Healthcare Data Governance & Anonymization Project

Presenting By
Rajkumar K P

Self Introduction

Hi I am Rajkumar, I completed MSc Biotechnology in 2022 and I am a detail-oriented Data Analyst with 2 years of experience in healthcare data analysis, EHRs, and claims data. Proficient in Python, SQL, Power BI and Tableau for data analytics, reporting and interactive dashboarding. Strong understanding of data privacy regulations including HIPAA, CCPA, and GDPR. Well-versed in machine learning and AI-driven techniques - applying predictive modeling, pattern recognition, and intelligent automation to unlock actionable insights and support data-informed decision-making.

Overview

- This project showcases a practical application of Data Governance and Privacy Compliance using a synthetic California hospital dataset. We focused on detecting Personally Identifiable Information (PII), Protected Health Information (PHI), and sensitive fields in patient records, followed by anonymizing them using standard techniques to meet HIPAA and CCPA guidelines.

- The goal is to simulate how real-world healthcare organizations must handle patient data responsibly - ensuring data privacy, regulatory compliance and ethical data usage.

Why We Must Follow Data Governance Policies like HIPAA & CCPA

Reason	Explanation
1. Legal Compliance	Both HIPAA and CCPA are mandatory regulations. Non-compliance can result in significant fines, legal actions, and loss of licenses.
2. Protection of Personal Data	These policies ensure that personally identifiable information (PII) and protected health information (PHI) are handled securely and ethically.
3. Builds Trust with Stakeholders	Patients and customers are more likely to engage with organizations that respect their data rights. This enhances brand reputation and user retention.
4. Prevents Data Breaches	By enforcing data access controls, audit trails, and encryption policies, governance frameworks help prevent data leaks and unauthorized access.
5. Facilitates Transparency	CCPA requires businesses to be transparent about what data they collect and how it's used—promoting accountability.
6. Enhances Operational Efficiency	Good governance ensures consistent data definitions, quality standards, and data lifecycle management, reducing duplication and improving analysis.
7. Readiness for Future Regulations	Organizations following HIPAA/CCPA are better positioned to comply with future privacy laws like GDPR, CPRA, etc.
8. Ethical Responsibility	Beyond legal needs, protecting someone’s personal or medical data is an ethical obligation—especially when dealing with vulnerable populations.

Importance of Anonymization in Healthcare

Reason	Explanation
1. Enables Safe Data Sharing for Research	Anonymized patient data can be shared with researchers and public health agencies without violating privacy—accelerating medical discoveries, epidemiology, and treatment optimization.
2. Reduces Re-identification Risk	Anonymization techniques minimize the chances of individuals being re-identified, even when datasets are combined with external sources.
3. Encourages Open Collaboration	Organizations can collaborate across departments, vendors, and academic institutions when data is anonymized—supporting cross-disciplinary healthcare innovation.
4. Facilitates AI/ML Model Training	High-quality, anonymized data is essential for training predictive models (e.g., for disease diagnosis or risk scoring) while respecting patient confidentiality.
5. Meets Ethical Standards in Trials	In clinical trials, anonymization ensures blinding and maintains the integrity of the study without exposing sensitive patient info.
6. Prevents Internal Misuse	Not every employee needs access to identifiable data. Anonymized data helps in maintaining need-to-know access boundaries, preventing intentional or accidental misuse.
7. Supports Public Health Reporting	Governments often require de-identified data for tracking outbreaks or planning resources without infringing on privacy.
8. Cost-effective Regulatory Strategy	Using anonymized data lowers the burden of regulatory audits, since non-identifiable data is often exempt from stricter compliance measures.

What Anonymization Techniques Were Used

Column	Anonymization Strategy Used
first_name & last_name	Masked with first letter only (e.g., R****)
dob	Converted into age groups (e.g., 21–30)
phone	Masked all but last 3 digits (e.g., ****789)
email	Obfuscated domain (e.g., ra****@***.com)
zip	Truncated to 3-digit region code
address	Dropped entirely for safety

Before & After: Patient Data Anonymization in Action

Original Patient Data (Raw / Identifiable)

patient_id	first_name	last_name	dob	age	gender	ethnicity	insurance_type	marital_status	address	city	state	zip	phone	email	registration_date
PAT000001	Danielle	Johnson	14-05-1940	85	Female	Asian	UHC	Married	32181 Johnson Course Apt. 389	Bakersfield	CA	93301		danielle.johnson40@example.com	03-01-2025
PAT000002	Anna	Baldwin	04-03-2010	15	Female	White	Aetna	Married	79402 Peterson Drives Apt. 511	Bakersfield	CA	93301	(615)759 407	anna.baldwin294@example.com	26-02-2025
PAT000003	James	Jones	31-03-2021	4	Male	White	UHC	Married	1316 Chavez Village	Portland	OR	97035	(925)853 419	james.jones377@example.com	27-01-2025

Anonymized Patient Data (Post-Processing)

patient_id	first_name	last_name	gender	ethnicity	insurance_type	marital_status	city	state	zip	phone	email	registration_date	age_group
PAT000001	D****	J****	Female	Asian	UHC	Married	Bakersfield	CA	933		da****@***.com	03-01-2025	71+
PAT000002	A****	B****	Female	White	Aetna	Married	Bakersfield	CA	933	****407	an****@***.com	26-02-2025	0-20
PAT000003	J****	J****	Male	White	UHC	Married	Portland	OR	970	****419	ja****@***.com	27-01-2025	0-20

Real-World Impact of Non-Compliance:

- **HIPAA**: Up to **$1.5 million** per violation per year.
- **CCPA**: Up to **$7,500 per intentional violation**.
- Reputation damage and **loss of customer trust** are even more costly.

Thank You !