Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01429-0

Privacy preserving strategies for electronic health records in the era of large language models

Jitendra Jonnagaddala & Zoie Shui-Yee Wong

Check for updates

Electronic health records (EHRs) secondary usage with large language models (LLMs) raise privacy challenges. National regulations like GDPR and HIPAA offer protection frameworks, but specific strategies are needed to mitigate risk in generative AI. Risks can be reduced by using strategies like privacy-preserving locally deployed LLMs, synthetic data generation, differential privacy, and deidentification. Depending on the task, strategies should be employed to increase compliance with patient privacy regulatory frameworks.

The secondary use of electronic health records (EHRs) involves the utilization of EHR data for various purposes other than their original intent in clinical operations, such as clinical research, health systems and services research, patient registries, quality improvement, disease surveillance, and other areas beyond direct patient care[1]. This secondary use of electronic health records (EHRs) has significantly accelerated in the past decade with advances in artificial intelligence (AI), especially in large language models (LLMs)[2]. EHRs are increasingly used with LLMs for primary uses, such as clinical documentation, generation and summarization, as well as for secondary uses, such as information extraction, retrieval, identification of eligible patients for research, medical education, and outcome reporting. Safeguarding both the structured and unstructured sensitive health information (SHI) of patients from EHRs is essential, particularly when LLMs are used for secondary uses[3,4]. Most countries across the world have stipulated regulatory privacy acts, guidelines and frameworks to achieve this goal.

The Health Insurance Portability and Accountability Act (HIPAA) from the United States of America provides guidelines for preserving patients' SHI[5]. Similarly, in Australia, the Privacy Act of 1988 acted as a counterpart of the HIPAA by providing guidelines and a framework. The General Data Protection Regulation (GDPR) provides a framework of laws and regulations for the collection and use of the SHI of individuals living in the European Union[6]. In Japan, the Act on the Protection of Personal Information (APPI)[7] sets a legal framework for protecting individual privacy and personal information, sharing similarities with the data protection principles of the GDPR. Serving as one of the early data protection laws in Asia, the Personal Data (Privacy) Ordinance from Hong Kong applies to maintain and protect patient privacy in the Hong Kong[8]. Most of these regulatory frameworks have several commonalities but also differ in certain aspects. For example, the definition of SHI and what constitutes SHI vary. Additionally, most of these regulations do not distinguish between structured and unstructured EHR data. This differentiation is important because structured EHR data is relatively easy to process compared with unstructured EHR data, which requires robust information extraction methods and is often not 100% accurate. This difference in complexity between structured and unstructured data can impact downstream tasks, including compliance with privacy regulations.

These privacy regulations were developed primarily to provide comprehensive requirements and guidelines for securely storing and handling SHI. While some regulations, such as the GDPR, encompass a wide range of data protection principles beyond health data, they also establish specific frameworks for safeguarding SHI and ensuring that it is appropriately redacted before public disclosure. These regulations highlight the importance of well-designed statistical analysis to protect and maintain privacy. This involves the use of large sample sizes to ensure that the results of research are meaningful while minimizing the reidentification risk to individuals. Most privacy regulations prohibit data sharing involving fewer than five individuals to reduce the likelihood of reidentification. Even in the absence of direct identifiers, an individual can sometimes be recognized through the combination of data points such as dates, diagnoses, and treatments, which can increase the risk of reidentification.

Even when adhering to these guidelines, there are still chances of unique reidentification and group reidentification leading to privacy risks. Unique reidentification occurs when a record in the dataset matches exactly with an identified individual, whereas group reidentification occurs when one or a few records in the dataset correspond to a small group of identified individuals[9]. There are instances where identity recovery algorithms applied to deidentified data may be able to retrieve one-to-one, one-to-few and few-to-few identities[10].

Privacy-preserving strategies

We summarize a series of privacy-preserving strategies that can be applied (Fig. 1). LLMs are frequently employed by clinicians, researchers and educators for various purposes, such as documentation, guideline retrieval and summarization and clinical decision support. Regarding adhering to privacy regulations and guidelines, the application of LLMs to both structured and unstructured EHR data is still in the early phases of development[11,12]. Notably, LLMs are often used to extract structured data from unstructured inputs, emphasizing their wide-ranging utility. Different technical approaches can be employed to achieve compliance and preserve privacy, depending on the task for which the LLM is used.

To reduce the risk in structured EHR data, differential privacy (DP) can be implemented before feeding the input to LLMs. DP can be defined as the inability to distinguish between two output probability distributions[13]. Suppose we have two databases of SHI that differ by only one patient. DP ensures that, even with access to one of the databases, it is not possible to identify the differing patient with certainty. This uncertainty is achieved by

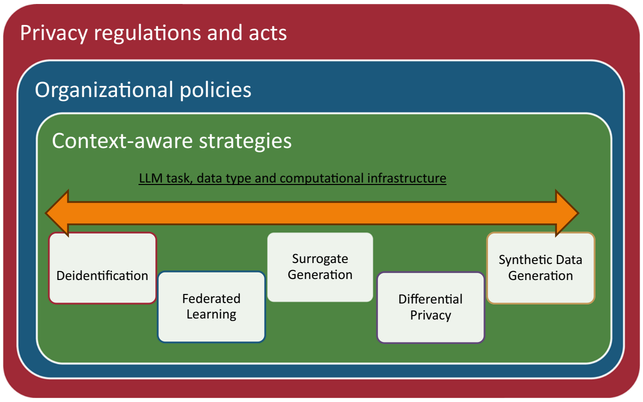

Fig. 1 | Context-aware privacy preserving strategies for secondary usage of EHRs. The figure summarizes how organizations can manage and address privacy issues through integrating regulatory compliance, developing policies and context-aware technical strategies to safeguard sensitive data and maintain privacy when using EHRs with LLMs. Strategies such as deidentification, surrogate generation, federated learning, and differential privacy need to be applied based on the LLM tasks, data types, and computational infrastructure.

introducing calibrated noise into the database query results, often using Laplace mechanism. Sophisticated deidentification solutions such as polymorphic encryption can also be employed to deidentify structured SHI while maintaining its usefulness for analytical purposes, which keeps the data secure and reduces the risk of reidentification. Some studies suggest the use of Federated Learning (FL) to increase model performance while ensuring data privacy[14]. FL is a decentralized approach that enables multiple organizations to train AI models collaboratively without sharing their raw data. By leveraging enormous volumes of data from many healthcare settings and maintaining data localization, this approach mitigates privacy concerns and risks. DP, deidentification, and FL are some of the techniques that are frequently applied to structured SHI data, but these techniques can also be applied to unstructured data.

Both structured and unstructured EHR data often capture and summarize SHI over different time points of disease progression, which are unique to patients. To reduce risk, stringent deidentification standards and robust measures must be adopted. One such example is through the deidentification and surrogation of the SHI. Deidentification has received much attention around the globe, and with improvements in LLM techniques, the accuracy of correctly identifying and removing personal identifiers while maintaining the data's utility for analysis and research has significantly increased[15]. Liu, J., et al. 2023 presented a hybrid deidentification pipeline for unstructured pathology reports[16,17], which incorporates both rule-based techniques and transformer models. The pipeline proved to be efficient in identifying SHI with an accuracy of 95%. Once the SHI is identified, the reports can be deidentified by removing or surrogating the identifiers. Additionally, earlier versions of this pipeline have been deployed in an Australian hospital[3]. Furthermore, the deidentified SHI can be further surrogated before being input into the LLMs. Surrogation maintains the syntactic and synoptic representations of the data[18]. Surrogate generation involves generating placeholder data that mimic the statistical properties of the original data without compromising individual privacy. This process of replacement with placeholders allows the deidentified data, whether unstructured or structured, to preserve the integrity of the data's structure and meaning. For example, a name [Mr. John Doe] is identified and anonymized as [Patient Name: (REDACTED)] followed by replacement with a surrogated name [Patient Name: Andy Ray]. However, this may have an impact on data utility. The impact depends on the level of granularity required in the analysis and the type of downstream task. For types of analyses, where identifiers are not crucial, the impact is minimal. However, for analyses that focus on specific identifiers such as locations and time periods, surrogation has a great impact on data utility. In such cases, other techniques, such as DP can be employed. As the techniques employed are subject to various performance and methodological limitations, it is important to carefully balance the trade-offs between maintaining data privacy and utility. Appropriate strategies should be employed based on the context (Fig. 1). While less aggressive methods can preserve data utility, they may also result in an increased risk of reidentification.

Studies have assessed the feasibility of employing lightweight, open, and offline LLM locally in secure environments to overcome some of the privacy concerns associated with using third-party LLM[19–21]. This strategy can be employed for both structured and unstructured data. Local deployment of LLM might also be computationally limited and resource intensive with the increasing number of EHRs captured. To address these limitations, one approach is to use lightweight locally deployed LLM or traditional rule-based and machine learning models to deidentify and generate surrogates before passing the information to proprietary or third-party LLMs. Another approach is to use synthetic data to fine-tune, pretrain or assess LLM performance on certain tasks. However, this also introduces the risk of misinformation and hallucinations. It is important to recognize that these approaches apply to both primary and secondary uses of EHR data.

Without proper security measures, whether the LLM is used in the cloud or locally, data breaches are possible, which can have long-term consequences. For example, a recent Australian data breach at MediSecure, an Australian prescription delivery service company, occurred in April 2024, exposing private and health-related information on the dark web[22]. Consequently, organizations need to employ several strategies to reduce the risk of reidentification in the context of these data leaks, wherein the leaked information can be used to reidentify patients in other datasets[4].

Employing these strategies can become complex, especially for patients who interact frequently with the health system. This also urges the identification and generation of surrogates without losing temporal information. However, there is always a trade-off between risks and benefits and context, such as sharing data publicly on the internet or through secure research environments. Depending on the type of SHI, whether structured, unstructured, or a combination of both, and the type of LLM employed, whether open or closed, and cloud-based or on-premises deployments, the associated risks might vary from minimal to concerning. To achieve the best outcomes, most of these strategies must be combined in the context of regulations and organizational policies. In addition, it is also important to understand the theoretical and practical limitations of various privacy-preserving techniques. As such, a context-aware approach that factors in the specific LLM tasks, the data type being processed, and the computational infrastructure is required to mitigate the privacy risks associated with LLMs. For example, tasks such as identifying patients of interest for clinical research involving the secondary use of unstructured EHRs require more robust privacy strategies and appropriate computational infrastructure than the use of structured data for the primary purpose of providing care. Furthermore, governance structures should include regular quality improvement plans to assess the effectiveness of deidentification methods as AI technology continues to evolve. Although we focused primarily on routinely collected EHR data in both unstructured and structured formats, the strategies and techniques discussed here can also be applied to other modalities, such as imaging and omics data.

Conclusion

LLMs are frequently employed by clinicians, researchers, and educators for various purposes. However, the regulatory frameworks and guidelines for managing privacy while using structured and unstructured EHR data with LLMs are still in the early stages and require timely reforms to address evolving data modalities and applications of LLMs. Unique or group reidentification risks are associated with the use of LLMs. From the perspective of data custodians and users, identifying and understanding the risks associated with LLMs, followed by employing appropriate strategies and governance frameworks within organizations, can not only reduce privacy risks but also provide an opportunity to address uncertainties.

Data availability

No datasets were generated or analysed during the current study.

Jitendra Jonnagaddala[1,2,3] ✉ **& Zoie Shui-Yee Wong**[4,5,6,7]
[1]School of Population Health, UNSW Sydney, Kensington, NSW, Australia. [2]NMC Royal Hospital, Khalifa City, Abu Dhabi, United Arab Emirates. [3]SREDH Consortium, Sydney, NSW, Australia. [4]Graduate School of Public Health, St. Luke's International University, Tokyo, Japan. [5]The Kirby Institute, UNSW Sydney, Kensington, NSW, Australia. [6]School of Medical Sciences, The University of Sydney, Sydney, NSW, Australia. [7]School of Public Health, The University of Hong Kong, Hong Kong, China.
✉e-mail: jitendra.jonnagaddala@unsw.edu.au

Received: 28 October 2024; Accepted: 2 January 2025;
Published online: 16 January 2025

References

1. Seymour, T., Frantsvog, D. & Graeber, T. Electronic health records (EHR). *Am. J. Health Sci.* **3**, 201 (2012).
2. Christian Rose, J. H. C. Learning from the EHR to implement AI in healthcare. *npj digital medicine* **7**, https://doi.org/10.1038/s41746-024-01340-0 (2024).
3. Liu, J. et al. OpenDeID pipeline for unstructured electronic health record text notes based on rules and transformers: deidentification algorithm development and validation study. *J. Med. Internet Res.* **25**, e48145 (2023).
4. Basil, N. N., Ambe, S., Ekhator, C. & Fonkem, E. Health records database and inherent security concerns: a review of the literature. *Cureus* **14**, e30168 (2022).
5. America, U. S. o. *Health insurance portability and accountability act of 1996. Public Law 104–191*, 1996).
6. Union, E. *General data protection regulation (GDPR)*, https://gdpr-info.eu (2016).
7. Japan, G. o. *Act on the protection of personal information, Act No. 57 of 2003*, https://www.japaneselawtranslation.go.jp/en/laws/view/4241/en (2023).
8. Government, D. o. J. o. t. H. K. *Personal Data (Privacy) Ordinance, Chapter 486 of the Laws of Hong Kong (PDPO)*, http://www.hklii.org.hk/hk/legis/ord/486/index.html (1994).
9. Sweeney, L. et al. Re-identification risks in HIPAA safe harbor data: a study of data from one environmental health study. *Technol. Sci.* **2017**, 2017082801 (2017).
10. The Office of the Victorian Information Commissioner. The Limitations of De-Identification Protecting Unit-Record Level Personal Information.https://ovic.vic.gov.au/privacy/resources-fororganisations/the-limitations-of-de-identification-protecting-unit-record-level-personal-information/ (2018).
11. Meskó, B. & Topol, E. J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digit. Med.* **6**, 120 (2023).
12. Raza, M. M., Venkatesh, K. P. & Kvedar, J. C. Generative AI and large language models in health care: pathways to implementation. *npj Digit. Med.* **7**, 62 (2024).
13. Siqi Zhang, X. L. Differential privacy medical data publishing method based on attribute correlation. *Sci. Rep.* https://doi.org/10.1038/s41598-022-19544-3 (2022).
14. Peng, L. et al. An in-depth evaluation of federated learning on biomedical natural language processing for information extraction. *NPJ Digit. Med.* **7**, 127 (2024).
15. Jonnagaddala, J., Dai, H. J. & Chen, C. T. *Large language models for automatic deidentification of electronic health record notes*. 1 edn, Vol. 2148 (Springer Nature, 2025).
16. Jonnagaddala, J., Chen, A., Batongbacal, S. & Nekkantti, C. The OpenDeID corpus for patient de-identification. *Sci. Rep.* **11**, 19973 (2021).
17. Gupta, S., Liu, J., Wong, Z. S. & Jonnagaddala, J. Preliminary evaluation of fine-tuning the OpenDeLD deidentification pipeline across multi-center corpora. *Stud. Health Technol. Inf.* **316**, 719–723 (2024).
18. Chen, A., Jonnagaddala, J., Nekkantti, C. & Liaw, S. T. Generation of surrogates for de-identification of electronic health records. *Stud. Health Technol. Inf.* **264**, 70–73 (2019).
19. Tai, I. C. Y. et al. Exploring offline large language models for clinical information extraction: a study of renal histopathological reports of lupus nephritis patients. *Stud. Health Technol. Inf.* **316**, 899–903 (2024).
20. Wiest, I. C. et al. Privacy-preserving large language models for structured medical information retrieval. *npj Digit. Med.* **7**, 257 (2024).
21. Chua, C. E. et al. Integration of customised LLM for discharge summary generation in real-world clinical settings: a pilot study on RUSSELL GPT. *Lancet Reg. Health – West. Pac.* **51**, https://doi.org/10.1016/j.lanwpc.2024.101211 (2024).
22. InSight. *MediSecure breach: implications for health care services and patients*, https://insightplus.mja.com.au/2024/37/medisecure-breach-implications-for-health-care-services-and-patients/ (2024).

Acknowledgements

J.J. was funded by the Australian National Health and Medical Research Council (grant GNT1192469). J.J. also acknowledges the funding support received through the Research Technology Services at the University of New South Wales Sydney, Google Cloud Research (award GCP19980904), and NVIDIA Academic Hardware grant programs.

Author contributions

J.J. and Z.S.Y.W. were responsible for conceptualizing the initial concept for this manuscript. Both analyzed the literature; J.J. was responsible for drafting and revising the manuscript. Z.S.Y.W. provided critical intellectual input and revisions. Both authors approved the manuscript.

Competing interests

J.J. and Z.S.Y.W. are both Guest Editors of Collection on Natural Language Processing of npj Digital Medicine. Z.S.Y.W. is also serving the npj Digital Medicine as an Associate Editor.

Additional information

Correspondence and requests for materials should be addressed to Jitendra Jonnagaddala.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.