10.1017/cts.2024.521

Compliance Considerations in the Geo-enrichment of an EHR Data Warehouse with Social & Environmental Determinants of Health

Maryam Abdallah*[1, 2], Neil Bahroos*[1, 2], Praveen Angyan[1, 2], Beau MacDonald[3], Camilla Catignas[2], Daniella Garofalo[1, 2], Amy Chuang[1, 2], Hakob Abajian[1, 2], John Wilson[3]

[1]Keck Medicine of USC, University of Southern California, Los Angeles, CA
[2]Southern California Clinical and Translational Science Institute, Los Angeles CA
[3]Spatial Sciences Institute, University of Southern California, Los Angeles CA

*Dual first authors

Corresponding Author: Neil Bahroos, neil.bahroos@med.usc.edu, 2250 Alcazar St, Suite 134B, Los Angeles CA 90033, T: 773.420.8400

Conflicts of Interest: None to disclose.

Abstract

Social and environmental determinants of health (SEDoH) are crucial for achieving a holistic understanding of patient health. In fact, geographic factors may have more influence on health outcomes than patients’ genetics. Integrating SEDoH into the electronic health record (EHR), however, poses notable technical and compliance-related challenges. We evaluated barriers to the integration of SEDoH in the EHR and developed a privacy-preserving strategy to mitigate risk of PHI exposure. Using coded identifiers for patient addresses, the strategy evaluates an alternative approach to ensure efficient, secure geocoding of data while preserving privacy throughout the data enrichment processes from numerous SEDoH data sources.

Introduction

Electronic health records (EHRs) are inherently limited in providing valuable information for social and environmental determinants of health (SEDoH), though such data is critical for comprehensive patient history and precision medicine initiatives. An individual’s zip code may be linked more to influencing health outcomes or issues than their genetics[1]. Improving our ability to capture SEDoH can bridge the gap in health disparities and improve outcomes for marginalized populations[2, 3]. Although some healthcare organizations have integrated patient-reported social determinants forms in their EHRs, data is often sparse.[4] While publicly accessible neighborhood-level SEDoH data exists, seamlessly integrating this information in the patient EHR is complex and presents compliance-related challenges.

Collecting SEDoH data begins with geocoding, or translating, an address or Census tract to its latitudinal and longitudinal coordinates.[5] Geocoded addresses are then geo-enriched with SEDoH data retrievable through extensive and publicly available datasets, but accurately linking the variables to patient data requires disclosing individual geographic identifiers. Once addresses are geocoded, they can be linked to corresponding neighborhood and community-level SEDoH variables derived from a multitude of datasets. Datasets are available via publicly hosted files, public APIs, and commercial APIs that are behind a paywall. These datasets often utilize geolocations defined by the US Census Bureau to report on various SEDoH,[4] and some initiatives have combined multiple data sources to create composite SEDoH indices.[6] Geographic identifiers beyond the first three digits of some zip codes are considered protected health information (PHI). Use and disclosure of PHI beyond the scope of providing patient care is restricted based on the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule.[7]

Geographic information system (GIS) software can geocode and enhance addresses with geospatial data.[1] The SEnDAE (Social and Environmental Determinants Address Enhancement) toolkit[8] employs an innovative strategy whereby an intermediate server separates the requesting health provider organization’s (HPO) IP address when transmitting deidentified patient addresses to a cloud-based geocoding service.[9] While this significantly reduces the risk of accidental PHI disclosure, further safeguards can be put in place by carrying out in-house geocoding within a self-contained GIS application. Conservative arguments trust that a self-contained approach may

https://doi.org/10.1017/cts.2024.521 Published online by Cambridge University Press

be the only HIPAA-compliant method to protect PHI during external data transfer.[10] This paper explores these compliance challenges and offers recommendations for integrating SEDoH with EHR data while minimizing risk.

Materials & Methods

To retrieve SEDoH variables for any individual patient, their home address must be translated, or geocoded, from its standard format into the specific latitudinal and longitudinal coordinates. The geocoded location can then be linked, or geo-enriched, with its corresponding social and environmental data points. This process can be completed either by sending location data to a web service or purchasing a local geocode database. The former option simplifies the process of geocoding, as there is no server set up, installation or maintenance required. Several such services exist, including a free service provided by the US Census Bureau.[4] However, utilizing web services involves risks associated with the disclosure of PHI to a remote server external to one's institution. The alternate option is to instantiate a local geocode database and service, which can be purchased from several companies. While this requires the additional steps of setup and maintenance of the server, and keeping the software up to date, it eliminates the need for external disclosure of PHI.

The current project purchased Esri's ArcGIS Pro 3.X with the Business Analyst Extension. The software was installed locally on a secure server created specifically for geocoding and geo-enrichment purposes. We designed a workflow, illustrated in Figure 1, to ensure the local GIS enhancement server contained only the minimum-necessary PHI required for geocoding and geo-enrichment. First, a randomized, deidentified ID is assigned to each patient, yielding a code key that is stored within our HIPAA-compliant Research Enterprise Data Warehouse (EDW). Second, deidentified patient IDs and their corresponding geographical addresses are loaded onto the local GIS server. Third, addresses are geocoded and subsequently assigned a randomized, deidentified address ID, yielding a code key that is stored within the local GIS server. Finally, deidentified address IDs are linked to corresponding Census Tract IDs and exported from the server. Census Tracts do not contain fixed individual geographic identifiers and are considered less specific geographic subdivisions than latitude and longitude, or even Census Block groups,[5] further minimizing PHI risk.

https://doi.org/10.1017/cts.2024.521 Published online by Cambridge University Press

Results

All patients (n=554,562) within the university's EHR who opted-in to participating in research and had valid addresses were included in this project. Full addresses and deidentified patient IDs were loaded onto the local GIS server. All current and previous patient addresses were included in the data, such that some patient IDs corresponded to multiple addresses. All addresses were geocoded and assigned a randomized address ID.

Datasets were selected from six data sources (Table 1). These sources were determined to contain valuable social and environmental variables and had the necessary geographic identifiers needed for linkage. All datasets were downloaded and stored locally, allowing geo-enrichment efforts to be completed on the local GIS server. Five of the six datasets were directly downloaded from the source and stored on the local server, which took about 10 seconds per dataset. The remaining dataset, the US Census American Community Survey (ACS), was only accessible through API calls and could not be directly downloaded. The ACS API was called with broad arguments to collect data for all addresses across the entire United States. ACS data was downloaded once a month, although it can be refreshed at any frequency as feasible for an institution. A free API key was registered, which the US Census requires for IP addresses that exceed 500 daily queries. Loading all ACS data on the local server via API calls took 8.28 minutes. While this meant that live data was not being obtained through the API, it allowed us to maintain the same level of privacy as downloadable, locally stored datasets.

Once all addresses were geo-enriched with corresponding variables from all datasets, deidentified patient IDs and linked geospatial data were exported from the local GIS server and loaded back into the Research EDW. Patient IDs were reidentified using the code key maintained within the Research EDW, and the newly geo-enriched patient data was integrated with our existing EHR data warehouse. The data was integrated with our OMOP (Observational Medical Outcomes Partnership) Common Data Model (CDM)[11] by modeling SEDoH data on our local concepts for extending the OMOP CDM and creating measurement tables. We also converted the SEDoH data from OMOP into observations in our local instance of i2b2[12], a self-service cohort discovery tool, using the SEnDAE ontology extension framework.[8] These efforts allowed for geo-enriched EHR data to be readily available for researchers and clinicians to query and extract. The toolkit and OMOP CDM are publicly available at https://github.com/scctsi/gis-toolkit.

https://doi.org/10.1017/cts.2024.521 Published online by Cambridge University Press

Discussion

This project was successful in geocoding and geo-enriching an EHR data warehouse in a secure, compliant manner. Utilizing the Esri database provided a minimal-cost solution to support this project. While local installation of the Esri database prevented external PHI transfer, there are limitations to this method. The setup, installation, and maintenance of such a server can be a burden to organizations. Due to the time-consuming nature of the in-house geocoding process and quality validation, our organization currently completes geocoding and geo-enrichment on an ad-hoc basis as a consultation service for research projects and once every two years for our entire patient population.

To streamline the geocoding process, we plan to transition to a secure process that utilizes APIs to an external service for geocoding. This has been reviewed and approved by our university's Compliance Department. The process involves setting up a server with a random hostname specifically for geocoding patient addresses. When an API call is made to an external geocoding service, the service may store patient addresses and referrer hostnames for auditing purposes, posing additional risk of patient reidentification. Utilizing a random hostname anonymizes the call such that our organization cannot be identified and linked to the patient addresses sent. This process takes an average of 0.27 seconds per address to geocode a sample set of addresses. Once this new process is fully implemented, we will geocode and geo-enrich our EHR data once a week.

Ethical issues remain inherent with the use of patient geographical data, and geolocation data is an element of PHI when linked to patients or HPOs.[13] We recommend becoming familiar with decisions associated with the geocoding process,[14] variability of positional accuracy, geocoding methods,[15] the use of different geographical units when matching address[16], as well as published practices and protocols for internet geolocation.[9, 10, 17] Institutional interpretation of HIPAA and privacy policies varies, and patient geolocation approaches should be evaluated by appropriate officials prior to implementation. [9]

Employing secure strategies to geocoding EHR data allows for the benefits of geo-enrichment of patient data while minimizing privacy and security risks. Once securely geocoded, data can be safely enriched with any place-based measures to study the impacts of SEDoH and design and prescribe interventions to yield better health outcomes. We have outlined a framework for

https://doi.org/10.1017/cts.2024.521 Published online by Cambridge University Press

secure, compliant geo-enrichment of patient data that can be adapted and implemented at other institutions. Increasing consideration of SEDoH in both research and clinical practice can ultimately reduce health inequities and improve outcomes for marginalized populations.

Acknowledgements: This effort was supported in part by the Southern California Clinical and Translational Science Institute (SC-CTSI) (UL1TR001855) and USC Keck's Health Data Innovation Program (HDIP). We thank Jane Choi and Masi Thomason for help and support with copy-editing the article.

Author Contributions:
Conceptualization: MA, NB, PA, BM, DG, JW; Writing - Original Draft: MA, NB, PA, CC, DG; Writing - Review & Editing: MA, BM, CC, DG, JW; Methodology: NB, BM, AC; Resources: MA, NB, CC; Software: PA, BM, HA; Validation: BM, HA; Data Curation: BM, AC; Formal Analysis: BM; Visualization: MA; Investigation, Supervision, Project Administration & Funding Acquisition: NB

Financial Support: This work was supported by grants UL1TR001855 from the National Center for Advancing Translational Science (NCATS) of the U.S. National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Disclosures: We have nothing to disclose.

https://doi.org/10.1017/cts.2024.521 Published online by Cambridge University Press

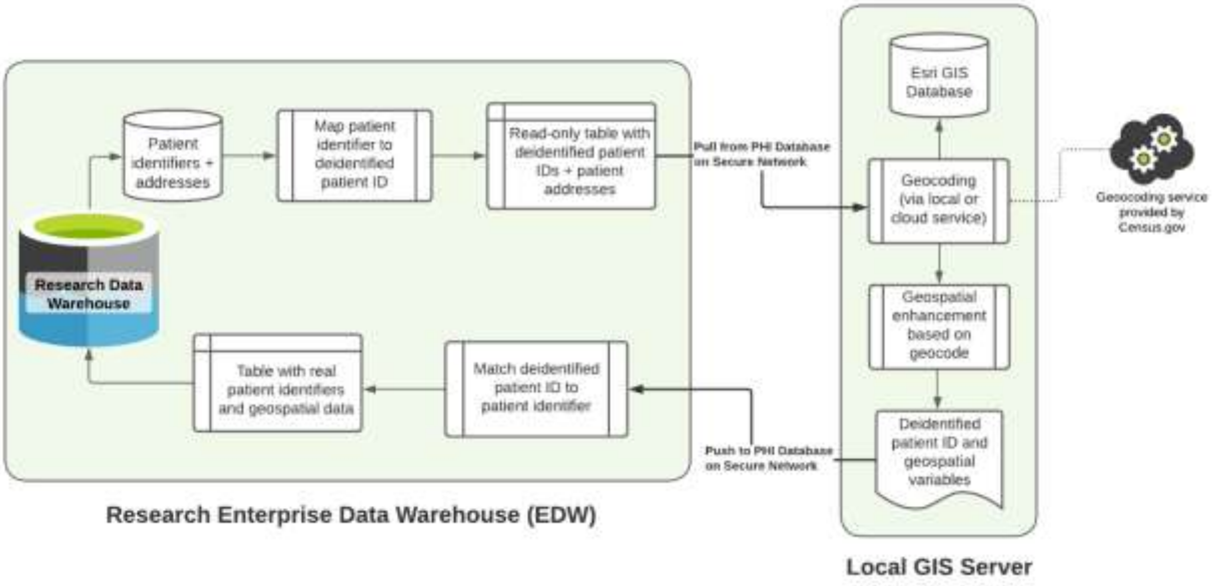
Patient identifiers + addresses
Map patient identifier to deidentified patient ID
Read-only table with deidentified patient IDs + patient addresses
Pull from PHI Database on Secure Network
Esri GIS Database
Geocoding (via local or cloud service)
Geocoding service provided by Census.gov
Research Data Warehouse
Geospatial enhancement based on geocode
Table with real patient identifiers and geospatial data
Match deidentified patient ID to patient identifier
Push to PHI Database on Secure Network
Deidentified patient ID and geospatial variables
Research Enterprise Data Warehouse (EDW)
Local GIS Server

Figure 1. Diagram of workflow utilized for geocoding and geo-enrichment.

Table 1. Sample data sources included in geo-enrichment of patient addresses.

Data Source	**Access Methodology**	**Links**
US Census American Community Survey, 5-year detailed tables (B), subject tables (S), and data profiles (DP)	API	https://www.census.gov/data/developers/data-sets/acs-5year.2018.html
Office of Environmental Health Hazard Assessment (OEHHA), on behalf of the California Environmental Protection Agency (CalEPA)	File	https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-40
Centers for Disease Control and Prevention. Agency for Toxic Substances and Disease Registry	File	https://www.atsdr.cdc.gov/placeandhealth/svi/
NASA Air Quality files	File	https://sedac.ciesin.columbia.edu/data/collection/aqdh/sets/browse
Southern California Environmental Health Sciences Center	File	https://scehsc.usc.edu/
US Department of Agriculture, Economic Research Service	File	https://www.ers.usda.gov/data-products/food-access-research-atlas/

https://doi.org/10.1017/cts.2024.521 Published online by Cambridge University Press

References

1. Harris DR. Geographic Information Systems as Data Sharing Infrastructure for Clinical Data Warehouses. *Journal of the Society for Clinical Data Management. 2023;3(4).* doi.org/10.47912/jscdm.240

2. Ford-Gilboe, M, Wathen, C.N, Varcoe, C, et al. How equity-oriented health care affects health: Key mechanisms and implications for Primary Health Care Practice and policy. *Milbank Q*. 2018;96(4):635-671. doi:10.1111/1468-0009.12349

3. Hatef E, Searle KM, Predmore Z, et al. The Impact of Social Determinants of Health on Hospitalization in the Veterans Health Administration. *Am J Prev Med*. 2019;56(6):811-818. doi:10.1016/j.amepre.2018.12.012

4. Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. *J Am Med Inform Assoc*. 2021;29(1):187-196. doi:10.1093/jamia/ocab199

5. Rana MKZ, Song X, Islam H, et al. Enrichment of a Data Lake to Support Population Health Outcomes Studies Using Social Determinants Linked EHR Data. *AMIA Jt Summits Transl Sci Proc*. 2023;2023:448-457. Published 2023 Jun 16. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10283101/

6. Bazemore AW, Cottrell EK, Gold R, et al. "Community vital signs": incorporating geocoded social determinants into electronic records to promote patient and population health. *J Am Med Inform Assoc*. 2016;23(2):407-412. doi:10.1093/jamia/ocv088

7. Office for Civil Rights (OCR). Methods for de-identification of PHI. HHS.gov. February 22, 2023. Accessed December 14, 2023. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard.

8. Kingsbury P, Abajian H, Abajian M, et al. SEnDAE: A resource for expanding research into social and environmental determinants of health. *Comput Methods Programs Biomed*. 2023;238:107542. doi:10.1016/j.cmpb.2023.107542

9. Rivera, Brian and Mark A. Hoffman. "Technical Strategies for Real-time Geocoding in Healthcare." *2018 IEEE International Smart Cities Conference (ISC2)* (2018): 1-5. doi: 10.1109/ISC2.2018.8656931

https://doi.org/10.1017/cts.2024.521 Published online by Cambridge University Press

10. Rundle AG, Bader MDM, Mooney SJ. The Disclosure of Personally Identifiable Information in Studies of Neighborhood Contexts and Patient Outcomes. *J Med Internet Res*. 2022;24(3):e30619. Published 2022 Mar 17. doi:10.2196/30619

11. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform*. 2015;216:574-578

12. *Informatics for Integrating Biology to the Bedside, Partners Healthcare Systems*. Version v1.7.12a. i2b2 TranSMART Foundation; 2023. www.i2b2.org

13. Goodchild M, Appelbaum r, Crampton J. *A White Paper on Locational Information and the Public Interest*. American Association of Geographers; 2022:10-40. Accessed December 15, 2023. doi:10.14433/2017.0113

14. Goldberg D, Wilson J, Knoblock C. Exploring the USEOF gazetteers and geocoders for the analysis and interpretation of a dynamically changing world . In: *Understanding Dynamics of Geographic Domains*. CRC Press; 2008:51-74. Accessed December 15, 2023. 10.1201/9781420060355.pt2

15. Jones RR, DellaValle CT, Flory AR, et al. Accuracy of residential geocoding in the Agricultural Health Study. *Int J Health Geogr*. 2014;13(37). doi:10.1186/1476-072X-13-37

16. Zandbergen PA. A comparison of address point, parcel and street geocoding techniques. *Comput Environ Urban Syst*. 2008;32(3):214-232. doi:10.1016/j.compenvurbsys.2007.11.006

17. Bader MD, Mooney SJ, Rundle AG. Protecting Personally Identifiable Information When Using Online Geographic Tools for Public Health Research. *Am J Public Health*. 2016;106(2):206-208. doi:10.2105/AJPH.2015.302951

https://doi.org/10.1017/cts.2024.521 Published online by Cambridge University Press