NAME : HARSHIT GADHIYA(23981757)

GROUP: De- Identifying Patient Medical Records(Group 4)

Github: https://github.com/harshit-3/Capstone-project-group-4.git

INTRODUCTION:

The difficulty of de-identifying sensitive healthcare data while maintaining patient privacy and data usability for research was the focus of this study. Ensuring adherence to laws such as GDPR, HIPAA, and the Australian Privacy Principles (APPs) under the Privacy Act 1988 is essential given the growing concern over data privacy. Using a variety of de-identification techniques, including k-anonymity, l-diversity, t-closeness, and proprietary de-identification methodologies, the research aimed to identify and mitigate the hazards associated with quasi-identifiers. To reduce the possibility of re-identification, unique identification analysis was also given particular attention.

MY TASK:

I focus on Finding high-risk quasi-identifiers in the dataset (like date of birth, clinic location, clinic name, and test results), protecting the data with anonymization techniques, and assessing the outcomes for compliance with privacy laws were the main responsibilities of the project. To improve privacy while maintaining data utility, I specifically contributed by conducting Unique Identification Analysis and putting custom de-identification techniques into place.

PROBLEM FORMULATION:

This project's challenge was to strike a balance between healthcare data's utility and privacy requirements. I was able to identify quasi-identifiers that presented a danger of re-identification by drawing on our prior knowledge and expertise in data privacy laws and privacy-preserving strategies. Since certain characteristic combinations, such a person's date of birth, clinic location, and illness, may be used to uniquely identify people, my strategy needed to manage this problem with the least amount of disruption to data processing. Iterative testing and modification were necessary throughout this procedure to guarantee data utility and privacy protection.

PROBLEM SOLVING:

To reduce the likelihood of re-identification in the dataset, the problem-solving approach concentrated on using specialized de-identification techniques. The main goal of this

strategy was to maintain the dataset's usefulness while guaranteeing patient confidentiality by changing important quasi-identifiers like birthdates, localities, and clinic names.

Birthdate De-Identification: The patient's Date of Birth (DOM) was one of the primary quasi-identifiers in the dataset. Precise dates of birth presented a substantial risk of re-identification, particularly when paired with other factors. I used a unique technique to address this, which involves randomizing the month and day and generalizing birthdates within the same year. A patient who was born on December 3, 1980, for instance, might have their birthdate changed to March 11, 1980. This approach protected people's identities while maintaining age-related data that could be used for study by preserving the broad period without disclosing the precise birthday. when there are not many patients in a certain birth year.

Generalization of Clinic Names: Another delicate quasi-identifier that needed to be carefully de-identified was the name of the clinic. For instance, when paired with other identifiers, certain clinic names, like "London Oncology Clinic," may result in re-identification. To counteract this, I renamed each clinic using a generic identification associated with its location using a proprietary pseudonymization technique. London clinics were renamed "London Clinic 1," "London Clinic 2," and so on, for instance. In this manner, the specific clinic data was eliminated, but the information was still useful for analysing regional healthcare trends and clinic effectiveness. I made sure that confidential information was hidden while still giving the data a useful structure by using generic clinic names.

ETHICAL RESPONSIBLE AI, and BROADER SOCIAL IMPACT:

Inadequate data anonymization can have serious ethical and social repercussions, particularly in the healthcare industry. Data breaches can result in monetary loss, psychological harm to people, and a decline in public trust. To address privacy issues, my work on Unique Identification Analysis made sure that no individual could be re-identified from the data. I also considered the possible harm to society from biased uses of anonymised data in this study. I assisted in reducing the possibility of bias in subsequent studies using this dataset by making sure that our anonymization methods did not disproportionately impact groups.

Over-generalization is one of the possible dangers of de-identification, which could lessen the usefulness of the data. I used strategies that maintained important statistical patterns while maintaining anonymity to lessen this. Furthermore, to prevent the ethical and legal repercussions of data misuse, compliance with privacy rules such as GDPR and APPs was crucial.

Personal Reflection:

The project's most technological difficulty was striking a balance between data utility and privacy. An innovative way for managing sensitive healthcare data was the modification of de-identification techniques (such as randomizing birthdates and generalizing clinic names). These techniques significantly decreased the chance of re-identification while guaranteeing that the data would continue to be valuable for study. It was initially challenging to decide how much to generalize without losing important data. Iterative analysis and testing, however, assisted in improving the strategy. I was especially pleased with how the clinic names were custom pseudonymized to preserve the dataset's research usefulness while guaranteeing adherence to privacy laws such as Australia's Privacy Act 1988 and the GDPR.

If I had more time, I would dedicate myself to the study of different technique for unique fields like UMRN instead of hash UMRN I would focus on identify in such a way that it resembles the original data. Custom method would be hard to reverse engineer over any traditional method available currently, so it would be possible to provide datasets for public use while fully protecting patient privacy, which would be beneficial for more extensive studies.

In conclusion, the study demonstrated the value of sophisticated de-identification strategies for healthcare datasets, emphasizing the necessity of approaches like custom generalization and UID analysis to reduce the danger of re-identification. These methods were crucial for securing sensitive patient data, producing anonymized data with analytical value, and adhering to GDPR, HIPAA, and Australia's Privacy Act 1988.

REDERENCE:

1) GDPR - General Data Protection Regulation:European Union. (2016). 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, L119, 1–88. Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679
2) HIPAA - Health Insurance Portability and Accountability Act: U.S. Department of Health & Human Services. (1996). *Health Insurance Portability and Accountability Act (HIPAA).* Retrieved from: https://www.hhs.gov/hipaa/index.html
3) Australian Privacy Principles (APPs): Office of the Australian Information Commissioner (OAIC). (2014). *Australian Privacy Principles Guidelines.* Retrieved from: https://www.oaic.gov.au/privacy/australian-privacy-principles
4) DATA BRICK: Lessons Learned from Deidentifyinf 700 Million Patient Notes: