Privacy and Anonymization

Michael Yao, Allison Chae

Learning Objectives

1. **Define** privacy and anonymity, and **describe** the techniques that can be used to anonymize patient medical data.
2. **Identify** key reasons why anonymization does not preserve patient identities in the real-world.
3. **Analyze** how current data acquisition practices and anonymization techniques may inadvertently harm minority patient populations.

Food for Thought: What does data privacy mean to you?

Some potential answers might include:

1. **Control of Access**: You should be able to control who accesses your data.
2. **Control of Use**: You should be able to have a say on how your data is used and for what purpose.
3. **Knowledge of Access/Use**: You should know when your data is used or accessed.
4. **Opt In (and Out)**: You should be able to add more data or remove your data at any point in time.
5. **Anonymity**: Your identity should remain private regardless of whatever you or others do with your data.

What other answers did you think of? Do you think these notions of privacy are currently satisfied in academic research using patient data?

Overview

As clinicians, we deal with patient data every day, and have an ethical (and legal) responsibility to protect patient privacy and confidential information. At the same time, we often work alongside scientists to use patient data to advance our understanding of science. How can we gain meaningful insights from data while still protecting patient identity?

💡 When you release your genomic data to the public, is the privacy of your parents and grandparents still preserved? What about the future privacy of your children and grandchildren?

According to the Health Insurance Portability and Accountability Act (HIPAA), one way to accomplish this is through **data anonymization**. In general, there are two main ways that researchers anonymize data:

1. **Coarsening** means we decrease the granularity of the features. For example, using 5-digit zip codes may make it too easy to identify individuals from a dataset, so we might instead choose to coarsen the zip codes by removing the last two digits of each zip code. Instead of including the exact ages of patients, we often coarse the data to only include the decade of the age of the patient.
2. **Reduction** means we remove entire features altogether. For example, we might choose to remove all patient names and medical record numbers from a dataset before making it accessible to researchers.

How can we be certain that a dataset is anonymized "*enough*"? Formally, a dataset is defined as ***k*-anonymous** if there are at least k copies of any given row in a dataset. The concept of k-anonymity is based in the idea of *anonymity in numbers* - if k is sufficiently large, then it should (hopefully) be impossible to identify any singular individual as a particular row of the dataset because the patient could be any of at least k rows.

A Detailed Look: HIPAA PHI

Let's take a look at the official list of HIPAA-protected attributes from the Health and Human Services Department, which are called **protected health information (PHI)**:

1. Names.
2. All geographical subdivisions smaller than a state (e.g., street address, city, county, precinct, ZIP code except for the initial three digits of a ZIP code).
3. All dates (except year) directly related to an individual (e.g., birth date, admission date, exact ages in years over the ages of 90).
4. Phone numbers, fax numbers, email addresses.
5. Social security numbers, health plan beneficiary numbers, driver license numbers, medical record numbers, etc.
6. License plates
7. IP addresses
8. Biometric identifiers (i.e., finger prints, voice recordings, genomic data)
9. Full-face photographic images
10. Any other unique identifying number, characteristic, or code

💡 The initial three digits of a zip code is still considered PHI by HIPAA if the number of individuals residing in all zip codes with those initial three digits is less than 20,000. Why do you think this is the case? How do you think the cutoff of 20,000 individuals was determined?

💡 Why are ages over 90 years-old considered PHI, but not younger ages?

Are there any attributes listed that you didn't expect? How about attributes that aren't listed above but *should* be included?

Hands-On Tutorial

For this exercise, take a look at the following table:

Table 1: **A Sample Patient Dataset**

PATIENT_ID	AGE	GENDER	BP	HIV_STATUS
P001	45	M	120/80	Negative
P002	60	F	135/85	Positive
P003	33	M	128/82	Negative
P004	50	F	142/90	Negative
P005	27	M	110/70	Positive
P006	38	F	125/78	Negative
P007	55	M	138/88	Negative
P008	43	F	132/84	Positive
P009	29	M	118/76	Negative

PATIENT_ID	AGE	GENDER	BP	HIV_STATUS
P010	61	F	145/92	Negative

Imagine that you're a student working in a research lab and are tasked with analyzing this dataset of patients from the Philadelphia area. Your research mentor tells you that this dataset contains all of the inpatient admissions to HUP from the past week.

Is Table 1 properly anonymized according to HIPAA regulations?

Separately during your lunch break, you hear on the news that a famous celebrity - a 50 year-old female (in this hypothetical situation) - was recently admitted to HUP last week for a hypertensive crisis, and was just recently discharged from the hospital.

Given this information, can you identify which patient ID corresponds to the famous celebrity?

The only patient in the table above corresponding to a 50 year-old female with hypertension is patient `P004`.

In case it wasn't already clear, this dataset and scenario is entirely fictional, and was actually generated using ChatGPT! You can take a look at the generation process here if you're interested.

Ignoring the fact that this was a small toy example, how difficult was it to *re-identify* a patient (namely, the famous celebrity) from the dataset? As a result of the successful re-identification of the patient, were you able to learn anything new about the patient (i.e., take a look at the `HIV_STATUS` column).

It turns out that a very similar re-identification strategy was used by Dr. Latanya Sweeney in 1997 where she successfully re-identified the then Governor of Massachusetts using publicly accessible, anonymized medical records released by the state of Massachusetts.

Why were we and Dr. Sweeney able to re-identify patients from an anonymized dataset? The main reason is that in both situations, we correlated the information in the table with outside knowledge and other datasets **in order to gain new, privileged information** about patients by synthesizing datasets together. There are countless other examples of re-identifying individuals from anonymized datasets, from identifying Netflix

Sweeney is an excellent writer and researcher, and we encourage you to check out two of her publications on this topic: [1] Sweeney L. Only you, your doctor, and many others may know. Tech Sci. (2015). Link to article; [2] Sweeney L. *k*-Anonymity: A model for protecting privacy. Int J Uncertainty, Fuzziness, and Knowledge-based Systems 10(5): 557-70. (2002). Link to article users from anonymized movie ratings to even finally catching the notorious *Golden State Killer*.

In summary, there are two key points that we hope you take away from this exercise:

Problems with Anonymization

1. Anonymization is ***not*** an effective tool to preserve patient privacy.
2. The *reason* why anonymization fails is that it assumes there are no other datasets or sources of information in the world to cross-reference (which is obviously not true).

Evidence-Based Medicine Discussion

Do current HIPAA-compliant anonymization standards effectively protect minorities and people of color?

1. Overview Article

All of Us Research Program Overview. National Institutes of Health. Accessed 19 May 2024. Link to article
tl;dr: *All of Us* is an NIH initiative to build a diverse database of Americans from all backgrounds in order to inform and power thousands of future studies on a variety of different health conditions. The overarching goal of the *All of Us* initiative is to power future advancements in precision medicine.

ℹ There's a great 2-minute intro video to the *All of Us* Research program here.

2. Yes, the anonymization achieved by the All-of-Us Research Program is sufficient.

Xia W, Basford M, Carroll R, Clayton EW, Harris P, Kantacioglu M, Liu Y, Nyemba S, Vorobeychik Y, Wan Z, Malin BA. Managing re-identification risks while providing access to the *All of Us* research program. J Am Med Inf Assoc 30(5): 907-14. (2023). doi:

10.1093/jamia/ocad021. PMID: 36809550
tl;dr: Cross-sectional study using the All of Us database containing data from over 300,000 participants at the time of the study. The authors used computational techniques to compute the re-identification risk for any given individual in the dataset. A large re-identification risk means that a given individual is unique in dataset and therefore more likely to be re-identified. The 95th percentile re-identification risk across all participants satisfies current government guidelines.

3. No, the All-of-Us Research Program hurts people of color.

Kaiser J. Million-person U.S. study of genes and health stumbles over including Native American groups. Science. (2019). Link to article
tl;dr: Native Americans have historically been mistreated by researchers and the US government, and are skeptical of participating in *All of Us*. Because so few Native Americans currently participate in *All of Us*, any new individual participant from a small tribe will have a high re-identification risk in spite of data safeguards. Tribes are seeking to be able to approve publications on their group and an opportunity to bless biological samples before disposal.

Summary

Anonymization is a common technique used to ensure that publicly released medical datasets are HIPAA-compliant and protect patient identities. Unfortunately, there is a growing body of evidence that shows that anonymization is no longer an effective technique for protecting patient data, and cannot provide any provable guarantees for patient privacy. At the end of the day, robustly guaranteeing patient privacy is a difficult task and requires conscious efforts from both clinicians and researchers alike.

ℹ There are other problems involving the All of Us research program, including a recent study inadvertently using “objective” mathematical techniques that inappropriately validates racist and xenophobic ideologies.[1] Even well-established data analysis techniques must be used and presented carefully!

Additional Readings

1. Gille F, Brall C. Limits of data anonymity: Lack of public awareness risks trust in health system activities. Life Sciences, Society and Policy 17(7). (2021). doi: 10.1186/s40504-021-00115-9
2. Savage N. Privacy: The myth of anonymity. Nature 537: S70-2. (2016). doi: 10.1038/537S70a. PMID: 27602747
3. Kapoor S. Revisiting HIPAA - Privacy concerns in healthcare tech. Berkeley Technology Law Journal. (2023). Link to article
4. Ohm P. Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA Law Review 57: 1701. (2010). Link to article
5. Pool J, Akhlaghpour S, Fatehi F, Burton-Jones A. A systematic analysis of failures in protecting personal health data: A scoping review. Int J Inf Manag 74: 102719. (2024). doi: 10.1016/j.ijinfomgt.2023.102719