

## Evaluating PHI Leakage Risks and Governance Mechanisms in Healthcare Generative AI Pipelines

Vinod Rufus Motani

Submitted: 05/03/2023

Revised: 12/04/2023

Accepted: 20/04/2023

**Abstract:** There has been an increased use of Generative Artificial Intelligence (GenAI) in healthcare to assist in documentation, diagnosis, communication with patients, and decision-making in administration. However, the increase in the use of this technology has resulted in some major risks with respect to PHI leakage through AI pipelines. In this research, PHI leakage risks and governance strategies in healthcare generative AI pipelines have been investigated using a qualitative research approach. Secondary data were gathered from peer-reviewed journal articles, healthcare regulations, policy papers, and AI governance literature. Inductive research approach and thematic analysis were applied to recognize major privacy risks and governance strategies across various stages of the AI pipeline. As a result of this analysis, the risks of PHI leakage at stages of data collection, preprocessing, model training, fine-tuning, and inference have been identified. Major vulnerabilities include insecure cloud storage, weak access control, prompt-based PHI leakage, membership inference attacks, data memorisation, and risks of retrieval-augmented generation. Moreover, organizational procedures such as continuous monitoring of models, staff education, and standardized governance framework of AI contribute to privacy compliance and security of AI implementation. In conclusion, privacy protection of PHI requires governance measures throughout the whole generative AI life cycle instead of only at one particular phase of development. Combining technical controls and regulatory and governance measures can lead to lowering privacy threats and fostering secure and trustworthy adoption of generative AI in healthcare organizations.

**Keywords:** *Generative Artificial Intelligence (GenAI), Protected Health Information (PHI), Healthcare AI, Data Privacy, PHI Leakage, AI Governance, HIPAA Compliance, Large Language Models (LLMs), Healthcare Data Security, Privacy-Preserving AI.*

### Introduction

Generative Artificial Intelligence (GenAI) has revolutionized the field of healthcare by assisting in documenting clinical practices, making decisions for doctors, communicating with patients, and administering operations. The use of large language models can allow the processing and creation of human-like text, thereby increasing efficiency and minimizing the burden on healthcare personnel. Nevertheless, there are many privacy and security challenges associated with such usage. This is because of Protected Health Information (PHI), which is the sensitive information of patients that should be kept secret under the guidelines of healthcare privacy laws. There are chances of leaking PHI during any stage of data collection, model training, inference, storage, or even third-party integrations. Effective governance mechanisms thus become necessary in order to deploy AI responsibly within healthcare settings (Truby, 2020). The current research assesses PHI leaks from generative AI processes in healthcare as well as governance mechanisms aimed at enhancing privacy, compliance, and security within

AI deployment in healthcare.

### Methods

The qualitative research approach was chosen for this study. The design perfectly fit the study objectives. The research investigated privacy threats and governance processes in healthcare GenAI pipelines (Agrawal *et al.*, 2021). The emphasis was made on conceptual understanding, analysis of the policies, and the available literature. Qualitative design allowed providing in-depth explanation rather than quantitative results. In addition, the secondary data collection technique was used in the study. The use of secondary data was appropriate for the given research problem. There is plenty of information related to healthcare AI. The research analyzed academic journals, industry reports, healthcare regulations, and policies. Besides, healthcare guidelines were considered as a source of information. The research process followed an inductive approach. The inductive approach helped to discover patterns from the available evidence. The research did not involve testing any hypothesis. Findings were established from the evidence available in the literature review. Common themes were discovered by comparing various sources. The inductive approach enabled a better comprehension of

---

*AI/ML Engineer, Independent Researcher, USA*

*motanivinodrufus18@gmail.com*

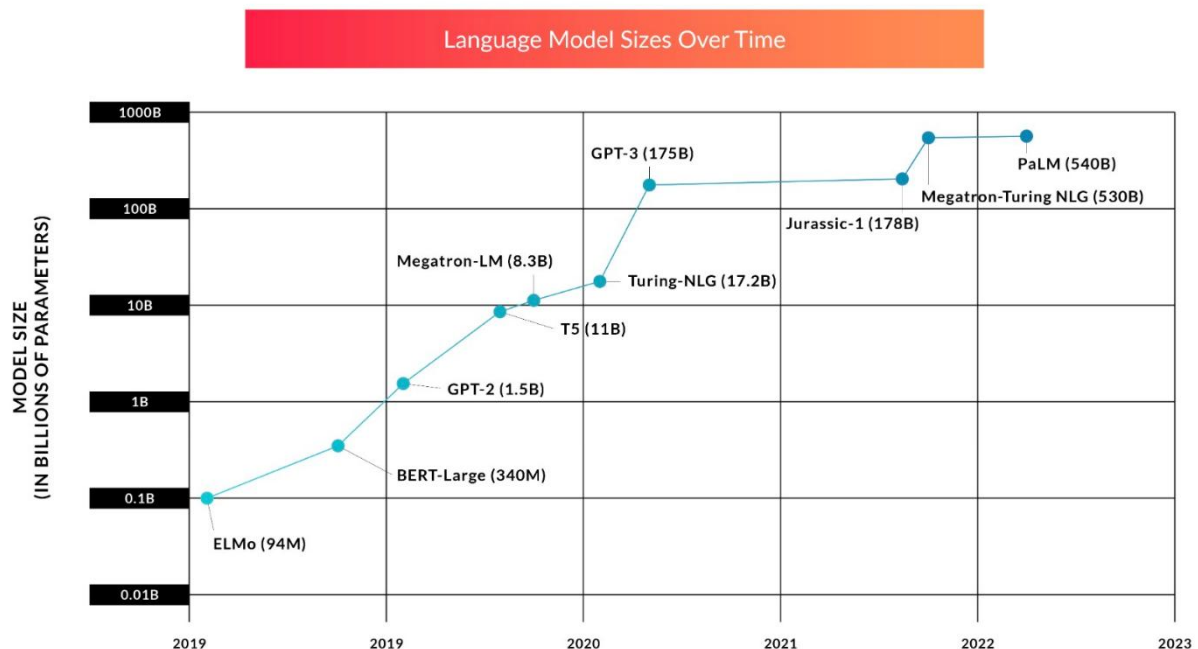
*ORCID: 0009-0005-3459-0859*

privacy problems. Thematic analysis technique was adopted to analyse the gathered evidence (Xu & Zammit, 2020). The thematic analysis involved organizing the evidence into meaningful themes. Some of the issues were PHII leaks, access control, consent management, auditing, and compliance. Theme was useful for making comparisons between various governance mechanisms. Theme also pointed out the gaps existing in the existing GenAI pipelines for healthcare. This method was selected since it was appropriate for the aims of the study (Labra *et al.*, 2020). The study required that the evidence be collected from various healthcare settings. It was also required to compare the governance mechanisms through various sources. Qualitative analysis helped in making an interpretation of the privacy risks. Thematic analysis provided organized results through various sources of evidence.

## Results

### *PHI Leakage Risks During Data Collection and Preprocessing*

It is evident from the reviewed literature that data acquisition and data preparation are the riskiest stages for the disclosure of patients' personal health information in healthcare generative AI pipelines. Electronic Health Records (EHR), clinical notes, lab results, imaging metadata, and patient IDs are fed into AI systems at these stages. Ineffective security mechanisms can make private information visible prior to applying privacy protections (Hu *et al.*, 2021). Insecure data acquisition systems are one of the major risks. Vulnerabilities in cloud storage configurations, insecure IAM policies, and unprotected ETL pipelines can lead to unauthorized access to raw data of patients.



**Figure 1: Language models increase in number of parameters**

(Source: Ramponi & O'Connor, 2023)

In addition, third-party AI vendors contribute significantly to increased risk due to sharing patients' information without adequate BAAs and processing controls. Free text clinical notes also pose an extra risk since they may include hard-to-detect identifiers. Healthcare data breach statistics prove the magnitude of the issue. According to the U.S. Office for Civil Rights, 686 healthcare data breaches of 500 or more records occurred in 2021, resulting in 44.9 million patient records being exposed. Network server weaknesses and cloud security issues comprised the most significant sources of breached PHI. Preprocessing procedures add additional vulnerabilities related to personal data (Dokuchaev, Maklachkova & Statev, 2020).

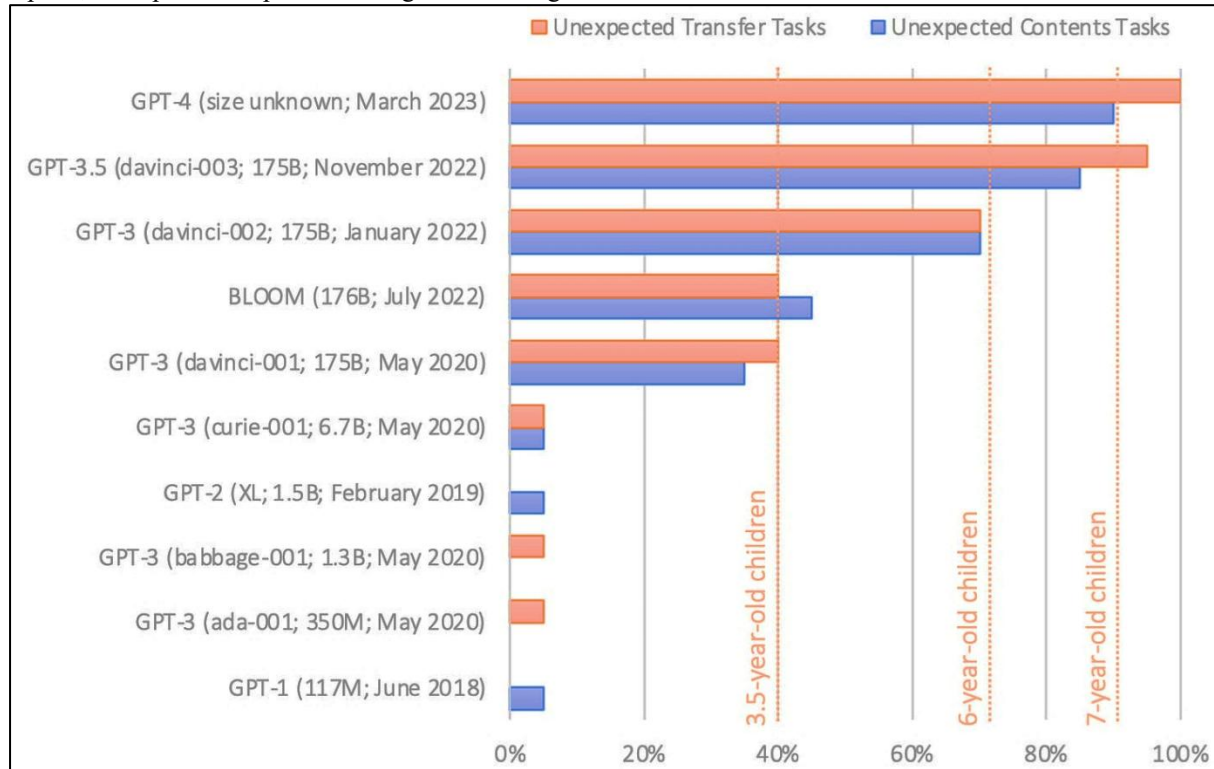
Numerous companies conduct normalization, feature extraction, or oversampling before dividing their datasets into training and testing. This procedure is called data leakage since the test data impacts the training process and raises the risk of memorization. Moreover, several studies state that deleting 18 HIPAA direct identifiers will not provide adequate protection as patients can be re-identified through quasi-identifiers such as age, ZIP code, medical history, and timestamps. The above-reviewed literature indicates the existence of four primary vectors for preprocessing attacks including re-identification attacks through quasi-identifiers, unauthorized access through weak APIs and file

permissions, PHI leaks through ETL logs, and data poisoning through altered HL7/FHIR records.

### **Privacy Vulnerabilities in Large Language Model Training and Fine-Tuning**

According to the study results, fine-tuning of healthcare LLMs contributes to the increased likelihood of PHI leakage due to the usage of clinical notes, discharge summaries, diagnostic reports, and patients' speech. Though fine-tuning

leads to better completion of medical tasks, it increases the likelihood of learning and repeating PHI during the inference stage (Lehmkuhl *et al.*, 2021). The important finding related to fine-tuning is the effect of duplicate data in the process of training. It has been found that healthcare-related text which occurs 30 or more times in the dataset has over 60% probability of being repeated in optimized extraction attack.



**Figure 2: GPT-1 and GPT-2 flunked text of LLM**

(Source: Lee &Trott, 2023)

The research findings have also noted backdoor attacks as privacy weaknesses. Attackers add secret triggers to fine-tuning datasets which enable PHI sharing when certain medical terms or pattern prompts are applied. Such attacks are still hard to discover because of high model accuracy. Another major finding relates to privacy collapse in fine-tuning cases. More training using specialized healthcare datasets leads to the weakening of current privacy defenses and causes unintended PHI sharing (Topaloglu *et al.*, 2021). The risk of such a problem is higher in specialized clinical language models trained with small institutional datasets. In general, the study reveals that duplication of clinical data, poisoned training datasets, and fine-tuning contribute to PHI leakages.

### **Inference-Time PHI Exposure Through Prompts and Model Outputs**

As indicated by the findings, inference time has been found to be a key phase in which there is exposure to PHI as generative AI performs tasks such as processing user inputs, retrieval of clinical data, and response generation in real time (Ahmed *et al.*, 2020). Inference time exposure is different from training-time leakage as it happens during interactions between healthcare users and the deployed AI model. There have been identified three key avenues of PHI exposure from the findings.

| Finding                                    | Numeric Data |
|--------------------------------------------|--------------|
| HIPAA PHI identifier categories            | 18           |
| Main PHI exposure avenues during inference | 3            |

|                                              |                                |
|----------------------------------------------|--------------------------------|
| Inference stage analysed                     | 1 (deployment/inference phase) |
| PHI exposure source: User prompts            | 1 of 3 avenues                 |
| PHI exposure source: Clinical data retrieval | 1 of 3 avenues                 |
| PHI exposure source: Model-generated outputs | 1 of 3 avenues                 |

**Table 1: Numeric Findings on PHI Exposure During Inference Time**

Firstly, healthcare providers and their patients tend to input unredacted information into the prompts. The 18 categories of PHI under HIPAA include information such as names, date of birth, address, medical record number, and biometric information. In cases where masked documents are not fed to the context window, the system has the potential to use sensitive clinical data in its response (Khan & Alotaibi, 2020). In such an event, there is even more risk to privacy in processes of diagnosing summary, prescription check, and clinical notes. Thirdly, PHI may be present in the outputs generated by the model. The prompt injection and extraction attacks have the ability to make the model divulge patient's sensitive data or produce

sensitive clinical text through prompting. Output filtering has been found to be very efficient in protecting data.

#### ***Governance Mechanisms for PHI Protection in Healthcare AI Pipelines***

As seen from the results, it is clear that governance measures have a crucial role to play in minimizing the risks of PHI leaks in healthcare generative AI pipelines. It is evident that those organizations that practice HIPAA-compliant governance ensure that privacy measures are used in the processes of data collection, training, inference, and storing (Eleanor, 2021).

| <b>Finding</b>    | <b>Evidence</b> | <b>Governance</b>  | <b>Protection</b>     | <b>Risk</b>            | <b>Outcome</b>  |
|-------------------|-----------------|--------------------|-----------------------|------------------------|-----------------|
| HIPAA Compliance  | 18 Identifiers  | Privacy Controls   | PHI Security          | Leak Reduction         | Secure Pipeline |
| Data Minimization | Less Data       | Identifier Removal | Privacy Protection    | Lower Exposure         | Reduced Risk    |
| De-identification | Synthetic Data  | Pseudonymization   | Identity Masking      | Re-identification Risk | Safer Training  |
| Cloud Governance  | 88% Adoption    | Cloud Controls     | Access Security       | Cloud Breaches         | Protected PHI   |
| Breach Prevention | US\$10.1M Cost  | Encryption BAAs    | Compliance Monitoring | Breach Losses          | Risk Mitigation |

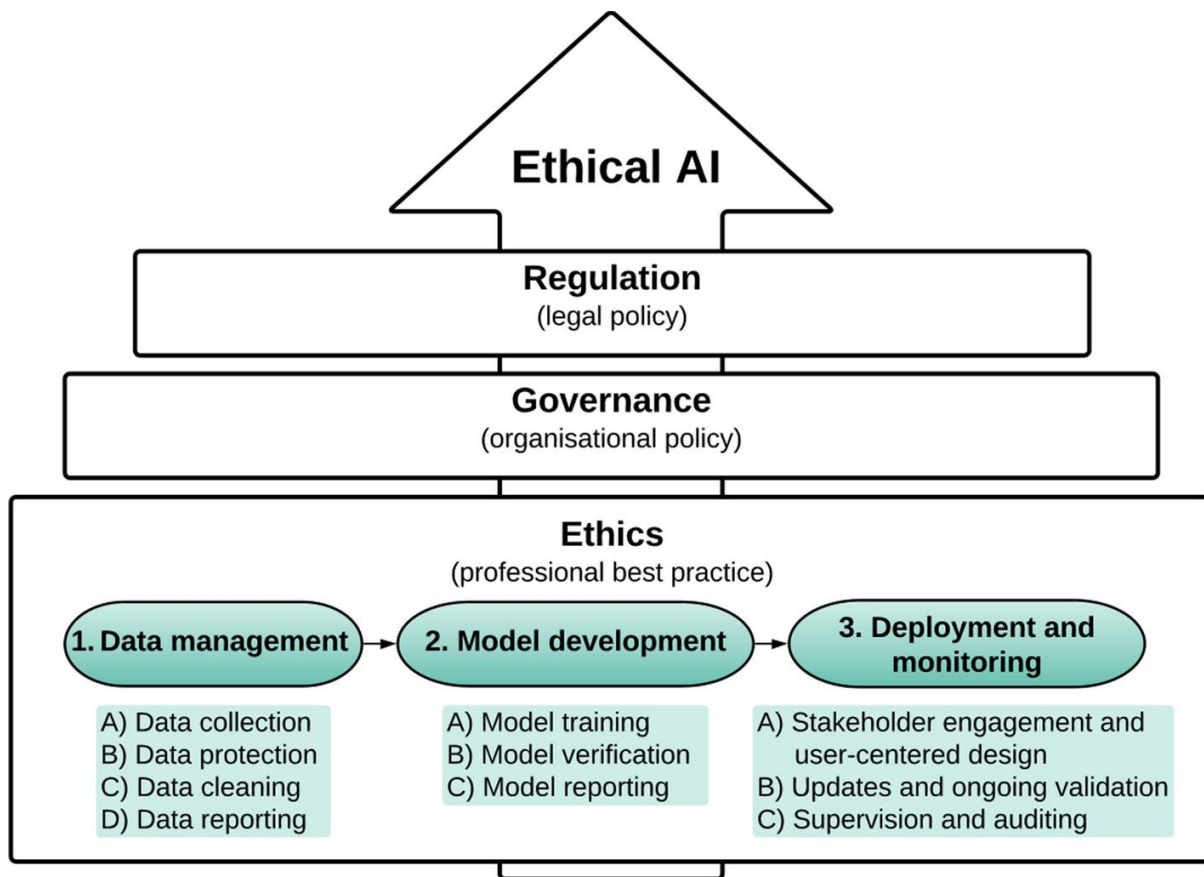
**Table 2: PHI Governance Findings in Healthcare Generative AI**

HIPAA calls for ensuring the protection of 18 types of PHI identifiers, including the patient name, address, medical record number, dates, and biometric information before the data gets into the AI system. Another result is that data minimization and de-identification are the most common governance measures used by healthcare organizations. They ensure the reduction of privacy risks through eliminating any unnecessary identifiers and using pseudonymous or synthetic data in the process of AI generation (Azizi *et al.*, 2021). The reviewed evidence revealed that 88% of healthcare firms had employed cloud-based AI applications, meaning that cloud governance was one of the crucial requirements to secure PHI. The average cost incurred by healthcare data breaches was estimated at US\$10.1 million. These findings provided support for the use of encryption, BAAs,

and compliance monitoring of healthcare AI pipelines.

#### ***Regulatory Compliance and Privacy Frameworks for Healthcare Generative AI***

According to the literature review, HIPAA is the most commonly utilized privacy regime within the field of generative AI in healthcare research. The scope of protection of HIPAA is the 18 categories of patient health information identifier, which includes patients' names, medical record number, addresses, dates, contact information, biometric information while processing and storing them using AI technologies. The analyzed literature reveals a fast-growing trend of interest in AI governance in the healthcare area (Guidance, 2021).



**Figure 3: Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers**  
 (Source: Solanki, Grundy & Hussain, 2023)

In 2020, there were 5,885 articles in PubMed related to the AI in the healthcare area, and 4,164 articles were published in 2021. There were 3,232 articles left according to inclusion criteria in 2020, in comparison with 2,182 articles in 2021, where 134 articles dealt with AI and COVID-19. Of those organizations experiencing high benefits from the application of AI, 57% had adopted end-to-end AI governance solutions, 53% had formal data governance processes, 49% had sought external validation and monitoring services, and 43% had adoptable explainable AI. Compared to others, these levels of adoption were between 21% and 32% (Personal Data Protection Commission, 2020).

#### **Organizational Challenges and Best Practices for Secure GenAI Deployment**

**Inadequate Workforce Training and AI Awareness:** Limited employee training was among the key organizational problems that affected the deployment of GenAI in a secure manner (Wannamaker, 2021). While only 34% of the surveyed organizations provided training to users about the working of AI algorithms, 39% of the

organizations included users during the designing and deployment process of AI.

**Weak Data Governance and Access Management:** Poor governance contributed to PHI vulnerability in the healthcare AI pipeline (Prado et al., 2020). It was discovered that 53% of the high-performing firms had a clear data governance process, while in other firms, the proportion was 27%. Likewise, 48% of the high performers had a scalable data labeling process against 22% of the other firms.

**Limited AI Model Validation and Monitoring:** The application of continuous validation was not applied uniformly throughout healthcare organizations. For example, 49% of high-performing organizations relied on third-party vendors for testing, verification, and monitoring of AI models compared to 23% of other organizations. Further, 43% utilized explainable AI, while adoption by other organizations was just 16%.

**Best Practices for Secure Healthcare GenAI Deployment:** The analysis showed that standardized governance was the most prevalent best practice. While 57% of top performing firms

had the end-to-end AI platform in place for data science, engineering, and deployment purposes, 45% of top firms created reusable AI models and another 45% had the habit of refreshing their AI technology stack (Bosch, Olsson & Crnkovic, 2021). This ensured improved security monitoring and PHI protection.

## Discussion

In the current research, risks of the PHI leakage and governing mechanisms in the generative AI pipeline in healthcare were studied using secondary literature. It was found that PHI exposure is present in different phases of AI operation, such as data gathering, data preprocessing, data modeling, fine-tuning, and inference. Healthcare data is very susceptible to such threats due to the nature of their usage as electronic health records, clinical data, diagnoses, and patients' communication that have sensitive information identifiers (Shah & Khan, 2020). From the perspective of technical vulnerabilities, increased risk emerges due to the utilization of unmasked clinical data by healthcare organizations, poor access control, insecure cloud data storage, and prompt-based generation systems. These results indicate that privacy threats exist not only during model development but also persist during real-time AI use. This research also reveals that governance solutions lower privacy threats when they are used in the context of the entire AI process. HIPAA protections, data minimization, de-identification, encryption, role-based access control, audit log, and vendor governance offer numerous layers of protection for patient health information (Xiang & Cai, 2021). Organizations that have developed AI governance practices demonstrate better monitoring and validation of models and increased model explainability in comparison with organizations that do not have many governance practices. In general, a successful implementation of healthcare generative AI is possible with the use of privacy-preserving technologies together with regulatory and organizational governance practices.

## Conclusion

There are considerable advantages of generative AI in healthcare with regard to the process of making decisions as well as efficient operation. However, there is an enormous PHI exposure risk associated with such technologies that span data acquisition, preprocessing, training, fine-tuning, and inference processes. The research revealed data exposure, model memorization, prompt exposure, and ineffective governance as principal privacy concerns. The key to securing health information lies in the following measures: HIPAA compliance, data minimization, de-identification, encryption, access control, logging, and monitoring.

## Reference List

- [1] Agrawal, N., Binns, R., Van Kleek, M., Laine, K., & Shadbolt, N. (2021, May). Exploring design and governance challenges in the development of privacy-preserving computation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-13). <https://dl.acm.org/doi/abs/10.1145/3411764.3445677>
- [2] Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database, 2020*, baaa010. <https://academic.oup.com/database/article-abstract/doi/10.1093/database/baaa010/5809229>
- [3] Azizi, Z., Zheng, C., Mosquera, L., Pilote, L., El Emam, K., & GOING-FWD Collaborators. (2021). Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ open, 11*(4), e043497. <https://bmjopen.bmj.com/content/11/4/e043497.abstract>
- [4] Bosch, J., Olsson, H. H., & Crnkovic, I. (2021). Engineering ai systems: A research agenda. *Artificial intelligence paradigms for smart cyber-physical systems*, 1-19. <https://www.igi-global.com/chapter/engineering-ai-systems/266130>
- [5] Dokuchaev, V. A., Maklachkova, V. V., & Statev, V. Y. (2020). Classification of personal data security threats in information systems. *T-Comm-Телекоммуникации и Транспорт, 14*(1), 56-60. <https://cyberleninka.ru/article/n/classification-of-personal-data-security-threats-in-information-systems>
- [6] Eleanor, H. (2021). Modernizing data security: Best practices for compliance with US and international privacy regulations. *International Journal of Trend in Scientific Research and Development, 5*(4), 1881-1894. <http://eprints.umsida.ac.id/16095/>
- [7] Guidance, W. H. O. (2021). Ethics and governance of artificial intelligence for health. *World Health Organization, 1*. <https://journals.sagepub.com/doi/abs/10.3233/SHTI200542>
- [8] Hu, Y., Kuang, W., Qin, Z., Li, K., Zhang, J., Gao, Y., ... & Li, K. (2021). Artificial intelligence security: Threats and countermeasures. *ACM Computing Surveys (CSUR), 55*(1), 1-36. <https://dl.acm.org/doi/abs/10.1145/3487890>
- [9] Khan, Z. F., & Alotaibi, S. R. (2020). Applications of artificial intelligence and big data analytics in m-health: A healthcare system

- perspective. *Journal of healthcare engineering*, 2020(1), 8894694. <https://onlinelibrary.wiley.com/doi/abs/10.1155/2020/8894694>
- [10] Labra, O., Castro, C., Wright, R., & Chamblas, I. (2020). Thematic analysis in social work: A case study. *Global social work-cutting edge issues and critical reflections*, 10(6), 1-20. <https://books.google.com/books?hl=en&lr=&id=UBT8DwAAQBAJ&oi=fnd&pg=PA183&dq=Thematic+analysis+technique+was+adopted+to+analyse+the+gathered+evidence&ots=ASmUBBV2HI&sig=dBhCq9-yxyzgtC3nQrYKLv6gs1Y>
- [11] Lee, T., B., & Trott, S., 2023. Large language models, explained with a minimum of math and jargon. <https://www.understandingai.org/p/large-language-models-explained-with>
- [12] Lehmkuhl, R., Mishra, P., Srinivasan, A., & Popa, R. A. (2021). Muse: Secure inference resilient to malicious clients. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 2201-2218). <https://www.usenix.org/conference/usenixsecurity21/presentation/lehmkuhl>
- [13] Personal Data Protection Commission. (2020). Model AI Governance Framework (2020). <https://openresearch-repository.anu.edu.au/bitstreams/00df4e3b-8176-4497-8384-a55e72c20f56/download>
- [14] Prado, M. D., Su, J., Saeed, R., Keller, L., Vallez, N., Anderson, A., ... & Pazos, N. (2020). Bonseyes ai pipeline—bringing ai to you: End-to-end integration of data, algorithms, and deployment tools. *ACM Transactions on Internet of Things*, 1(4), 1-25. <https://dl.acm.org/doi/abs/10.1145/3403572>
- [15] Ramponi, M., & O'Connor, R., 2023. The Full Story of Large Language Models and RLHF. <https://www.assemblyai.com/blog/the-full-story-of-large-language-models-and-rlhf>
- [16] Shah, S. M., & Khan, R. A. (2020). Secondary use of electronic health record: Opportunities and challenges. *IEEE access*, 8, 136947-136965. <https://ieeexplore.ieee.org/abstract/document/9146114/>
- [17] Solanki, P., Grundy, J., & Hussain, W. (2023). Operationalising ethics in artificial intelligence for healthcare: A framework for AI developers. *AI and Ethics*, 3(1), 223–240. <https://doi.org/10.1007/s43681-022-00195-z>
- [18] Topaloglu, M. Y., Morrell, E. M., Rajendran, S., & Topaloglu, U. (2021). In the pursuit of privacy: the promises and predicaments of federated learning in healthcare. *Frontiers in Artificial Intelligence*, 4, 746497. <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.746497/full>
- [19] Truby, J. (2020). Governing artificial intelligence to benefit the UN sustainable development goals. *Sustainable Development*, 28(4), 946-959. <https://onlinelibrary.wiley.com/doi/abs/10.1002/sd.2048>
- [20] Wannamaker, K. (2021). Situated Self-Tracking: Ideating, Designing, and Deploying Dedicated User-driven Personal Informatics Systems. <https://ucalgary.scholaris.ca/items/77a91ad5-e560-4ae2-a5a6-b18388c64723>
- [21] Xiang, D., & Cai, W. (2021). Privacy protection and secondary use of health data: strategies and methods. *BioMed Research International*, 2021(1), 6967166. <https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/6967166>
- [22] Xu, W., & Zammit, K. (2020). Applying thematic analysis to education: A hybrid approach to interpreting data in practitioner research. *International journal of qualitative methods*, 19, 1609406920918810. <https://journals.sagepub.com/doi/abs/10.1177/1609406920918810>