Beyond Safe Harbor: Risk of Exposing Location in De-Identified Clinical Data

Alfred Jerrod Anzalone[1], Carol R. Geary[2], and James C. McClay[3]

1. Department of Neurological Sciences, University of Nebraska Medical Center (UNMC), Omaha, NE 68198; 2. Department of Pathology & Microbiology, UNMC, Omaha, NE 68198; 3. Department of Emergency Medicine, UNMC, Omaha, NE 68198

Introduction

De-identified clinical research data warehouses (DI-CRDW) are integrated repositories of data extracted from from numerous source system, including electronic health records, with identifiers removed. According to the HIPAA Privacy Rule[1], de-identification can be achieved through two methods: Expert Determination and Safe Harbor.[2]

The Safe Harbor rule requires strict exclusion of the 18 HIPAA identifiers, including non-aggregated location units smaller than the state.[2] It is commonplace to enhance DI-CRDW data through linkage with external data sources to create a more comprehensive and accurate patient profile. These external data sources include location-based indices, such as the Rural-Urban Commuting Area (RUCA) Codes.

Inherent Problem

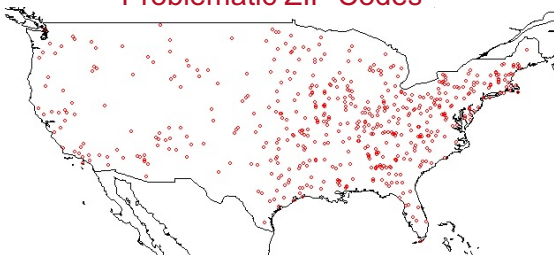
DI-CRDW each operate under their own Institutional Review Board discretion, with the highest priority being to maintain patient privacy. The Common Data Models most deployed in these environments (OMOP, PCORnet, and ACT) have varying standards for available location-based information. Most institutions provide investigators with access to 3-digit ZIP Codes and state, which is compliant with HIPAA Privacy Rules. As more research relies on de-identified data (both for security and expediency) for clinical and translational research, additional SDOH measures are being supplied to support these needs. While that practice does not directly expose PHI, providing multiple location-based measures does.

RUCA Codes

Rural-Urban Commuting Area (RUCA) codes are a classification scheme for measuring population density, urbanization, and commuting trends.[3]

Primary RUCA Code	Classification description
1	Metropolitan area core: primary flow within an urbanized area (UA)
2	Metropolitan area high commuting: primary flow 30% or more to a UA
3	Metropolitan area low commuting: primary flow 10% to 30% to a UA
4	Micropolitan area core: primary flow within an urban cluster of 10,000 to 49,999 (large UC)
5	Micropolitan high commuting: primary flow 30% or more to a large UC
6	Micropolitan low commuting: primary flow 10% to 30% to a large UC
7	Small town core: primary flow within an urban cluster of 2,500 to 9,999 (small UC)
8	Small town high commuting: primary flow 30% or more to a small UC
9	Small town low commuting: primary flow 10% to 30% to a small UC
10	Rural areas: primary flow to a tract outside a UA or UC

Problematic ZIP Codes



This map represents the more than 500 5-digit ZIP Codes that would be re-identifiable using the approach for supplying investigators with Primary RUCA Code as described in this project in the presence of a 3-digit ZIP Code, which is common practice in DI-CDRW.

Re-Identification Matrix Using De-Identified Data Elements

3-Digit ZIP Code	5-Digit ZIP Code	State	Primary RUCA Code
…	…	…	…
013	01351	MA	4
013	01360	MA	5
013	01362	MA	2
013	01366	MA	5
013	01364	MA	4
013	01366	MA	3
013	01367	MA	6
013	01368	MA	1
013	01370	MA	10
013	01373	MA	2

Two-point Vector to Reidentify 5-Digit ZIP Codes

Here we demonstrate the potential risk of re-identification when using a HIPAA Safe Harbor compliant method of aggregating ZIP Codes into 3-digit clusters of more than 20,000 people.[3] Within a single aggregate 3-digit ZIP Code cluster (013 in Massachusetts) we can reidentify two 5-digit ZIP Codes. Institutions need to demonstrate caution when releasing multiple, location-based measures to researchers for de-identified studies.

Solution for Rural Health Studies

We have a large contingency of rural health investigators, so identifying validated measures of rurality is of high interest to support local research. To enable this while maintaining patient privacy, we deploy an aggregated measure of rurality above the level of Primary RUCA Code (urban, urban-adjacent rural, and nonurban-adjacent rural) while removing 3-Digit ZIP Codes from data released to investigators.

If an investigator requests additional, location-based measures for an individual study we perform an individual vulnerability assessment using a standardized protocol to support these requirements while ensuring no PHI re-identification is possible.

Conclusions

While research is important, the preservation of patient privacy needs to be built into institutional models for expanding research using real-world data. As demonstrated here, this is not an isolated problem. Use of location-based indices in de-identified research pose inherent privacy concerns for all institutions. While researchers have a vested interest in preserving patient privacy, system architecture needs to reflect the reality that re-identification is possible and prevent that possibility through considerate data release practices.

References

1. U.S. Dept. of Health and Human Services. Standards for privacy of individually identifiable health information, Final Rule. Federal Registrar 2002; 45 CFR, Parts 160–4.
2. Health Information Technology: Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology, 45 CFR part 170 (2014).
3. U.S. Department of Agriculture, Economic Research Service. Rural-Urban Commuting Area Codes. https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/

The project described utilizes the UNMC Clinical Research Analytics Environment (CRANE). CRANE is supported funding from the National Institute of General Medical Sciences, U54 GM115458 and the Patient Centered Outcomes Research Institute, PCORI CDRN-1306-04631. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or PCORI.