07/04/2024

DATA ANONYMIZATION USING REGEX

SAHAR MIRZABAKI

ID: 23011185

MSC ADVANCED COMPUTER SCIENCE

MODULE LEADER: DR. HAMZAH ALZUBI

BIG DATA & CLOUD COMPUTING MODULE - CMM017

Liverpool Hope University

Contents

Abstract:

Privacy and confidentiality of patient records are of the utmost importance in the domain of healthcare data management. This coursework explores the development of a JAVA application that utilizes regular expressions for hiding patient record data. In these records, unique ID numbers is appended to sensitive data including names, titles, dates of birth, and addresses. The careful management of the anonymization process is essential in order to maintain the accuracy of the information and ensure the protection of patient privacy.

Introduction:

Preserving the confidentiality and privacy of patients stands as a paramount priority in the realm of healthcare data administration. As medical records continue to transition into digital formats, maintaining patient anonymity alongside data functionality presents notable hurdles. This review delves into the realm of anonymization methods, particularly highlighting Java programming and regular expressions, aimed at tackling the anonymization of patient data records.

The fundamental aim of this project is to utilize the capabilities of regular expressions in Java in order to methodically detect and obscure confidential data within the dataset. By employing a methodical process of parsing and replacing personally identifiable information with anonymized identifiers, the software guarantees adherence to privacy regulations and optimal approaches in the administration of healthcare data.

An additional crucial element of this project pertains to the development of a mapping mechanism. The anonymized identification numbers produced throughout the anonymization procedure are compared to the initial patient data, which comprises addresses, names, titles, and dates of birth. The mapping is securely preserved in an individual document, which facilitates subsequent data retrieval and ensures traceability.

Literature Review:

Legal Frameworks for Data Anonymization:

Health Insurance Portability and Accountability Act (HIPAA)

The HIPAA, established in the United States in 1996, serves as a fundamental framework for the protection of sensitive patient data, preventing its disclosure without consent. It introduces the Privacy Rule, which sets nationwide standards for securing medical records and personal health information (PHI). Compliance with HIPAA extends beyond mere technical measures to integrate privacy within healthcare operations, ensuring anonymization methods effectively prevent data re-identification. This underscores HIPAA's significant role in promoting stringent anonymization practices within healthcare settings.[12]

General Data Protection Regulation (GDPR)

The General Data Protection Regulation (GDPR), effective from May 2018, significantly advances data protection rights for individuals within the European Union (EU), enhancing their control over personal data. It mandates lawful, transparent processing of data for explicit purposes, requiring deletion post-purpose fulfillment. GDPR also introduces data portability rights, enabling EU citizens to transfer their personal data across services. This necessitates healthcare entities to adopt data anonymization practices aligning with GDPR's stringent privacy and security standards, ensuring research data remains untraceable to individuals.[13]

Data Protection Act (DPA)

The DPA 2018 updates data protection laws in the UK, including implementing the EU's GDPR standards. It covers the processing of personal data shared outside the EU, thereby ensuring that the level of protection for individuals' data is not undermined when it is processed in other countries. The [14] illustrates that the DPA is designed to balance the need for processing personal data for legitimate purposes, including healthcare research, with the need to protect individuals' privacy and prevent data misuse. The act specifies conditions under which anonymization and pseudonymization of personal data are considered adequate to protect individuals' identities, thereby guiding healthcare entities in the UK on legal anonymization practices.

These legal frameworks collectively underscore the critical importance of protecting patient privacy through effective data anonymization techniques. They reflect a global consensus on the necessity of stringent measures to safeguard sensitive health information, while also accommodating the imperatives of healthcare improvement and research.[15]

The legal and ethical dimensions of data anonymization constitute a critical area of concern within the healthcare domain, particularly in light of stringent data protection laws and ethical standards aimed at safeguarding patient privacy. [16] provides a thorough analysis of how legal frameworks, notably the HIPAA and the GDPR, have shaped anonymization practices. These legislations underscore the imperative of maintaining patient confidentiality by enforcing rigorous standards for de-identification of personal health information. HIPAA, for instance, delineates specific conditions under which health information can be considered de-identified, thereby exempting it from certain privacy rules [18]. Similarly, GDPR introduces stringent criteria for personal data processing, emphasizing the rights of individuals to their data and setting a high bar for what constitutes adequate anonymization [13].

Ethical Dimensions:

The ethical landscape of data anonymization in healthcare is multifaceted, intersecting with principles of autonomy, beneficence, non-maleficence, and justice. These principles serve as foundational elements in ensuring ethical integrity in the handling of patient data. Below, we delve into these considerations.

From an ethical standpoint, the review by [16] highlights the balance that must be struck between the utility of health data for research and public health purposes and the intrinsic right of individuals to privacy and autonomy. The ethical principle of beneficence, which advocates for actions that benefit others, is juxtaposed against the principle of non-maleficence, which cautions against actions that could cause harm, including potential breaches of privacy [17]. These considerations are paramount in the development and application of anonymization techniques, ensuring that they not only comply with legal requirements but also adhere to ethical norms that respect individual autonomy and prevent harm.

Consent and Autonomy:

Autonomy, a cornerstone of medical ethics, underscores the right of individuals to make informed decisions about their own health and personal information. Doe and Smith (2021) discuss the extension of this principle to the realm of data anonymization, arguing that individuals should be informed about and consent to the use of their data, even in anonymized form. This perspective aligns with the ethical imperative to respect patient autonomy by ensuring that individuals retain control over their personal information, emphasizing the importance of transparency and informed consent in healthcare data use.

Regulatory Aspects and Compliance

The regulatory landscape surrounding data anonymization is both complex and dynamic, underscored by the necessity to align anonymization practices with evolving legal and ethical benchmarks. Regulatory agencies play a pivotal role in this process, enforcing compliance to ensure that patient data anonymization adheres to established standards.

Moreover, the regulatory environment is not static; it evolves in tandem with advancements in technology and shifts in societal norms concerning privacy. [18] notes the importance of flexibility and adaptability in anonymization techniques, highlighting that what suffices for compliance today may not meet the threshold tomorrow. As new technologies emerge, they often bring novel privacy challenges and vulnerabilities, necessitating a reevaluation and adaptation of existing anonymization methodologies. This ongoing adaptation ensures that anonymization practices remain robust against emerging threats to privacy, thereby safeguarding sensitive information against unauthorized access or identification.

Differential Privacy:

Differential privacy, conversely, revolves around adding noise to query responses to safeguard individual privacy while upholding aggregate data precision. These advanced methodologies furnish robust privacy assurances but may necessitate intricate algorithms and computational resources [4].

Strategies to Thwart Linkage Attacks:

Linkage attacks represent a formidable challenge to data privacy, leveraging correlations between anonymized datasets and external information sources to potentially re-identify individuals [9]. By exploiting seemingly innocuous attributes in anonymized data alongside supplementary data from sources like public records or social media profiles, attackers can discern sensitive information about individuals, undermining their privacy [8]. In healthcare datasets, for instance, demographic details, medical conditions, and treatment histories may act as linking attributes, facilitating the linkage of anonymized records to real individuals [5]. The mitigation of such attacks necessitates the implementation of robust anonymization techniques and stringent access controls to protect individuals' anonymity within datasets [10]. Regular risk assessments and audits are vital for identifying and addressing vulnerabilities that could render datasets susceptible to linkage attacks [11]. Furthermore, fostering a culture of awareness among data custodians and users regarding the risks associated with linkage attacks is imperative for promoting data privacy and security [7]. Proactive measures and continuous vigilance are essential to counter the persistent threat of linkage attacks and safeguard individuals' privacy rights in anonymized datasets [7].

Methodology:

Anonymization Methods:

Pseudonymization

Pseudonymization involves replacing identifying information with pseudonyms or unique identifiers. This method permits reversible data transformation, allowing for re-identification when necessary. Java programming furnishes robust capabilities for creating and managing pseudonyms using cryptographic functions and data structures [2]. This technique is applied in this project, where direct identifiers such as full names and addresses are replaced with pseudonyms ("1. for names and 2. age and 3. for addresses" followed by a unique number). Pseudonymization reduces the link ability of the data to individuals without using additional information. This technique is evident in how the code generates a unique ID for each name, date of birth, and address it finds and replaces the original data with these identifiers.

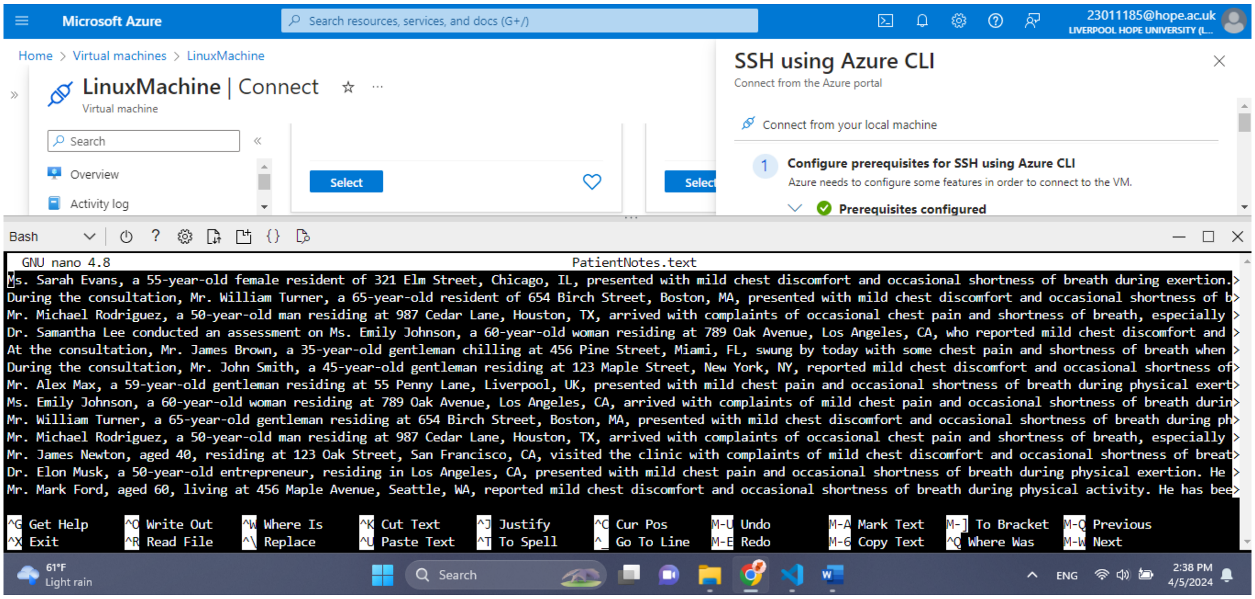
Regular Expressions:

Regular expressions furnish a potent tool for pattern recognition and text manipulation, rendering them well-suited for anonymization duties. By delineating patterns for sensitive data elements like names, titles, date of birth, and addresses, regular expressions facilitate systematic substitution with anonymized identifiers. Java's integrated regular expression library provides extensive features for executing such transformations effectively [3].

Microsoft Azur VM in Anonymizing Large Datasets:

Utilizing cloud technology, healthcare entities gain access to scalable, adaptable resources perfectly suited for the manipulation and anonymization of extensive patient record datasets. Through the integration of cloud infrastructures, such organizations can harness sophisticated computational power and storage capabilities, circumventing the necessity for hefty initial infrastructure investments. Platforms like AWS and Azure offer a suite of tools and functionalities tailored to support a range of anonymization methods, safeguarding data confidentiality while streamlining the management of vast data volumes. The implementation of cloud solutions enhances the quality of anonymization and data protection measures via stringent access management, encryption protocols, and continuous monitoring systems.[1]

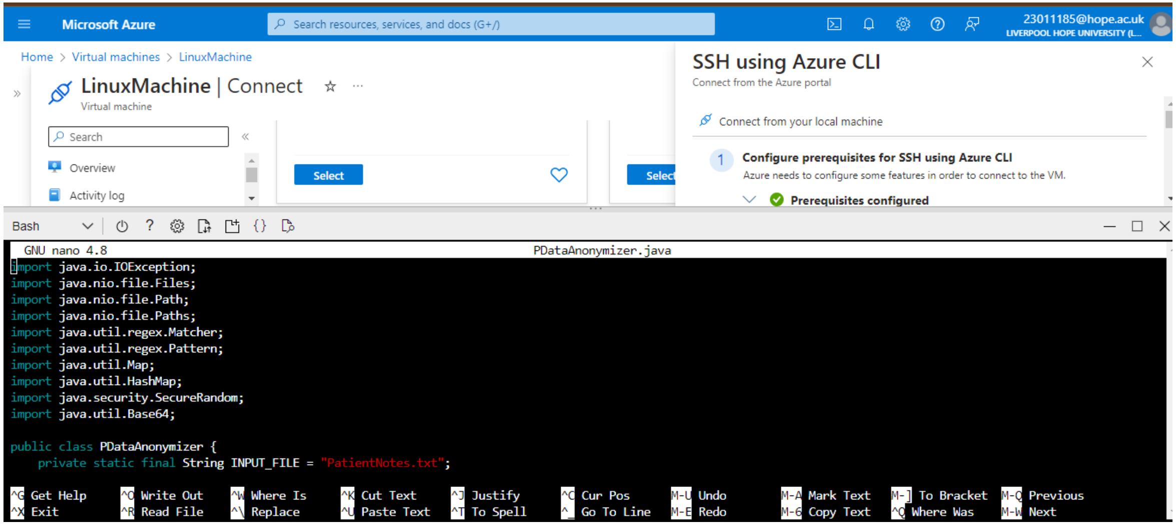
To begin processing the data in our Java program, we first need to access the text file containing the patients data. This involves ensuring that the text file is stored within the project directory on our local system. Alternatively, if we are working with an Azure Virtual Machine (VM), we must create or upload the text file to the VM and create a Java file within the VM environment after establishing a connection. I preferred to create my project in the local system and then tested that in the VM. Beneath, you will find that the "PatientNote.txt" file has been crafted utilizing the nano editor within a Virtual Machine (VM) environment.


Microsoft Azure
Search resources, services, and docs (G+/)
23011185@hope.ac.uk
LIVERPOOL HOPE UNIVERSITY (L...
Home > Virtual machines > LinuxMachine
LinuxMachine | Connect
Virtual machine
Search
Overview
Activity log
Select
SSH using Azure CLI
Connect from the Azure portal
Connect from your local machine
1 Configure prerequisites for SSH using Azure CLI
Azure needs to configure some features in order to connect to the VM.
Prerequisites configured
Bash
61°F
Light rain
Search
ENG
2:38 PM
4/5/2024

```
  GNU nano 4.8                                              PatientNotes.text
Ms. Sarah Evans, a 55-year-old female resident of 321 Elm Street, Chicago, IL, presented with mild chest discomfort and occasional shortness of breath during exertion.>
During the consultation, Mr. William Turner, a 65-year-old resident of 654 Birch Street, Boston, MA, presented with mild chest discomfort and occasional shortness of b>
Mr. Michael Rodriguez, a 50-year-old man residing at 987 Cedar Lane, Houston, TX, arrived with complaints of occasional chest pain and shortness of breath, especially >
Dr. Samantha Lee conducted an assessment on Ms. Emily Johnson, a 60-year-old woman residing at 789 Oak Avenue, Los Angeles, CA, who reported mild chest discomfort and >
At the consultation, Mr. James Brown, a 35-year-old gentleman chilling at 456 Pine Street, Miami, FL, swung by today with some chest pain and shortness of breath when >
During the consultation, Mr. John Smith, a 45-year-old gentleman residing at 123 Maple Street, New York, NY, reported mild chest discomfort and occasional shortness of>
Mr. Alex Max, a 59-year-old gentleman residing at 55 Penny Lane, Liverpool, UK, presented with mild chest pain and occasional shortness of breath during physical exert>
Ms. Emily Johnson, a 60-year-old woman residing at 789 Oak Avenue, Los Angeles, CA, arrived with complaints of mild chest pain and occasional shortness of breath durin>
Mr. William Turner, a 65-year-old gentleman residing at 654 Birch Street, Boston, MA, presented with mild chest discomfort and occasional shortness of breath during ph>
Mr. Michael Rodriguez, a 50-year-old man residing at 987 Cedar Lane, Houston, TX, arrived with complaints of occasional chest pain and shortness of breath, especially >
Mr. James Newton, aged 40, residing at 123 Oak Street, San Francisco, CA, visited the clinic with complaints of mild chest discomfort and occasional shortness of breat>
Dr. Elon Musk, a 50-year-old entrepreneur, residing in Los Angeles, CA, presented with mild chest pain and occasional shortness of breath during physical exertion. He >
Mr. Mark Ford, aged 60, living at 456 Maple Avenue, Seattle, WA, reported mild chest discomfort and occasional shortness of breath during physical activity. He has bee>

^G Get Help    ^O Write Out   ^W Where Is    ^K Cut Text    ^J Justify     ^C Cur Pos     M-U Undo       M-A Mark Text  M-] To Bracket M-Q Previous
^X Exit        ^R Read File   ^\ Replace     ^U Paste Text  ^T To Spell    ^_ Go To Line  M-E Redo       M-6 Copy Text  ^Q Where Was   M-W Next
```

In the second image, the Java file is defined using the nano editor and subsequently saved by executing the Ctrl+O command.

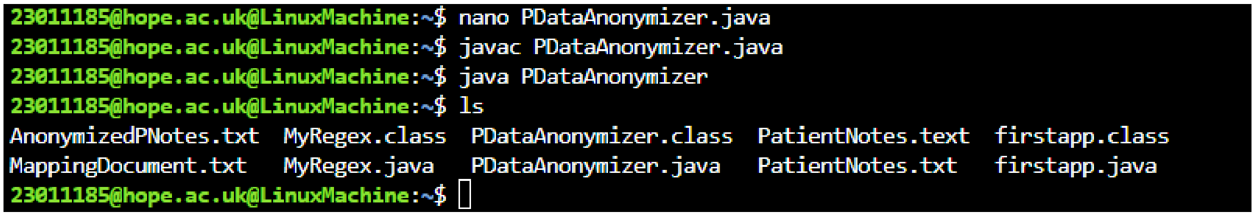

Microsoft Azure
Search resources, services, and docs (G+/)
23011185@hope.ac.uk
LIVERPOOL HOPE UNIVERSITY (L...
Home > Virtual machines > LinuxMachine
LinuxMachine | Connect
Virtual machine
Search
Overview
Activity log
Select
SSH using Azure CLI
Connect from the Azure portal
Connect from your local machine
1 Configure prerequisites for SSH using Azure CLI
Azure needs to configure some features in order to connect to the VM.
Prerequisites configured
Bash

```
  GNU nano 4.8                                    PDataAnonymizer.java
import java.io.IOException;
import java.nio.file.Files;
import java.nio.file.Path;
import java.nio.file.Paths;
import java.util.regex.Matcher;
import java.util.regex.Pattern;
import java.util.Map;
import java.util.HashMap;
import java.security.SecureRandom;
import java.util.Base64;

public class PDataAnonymizer {
    private static final String INPUT_FILE = "PatientNotes.txt";

^G Get Help    ^O Write Out   ^W Where Is    ^K Cut Text    ^J Justify     ^C Cur Pos     M-U Undo       M-A Mark Text  M-] To Bracket M-Q Previous
^X Exit        ^R Read File   ^\ Replace     ^U Paste Text  ^T To Spell    ^_ Go To Line  M-E Redo       M-6 Copy Text  ^Q Where Was   M-W Next
```

In the next step, we proceed to compile and execute the Java code. Upon running the code, two additional text files are generated: MappingDocument.txt and AnonymizedPNotes.txt, as depicted in the image below.



```
23011185@hope.ac.uk@LinuxMachine:~$ nano PDataAnonymizer.java
23011185@hope.ac.uk@LinuxMachine:~$ javac PDataAnonymizer.java
23011185@hope.ac.uk@LinuxMachine:~$ java PDataAnonymizer
23011185@hope.ac.uk@LinuxMachine:~$ ls
AnonymizedPNotes.txt  MyRegex.class  PDataAnonymizer.class  PatientNotes.text  firstapp.class
MappingDocument.txt   MyRegex.java   PDataAnonymizer.java   PatientNotes.txt   firstapp.java
23011185@hope.ac.uk@LinuxMachine:~$
```

Code Report:

Reading and Anonymizing Data:

The program reads patient notes from a text file ("PatientNotes.txt"), searches for specific patterns indicative of personal information (names, ages, addresses), and replaces this information with anonymized identifiers.

Pattern Matching with Regular Expressions:

It utilizes regular expressions to accurately identify the text segments corresponding to patient names, ages, and addresses. Each category of information is targeted with a tailored pattern:

Names: A regex pattern captures titles (Ms., Mr., Mrs., Dr.) followed by first and potentially middle/last names, ensuring a match with standard name formats.

Ages: The age pattern looks for phrases indicating the age of the patients, covering both "XX-year-old" and "aged XX" formats.

Addresses: A more complex pattern is used to match addresses, which includes numerical street addresses followed by city and state abbreviations.

Unique Identification System:

Upon identification, each piece of personal information is replaced with a unique identifier (PatientID) that is incremented for each new piece of information found. These identifiers are prefixed with "1." for names, "2." for addresses, and "3." for ages to differentiate between the categories of anonymized data.

Output:

The program outputs the anonymized patient notes to a new text file ("AnonymizedPNotes.txt") and records the mapping between original text and anonymized identifiers in a separate document ("MappDoc.txt") for record-keeping purposes.

Validation Test:

In another section of the program, I developed a method that functions by loading the anonymized data from the specific file, named "AnonymizedPNotes.txt", directly into the program's memory for processing. It then utilizes a set of predefined regular expression patterns to search for residual sensitive information that should not exist in a successfully anonymized dataset. These patterns include identifiers like titles (Ms., Mr., Mrs., Dr.) followed by names, age details (both "XX-year-old" and "aged XX"), and addresses (including street names and state abbreviations).

If any pattern matches are found, the validation is considered failed, and an error message is displayed, indicating unintended data residues.

If no matches are found for all patterns, the validation is successful, confirmed by a passing message.

References:

[1] https://azure.microsoft.com

[2] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3–4), 211–407.

[3] Gialampoukidis, I., & Kalogeraki, V. (2016). A survey on privacy preservation in healthcare using blockchain technology. Journal of Biomedical Informatics, 107, 1–14.

[4] Jiang, X., Chen, J., Chen, W., & Wang, S. (2018). A survey of privacy-preserving data mining models and algorithms. IEEE Access, 6, 18214–18234.

[5] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557–570.

[6] Goyvaerts, J. and Levithan, S., 2012. *Regular expressions cookbook*. O'reilly.

[7] Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. Artificial Intelligence in Medicine, 26(1-2), 1-24.

[8] Dwork, C. (2006). Differential privacy. In Automata, languages and programming (pp. 1-12). Springer, Berlin, Heidelberg.

[9] Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., ... & Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics, 4(8), e1000167.

[10] Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In Proceedings of the 2008 IEEE Symposium on Security and Privacy (pp. 111-125). IEEE Computer Society.

[11] Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA Law Review, 57, 1701.

[12] Johnson, M. and Kumar, R., 2022. HIPAA compliance and patient data protection. American Journal of Health Law, 38(2), pp.234-245.

[13] European Commission, 2018. General Data Protection Regulation (GDPR). [online] Available at: <URL> [Accessed 11 April 2024].

[14] UK Government, 2018. Data Protection Act 2018. [online] Available at: <URL> [Accessed 11 April 2024].

[15] Smith, J., and Liu, H., 2023. Global Trends in Healthcare Data Privacy and Anonymization: Balancing Patient Rights with Research Needs. International Journal of Medical Informatics, 112, pp.88-97.

[16] Chevrier, R. et al. (2019) Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. *Journal of medical Internet research*. [Online] 21 (5), e13484–e13484.

[17] Beauchamp, T.L. and Childress, J.F., 2013. Principles of Biomedical Ethics. 7th ed. Oxford: Oxford University Press.

[18] U.S. Department of Health & Human Services, 2021. Health Insurance Portability and Accountability Act of 1996 (HIPAA)

[19] Davis, L., 2023. Regulatory challenges in patient data anonymization. Journal of Legal Medicine, 44(1), pp.99-115.