Anonymous Record Linkage Between EPR and CDW-H: Toward Development of a Federated Genotype-Phenotype System

Dongqiuye Pu, M.S.[1,2], Stavros Garantziotis, M.D.[3], Javed Mostafa, Ph.D.[1,2]

[1]Laboratory of Applied Informatics Research, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA
[2]The North Carolina Translational & Clinical Science (TraCS) Institute, Chapel Hill, NC, USA
[3]Clinical Research Unit, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA

Abstract:

Environmental Polymorphisms Registry (EPR) is a large-scale phenotype-by-genotype registry developed by National Institute of Environmental Health Sciences to facilitate translational research. The link between personal identity and collected genomic data was preserved in EPR which creates opportunities for EPR to be linked to phenotype-rich databases, such as the Carolina Data Warehouse for Health (CDW-H) located at the University of North Carolina hospital system. CDW-H contains clinically-relevant data for patients who have been admitted to UNC healthcare system. To validate the feasibility of linking EPR with CDW-H, the number of matching records between the two databases had to be established. To that end, combinations of subjects' demographic identifiers from both databases were converted to anonymized hash codes, which were then matched to determine the number of overlapping records. Preliminary results showed that combination of last name, gender, data of birth and zip code would generate over 2,700 matches between the two databases.

Introduction & Background:

Maintaining patients' privacy while exchanging PHI (personal health information) across institutions is one of major challenges in building cross-site registries. In most cases, direct exchange of patients' information is considered much too risky and therefore is typically avoided. Exchange of anonymized information derived from subjects' personal information between agencies is considered appropriate in these circumstances and also conforms to the minimum necessary principle of the HIPAA privacy rule. In this study, we have performed research on anonymized record linkage between Carolina Data Warehouse for Health (CDW-H) and Environmental Polymorphisms Registry (EPR), serving as feasibility validation for future joint efforts.

EPR[1,2] is a large-scale phenotype-by-genotype registry developed by National Institute of Environmental Health Sciences (NIEHS) with the motivation of facilitating translational research of complex human diseases. It consists of over 15,000 individuals of diverse sex, age race and ethnicity. The EPR is a linked DNA biorepository, which differentiate itself from previous anonymous biorepository where individual identities were destroyed after data collection. This greatly increases its likely application along two dimensions: 1) the individual could be recalled for phenotyping depending on the specific genotype-phenotype relationship of interest; 2) EPR could be linked to phenotype-rich data sources to facilitate large-scale genotype-phenotype research could be conducted.

CDW-H is an enterprise-wide data warehouse developed by the University of North Carolina (UNC) Health Care System to meet the dual challenges of enhancement of quality of care and promote clinical research with UNC hospital patient population. It is a central repository containing clinical, research and administrative data ranging from billing and insurance to diagnosis and medication domains.

To establish the feasibility of joint studies between NIEHS and UNC using EPR and CDW-H data, the number of overlapping records has to be established. The EPR consists entirely of genomic and environmental data of over 15,000 subjects who have volunteered to participate in long term genetic research. CDW-H system consists of medical history of over two million patients registered at UNC hospitals.

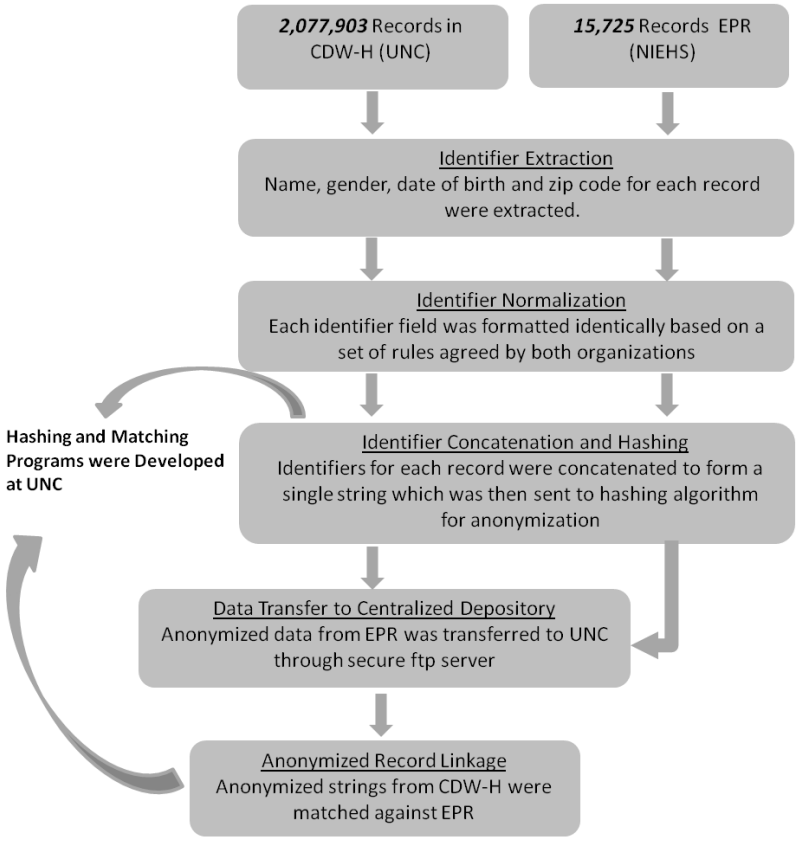

Figure 1. Project workflow for anonymized record linkage

Methods:

Obtaining Data Access Permission

To conform with research and data access regulations, IRB permission was obtained at UNC with NIEHS as supporting agency. Under the IRB permitted study protocol, research staff involved in this project were given full records of each respective database only for the purpose of generating hash codes which were to be shared across institutions. Data were to be extracted from respective databases by licensed administrative staff at each agency. Also, additional IRB stipulation required that the anonymized data be destroyed at the end of the study period.

Project Workflow

Figure 1 displays the coordination between the two agencies in performing this research. Note that, to ensure research staff will not have access to data from the other organization, the hashing application developed at UNC was delivered to NIEHS so that the hashing process could be performed separately at NIEHS. Hash codes from NIEHS were then delivered to a centralized repository at UNC through a secure ftp server. The anonymized record linkage step was performed at UNC and data was immediately destroyed upon completion.

Data Preparation

Both institutions agreed on the following fields to be extracted for data matching: last name, date of birth, gender and zip code. Only last names were included for name field because the registration of first name, prefix and middle name are not consistent across agencies as stated elsewhere[4].

Data extracted from each respective database were normalized and formatted to be represented in identical formats. Two levels of format normalization were conducted. First level was for data extraction. Because it involved institutions performing hashing task using the same computer program but at separate locations, it is important that input data be in identical format. Table 1 shows the data extraction format that was agreed upon by both parties.

Table 1. Data extraction format agreed by both parties

Last Name	Date of Birth	Gender	Zip Code
No special requirement	• Formatted as mm/dd/yyyy, slashes should be retained	• Fully spelled words, i.e., male or female	• 5-digit number

Second level of data formatting took place just before the hashing procedure. Table 2 shows the rules of normalization during this step.

Table 2. Data normalization rules. All white spaces were removed before normalization.

Last Name	Date of Birth	Gender	Zip Code
• To lower case • Remove special characters	• Remove slashes • Date format is mmddyyyy	• To lower case	• 5-digit number

Hashing

The normalized data were used as input into the hashing procedure. The normalization steps ensured the probability of two identical records being mapped to identical hash codes was relatively high. Each field was normalized separately in data preparation step and was concatenated with each other to generate different combinations, which were then fed into the hashing program to generate anonymized hash codes.

Complexity of hashing algorithm is constantly evolving to make it difficult for hash codes to be inversely engineered to original information. Selecting best hashing algorithm is the art of balancing security level and performance. In this case, we chose secure hashing algorithm version 2 (with 256 bit security level) which according to NIST standard is the most widely approved and adopted algorithm to date[5].

Record Linking

Since there are over two million health records in CDW-H and over 15,000 records in EPR, it would have been computationally expensive if we tried to match them directly. To decrease the running time, the anonymized hash codes were sorted for both datasets in increasing order and then compared according to the pseudo code illustrated in figure 2. To ensure the accuracy of the matching process, at least one of the datasets should be duplicate-free. We removed duplicate hash codes for EPR dataset since it had smaller cardinality.

```
Assume each list was sorted a priori in increasing order and hashcode.EPR is duplicate free
Set the counter to 0
While neither list has reached the end
        if hashcode.CDWH == hashcode.EPR
                Move to the next hashcode on the list for both CDWH and EPR
                Add one to the counter
        elseif hashcode.CDWH < hashcode.ERP
                Move to the next hashcode on the list for CDWH only
        else
                Move to the next hashcode on the list for EPR only
```

Figure 2. Pseudo code for anonymized record matching.

Results & Discussion

Table 3 shows the matching results for four combinations of personal identifiers. Not surprisingly, gender doesn't have much discriminative power since it only has two possible values. Based on previous research[3], last name and date of birth fields can have high discriminative power for uniquely identifying an individual. Therefore, it is reasonable to believe that the number of records in common between CDW-H and EPR would be at least 2,700. In terms of sample size, this result implies that future research based on joint dataset between CDW-H and EPR would have the potential of generating results of statistical significance. The presented results in this report are preliminary and validation of the matching result would require full access to patients' personal information. A second round of IRB application is in progress and validation would be performed in the near future.

Table 3. Anonymized record linking results using 4 combinations of identifiers

Combinations	Unique Records in EPR	Matches
Last name + DOB + zip + Gender	15705	2746
Last name + DOB + zip	15705	2754
Last name + DOB	15691	4071
Last name + zip	14143	7081

Conclusion:

Due to institutional regulations, sometimes a case for sharing of data has to be established (i.e., a feasibility analysis) without actually transmitting identifiable record elements across institutional boundaries. The method we developed is likely to be helpful to studies that involve cross-institutional sharing of data in a secure way for feasibility analysis and beyond. Approval will be sought on an updated institutional review board application to access all the key data elements associated with subjects for the validation step. We are fairly confident that the method developed in this study, upon execution of additional validation steps, would provide a sound foundation for engaging in research effort between the two institutions currently exploring for a collaborative approach for conducting genotype-phenotype linkage projects.

Acknowledgement:

The assistance from the TraCS institute (grant 5-31282) and the Health and Human Services Office of the National Coordinator for Health Information Technology (grant 5-43726) are gratefully acknowledged. We also thank Chris Wachtstetter (NIEHS) for his generous support in sharing data with us and David Hurley from the TraCS institute for his effort as a project coordinator.

References

1. Chulada, P. C., Vahdat, H. L., Sharp, R. R., DeLozier, T. C., Watkins, P. B., Pusek, S. N., & Blackshear, P. J. (2008). The Environmental Polymorphisms Registry: a DNA resource to study genetic susceptibility loci. *Human genetics*, *123*(2), 207-14.

2. Chulada, P. C., Vainorius, E., Garantziotis, S., Burch, L. H., Blackshear, P. J., & Zeldin, D. C. (2011). The Environmental Polymorphism Registry : A Unique Resource that Facilitates Translational Research of Environmental Disease, *1*(11), 1523-1527.

3. Weber, S. C., Lowe, H., Das, A., & Ferris, T. (2012). A simple heuristic for blindfolded record linkage. *Journal of the American Medical Informatics Association : JAMIA*, 1-6.

4. Arellano, M. G., & Weber, G. I. (1998). Issues in identification and linkage of patient records across an integrated delivery system. *Journal of healthcare information management : JHIM*, *12*(3), 43-52.

5. Dang, Q. (2009). Recommendation for Applications Using Approved Hash Algorithms. *NIST Special Publication 800-107*.