

Privacy Engineering: Beyond the HIPAA Checklist

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 3 of "Healthcare Analytics and AI: Building Systems That Actually Work"

The essentials: what to know, why it matters at the bedside, and how models fail.

One-sentence thesis. Removing the obvious identifiers is a legal step, not a privacy guarantee, so privacy in healthcare is an engineering discipline that has to build a mathematical ceiling above HIPAA's legal floor.

PART 1: The Legal Floor and Why It Leaks

The 18 Identifiers and Safe Harbor

HIPAA's Privacy Rule defines Protected Health Information and lists 18 identifier categories (names, dates, geography smaller than a state, medical record numbers, biometrics, and more). Safe Harbor de-identification means stripping all 18; Expert Determination instead has a statistician certify that the re-identification risk is "very small." Safe Harbor is essentially taking the name tag off a record, and the record can still be recognized by everything else it is wearing. The list was codified around 2000, when data linkage was expensive, so it treats identifiers as discrete items to delete and ignores the combinatorial explosion when "non-identifiable" fields are joined across datasets. A note describing a "42-year-old male marathon runner with a history of aortic dissection repair" is technically PHI-free yet may match exactly one person in a city.

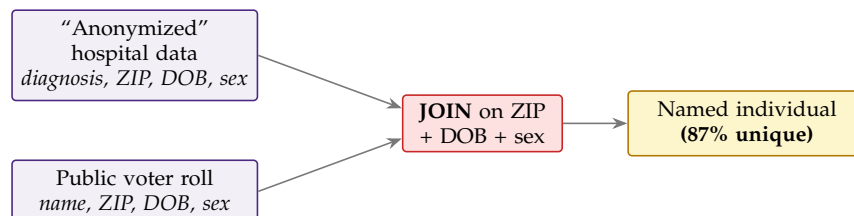
The Sweeney Effect

The most famous proof that "de-identified" does not mean anonymous is a bet a graduate student won

for twenty dollars. In 1997 Latanya Sweeney bought Cambridge voter rolls, cross-referenced three fields (5-digit ZIP, date of birth, and sex) against "anonymized" hospital data, and re-identified the governor's own medical records. She later showed those three ordinary attributes uniquely pin down most of the country:

$\{\text{ZIP, DOB, sex}\} \Rightarrow 87\% \text{ of Americans}$

By 2019, Rocher and colleagues showed machine learning could re-identify **99.98 percent** of people from 15 demographic attributes, many of which Safe Harbor permits keeping. Re-identification risk scales with the richness of available external data, and in 2026 external data (social media, consumer genomics, wearables, location trails) is everywhere, so a dataset can be compliant on paper and trivial to triangulate in practice.



The Sweeney linkage attack. Two "de-identified" sources share three quasi-identifiers; joining on them re-identifies 87 percent of Americans.

Three Canonical Attacks

Three attack families break de-identified data. Linkage cross-references quasi-identifiers with outside data, the mechanism Sweeney used. Homogeneity ex-

ploits equivalence classes where every member shares the sensitive attribute: if all 30-year-old males in a zip code have an HIV diagnosis, the diagnosis leaks without naming a record. Background knowledge uses

individual-specific facts, so knowing a colleague was hospitalized on a given date in a given city can collapse a redacted dataset to one row. These are exactly

the gaps the 18-identifier checklist cannot see, and defending against them requires the mathematical frameworks below, not a longer redaction list.

k-Anonymity, l-Diversity, t-Closeness

If the danger is being the only 42-year-old marathon runner in the file, the first defense is to make sure you are never alone in it. That is the whole intuition behind k-anonymity: generalize and suppress the quasi-identifiers (age 34 becomes 30-39) until every combination of them is shared by a crowd of at least k records. In symbols:

k-anon: each QI combo shared by $\geq k$ records

l-Diversity adds that each equivalence class holds at least l distinct sensitive values, defeating homogene-



k=5, homogeneous
all share one dx $\Rightarrow$ leaks



k=5, *l*-diverse
varied dx $\Rightarrow$ protected

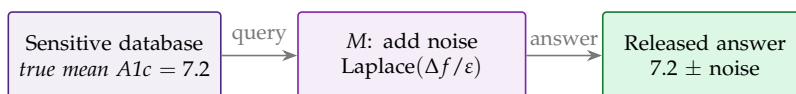
Why *k*-anonymity alone is not enough. Both classes hide each person among $k = 5$, but if all share one diagnosis it still leaks; *l*-diversity requires varied sensitive values.

Differential Privacy

Every syntactic method shares one weakness: it defends only against the attacks you anticipated. Differential privacy changes the question entirely. Instead of asking whether a record can be re-identified, it guarantees that whether or not any single person is in the dataset barely changes what a query can return, so an attacker learns almost nothing about any individual no matter how much side data or computing power they bring. Formally, for any two datasets differing by one record, the output distributions stay within a factor of e^ϵ :

$$\Pr[M(D_1) \in S] \leq e^\epsilon \Pr[M(D_2) \in S] + \delta$$

Smaller epsilon means stronger privacy ($\epsilon \leq 1$ is



smaller $\epsilon \Rightarrow$ more noise $\Rightarrow$ stronger privacy

ity, and t-closeness further requires that each class's sensitive distribution stay within t of the overall distribution, defeating skewness. **Watch out.** All three are syntactic: they make data look private under a specific definition but do not stop an attacker with enough auxiliary information. A dataset generalized to "age 0-100" is perfectly k-anonymous and perfectly useless, and this limitation is exactly what motivated differential privacy.

ANALOGY

k-anonymity is hiding in a crowd of k people; l-diversity makes sure the crowd is not all wearing the same diagnosis.

strong, $\epsilon > 10$ is weak); the Laplace mechanism adds noise scaled to $\Delta f / \epsilon$, where Δf (sensitivity) is the most one record can change the answer. Crucially, DP makes queries private, not data anonymous: the raw data still exists and must be secured. For a 1,000-patient A1c registry, sensitivity is 0.01, so at $\epsilon = 1.0$ the published average moves only to about 7.19-7.21 (clinically indistinguishable from 7.2), but at $\epsilon = 0.01$ it lands anywhere from 6.2 to 8.1 and is useless. Training with DP-SGD at $\epsilon = 1$ typically costs 5 to 15 percent accuracy, and hits minority subgroups hardest.

ANALOGY

Differential privacy is a frosted-glass window: you can tell the room is occupied and roughly how many people are inside, but you cannot identify any one face.

Differential privacy protects the query, not the data: calibrated Laplace noise makes any one record's presence undetectable in the released answer.

Epsilon	Noise scale (Laplace)	Reported A1c range	Clinical usability
1.0	~0.01	7.19–7.21%	Indistinguishable from truth
0.1	~0.1	6.9–7.5%	Population signal survives, precision lost
0.01	~1.0	6.2–8.1%	Useless for clinical decisions

Synthetic Data

Rather than protect real records, synthetic data generation replaces them with artificial records that preserve distributions, correlations, and temporal patterns without matching any real person, like a portrait painted in the style of a crowd, statistically faithful to the group but a likeness of no one in it. Tools include CTGAN (mixed data types), TimeGAN (longitudinal structure), medGAN/medWGAN (discrete codes),

and newer diffusion models (DDPMs) that now beat GANs on fidelity and privacy; by 2025, roughly **60 percent** of AI training data globally was synthetic. The tension is irreducible: a perfect copy is maximally useful and maximally privacy-violating, so evaluate both utility (downstream AUC, plausibility) and privacy (membership-inference resistance, nearest-neighbor distance to real records). PATE-GAN and DP-GAN add calibrated noise for formal guarantees at the cost of realism.

Technique	Guarantee	Protects against	Utility cost
k -anonymity	Syntactic	Linkage	Generalization, suppression
ℓ -diversity / t -closeness	Syntactic	Homogeneity, skewness	More suppression
Differential privacy	Provable	Any attacker, any side data	Noise scales as $1/\epsilon$
Synthetic data	Empirical	Membership inference (variable)	5–15% downstream AUC loss

Consumer AI Lives Outside HIPAA

HIPAA binds covered entities (plans, clearinghouses, providers) and their business associates. OpenAI's consumer product is neither, so a patient who types a cancer diagnosis and chemo regimen into ChatGPT is sharing PHI with a company that has no HIPAA obligation: no BAA, no minimum security standard, no breach-notification duty, no right to an accounting of disclosures. The scale question is settled: ChatGPT Health launched as a limited, waitlist-gated product on January 7, 2026 and opened to all U.S. users eighteen and older on July 23, 2026, by which point OpenAI reported roughly **300 million** health-related queries per week. Its "enhanced privacy" (a separate chat history, no use for foundation-model training) is an internal policy commitment, not a federal mandate, and a policy can be changed unilaterally in a way a BAA cannot. ECRI named AI chatbot misuse the number one health technology hazard for 2026, with tens of millions of people asking health questions daily. State laws are filling the gap: Washington's My Health My Data Act (a private right of action) and California's CCPA/CPRA apply to any entity handling consumer health data, together covering about 50 million people. HIPAA compliance is now necessary but no longer sufficient.

Shadow AI Inside the Hospital

A January 2026 Wolters Kluwer survey found more than 40 percent of healthcare workers aware of colleagues using unauthorized AI tools, and 20 percent admitting personal use, mainly for faster workflow. The pattern is familiar: a resident pastes a discharge summary (with the patient's name on the screen) into ChatGPT, gets a polished draft in 30 seconds, and the PHI now sits on a consumer server outside the audit trail, an unlogged side door in a building with an otherwise monitored entrance. **Watch out.** The institution, not the individual, bears primary enforcement liability, and there is no audit trail, no clinical

validation, and no accountability. Banning consumer AI fails because of personal devices; the fix is providing governed alternatives (Health Spaces: HIPAA-compliant enterprise environments from OpenAI, Anthropic, and Epic's named agents) fast and capable enough to remove the incentive to go around the system.

Case Study: The \$4.75M Montefiore Fine

In February 2024 OCR fined Montefiore **\$4.75 million**. The hospital had a privacy officer, EHR encryption, and training, but could not map where ePHI lived; an employee stole the records of 12,517 patients and sold them to an identity-theft ring, and the theft ran undetected from 2013 to 2015. In 2026 ePHI flows through training pipelines, inference logs, model weights, embeddings, and every third-party API call, so a model trained on ePHI is itself a container of ePHI via memorization, and membership-inference and model-inversion attacks can recover it. Comprehensive asset inventories, once best practice, are mandatory under the proposed 2026 HIPAA Security Rule update.

Should You Build This?

Build it when health data must leave its original care context (research, model training, a vendor integration) and you can commit to privacy as engineering: a quantified re-identification risk, differential privacy or synthetic data where the use case tolerates the utility cost, and an ePHI map that includes training pipelines and inference logs. **Think twice when** the plan is Safe Harbor identifier-stripping plus good intentions, since three attributes unmasked a governor, or when the governed tool you offer clinicians is slower than the consumer chatbot it has to displace, because shadow AI fills every convenience gap you leave open. **Who bears the risk.** Patients, and irreversibly: a privacy failure is permanent for the person whose diagnosis leaked and a budget line for the entity that leaked it.

Quick Self-Test

1. Which three attributes let Sweeney uniquely identify 87 percent of Americans, and why does Safe Harbor fail to stop this?
2. A dataset is $k = 10$ anonymous but every member of a class has the same diagnosis. Which attack still

works, and which property fixes it?

3. In the differential privacy definition, what does epsilon control, and what happens to the A1c registry as epsilon drops from 1.0 to 0.01?
4. Why is differential privacy described as making queries private rather than data anonymous?
5. Why does a patient's ChatGPT health conversation fall outside HIPAA, and what fills the gap?
6. Montefiore had encryption and training. What single failure produced the \$4.75M fine, and why does it matter for a model's weights and logs?

If You Remember One Thing

De-identification is a legal checkbox; privacy is an engineering practice. The 18 identifiers are a floor that leaks the moment richer external data exists, which in 2026 is always. Layer real defenses: k-anonymity against linkage, differential privacy for a provable guarantee, synthetic data to remove real records from development, and comprehensive ePHI mapping that now must include training pipelines, inference logs, and model weights. Compliance is necessary; it was never sufficient.

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552