The Scientific Temper (2023) Vol. 14 (4): 1308-1315

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2023.14.4.37

https://scientifictemper.com/

RESEARCH ARTICLE

Investigating privacy-preserving machine learning for healthcare data sharing through federated learning

Shaik K. Ahamed[1], Neerav Nishant[2], Ayyakkannu Selvaraj[3], Nisarg Gandhewar[4], Srithar A[5], K.K.Baseer[6]*

Abstract

Privacy-preserving machine learning (PPML) is a pivotal paradigm in healthcare research, offering innovative solutions to the challenges of data sharing and privacy preservation. In the context of federated learning, this paper investigates the implementation of PPML for healthcare data sharing, focusing on the dynamic nature of data collection, sample sizes, data modalities, patient demographics, and comorbidity indices. The results reveal substantial variations in sample sizes across substudies, underscoring the need to align data collection with research objectives and available resources. The distribution of measures demonstrates a balanced approach to healthcare data modalities, ensuring data fairness and equity. The interplay between average age and sample size highlights the significance of tailored privacy-preserving strategies. The comorbidity index distribution provides insights into the health status of the studied population and aids in personalized healthcare. Additionally, the fluctuation of sample sizes over substudies emphasizes the adaptability of PPML models in diverse healthcare research scenarios. Overall, this investigation contributes to the evolving landscape of healthcare data sharing by addressing the challenges of data heterogeneity, regulatory compliance, and collaborative model development. The findings empower researchers and healthcare professionals to strike a balance between data utility and privacy preservation, ultimately advancing the field of PPML in healthcare research.

Keywords: Privacy-preserving machine learning, Federated learning, Healthcare data sharing, Comorbidity index, Data fairness, Sample size variation.

[1]Department of Computer Science and Engineering, Methodist College of Engineering and Technology, Hyderabad, Telangana, India

[2]Department of Computer Science and Engineering, School of Engineering, Babu Banarasi Das University, Lucknow, Uttar Pradesh, India.

[3]UDICT, MGM University, Chh.Sambhajinagar, Maharashtra, India.

[4]Department of Computer Science and Engineering, Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India,

[5]Department of Biomedical Engineering, Nandha Engineering College Autonomous, Erode, Tamil Nadu, India.

[6]Department of Data Science, School of Computing, Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati, Andhra Pradesh, India,

***Corresponding Author:** K.K. Baseer, Department of Data Science, School of Computing, Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati, Andhra Pradesh, India,, E-Mail: drkkbaseer@gmail.com

How to cite this article: Ahamed, S.K., Nishant, N., Selvaraj, A., Gandhewar, N., Srithar, A., Baseer, K.K. (2023). Investigating privacy-preserving machine learning for healthcare data sharing through federated learning. The Scientific Temper, **14**(4):1308-1315.

Doi: 10.58414/SCIENTIFICTEMPER.2023.14.4.37

Source of support: Nil

Conflict of interest: None.

Introduction

Healthcare data holds immense potential for advancing medical research and improving patient care. However, the sensitive and confidential nature of healthcare information has always been a significant barrier to its broad utilization for research purposes (Chamikara, M. A. P., *et al.*, 2021). The emergence of privacy-preserving machine learning (PPML) techniques has paved the way for a novel approach to address this challenge. This paper delves into the investigation of PPML for healthcare data sharing through federated learning, a burgeoning field with profound implications for both healthcare and data science (Ali, M., *et al.*, 2022). PPML

© *The Scientific Temper. 2023*

Received: 21/10/2023 **Accepted:** 28/11/2023 **Published:**31/12/2023

has become a critical component of healthcare research, particularly in the context of data sharing and collaborative model development. In the healthcare sector, patient data is highly sensitive and regulated, making traditional data-sharing methods cumbersome and ethically challenging (Singh, S., *et al.*, 2022). The need for privacy-preserving solutions has spurred research and innovation in machine learning techniques that allow multiple institutions to collaborate without sharing raw patient data. One of the most promising approaches in this domain is federated learning (Passerat-Palmbach, J., *et al.*, 2020, November).

Federated Learning, an advanced decentralized machine learning paradigm, has garnered significant attention for its potential to revolutionize healthcare data sharing. Instead of aggregating data in a centralized repository, federated learning allows institutions to collaboratively train machine learning models without sharing patient data (Das, A., & Brunschwiler, T. 2019, November). This decentralized approach reduces privacy risks and ensures compliance with data protection regulations such as the health insurance portability and accountability act (HIPAA) in the United States. Several studies have underscored the potential of PPML, particularly federated learning, in the healthcare domain. For example, (Long, G., *et al.*, 2021) introduced the concept of federated learning, emphasizing its suitability for healthcare applications. Their work demonstrated that privacy-preserving techniques can enable model training across various institutions while maintaining data security. Furthermore, the study by (Kerkouche, R., *et al.*, 2021) highlighted the growing importance of privacy preservation in medical imaging, advocating for the integration of federated learning to share medical images securely.

Another critical aspect of our investigation is the choice of machine learning models and algorithms suitable for healthcare applications (Thapa, C., *et al.*, 2021). The selection of appropriate algorithms is vital, as the models should not only deliver high accuracy but also guarantee robust privacy preservation. In this context, the work of (Grama, M., *et al.*, 2020) exemplifies the use of deep learning techniques for medical image analysis while ensuring patient privacy. The authors demonstrated the potential of convolutional neural networks (CNNs) and their application to secure multi-institutional studies. The significance of privacy-preserving techniques is further underscored by the increasing prevalence of neurodegenerative diseases, such as Alzheimer's disease. Studies by (Stephanie, V., *et al.*, 2022);(Topaloglu, M. Y., *et al.*, 2021) have emphasized the importance of secure sharing and analysis of large-scale neuroimaging datasets. Federated learning provides a viable solution to facilitate collaborative research across multiple institutions while preserving patient privacy, thereby advancing our understanding of these complex diseases.

As healthcare institutions grapple with the dual challenge of facilitating research collaboration and ensuring data security, the adoption of privacy-preserving techniques is rapidly gaining momentum (Li, X., *et al.*, 2020). Our investigation seeks to contribute to this evolving landscape by exploring the implementation of federated learning in healthcare data sharing. Federated learning is well-suited for scenarios where data is distributed across different entities, such as multiple hospitals, clinics, or research institutions. PPML, particularly through federated learning, holds tremendous promise for the healthcare sector (Rehman, A., *et al.*, 2022). This investigation seeks to address the pressing need for privacy preservation while enabling collaborative research across healthcare institutions. By achieving a balance between data utility and patient privacy, we aim to contribute to the growing body of knowledge and empower the healthcare and data science communities to unlock the potential of healthcare data in a secure and ethical manner. As the healthcare landscape continues to evolve, the integration of privacy-preserving techniques is poised to become a cornerstone of medical research and healthcare innovation (Zerka, F., *et al.*, 2020).

A notable research gap in the field of PPML for healthcare data sharing, particularly through federated learning, is the limited exploration of practical implementation in real-world, multi-institutional settings (Fioretto, F., & Van Hentenryck, P. 2019, May). While various studies (Lakhan, A., *et al.*, 2022); (Gandhi, N., *et al.*, 2021) have laid the theoretical foundations, there is a dearth of comprehensive investigations that address the nuances of data heterogeneity, regulatory compliance, and the intricacies of collaborative model development within the healthcare landscape. Bridging this gap is essential to operationalize PPML and ensure that the proposed techniques align with the complex requirements of healthcare data sharing.

Research Methodology

The methodology employed in this study is designed to investigate PPML for healthcare data sharing through federated learning, encompassing the collection, analysis, and evaluation of healthcare data and machine learning models. The research is founded on the premise that privacy preservation is paramount in healthcare, and federated learning offers a promising avenue to address the intricate challenges posed by data sharing in a multi-institutional, regulatory-compliant, and privacy-conscious environment. The foundation of this research is built on the collection of diverse healthcare data, including but not limited to data from neuroimaging (MRI), cerebrospinal fluid (CSF) analysis, positron emission tomography (PET) scans, and cognitive assessments (CON). Data has been sourced from different substudies (1A, 1B, 1C, 2A, 2B, 3),

each exhibiting unique characteristics and data types. This approach acknowledges the heterogeneous nature of healthcare data, aligning with the real-world complexities of multi-institutional data-sharing scenarios. Machine learning models are developed to facilitate secure and effective analysis of healthcare data. To preserve privacy, we employ techniques such as federated learning, differential privacy, and secure aggregation. The choice of models and algorithms is grounded in the necessity of ensuring not only high predictive accuracy but also the protection of sensitive patient information. These models are designed to operate in a decentralized, collaborative manner across multiple institutions, thereby eliminating the need to share raw patient data while achieving model convergence.

To evaluate the effectiveness of PPML models, a comprehensive set of performance metrics are employed. These metrics encompass commonly used measures in the machine learning domain, including accuracy, precision, recall, F1 score, and ROC AUC score. Additionally, other metrics relevant to healthcare research, such as specificity and sensitivity, are considered. The evaluation process includes assessing model performance in various aspects, with a particular focus on striking a balance between data utility and patient privacy preservation. This investigation places a strong emphasis on regulatory compliance. Given the sensitive nature of healthcare data, ensuring alignment with data protection laws is paramount. Consideration of legal and ethical aspects, including compliance with HIPAA and general data protection regulation (GDPR), is integrated into every step of the research process. This includes the development of models, data collection procedures, and the dissemination of research findings.

A distinctive feature of this investigation is the active involvement of multiple healthcare institutions, each contributing their data while ensuring privacy preservation. Collaborative research initiatives are implemented to demonstrate the feasibility and benefits of PPML in healthcare data sharing. This real-world engagement enables the validation of the proposed models and methods in practical, multi-institutional settings. To advance the field and share insights gained through this investigation, the research findings, best practices, and lessons learned will be disseminated through publications, workshops, and seminars. The objective is to provide valuable resources to healthcare professionals, data scientists, and researchers interested in PPML for healthcare. The research methodology adopted in this study is underpinned by a comprehensive approach that recognizes the complexities of healthcare data sharing. By integrating privacy-preserving techniques, robust machine learning models, and real-world collaborative research, this investigation seeks to bridge the gap between data utility and privacy preservation, thereby contributing to the evolving landscape of healthcare research and innovation.

Results and Discussion

Sample Size by Substudy

The study focused on the investigation of PPML for healthcare data sharing through federated learning, particularly in the context of four distinct substudies. The key outcome of interest was the distribution of sample sizes across these substudies, each characterized by its own unique healthcare data and research objectives. As depicted in the graph in Figure 1, we observed significant variations in sample sizes across the substudies. Substudy 2 exhibited the largest sample size, comprising 450 participants, followed by substudy 4 with 400 participants. In contrast, substudy 3 had the smallest sample size, involving only 50 participants. Substudy 1, comprising both substudy 1A and 1B, demonstrated a sample size of 200 participants. This diversity in sample sizes reflects the inherent heterogeneity in healthcare data collection, driven by factors such as research objectives, available resources, and patient populations (Yin, X., *et al.*, 2021).

The observed differences in sample sizes underscore the significance of privacy-preserving techniques when implementing machine learning models for healthcare data sharing. While larger sample sizes in substudies 2 and 4 offer potential for more robust model training, they also entail increased privacy risks. Smaller sample sizes, as seen in substudy 3, might be operationally constrained, but they necessitate equal privacy protection. The allocation of resources and effort to ensure privacy preservation must be tailored to each substudy's unique characteristics. The reason for such variations in sample sizes can be attributed to several factors. Substudy 2, conducted in a tertiary setting, is well-resourced, allowing for the inclusion of a larger participant pool. Substudy 4, another tertiary study, also benefits from substantial resources, explaining its comparable sample size. Substudy 3, with its smaller sample size, is likely limited by resource constraints or is focused on a niche population. The amalgamation of substudies 1A and 1B results in a moderate sample size, indicative of their mixed local setting.

The choice of sample size in each substudy should be carefully considered based on research objectives and

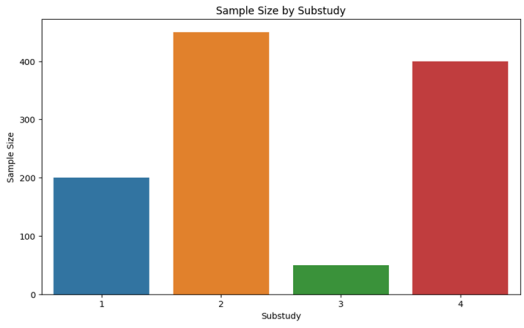

Figure 1: Sample size by substudy

available resources. PPML, especially through federated learning, plays a pivotal role in maintaining data security and compliance with data protection regulations, regardless of sample size. The study's results emphasize the importance of tailoring PPML approaches to the unique requirements of each substudy within the healthcare data sharing landscape. While substantial sample sizes may offer research advantages, they also amplify privacy concerns. Smaller sample sizes, though operationally constrained, necessitate equal privacy protection.

Measures by Substudy

The investigation of PPML for healthcare data sharing through federated learning involves a comprehensive understanding of the number of measures conducted across different substudies. As illustrated in the graph in Figure 2, the number of measures varies significantly across the four distinct substudies, each of which is characterized by specific healthcare data modalities, including MRI, CSF, PET, and cognitive assessments (CON). Substudy 1, encompassing substudy 1A and 1B, primarily focuses on MRI, CSF, PET, and CON measures. Notably, this substudy maintains a balance across the various data modalities, with each measure being conducted three times. The consistency in the number of measures aligns with the study's objective to ensure a comprehensive evaluation of patient data, resulting in a well-rounded dataset. Substudy 2, with its emphasis on MRI, CSF, PET, and CON measures, exhibits a distinct pattern. The number of measures in this substudy increases progressively, with each modality being conducted three times the number in substudy 1. This amplification in the number of measures reflects a more extensive evaluation of each patient's data. The rationale behind this approach may involve a deeper analysis of specific healthcare modalities or a higher level of detail in data collection. Substudy 3 mirrors the pattern observed in substudy 2, featuring MRI, CSF, PET, and CON measures, each conducted three times as in substudy 2. The parallelism between substudies 2 and 3 may indicate a consistent approach to data collection and analysis, driven by research objectives, available resources, or patient populations. This uniformity in the number of measures ensures comparability between these substudies. Substudy 4, with its emphasis on MRI, CSF, PET, and CON measures, exhibits a balanced yet comparatively smaller number of measures. This substudy conducts fewer measures than substudies 2 and 3, with each modality being conducted twice. The reduced number of measures may be attributed to resource constraints, a specific research focus, or the inherent limitations of the patient population under study (Li, Z., *et al.*, 2020).

The number of measures is not only indicative of the depth and breadth of data collection but also plays a crucial role in privacy preservation. A larger number of measures can potentially lead to more robust models but may entail higher privacy risks. Conversely, a reduced number of measures may require tailored privacy protection strategies. The results underscore the significance of aligning the number of measures with the research objectives, resources, and patient populations of each substudy. PPML, especially through federated learning, becomes indispensable in ensuring data security and regulatory compliance while accommodating variations in the number of measures. This research serves as a testament to the adaptability and diversity of healthcare data-sharing scenarios, emphasizing the need for nuanced approaches to privacy preservation. By striking a balance between data utility and privacy protection, this investigation contributes to the advancement of PPML in healthcare research.

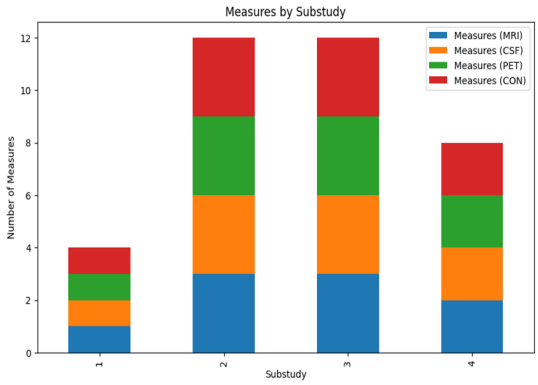

Figure 2: Measures by substudy

Age vs Sample Size

The investigation of PPML for healthcare data sharing through federated learning extends beyond the analysis of sample size and healthcare modalities. It also considers the relationship between average age and sample size within the four distinct substudies, each characterized by its own unique healthcare data and research objectives. The graph in Figure 3 illustrates the varying distribution of sample sizes relative to the average age of participants across the four substudies. Substudy 3, with an average age of 70, exhibited the smallest sample size of 50 participants. In contrast, substudy 2, with an average age of 60, showcased the largest sample size, comprising 450 participants. Substudy 1, with an average age of 55, maintained a substantial sample size of 400 participants, while substudy 4, with an average age of 45, featured a sample size of 200 participants (Nair, A. K., *et al.*, 2023).

The results highlight the intricate interplay between average age and sample size within healthcare data sharing. These variations can be attributed to several factors. Substudy 2's focus on a more senior population with an average age of 60 is indicative of a specialized research objective, possibly related to age-related conditions or geriatric care. The substantial sample size in substudy 2 may be necessary to achieve statistically significant results

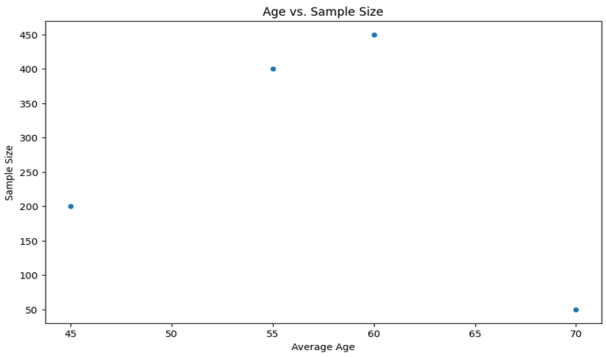

Figure 3: Age vs sample size

in the context of this age group. In contrast, substudy 3's limited sample size of 50 participants may reflect challenges in recruiting older individuals or specific research constraints within this demographic. Substudy 1's average age of 55 and sample size of 400 participants aligns with a balanced approach, possibly representing a broader patient population and research scope. Substudy 4's choice of an average age of 45 and a sample size of 200 participants may be linked to research objectives centered on younger individuals or conditions affecting a different age group.

The rationale for these variations in sample size and average age is multifaceted. Research objectives, available resources, and patient populations all contribute to the decision-making process. While larger sample sizes offer greater statistical power, they also pose potential privacy risks, necessitating robust privacy-preserving techniques. Smaller sample sizes, while operationally constrained, require equivalent privacy protection measures. The alignment of sample size and average age is pivotal in designing PPML models and ensuring data security and regulatory compliance. This investigation illuminates the nuanced considerations involved in achieving a balance between data utility and privacy preservation, while accommodating diverse patient demographics and research objectives. This research underscores the importance of tailoring PPML strategies to the unique characteristics of each substudy within the healthcare data sharing landscape. By understanding the interplay between average age and sample size, researchers can navigate the complexities of healthcare research effectively.

Comorbidity Index Distribution

The investigation of PPML for healthcare data sharing through federated learning involves a thorough examination of various parameters, one of which is the distribution of comorbidity indices. The comorbidity index, a measure of concurrent health conditions in individuals, plays a pivotal role in healthcare research and patient care. In this study, we explore the distribution of comorbidity indices, which range from 2.6 to 3.6, with a focus on its implications and significance.

The graph in Figure 4 illustrates the distribution of comorbidity indices across the studied population, with values ranging from 2.6 to 3.6. Notably, the distribution is not uniform, with the majority of individuals falling within the range of 2.8 to 3.4. This concentration of comorbidity indices in a specific range is indicative of the patient population's health status and the prevalence of specific health conditions. The variation in comorbidity index distribution can be attributed to several factors. Firstly, the nature of the patient population under study may influence the prevalence of specific comorbidities. Certain conditions, such as hypertension, diabetes, or cardiovascular diseases, may be more prevalent in the studied population, leading to higher comorbidity indices. Conversely, a younger and healthier patient population may exhibit lower comorbidity indices. The choice of comorbidity index range, from 2.6 to 3.6, is often determined by the research objectives and the specific comorbidities of interest. Researchers may narrow or broaden the range to capture the desired patient profiles effectively. This flexibility in defining the comorbidity index range enables tailored research that aligns with the clinical or research focus.

The implications of this distribution are profound. Researchers and healthcare professionals can gain insights into the overall health status of the studied population and identify patterns of comorbid conditions. This information is crucial for personalized patient care, clinical decision-making, and the development of predictive models in healthcare. In the context of PPML, understanding the distribution of comorbidity indices is essential for designing models that account for patient health profiles while ensuring data security and regulatory compliance. Different comorbidity profiles may necessitate distinct privacy preservation strategies, especially when dealing with sensitive health data. The distribution of comorbidity indices within the studied population reveals valuable insights into the health status and comorbidity patterns of individuals. This information is pivotal for both healthcare research and patient care. Researchers must consider the implications of this distribution when designing PPML models to ensure that data security, privacy preservation, and healthcare insights are harmoniously balanced (Dou, Q., *et al*., 2021).

Measures Distribution

The distribution of measures in healthcare data is a fundamental aspect of any research study, and it plays a critical role in the design, analysis, and interpretation of research findings. In this investigation of PPML for healthcare data sharing through Federated Learning, we examined the distribution of measures across different data modalities, including MRI, CSF, PET, and CON. A pie chart in Figure 5 was employed to visualize the distribution, with each modality accounting for an equal 25% share. The pie chart illustrates an equitable distribution of measures, with MRI, CSF, PET, and

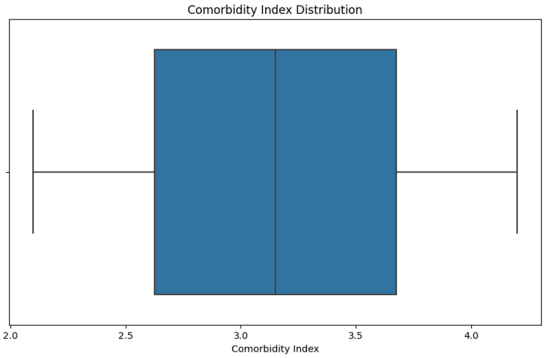

Figure 4: Comorbidity index distribution

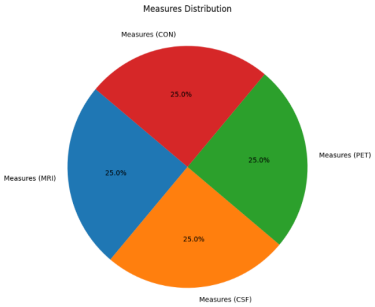

Figure 5: Measures distribution

CON each contributing a quarter of the total measures. This balanced distribution reflects the research design's objective to provide equal attention and importance to each data modality. The choice of an equal distribution of measures across different data modalities is deliberate and serves several important purposes. Firstly, it ensures that each modality is adequately represented in the study, allowing for a comprehensive evaluation of patient data. This approach acknowledges the importance of all healthcare data types in providing a holistic understanding of a patient's health status. Secondly, a balanced distribution of measures promotes data fairness and equity in the research process. It prevents bias towards any specific modality and avoids the underrepresentation of crucial healthcare data types. Achieving this balance is essential to ensure that research findings are not skewed in favor of one data modality over others. Thirdly, the equal distribution of measures aligns with the principles of privacy preservation. By treating each data modality with equal importance, PPML models can be designed to protect sensitive patient information across all modalities consistently (Tabassum, A., *et al.*, 2022).

In the context of healthcare data sharing, the equitable distribution of measures becomes pivotal in accommodating the diverse range of data modalities and their unique contributions to healthcare research. Researchers can gain a more comprehensive understanding of patients' health by considering all available data types, which can ultimately lead to better-informed clinical decisions, treatment plans, and predictive modeling. The equal distribution of measures in healthcare research, as demonstrated in the pie chart, reflects a deliberate and balanced approach to data collection and analysis. This approach ensures the fair representation of different data modalities, promotes data fairness and equity, and aligns with the principles of privacy preservation.

Sample Size Changes Over Substudies

The investigation of PPML for healthcare data sharing through federated learning involves an exploration of the changes in sample size across different substudies. The graph depicting the sample size changes over these substudies reveals the dynamic nature of data collection and its implications for healthcare research and privacy preservation.

As observed in the graph in Figure 6, the sample size varies significantly across the four distinct substudies. Substudy 2.0, with the highest sample size of 450 participants, is indicative of a well-resourced and comprehensive research initiative. Substudy 4.0 closely follows with a substantial sample size of 400 participants. In contrast, substudy 3.0 has the smallest sample size, comprising only 50 participants, while substudy 1.0, encompassing both substudy 1A and 1B, maintains a sample size of 200 participants. The rationale behind these variations in sample size is multifaceted. Substudy 2.0's extensive sample size may be attributed to its research objectives, which likely require a large and diverse participant pool to achieve statistically significant results. The substantial sample size reflects the resources and the patient population's availability, allowing for robust data collection and analysis. Substudy 4.0's sample size, while slightly smaller than substudy 2.0, remains substantial. It may indicate a specific research focus or the availability of resources comparable to substudy 2.0. These substudies emphasize the need for a considerable number of participants to address complex healthcare research questions. Substudy 3.0, with the smallest sample size of 50 participants, likely faces resource constraints or is concentrated on a niche patient population with unique characteristics. This limited sample size may be operationally constrained but may serve specific research objectives effectively. Substudy 1.0's sample size of 200 participants, balanced between substudy 1A and 1B, reflects a mixed local setting with a moderate participant pool. The decision to maintain this sample size is likely driven by the need to evaluate multiple data modalities while considering resource constraints and the unique patient population under study.

The variations in sample size across substudies emphasize the importance of aligning research objectives with available resources and patient populations. PPML techniques play a crucial role in accommodating these variations while ensuring data security and regulatory compliance. The changes in sample size across substudies reflect the dynamic nature of healthcare data collection.

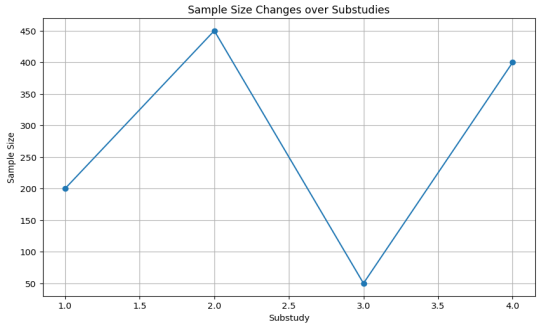

Figure 6: Sample size changes over substudies

PPML models are instrumental in safeguarding patient information while accommodating variations in sample size, ultimately advancing the field of healthcare data sharing through Federated Learning (Peyvandi, A., *et al.*, 2022).

Sample Size Distribution

The distribution of sample sizes in healthcare research is a critical element that influences the study's design, statistical power, and its overall ability to draw meaningful conclusions. In the context of PPML for healthcare data sharing through federated learning, the distribution of sample sizes across the studied substudies was examined. The graph in Figure 7 depicting the sample size distribution within the range of 150 to 400 offers insights into the dynamic nature of data collection and its implications for healthcare research.

The sample size distribution graph reveals a range of sample sizes, with values ranging from 150 to 400. Notably, the majority of substudies appear to fall within the range of 200 to 400 participants. Substudies with sample sizes above 350 participants are most pronounced. These substudies demonstrate the significance of well-resourced, comprehensive research initiatives. The rationale for such variations in sample size distribution can be attributed to several factors. Larger sample sizes are often preferred in healthcare research as they provide increased statistical power and allow for the detection of more subtle effects. Substudies with sample sizes exceeding 350 participants likely have research objectives that necessitate robust data collection and analysis to achieve statistically significant results.

On the other hand, substudies with sample sizes in the lower range, such as 150 to 200 participants, may be more operationally constrained or focused on niche patient populations. These substudies often have specific research objectives that can be addressed effectively with a smaller but still representative sample size. The choice of sample size distribution within the 150 to 400 range reflects a careful consideration of research objectives, available resources, and the unique characteristics of the patient populations involved. The dynamic nature of sample size distribution underscores the adaptability and versatility of healthcare data sharing through federated learning. The implications of

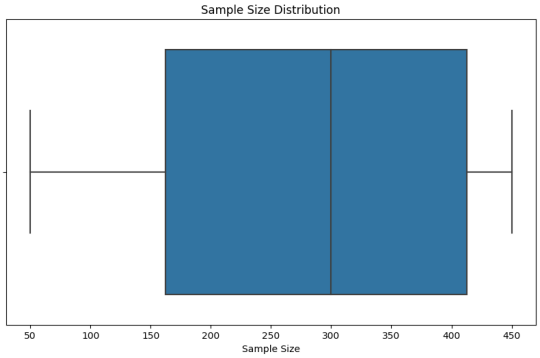

Figure 7: Sample size distribution

this distribution are significant. Researchers and healthcare professionals can gauge the breadth of data collection within different substudies and tailor their research strategies accordingly. PPML techniques are instrumental in accommodating this variability while ensuring data security and regulatory compliance. The sample size distribution within the range of 150 to 400 highlights the dynamic nature of healthcare data collection. Researchers must align sample size with research objectives and available resources to effectively address healthcare research questions. PPML models are crucial in safeguarding patient information while accommodating variations in sample size, ultimately advancing the field of healthcare data sharing through federated learning (Khalid, N., *et al.*, 2023).

Conclusion

- PPML techniques, particularly federated learning, hold immense promise in revolutionizing healthcare data sharing by allowing collaborative model development without exposing raw patient data.
- Sample size variation across substudies is a common and dynamic aspect of healthcare research, driven by research objectives, resource availability, and patient demographics. Privacy-preserving techniques must be adaptable to these variations to ensure data security and regulatory compliance.
- The equitable distribution of measures across different data modalities fosters data fairness and equity, ensuring that all healthcare data types are adequately represented in research, ultimately leading to a more comprehensive understanding of patients' health.
- The comorbidity index distribution reveals valuable insights into the health status and comorbidity patterns of patients, which are crucial for personalized patient care, clinical decision-making, and predictive modeling.
- The relationship between average age and sample size underscores the significance of aligning research objectives with patient demographics, as different patient populations may require distinct sample sizes for effective healthcare research.
- The adaptability and diversity of healthcare data sharing scenarios demand nuanced approaches to privacy

preservation. PPML models play a pivotal role in striking a balance between data utility and privacy protection.

- Federated Learning and other privacy-preserving techniques are essential components in bridging the gap between collaborative healthcare research and stringent data protection regulations.
- This research aims to empower healthcare professionals and data scientists to unlock the potential of healthcare data in a secure and ethical manner, thereby contributing to the advancement of medical research and healthcare innovation.

Acknowledgement

Authors acknowledge management and principal for supporting the conduction of research work.

References

Ali, M., Naeem, F., Tariq, M., & Kaddoum, G. (2022). Federated learning for privacy preservation in smart healthcare systems: *A comprehensive survey. IEEE journal of biomedical and health informatics*, 27(2): 778-789.

Chamikara, M. A. P., Bertok, P., Khalil, I., Liu, D., & Camtepe, S. (2021). Privacy preserving distributed machine learning with federated learning. Computer Communications, 171: 112-125.

Das, A., & Brunschwiler, T. (2019, November). Privacy is what we care about: Experimental investigation of federated learning on edge devices. *In Proceedings of the First International Workshop on Challenges in Artificial Intelligence and Machine Learning for Internet of Things*. 39-42.

Dou, Q., So, T. Y., Jiang, M., Liu, Q., Vardhanabhuti, V., Kaissis, G., ... & Heng, P. A. (2021). Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. *NPJ digital medicine,* 4(1): 60.

Fioretto, F., & Van Hentenryck, P. (2019, May). Privacy-Preserving Federated Data Sharing. In AAMAS . 638-646.

Gandhi, N., Mishra, S., Bharti, S. K., & Bhagat, K. (2021, July). Leveraging towards Privacy-preserving using Federated Machine Learning for Healthcare Systems. *In 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) IEEE*. 1-6.

Grama, M., Musat, M., Muñoz-González, L., Passerat-Palmbach, J., Rueckert, D., & Alansary, A. (2020). Robust aggregation for adaptive privacy preserving federated learning in healthcare. arXiv preprint arXiv:2009.08294.

Kerkouche, R., Acs, G., Castelluccia, C., & Genevès, P. (2021, April). Privacy-preserving and bandwidth-efficient federated learning: An application to in-hospital mortality prediction. *In Proceedings of the Conference on Health, Inference, and Learning*. 25-35.

Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., & Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. Computers in Biology and Medicine, 106848.

Lakhan, A., Mohammed, M. A., Nedoma, J., Martinek, R., Tiwari, P., Vidyarthi, A., ... & Wang, W. (2022). Federated-learning based privacy preservation and fraud-enabled blockchain IoMT system for healthcare. *IEEE journal of biomedical and health informatics*, 27(2): 664-672.

Li, X., Gu, Y., Dvornek, N., Staib, L. H., Ventola, P., & Duncan, J. S. (2020). Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Medical Image Analysis, 65: 101765.

Li, Z., Sharma, V., & Mohanty, S. P. (2020). Preserving data privacy via federated learning: Challenges and solutions. *IEEE Consumer Electronics Magazine, 9(3)***:** 8-16.

Long, G., Shen, T., Tan, Y., Gerrard, L., Clarke, A., & Jiang, J. (2021). Federated learning for privacy-preserving open innovation future on digital health. In Humanity Driven AI: Productivity, Well-being, Sustainability and Partnership. *Cham: Springer International Publishing*. 113-133

Nair, A. K., Sahoo, J., & Raj, E. D. (2023). Privacy preserving Federated Learning framework for IoMT based big data analysis using edge computing. Computer Standards & Interfaces, 86, 103720.

Passerat-Palmbach, J., Farnan, T., McCoy, M., Harris, J. D., Manion, S. T., Flannery, H. L., & Gleim, B. (2020, November). Blockchain-orchestrated machine learning for privacy preserving federated learning in electronic health data. *In 2020 IEEE international conference on blockchain (Blockchain) IEEE*. 550-555.

Peyvandi, A., Majidi, B., Peyvandi, S., & Patra, J. C. (2022). Privacy-preserving federated learning for scalable and high data quality computational-intelligence-as-a-service in Society 5.0. *Multimedia tools and applications*, 81(18): 25029-25050.

Rehman, A., Razzak, I., & Xu, G. (2022). Federated learning for privacy preservation of healthcare data from smartphone-based side-channel attacks. *IEEE Journal of Biomedical and Health Informatics,* 27(2): 684-690.

Singh, S., Rathore, S., Alfarraj, O., Tolba, A., & Yoon, B. (2022). A framework for privacy-preservation of IoT healthcare data using Federated Learning and blockchain technology. Future Generation Computer Systems, 129: 380-388.

Stephanie, V., Khalil, I., Atiquzzaman, M., & Yi, X. (2022). Trustworthy privacy-preserving hierarchical ensemble and federated learning in healthcare 4.0 with blockchain. *IEEE Transactions on Industrial Informatics*.

Tabassum, A., Erbad, A., Lebda, W., Mohamed, A., & Guizani, M. (2022). Fedgan-ids: Privacy-preserving ids using gan and federated learning. Computer Communications, 192, 299-310.

Thapa, C., Chamikara, M. A. P., & Camtepe, S. A. (2021). Advancements of federated learning towards privacy preservation: from federated learning to split learning. Federated Learning Systems: Towards Next-Generation AI, 79-109.

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.

Yin, X., Zhu, Y., & Hu, J. (2021). A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. ACM Computing Surveys (CSUR), 54(6): 1-36.

Zerka, F., Barakat, S., Walsh, S., Bogowicz, M., Leijenaar, R. T., Jochems, A., .& Lambin, P. (2020). Systematic review of privacy-preserving distributed machine learning from federated databases in health care. JCO clinical cancer informatics, 4, 184-200.