FEDERATED LEARNING OVER ELECTRONIC HEALTH RECORD DATA: a Proposed Implementation

A RENCI Technical Report
TR-24-01

Karamarie Fecho
Renaissance Computing Institute, Chapel Hill, North Carolina, USA; Copperline Professional Solutions, LLC, Pittsboro, North Carolina, USA

Jackson Callaghan
Scripps Research Institute, La Jolla, California, USA

Gwênlyn Glusman
Institute for Systems Biology, Seattle, Washington, USA

Jennifer Hadlock
Institute for Systems Biology, Seattle, Washington, USA

Ashok Krishnamurthy
Renaissance Computing Institute, Chapel Hill, North Carolina, USA

Casey Ta
Columbia University Irving Medical Center, New York, New York, USA

Max Wang
CoVar, LLC, Durham, North Carolina, USA

Corresponding author: Karamarie Fecho, PhD, c/o Renaissance Computing Institute, 100 Europa Drive, Suite 540, Chapel Hill, North Carolina, USA 27517
Email: kfecho@copperlineprofessionalsolutions.com, kfecho@renci.org

renci

www.renci.org

RENCI Technical Report

Federated Learning over Electronic Health Record Data: a Proposed Implementation

Karamarie Fecho[1,2], Jackson Callaghan[3], Gwênlyn Glusman[4], Jennifer Hadlock[4], Ashok Krishnamurthy[1], Casey Ta[5], Max Wang[6]

[1]Renaissance Computing Institute, Chapel Hill, North Carolina, USA; [2]Copperline Professional Solutions, LLC, Pittsboro, North Carolina, USA; [3]Scripps Research Institute, La Jolla, California, USA; [4]Institute for Systems Biology, Seattle, Washington, USA; [5]Columbia University Irving Medical Center, New York, New York, USA; [6]CoVar, LLC, Durham, North Carolina, USA

Summary

"Federated learning" is a relatively new concept that has emerged from the fields of machine learning and networking, in which centralized machine learning models are applied to private decentralized data. As part of the Biomedical Data Translator ("Translator") program, we have developed three regulatory-compliant, open-source services for exposing insights derived from the electronic health records of three healthcare systems. Herein, we propose a two-phased approach for extending our existing services to support federated learning. In the first phase, we propose a proof-of-concept demonstration, leveraging the current Translator architecture and a new Translator service termed BioPack, which comprises a general workflow manager, a centralized server, and a subgraph retrieval service. BioPack will be used to run sequential queries across our three Translator clinical endpoints, thus mimicking federated machine learning. Those queries will take the form of a simple mathematical algorithm that performs a calculation across the results that are returned by each of the three Translator clinical endpoints for each sequential query. In the second phase, we propose to extend the phase one effort, leveraging a secure version of BioPack or a separate secure server for centralized queries, but directly targeting the private clinical databases that support the Translator clinical endpoints. We will prioritize privacy and security during both phases of the work. Collectively, our proposal will allow us to generate new clinical insights for contribution back to our institutions and the Translator program.

Federated Learning

The concept of "federated learning" emerged from the fields of machine learning and networking, with the term being formally introduced by McMahon and Ramage at Google in 2017[1]. The idea was driven by the rapid rise in mobile devices and the need to balance privacy concerns with the desire to leverage the huge amount of data stored on individual devices, rather than in a centralized local server or within a cloud environment[2,3]. With federated learning, a shared centralized machine learning model is applied to private decentralized data, without requiring centralization of the data, thus supporting learning and model improvement from multiple data sources, while allowing data owners to remain in control of their data. Data owners apply shared machine learning models to their data within their own secure local environment to train the models. As those models become updated, the updates to model weights and/or parameters, but not the data used to generate them, get pushed to an agreed-upon environment, which could be cloud-based or traditional server-based, and the process continues until model optimization. The net benefit for engaged

parties is to apply the learned federated knowledge to increase the value of their independent data assets, without compromising the security of those assets or losing ownership of them.

The general premise underlying federated learning is that data are valuable assets that must be protected. Foremost, data generators, including the owners of mobile devices and patients, value their personal data and are rightfully concerned with privacy, as should be researchers, institutions, and any secondary users of those data. In addition, the institutions that hold these data assets, including the financial, insurance, and healthcare industries[4,5], view their collected data as financially valuable assets and are understandably reluctant to share the data, at least not without a net gain on their part.

Federated learning offers a solution to these challenges, one that has several key advantages over traditional centralized approaches to the application of machine learning models[6]. In addition to allowing data owners control over their data assets, the approach supports multi-institutional collaboration, while preserving privacy and security, by separating the data from the models and allowing the models to come to the data rather than the other way around. Second, because federated learning leverages local computational resources, it is inherently more economical and efficient than centralized learning. Third, the approach is scalable, especially when attempting to apply machine learning models to very large datasets. Related, when numerous institutions are willing to participate in federated learning, with security and privacy concerns addressed, the trained models are often more generalizable than centralized models due to larger and more diverse training data.

Nonetheless, federated learning does not alleviate all security or privacy concerns, but rather shares certain security and privacy issues with centralized approaches. Additionally, there are distinct disadvantages with federated learning over traditional approaches. First, at the heart of federated learning is a foundation of trust, which itself builds upon a foundation of transparency and accountability. Indeed, all parties must be transparent and accountable when engaged in federated learning, and they must abide by agreed-upon terms of engagement. They also must acquire upfront approvals, including any necessary regulatory approvals such as those required by Institutional Review Boards[7] and any institutional approvals that may apply to the engaged parties. Approvals also are required for model and parameter updates, although those approvals need not necessarily involve regulatory bodies, but rather can take place via mutual software platforms such as GitHub, if that arrangement has been included and agreed upon in the rules of engagement. Second, time lags and delays, as well as overhead costs, may present challenges with federated learning, as it often takes time to update model weights and parameters with each round of training. This point is worth emphasizing, as federated learning carries inherent dependencies on external parties and their workflow timelines. Third, security breaches and data leakage are always possible, even when careful and considerate security protections have been put in place. Malicious actors thrive on opportunities to introduce security breaches, even when there is no financial or other reward[8]. In addition, human error can be a contributing factor to accidental security breaches[9]. Finally, federated learning may introduce issues related to model quality, particularly if participating institutions have not adopted the same quality control measures. Indeed, mathematically rigorous approaches to support differential privacy and the sharing of aggregate data often compromise data quality and usability[10].

Recognizing the advantages and disadvantages of federated learning, we propose a multi-phase framework and approach for applying federated learning to the clinical data systems available at our institutions, namely, the University of North Carolina at Chapel Hill (UNC), Columbia University Irving Medical Center, and the Institute for Systems Biology. Our proposal leverages our ongoing collaborative work as members of the Biomedical Data Translator ("Translator") Consortium, funded by the National Center for Advancing Translational Sciences[11,12]. Herein, we briefly describe the Translator program and the clinical data services that we have developed as a part of that program. We then provide an overview of our approach toward federated learning over clinical data. We close with next steps and final remarks.

The Integrated Clinical and Environmental Exposures Service (ICEES), Columbia Open Health Data (COHD) Service, and Multiomics Clinical Connections (MCC) Service

The Translator system is an open-source, knowledge graph (KG)–based system that aims to support the integration and cross-query analysis of hundreds of open clinical and biomedical data sources, applying advanced reasoning algorithms to derive new insights into human diseases and potential therapeutics. The overall goal of Translator is to shift the definition of disease from a symptom-based classification to a mechanistic-based classification[11,12]. To achieve data integration across the disparate data sources, Biolink Model[13] is leveraged for data representation, semantic harmonization, and identifier resolution. To date, the Translator system has integrated over 300 disparate data sources to derive clinically meaningful insights into asthma, multiple sclerosis, cyclic vomiting syndrome, Fanconi anemia, Ehlers-Danlos syndrome, and many other diseases and syndromes[14].

The fundamental tenets of the Translator Consortium include (1) a focus on clinical insights derived from clinical data such as electronic health record (EHR) data and (2) open team science and open-source software development[14]. The focus on clinical data differentiates the Translator system from other biomedical KG-based systems such as Causaly[15,16] and Elsevier's EmBiology[17]. However, that same focus conflicts with the focus on open team science and open-source software development. To address these challenges, we have developed regulatory-compliant approaches for openly exposing and querying knowledge derived from EHR data that have been represented in a manner that preserves patient privacy and minimizes security risks[18].

For example, ICEES openly exposes EHR data from UNC Health that have been integrated with a variety of publicly available sources of environmental exposures data such as airborne pollutant exposures and socioeconomic exposures[19]. The data are stripped of protected health information via the HIPAA Safe Harbor method[20,21] before being exposed in semi-aggregated form via an open application programming interface (openAPI)[22]. ICEES is cohort-specific in design and supports dynamic cohort creation and the application of basic statistical analyses such as Chi Square analysis and Fisher's Exact Test. COHD exposes patient counts and prevalence estimates for patient demographics, diagnoses, medications, and procedures, and the co-occurrences between them[23]. The EHR results are derived from Columbia University Irving Medical Center and include data on ~5M patients. The Open Health Data @ Carolina Service provides an instantiation of COHD but at UNC Health and includes UNC Health EHR data on ~6M patients[24]. Finally, the MCC Service provides insights on population-level risk factors for several chronic medical conditions using research insights derived from over 30 million EHR records from Providence Health Systems and Affiliates[25]. We note that COHD, Open Health Data @ Carolina, and MCC Service results are derived from the OHDSI-compatible databases at each institution[26], using OMOP as the common data model (CDM)[27]. This offers a tremendous opportunity, as the use of OMOP as a CDM, coupled with the adoption of Biolink Model[13], supports interoperability of the EHR data available at each institution and harmonizes the application of machine learning models across institutions via federated learning.

We have collectively queried these open clinical knowledge sources to rapidly generate insights into asthma and other diseases. For instance, we cross-queried ICEES and COHD to generate insights into the relationship between sex, obesity, diabetes, exposure to particulate matter, and asthma[28]. We've additionally queried the Translator clinical knowledge sources to rapidly derive insights into drug-induced liver injury, coronavirus infection, psoriatic arthritis, and other diseases[29]. Moreover, we've deployed demonstration instances of ICEES to support secure multiparty computation (SMC), which is an approach for secure data sharing that is relevant to federated learning[30]. In our demonstration SMC project, we applied SMC to securely calculate cross-institutional counts of theoretical patients with rare disease, the goal being to determine if the collective sample size was sufficient to support a multi-institutional, statistically valid, research study on rare disease[31]. Here, we propose to move beyond open queries of Translator clinical knowledge sources to the application of machine learning models in the context of federated learning.

Federated Learning Proposal: Implementation Plan

We envision a two-phase approach for implementation of federated learning using COHD, Open Health Data @ Carolina, and MCC Service (Figure 1). (Note that we will not include ICEES in this effort, as that service does not share OMOP as its CDM. However, we will leverage our prior SMC work with ICEES and the lessons learned under that effort.)

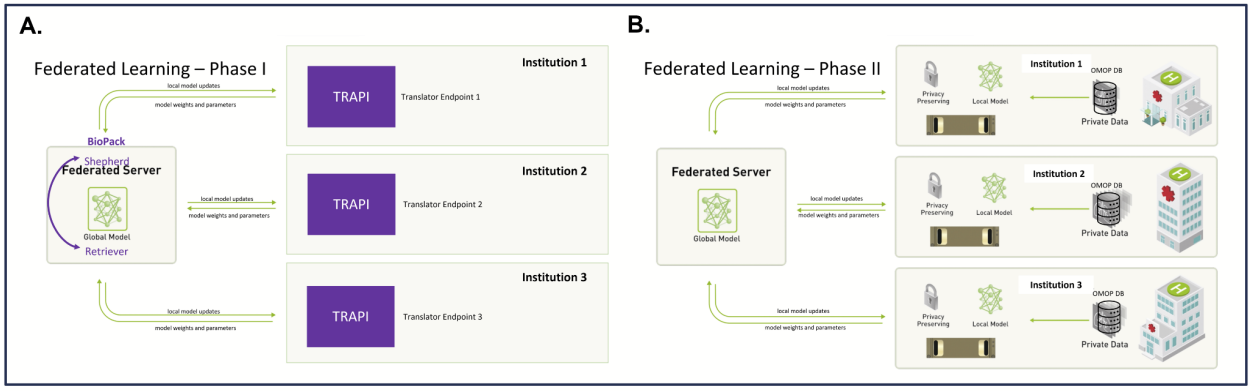

Figure 1. Proposed two-phase approach for federated learning over clinical data. (A) Phase I will involve using BioPack's Shepherd workflow management system and Retriever's subgraph retrieval service to run a federated query across COHD, Open Health Data @ Carolina, and the MCC Service, using our existing TRAPI endpoints. This work will leverage our ongoing Translator work and has all regulatory and institutional approvals in place to move forward. (B) Phase II will move beyond the Phase I work to implement federated learning. The Phase II work will leverage the Phase I work, but it will involve the use of a secure federated server (perhaps a secure version of BioPack) and apply federated queries across the OMOP databases at our three institutions. The Phase II work will require new regulatory and institutional approvals to move forward. TRAPI = Translator Reasoner Standard Application Programming Interface.

In the first phase of the proposed approach (Figure 1A), we will invoke the current Translator architecture to demonstrate proof-of-concept federated learning by running a simple mathematical algorithm[32] across our three Translator services, using existing Translator Reasoner Standard Application Programming Interfaces (TRAPI)[33]. We note that we have obtained all regulatory and institutional agreements to complete this phase of the proposed work. Specifically, we will leverage a new Translator component termed "BioPack". BioPack comprises two main components: "Shepherd" serves as a general workflow manager and centralized server; and "Retriever" is a subgraph retrieval service[34]. For the proposed application, Shepherd will direct a query to Retriever that asks for the sample size and the natural logarithm of the odds ratio for an observed association between a drug exposure and a disease outcome. Retriever will then query COHD, Open Health Data @ Carolina, and the MCC Service and return the two requested metrics back to Shepherd, which will then calculate the weighted average of the combined results, following an algorithm that we developed for Translator[32].

$$cohd: log\_odds\_ratio = OR1$$

$$total\_sample\_size = N1$$

$$weight = W1 = N1/(N1 + N2 + N3)$$

$$ohd@carolina: log\_odds\_ratio = OR2$$

$$total\_sample\_size = N2$$

$$weight = W2 = N2/(N1 + N2 + N3)$$

mcc: log_odds_ratio = OR3

total_sample_size = N3

weight = W3 = N3/(N1 + N2 + N3)

*clinical_information_score = (W1 * OR1 + W2 * OR2 + W3 * OR3) / (W1 + W2 + W3)*

A cap of 10 is imposed on the |log_odds_ratio| for each source to avoid skewing the overall score due to extremely high log_odds_ratio values, which are uncommon and often reflect very large sample sizes. Finally, the overall score is normalized [0,1) using the logistic normalization function in Python.

The key code block is:
import numpy as np
*def logistic_norm(x): return (1/(1+np.exp(-np.abs(x)))-0.5) * 2*

We will repeat this exercise a total of three times by querying three different drug-disease combinations, thus mimicking dynamic federated learning. We will verify our results by direct query of each source and comparison with our Translator results[32]. Importantly, the goal of the Phase I work is to demonstrate proof-of-concept federated learning, using existing infrastructure and approvals and leveraging our prior collaborative work.

For the second phase of the proposed work (Figure 1B), we propose to initially run the same algorithm used for the Phase I work, but targeting the OMOP databases directly, rather than through the open TRAPI endpoints. We will use either a secure version of the BioPack service or a secure centralized server to manage and execute the queries and algorithm updates. We note that unlike the Phase I work, the Phase II work will require new regulatory and institutional approvals, which is partially why the Phase I proof-of-concept work is critical. Nonetheless, we believe that our nearly ten-year history of successful research on and sharing of open clinical data will help to ensure that the proposed work moves forward. The goal of the Phase II work is to build the infrastructure to move toward implementation of more sophisticated machine learning models such as generalized linear models, random forest models, and causal network models. We note that our team has significant experience applying these models to EHR data[35–38], and that experience will additionally facilitate the proposed work. For instance, we may initially ask if sex, race, obesity, diabetes, and exposure to airborne particulate matter are predictive of prednisone use among patients with asthma. In this example, the goal would be to replicate our prior work[28] and position us to move toward more complex models and high-impact use cases.

Privacy and Security Considerations

Privacy and security concerns are always critical to consider when working with EHR data. The proposed work is no different. For the proposed Phase I effort, all regulatory and institutional requirements have been met, with fully executed agreements in place. Our institutions have safely exposed deidentified EHR-derived insights for the past several years, without any security or privacy breaches, thus demonstrating the reliability of our approach. However, the proposed Phase II effort will require new Institutional Review Board protocols and approvals, as well as new institutional approvals. To protect against security breaches during Phase II, we may hire a cybersecurity expert to evaluate the technical privacy and security risks posed by our proposed data/system flow. Additionally, we may obtain "Expert Determination[20,21]" using a service such as Datavant[39]. Given our history of successful human subjects protection and institutional trust, we are confident that the Phase II work will receive all necessary approvals to move forward. However, we recognize

that we may need to modify our proposed plan to ensure that patient privacy is protected and that our technical approach is secure.

Next Steps and Concluding Remarks

Having established the basic framework and approach for the proposed federated learning model, we are now ready to move forward with implementation of the Phase I work and initiate the regulatory and institutional discussions that will be required to begin the proposed Phase II work. We believe that our innovative approach to open multi-institutional sharing of EHR-derived insights data will prove valuable to our collaborating institutions, as well as the Translator program, and serve as a powerful exemplar for other institutions to adopt.

Acknowledgements

We gratefully acknowledge the contributions of all members of the ICEES, COHD, Open Health Data @ Carolina, MCC Service, and BioPack teams. In addition, we kindly thank the Biomedical Data Translator Consortium and NCATS leadership for their valuable input and support over the first two phases of the Translator program.

Funding Support

This work was supported by funds from the National Center for Advancing Translational Sciences (awards OT3TR002020, OT3TR002026, OT2TR003430, OT2TR003434, OT2TR003449).

References

1. McMahan B, Ramage D. Federated Learning: Collaborative Machine Learning without Centralized Training Data. April 6, 2017. https://research.google/blog/federated-learning-collaborative-machine-learning-without-centralized-training-data/

2. McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA y. Communication-Efficient Learning of Deep Networks from Decentralized Data. Published online January 26, 2023. http://arxiv.org/abs/1602.05629

3. Bonawitz K, Ivanov V, Kreuter B, et al. Practical Secure Aggregation for Federated Learning on User-Held Data. Published online 2016. doi:10.48550/ARXIV.1611.04482

4. Gill JK. Federated Learning Applications and Its Working | 2023. August 14, 2024. https://www.xenonstack.com/blog/federated-learning-applications

5. Rishimodi. Federated Learning will change the way companies use your data. April 25, 2024. https://medium.com/@rishimodi99/federated-learning-will-change-the-way-companies-use-your-data-403ecb923e37

6. Javatpoint. Federated Learning in Machine Learning. 2024. https://www.javatpoint.com/federated-learning-in-machine-learning

7. Lapid MI, Clarke BL, Wright RS. Institutional Review Boards: What Clinician Researchers Need to Know. Mayo Clinic Proceedings. 2019;94(3):515-525. doi:10.1016/j.mayocp.2019.01.020

8. FTC. The Most Common Reasons Hackers Hack. February 28, 2024. https://www.ftc.net/business/blog/the-most-common-reasons-hackers-hack/

9. Breachsense. Why are so many data breaches caused human error? 2024. https://www.breachsense.com/blog/data-breach-human-error/

10. Steinke T, Ullman J. Between Pure and Approximate Differential Privacy. The President and Fellows of Harvard College; 2015. https://privacytools.seas.harvard.edu/publications/between-pure-and-approximate-differential-privacy

11. The Biomedical Data Translator Consortium. Toward A Universal Biomedical Data Translator. Clinical Translational Sci. 2019;12(2):86-90. doi:10.1111/cts.12591

12. Fecho K, Thessen AE, Baranzini SE, et al. Progress toward a universal biomedical data translator. Clinical Translational Sci. 2022;15(8):1838-1847. doi:10.1111/cts.13301

13. Unni DR, Moxon SAT, Bada M, et al. Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. Clinical Translational Sci. 2022;15(8):1848-1855. doi:10.1111/cts.13302

14. The Biomedical Data Translator Consortium. The Biomedical Data Translator Program: Conception, Culture, and Community. Clinical Translational Sci. 2019;12(2):91-94. doi:10.1111/cts.12592

15. Kiachopoulos Y, Saudabayev A, The Causaly Team. Causaly. 2024. https://www.causaly.com/

16. Causaly: Generative AI for Life Sciences. Published online 2024. https://www.causaly.com/

17. EmBiology. Published online 2024. https://www.elsevier.com/products/embiology

18. Ahalt SC, Chute CG, Fecho K, et al. Clinical data: sources and types, regulatory constraints, applications. Clinical Translational Sci. 2019;12(4):329-333. doi:10.1111/cts.12638

19. Fecho K, Pfaff E, Xu H, et al. A novel approach for exposing and sharing clinical data: the Translator Integrated Clinical and Environmental Exposures Service. Journal of the American Medical Informatics Association. 2019;26(10):1064-1073. doi:10.1093/jamia/ocz042

20. US Department of Health and Human Services. HIPAA Deidentification Methods. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

21. Office for Civil Rights, US Department of Health and Human Services. Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.; 2024. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

22. OpenAPI. https://spec.openapis.org/oas/latest.html

23. Ta CN, Dumontier M, Hripcsak G, Tatonetti NP, Weng C. Columbia Open Health Data, clinical concept prevalence and co-occurrence from electronic health records. Sci Data. 2018;5(1):180273. doi:10.1038/sdata.2018.273

24. Biomedical Data Translator Consortium. Open Health Data @ Carolina. Published online 2024. https://github.com/ExposuresProvider/open-health-data-at-carolina

25. Hadlock J, Wei Q. Multiomics Clinical Connections Service. Published online 2024. https://github.com/NCATSTranslator/Translator-All/wiki/Multiomics-EHR-Risk-KP

26. OHDSI: Observational Health Data Sciences and Informatics. Published online 2024. https://www.ohdsi.org/

27. OHDSI. Standardized Data: The OMOP Common Data Model. Published online 2024. https://www.ohdsi.org/data-standardization/

28. Fecho K, Ahalt SC, Arunachalam S, et al. Sex, obesity, diabetes, and exposure to particulate matter among patients with severe asthma: Scientific insights from a comparative analysis of open clinical data sources during a five-day hackathon. Journal of Biomedical Informatics. 2019;100:103325. doi:10.1016/j.jbi.2019.103325

29. Fecho K, Bizon C, Issabekova T, et al. An approach for collaborative development of a federated biomedical knowledge graph-based question-answering system: Question-of-the-Month challenges. J Clin Trans Sci. 2023;7(1):e214. doi:10.1017/cts.2023.619

30. Baldin I, Chase J, Crabtree J, et al. ImPACT: A networked service architecture for safe sharing of restricted data. Future Generation Computer Systems. 2022;129:269-285. doi:10.1016/j.future.2021.11.026

31. Scott E. Secure Multiparty Computation. Published online 2024. https://github.com/RENCI-NRIG/impact-smc

32. Biomedical Data Translator Consortium. Translator Clinical Information Score. Published online 2024. https://github.com/NCATSTranslator/Translator-All/wiki/Translator-Clinical-Information-Score

33. Biomedical Data Translator Consortium. NCATS Biomedical Translator Reasoners Standard API. Published online 2024. https://github.com/NCATSTranslator/ReasonerAPI

34. Biomedical Data Translator Consortium. BioPack. Published online 2024. https://github.com/BioPack-team/shepherd and Retriever here: https://github.com/BioPack-team/retriever

35. Fecho K, Haaland P, Krishnamurthy A, et al. An approach for open multivariate analysis of integrated clinical and environmental exposures data. Informatics in Medicine Unlocked. 2021;26:100733. doi:10.1016/j.imu.2021.100733

36. Lan B, Haaland P, Krishnamurthy A, et al. Open Application of Statistical and Machine Learning Models to Explore the Impact of Environmental Exposures on Health and Disease: An Asthma Use Case. IJERPH. 2021;18(21):11398. doi:10.3390/ijerph182111398

37. Sharma P, Haaland P, Krishnamurthy A, et al. Evaluating robustness of a generalized linear model when applied to electronic health record data accessed using an Open API. Health Informatics J. 2023;29(2):146045822311708. doi:10.1177/14604582231170892

38. Sinha M, Haaland P, Krishnamurthy A, et al. Causal Analysis for Multivariate Integrated Clinical and Environmental Exposures Data. Published online December 21, 2022. doi:10.1101/2022.12.20.22283734

39. Datavant: Making the world's health data secure, accessible, and usable. Published online 2024. https://www.datavant.com/about