Lecture Summary

CB&B 7520 - Biomedical Data Science: Mining and Modeling

Lecture Title and Date

Biomedical Data Privacy

March 30, 2026

Guest Instructor: Hoon Cho

Objectives of the Lecture

By the end of this lecture, students should be able to:

• Explain why privacy is especially important for biomedical and genomic data. • Distinguish major privacy risks, including re-identification, membership inference, data linkage, phenotype inference, and data reconstruction.
• Describe why legal/regulatory protections and simple de-identification are often insufficient. • Compare core privacy-preserving approaches: data sanitization, differential privacy, and secure computation.
• Explain the privacy-utility tradeoff in biomedical data sharing and machine learning.

Key Concepts & Definitions

Term	Definition
Protected Health Information (PHI)	Individually identifiable health information held or transmitted by a covered entity or business associate. Under HIPAA, PHI is the core regulated category that privacy controls are designed to protect.
De-identification	The process of removing the association between a dataset and the data subject. In practice, de-identification can be done by an expert determination or by removing specified identifiers under Safe Harbor.
Anonymization	A stronger form of de-identification aimed at making re-identification infeasible or at least highly unlikely. In biomedical data, anonymization often requires both structural transformations and risk assessment.
Data Linkage	Joining datasets using quasi-identifiers or inferred genetic features, turning separate "anonymous" records into identifiable ones.

Differential Privacy (DP)	A formal privacy guarantee that limits how much the output of an analysis can change when a single individual is added or removed. DP is usually achieved by adding calibrated randomness to answers, gradients, or model updates.
Membership inference attack (MIA):	an attack that asks whether a target individual was part of a dataset or study cohort.
Homomorphic Encryption (HE)	Encryption that allows computation on ciphertexts without first decrypting them. This makes it possible to outsource analysis while keeping the raw data hidden from the computer performing the computation.
Secure Multiparty Computation (MPC)	A cryptographic protocol that lets multiple parties compute a joint function while keeping each party's input private. Participants learn only the final result, not the other parties' raw data.
Federated Learning (FL)	A distributed learning framework in which data stay on local sites and only model updates are shared. FL reduces raw-data centralization, but update leakage can still happen, so it is often paired with DP or cryptography.
Trusted Execution Environment (TEE)	A hardware-protected isolated execution area that can run code on sensitive data with reduced exposure to the host operating system. TEEs are useful, but they rely on hardware trust and careful attestation.

1. Why biomedical privacy matters

Biomedical data reveal deeply personal information about the body and mind, including physical health, mental health, genetic risk, and family relationships. The lecture emphasized that privacy breaches can lead to stigma, discrimination, and lost opportunities in employment or insurance, and that the consequences extend beyond the individual to biological relatives. Privacy is therefore important not only for preventing harm to study participants, but also for sustaining trust, enabling scientific collaboration, and supporting public-health impact.

Regulation is important but not sufficient. The Common Rule, HIPAA, GINA, and GDPR were presented as major safeguards, but the lecture notes that legal protections evolve slowly and often leave ambiguity around what counts as "de-identified" or safe to share. The deeper issue is structural: biomedical science needs data sharing, but sharing always carries some privacy loss.

Ex: 23andMe breach showed the potential damage that can occur if this data isn't handled correctly or isn't protected

2. Difference of Genomic Data

Genomic data are static, highly unique, and shared across relatives, so they can function almost like a persistent identifier. Classic work showed that only a small number of independent SNPs can be enough to identify a person, and genealogy databases can turn a partial genetic profile into identity by finding relatives and triangulating the target.

Ex: Golden State Killer and how they were caught through usage of genomic data and information

3. Laws and regulations

The lecture reviewed several major legal/regulatory safeguards: the Common Rule (1991) for IRB review and informed consent in human-subjects research, the HIPAA Privacy Rule (2003) for protected health information, GINA (2008) for preventing genetic discrimination in employment and health insurance, and GDPR (2016) for personal-data protection in the EU. However, these safeguards have important limitations: they evolve slowly, they leave ambiguity around what counts as truly de-identified or anonymized information, and they often reduce data sharing without resolving the underlying tension between privacy and scientific utility.

4. Membership inference and data linkage

Aggregate statistics can leak cohort membership. In a membership inference attack, an attacker compares a target genotype to released study statistics or summary information and asks whether that person was part of a sensitive cohort. The lecture also covered data linkage: records that seem anonymous can be matched across datasets using shared identifiers or inferred features. The classic example is Latanya Sweeney's re-identification of the Massachusetts governor using ZIP code, date of birth, sex, and voter-registration data.

Gene expression and other functional genomics profiles can leak genotype information through eQTLs; in a simple eQTL example, genotype classes correspond to distinct mean expression levels, so expression can be used to infer genotype. Recent single-cell work shows that pseudobulk profiles by cell type can substantially improve link age, because cell-type-specific signals reveal more of the hidden genotype.

4A. Phenotype inference and data reconstruction

Once identity is known, genomes can be used to infer traits such as disease risk, carrier status, pharmacogenomic response, ancestry, and physical features. The lecture emphasizes that many of these are probabilistic rather than deterministic, but still sensitive enough to create privacy and discrimination concerns.

The lecture extends classical linkage attacks to omics. Gene expression and other functional genomics profiles can leak genotype information through eQTLs; in a simple eQTL example, genotype classes correspond to distinct mean expression levels, so expression can be used to infer genotype. Recent single-cell work shows that pseudobulk profiles by cell type can substantially improve link age, because cell-type-specific signals reveal more of the hidden genotype.

An important example involved genotype-imputation servers. By crafting uniquely matching queries, an attacker may reconstruct haplotypes from the reference panel, and the lecture emphasized that even returning only discrete genotype predictions does not fully prevent leakage. The lecture then connected

This idea applies to modern machine learning: larger language models can memorize more training data, which creates privacy and confidentiality risks for biomedical LLMs that may see PHI during pretraining or fine tuning.

5. Computational approaches to privacy protection

The first protection strategy discussed was data sanitization. The lecture showed that directly masking sensitive variants is challenging because correlations such as linkage disequilibrium can still reveal the hidden region. In functional-genomics data, one proposed approach is to release privacy-sanitized alignments (pBAM) in which private genetic variants have been identified and removed.

5A Differential privacy

Differential privacy (DP) gives a mathematically precise promise about the effect of any single person on the output of an analysis. If two datasets differ in one individual's record, then a DP mechanism should produce nearly the same output distribution on both datasets. This means an analyst learns almost the same thing whether that person participated or not. The privacy guarantee is controlled by ε: smaller ε means stronger privacy, but usually more noise and less accuracy. This is one reason DP has become a standard tool for public statistics, biomedical releases, and private machine learning.

$$\varepsilon\text{-DP: } \Pr[M(D) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D') \in S]$$

For a function f, the lecture's core idea is to measure sensitivity,

$$D \sim D' \; \| f(D) - f(D') \| \,, \; \Delta f = \max$$

and then add calibrated noise. In the Laplace mechanism, one uses

while the Gaussian mechanism

$M(D) = f(D) + \mathrm{Lap}_{0, \frac{\Delta f}{\varepsilon}}$

$M(D) = f(D) + N(0, \sigma^2), \sigma \geq \Delta_2(f)$

for (ε, δ)-DP. One representative derivation step is that the likelihood ratio between neighboring datasets is bounded by the ratio of the two Laplace densities, which collapses to e^{ε} once the noise scale is set to $\Delta f / \varepsilon$.

6. Secure computation

The lecture also reviewed secure computation frameworks for settings where data should not be revealed at all during analysis. Homomorphic encryption allows computations to be carried out on encrypted data. Secure multiparty computation splits data into random-looking shares and computes jointly on those shares. Trusted execution environments use secure enclaves plus remote attestation to protect code and data during execution. These tools support privacy-preserving analytic services and multi-site biomedical studies. The closing example, secure federated GWAS, showed how cross-biobank genomic analysis can be performed with stronger privacy guarantees.

Discussion/Comments

- The lecture made a strong case that de-identification is not a reliable endpoint for genomic or omics data; inference and linkage attacks can often bypass it.
- One of the most important conceptual points is that privacy harms are relational: a person's genome can reveal information about relatives and population groups, not just the individual. • The most challenging practical issue is the privacy-utility tradeoff. Differential privacy offers a rigorous guarantee, but the amount of noise required for useful biomedical analyses may be task-dependent and sometimes costly.
- The discussion of biomedical LLMs was especially timely. It raises an unresolved question: how do we distinguish harmful memorization of sensitive data from useful domain knowledge that a model should retain?
- A good follow-up class discussion would be: when should researchers prefer access control, data sanitization, differential privacy, or secure computation, and what criteria should drive that choice?

Suggested Readings

- Cho et al., Annual Review of Biomedical Data Science, 2024 - broad survey of privacy-enhancing technologies in biomedicine.
- Naveed et al., ACM Computing Surveys, 2015 - overview of why genomic privacy is uniquely difficult. • Bonomi, Huang, and Ohno-Machado, Nature Genetics, 2020 - taxonomy of privacy attacks in genomic settings.

• Erlich et al., Science, 2018 - genealogical re-identification through relatives in consumer-genomics databases.
• Walker et al., Cell, 2024 - linkage attacks on single-cell RNA-seq data.
• Mosca and Cho, Genome Biology, 2023 - reconstruction attacks against genotype-imputation services. • Dwork et al., 2006 and Abadi et al., 2016 - foundations of differential privacy and DP-SGD. • Carlini et al., USENIX Security, 2021 and ICLR, 2023 - memorization and data extraction in large language models.

Are the readings useful? Yes. The most useful single reading is the Cho et al. 2024 survey because it connects the privacy risks, formal protection methods, and biomedical use cases in one place. For class purposes, the most useful subsections are those on attack models, differential privacy, and secure computation. If one additional reference were to be added, I would suggest Dwork and Roth, The Algorithmic Foundations of Differential Privacy, because it gives a clearer and more systematic explanation of the formal DP definitions.

Suggested References for Key Concepts

• Genomic privacy and re-identification: Naveed et al., 2015; Erlich et al., 2018. • Privacy Attack Taxonomy: Bonomi, Huang and Ohno-Machado, Nature Genetics, 2020 • Membership inference and data linkage: Homer et al., 2008; Sweeney, 1997; Walker et al., 2024. • Differential privacy: Dwork et al., 2006; Dwork and Roth, 2014; Abadi et al., 2016. • Data reconstruction and leakage from services/models: Mosca and Cho, 2023; Carlini et al., 2021; Carlini et al., 2023.

• Secure computation in biomedicine: Cho et al., 2024; Cho and Froelicher et al., Nature Genetics, 2025.