, participation in community programs, and other social determinants of health.
- CDC's Clinical and Community Data Initiative (CODI) deployed a pilot project in Denver, CO to share and link information on individuals between three healthcare providers and two community-based organizations ("Data Owners").
- CODI linked records using a version of the *anonlink* Privacy Preserving Record Linkage (PPRL) software, which allows a designated party to match records using obfuscated information.

Because PPRL ensures privacy, evaluating linkage quality can be a challenge compared to plain-text linkage

Objective/Purpose

Initial results from CODI's PPRL tool indicated lower-than-expected date of birth (DOB) concordance for patients matched across data owners, and further analyses also suggested issues with linkage quality.

- Expected DOB concordance: >90%
- Actual (after first PPRL run): **<60%** between some organizations

To better define and fix PPRL issues, the CODI Team developed a four-stage quality assurance (QA) toolkit, built into the PPRL process:

Stage 1: Data characterization pre-linkage
Stage 2: Examining initial hashes pre-linkage
Stage 3: Examining linkage patterns post-linkage
Stage 4: Analyzing demographic concordance of patients post-linkage

Stage 1: Data Characterization

Prior to linkage, Data Owners ran a data characterization script to assess the quality of their PII data compared to expected values. For example, for DOB the assessment checks: earliest and latest dates, number of missing dates, number of dates before the earliest expected. This was checked against what was known about the organization (e.g., the expected age of patients).

Results: Identified duplicate records; corrected a Data Owner issue with "sex" field; discontinued use of "parent email" in PPRL

Stage 2: Hash Assessment

After hashing, Data Owners compared:

1. **Count of unique hashes**: to ensure count matches the number of input records
2. **Exact hash matches across Data Owners:** to estimate Link ID matches post-linkage

Results: Identified duplicate records due to multiple, historical addresses in one Data Owner's database

Stage 3: Link Pattern Analysis

The Linkage Agent ran anonlink multiple times to find matches on a combination of name, sex, DOB, and a variable fourth characteristic. These matches were then aggregated to produce the final linkages. The frequency of matches across each combination of characteristics can inform data quality.

Results: After QA, the percentage of linkages based on disproportionately many matches was reduced from 8% to <1%, and the percentage of linkages without a match on all sets of characteristics was reduced from 27% to 9%.

Stage 4: Demographic Concordance

Once the CODI database is populated with linked data, we queried for DOB and sex to assess concordance across partners. DOB concordance of 90% implies high quality matching (concordance was <60% for some Data Owners after PPRL run 1).

Results: PPRL run 1 revealed an issue with a Data Owner's internal data tables that was degrading matches. After the issue was fixed prior to run 2, concordance increased to >92% for matches between all Data Owners (Table 1).

Key: N links R2 / % change from R1 / N (%) matching DOB R1 / N (%) matching DOB R2 / % DOB match change (R1-R2)	Clinical Data Owner 1	Clinical Data Owner 2	Clinical Data Owner 3	Community Data Owner 1	Community Data Onwer 2
Clinical Data Owner 1	Unique Link IDs: R1: 118,388 R2: 118,307				
Clinical Data Owner 2	37,982 -8.2% 35,523 85.8% 36,906 97.2% +11.3%	Unique Link IDs: R1: 379,637 R2: 380,379			
Clinical Data Owner 3	12,424 -39.7% 11,609 56.3% 11,697 94.1% +37.8%	52,697 -25.1% 49,735 70.7% 50,815 96.4% +25.8%	Unique Link IDs: R1: 506,972 R2: 187,187		
Community Data Owner 1	1,030 +1.0% 884 86.7% 957 92.9% +6.2%	2,396 +4.1% 2,128 88.8% 2,260 94.3% +5.5%	1,102 -9.1% 992 81.8% 1,040 94.1% +12.5%	Unique Link IDs: R1: 6,948 R2: 6,917	
Community Data Onwer 2	400 -4.8% 384 91.4% 388 97.0% +5.6%	227 -11.2% 268 85.9% 272 98.2% +12.3%	51 -57.9% 57 47.1% 49 96.1% +49%	3 -50.0% 4 66.7% 3 100.0% +33.3%	Unique Link IDs: R1: 698 R2: 697

>20% increase in DOB concordance
>10% increase in DOB concordance
5-9.9% increase in DOB concordance
0.1-4.9% increase in DOB concordance
No increase in DOB concordance
Reduction in DOB concordance

Table 1: Demographic concordance results from the first (R1) and second (R2) PPRL runs across each two-site combination of Data Owners. The green highlighted cells at the bottom of each box show the percent change in DOB concordance between the two rounds, once all four QA checks, tool modifications, and data quality fixes had been implemented.

Explore the PPRL solution:

(http://github.com/mitre/data-owner-tools)

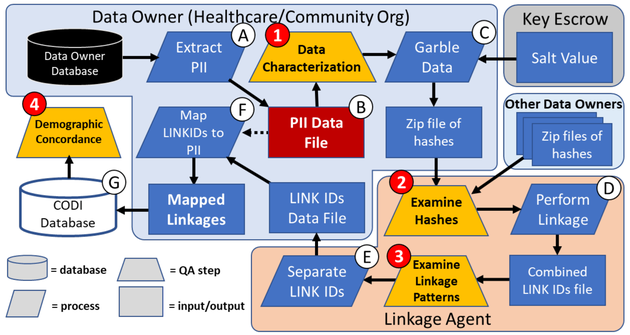

Figure 1: CODI PPRL Process flow. CODI Data Owners extract patient/participant data from their system (A) and create a file with PII needed for linkage (B). Data Owners then hash the data (C) using a salt value provided by a third-party Key Escrow. All Data Owners send their hashes to a Linkage Agent who creates deidentified LINK IDs (D) and separates by Data Owner (E). Data Owners then map LINK IDs back to PII (F) so that they can populate the CODI database with clinical or program data (G). The CODI team added data checks (1-4) to improve linkage quality.

Outcome: Modifications to anonlink-based PPRL tool

Based on QA findings, the CODI team updated the PPRL setup:

1. Removed use of parent email in matching
2. Removed dashes and "20" from the year digits of DOB
3. Implemented positional bigrams for DOB
4. Standardized zip codes to 5 digits

Conclusions

- To improve PPRL quality, each Data Owner must submit high quality data.
- Robust quality assurance techniques improved linkages in CODI by suggesting data quality improvements and PPRL tool configuration changes.
- As PPRL sees greater adoption as a health informatics tool, sound quality assurance measures are needed to ensure proper linkage.

University of Colorado | PUBLIC HEALTH INSTITUTE AT DENVER HEALTH | MITRE SOLVING PROBLEMS FOR A SAFER WORLD | CDC Centers for Disease Control and Prevention