# Considerations for De-identification of the OMOP Common Data Model

Jose D. Posada, Natasha Flowers, Priya Desai
Stanford Healthcare

## Background

De-identification, defined by NIST as "any process of removing the association between a set of identifying data and the data subject" [1], is essential for OMOP-CDM datasets because healthcare data is susceptible to privacy regulations worldwide that protect patients from disclosure of sensitive information that may affect their reputation or livelihood [29]. The primary driver for de-identification in the OHDSI community context is the secondary use of data for research purposes—using healthcare data for purposes other than direct patient care—which requires preserving patient privacy while enabling valuable research applications [30,31]. HIPAA's Safe Harbor method [3], which requires removal of 18 specific identifiers, has become the most common approach for de-identification in healthcare datasets.

The OMOP-CDM contains several types of Personal Identifiable Information that must be addressed during de-identification. Direct identifiers like person_source_value require random identifier generation or deterministic encryption with secure key management [12], while dates typically need consolidation or patient-level shifting to preserve event timelines [13]. Source_value fields throughout the OMOP-CDM tables present particular challenges as they may contain free-form text with embedded PII, especially in fully populated implementations where provenance data is preserved for concept_id mapping and measurement details [6].

## Recommendations

De-identification of string data types in OMOP-CDM requires careful handling of source_value fields and note_text columns that may contain free-text input from source systems [6]. The approach varies based on intended use, from removing all unmapped values for standardized network research to implementing extensive allow lists for source_values that preserve provenance while ensuring concept_id mapping integrity [6]. The note_text column presents particular challenges as it contains rich clinical narratives essential for research but requires sophisticated processing to maintain utility while removing embedded PII [9].

### Free-Text De-identification Strategies

Three primary strategies address free-text de-identification in OMOP-CDM implementations [9,10,11]. The simplest approach removes everything not mapped to concept_ids, with

exceptions like drug_exposure.sig fields that require preservation [6]. More comprehensive strategies utilize curated allow lists with human review of unique unmapped values, considering provenance metadata that provides context for interpreting source_values [6].

Advanced implementations employ natural language processing techniques including ensemble methods combining regular expressions, machine learning models, and redaction or Hidden in Plain Sight approaches [9,10]. Systematic reviews have analyzed 18 automated text de-identification systems, revealing two primary approaches: pattern matching using regular expressions and machine learning methods [9].

Deep learning has revolutionized text processing, with artificial neural network systems requiring no handcrafted features achieving superior performance through Recurrent Neural Networks with word embeddings [10]. However, cross-institutional generalizability remains challenging, with performance dropping from F1 0.9547 to 0.8568 when applied across institutions, requiring fine-tuning strategies for domain adaptation [11].

Evaluation occurs through either leakage rate assessment or full annotation processes to ensure effective de-identification while preserving the clinical data utility essential for OMOP-CDM research applications [13].

### Technical Implementation Approaches

**Date shifting preserves temporal relationships** while meeting HIPAA requirements through algorithms assigning unique shift values to each Subject ID, maintaining internally consistent durations between events within patient records [12,13]. Implementations typically shift dates by up to 364 days, with ±180-day shifts meeting Safe Harbor criteria [3].

**Deterministic encryption and hashing** enable consistent replacement algorithms maintaining referential integrity, as described in foundational work evaluated using Veterans Health Administration clinical documents [12,13]. These methods ensure that the same original identifier always maps to the same de-identified value across all tables and time periods.

**Privacy-preserving techniques** include federated learning through Personal Health Train architectures, where analysis comes to data rather than requiring data sharing [25,26,27,28]. These approaches have been successfully deployed for lung cancer research across 20,000+ patients while maintaining data sovereignty at source institutions.

## Assessment and Evaluation

**Multiple risk assessment approaches** are available including prosecutor risk (when attackers have background knowledge), journalist risk (random record selection with public information), and marketer risk (commercial database attacks) [14,36]. Research has demonstrated median risk reductions of 90.1% through adversarial modeling frameworks integrating realistic attacker capabilities [14].

**Statistical privacy models provide mathematical guarantees** through k-anonymity ensuring records are indistinguishable from at least k-1 others [32], l-diversity requiring well-represented sensitive attribute values [33], and t-closeness maintaining distribution similarity between equivalence classes and overall datasets [34]. These models have been successfully applied to OMOP-CDM implementations, with studies achieving 0.03% re-identification success rates in cloud computing environments [7].

**HIPAA Expert Determination requires systematic validation** through five-step processes: initial risk evaluation, statistical method application, implementation by data managers, final verification of "very small" risk standards, and comprehensive documentation of all methods and justifications [3].

# Emerging Technologies and Future Directions

**Large language models enable automated de-identification** as demonstrated by recent research using GPT-3 embeddings for semantic matching, achieving AUC of 0.9975 for clinical trial term mapping while maintaining accessibility for research teams without extensive informatics support [40].

**Synthetic data generation** provides privacy-preserving alternatives through advanced generative models validated for lung cancer prognostic models, demonstrating effectiveness throughout medical modeling pipelines while preserving data utility [19,20,21,22,41,42].

**Continuous risk monitoring** enables real-time assessment through automated validation systems providing online leakage assessment, query rejection mechanisms for low subject counts, and session-dependent randomization protecting against cross-researcher linking [14,36].

# Conclusion

OMOP-CDM de-identification requires integrating regulatory compliance with technical innovation and ongoing community collaboration. The evidence base spans federal standards [1,2,3], peer-reviewed research [7,8,9,10,11], and professional guidelines [15,17], providing a robust foundation for privacy-preserving observational health research while maintaining data utility for scientific discovery. Success depends on balancing the model's person-centric design with rigorous de-identification strategies, particularly for source values and unstructured clinical notes, supported by active engagement with evolving best practices and continuous privacy risk assessment.

# Disclaimer

The recommendations contained in this document are not legal advice [6]. Please refer to your local regulations and consult with legal counsel to ensure you meet the legal standard for de-identification of healthcare data in your jurisdiction.

---

# References


1. National Institute of Standards and Technology. NIST Special Publication 800-188: De-Identifying Government Datasets: Techniques and Governance. Gaithersburg, MD: NIST; 2023 Sep. Available from: https://csrc.nist.gov/pubs/sp/800/188/final
2. National Institute of Standards and Technology. NIST Internal Report 8053: De-Identification of Personal Information. Gaithersburg, MD: NIST; 2015 Oct. Available from: https://csrc.nist.gov/pubs/ir/8053/final
3. U.S. Department of Health and Human Services, Office for Civil Rights. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Washington, DC: HHS; 2012 Nov 26. Available from: https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html
4. European Parliament and Council. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union. 2016;L119:1-88.
5. OHDSI Collaborative. The Book of OHDSI: Observational Health Data Sciences and Informatics. 2019. Available from: https://ohdsi.github.io/TheBookOfOhdsi/
6. OHDSI Collaborative. Preserving Privacy in an OMOP CDM Implementation. OMOP Common Data Model Documentation. Available from: https://ohdsi.github.io/CommonDataModel/cdmPrivacy.html
7. Jeon H, Oh JH, Lee S, Lee Y, Lee JH, Lim JM, et al. Proposal and Assessment of a De-Identification Strategy to Enhance Anonymity of the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) in a Public Cloud-Computing Environment: Anonymization of Medical Data Using Privacy Models. J Med Internet Res. 2020;22(11):e19597. Available from: https://www.jmir.org/2020/11/e19597/
8. Tak WY, Lee S, Lee JH, Lee Y, Lim JM, Oh JH, et al. Perceived Risk of Re-Identification in OMOP-CDM Database: A Cross-Sectional Survey. J Korean Med Sci. 2022;37(26):e201. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9259248/
9. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol. 2010;10:70. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-10-70
10. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc. 2017;24(3):596-606.

11. Yang H, Garibaldi JM. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. BMC Med Inform Decis Mak. 2019;19(Suppl 5):232. Available from: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0935-4
12. Friedlin FJ, McDonald CJ. A software tool for removing patient identifying information from clinical documents. J Am Med Inform Assoc. 2008;15(5):601-10.
13. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Evaluating current automatic de-identification methods with Veteran's health administration clinical documents. BMC Med Res Methodol. 2012;12:109. Available from: https://link.springer.com/article/10.1186/1471-2288-12-109
14. Xia W, Heatherly R, Ding X, Li J, Malin BA. Enabling realistic health data re-identification risk assessment through adversarial modeling. J Am Med Inform Assoc. 2021;28(4):744-752. Available from: https://academic.oup.com/jamia/article/28/4/744/6101071
15. National Institute of Standards and Technology. NIST Privacy Framework Version 1.1: A Tool for Improving Privacy Through Enterprise Risk Management. Gaithersburg, MD: NIST; 2020 Jan. Available from: https://www.nist.gov/privacy-framework
16. National Institute of Standards and Technology. NIST Updates Privacy Framework, Tying It to Recent Cybersecurity Guidelines. Gaithersburg, MD: NIST; 2025 Apr. Available from: https://www.nist.gov/news-events/news/2025/04/nist-updates-privacy-framework-tying-it-recent-cybersecurity-guidelines
17. American Medical Informatics Association. AMIA's Code of Professional and Ethical Conduct 2018. J Am Med Inform Assoc. 2020;27(11):1803-8. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7646920/
18. Kurian AW, Ward KC, Howlader N, Deapen D, Hamilton AS, Mariotto A, et al. Genetic Testing and Results in a Population-Based Cohort of Breast Cancer Patients and Ovarian Cancer Patients. J Clin Oncol. 2019;37(15):1305-15.
19. Liu Q, Chen S, Jiang R, Wong WH. Simultaneous deep generative modelling and clustering of single-cell genomic data. Nat Mach Intell. 2021;3:536-44.
20. Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-preserving generative deep neural networks support clinical data sharing. Circ Cardiovasc Qual Outcomes. 2019;12(7):e005122.
21. McLachlan S, Dube K, Hitman GA, Fenton NE, Kyrimi E. Bayesian networks in healthcare: Distribution by medical condition. Artif Intell Med. 2020;107:101912.
22. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng. 2021;5(6):493-7.
23. Jordon J, Yoon J, van der Schaar M. PATE-GAN: Generating synthetic data with differential privacy guarantees. In: International Conference on Learning Representations; 2019.
24. El Emam K, Hoptroff R. The Synthetic Data Paradigm for Using and Sharing Data. Cambridge, MA: Academic Press; 2022.

25. Boenisch F, Dziedzic A, Schuster R, Shamsabadi AS, Shumailov I, Papernot N. When the curious abandon honesty: Federated learning is not private. In: 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P); 2023. p. 175-199.
26. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: Challenges, methods, and future directions. IEEE Signal Process Mag. 2020;37(3):50-60.
27. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell. 2020;2(6):305-11.
28. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3:119.
29. Institute of Medicine. Health Data is a Special Kind of Data. In: Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. Washington, DC: The National Academies Press; 2009.
30. Rothstein MA. Is deidentification sufficient to protect health privacy in research? Am J Bioeth. 2010;10(9):3-11.
31. Ohno-Machado L. To share or not to share: that is not the question. Sci Transl Med. 2012;4(165):165cm16.
32. Sweeney L. k-anonymity: A model for protecting privacy. Int J Uncertain Fuzziness Knowledge Based Syst. 2002;10(5):557-70.
33. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: Privacy beyond k-anonymity. ACM Trans Knowledge Discov Data. 2007;1(1):3.
34. Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd International Conference on Data Engineering; 2007. p. 106-115.
35. Dwork C. Differential privacy: A survey of results. In: Agrawal M, Du D, Duan Z, Li A, editors. Theory and Applications of Models of Computation. TAMC 2008. Lecture Notes in Computer Science, vol 4978. Berlin: Springer; 2008. p. 1-19.
36. El Emam K, Dankar FK, Issa R, Jonker E, Amyot D, Cogo E, et al. A globally optimal k-anonymity method for the de-identification of health data. J Am Med Inform Assoc. 2009;16(5):670-82.
37. Fung BC, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: A survey of recent developments. ACM Comput Surv. 2010;42(4):1-53.
38. Ji Z, Lipton ZC, Elkan C. Differential privacy and machine learning: a survey and review. arXiv preprint arXiv:1412.7584. 2014 Dec 23.
39. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, et al. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; 2016. p. 308-318.
40. Wake Forest School of Medicine. Breaking Digital Health Barriers Through a Large Language Model–Based Tool for Automated Observational Medical Outcomes Partnership Mapping: Development and Validation Study. J Med Internet Res. 2025;27(1):e69004. Available from: https://www.jmir.org/2025/1/e69004
41. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018;25(3):230-8.

42. Chen RJ, Lu MY, Williamson DFK, Chen TY, Lipkova J, Noor Z, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell. 2022;40(8):865-878.e6.
43. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH Harmonised Guideline: General Considerations for Clinical Studies E8(R1). Geneva: ICH; 2021.
44. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: WHO; 2021.
45. European Medicines Agency. Regulatory Science to 2025: Strategic reflection. Amsterdam: EMA; 2020.
46. U.S. Food and Drug Administration. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. Silver Spring, MD: FDA; 2021.
47. International Organization for Standardization. ISO/IEC 27001:2022 Information security management systems — Requirements. Geneva: ISO; 2022.
48. Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (1996).
49. 21st Century Cures Act, Pub. L. No. 114-255, 130 Stat. 1033 (2016).
50. California Consumer Privacy Act of 2018, Cal. Civ. Code § 1798.100 et seq. (2018).