Reconstruction Attacks and Privacy Defenses

An Empirical Study of Healthcare Data Privacy

Benjamin Tran

CS 5510 - Data Privacy and Security

October 29, 2025

Abstract

This report investigates reconstruction attacks on sensitive healthcare data and evaluates three defense mechanisms: rounding, Gaussian noise, and subsampling. Using a synthetic dataset of 100 patient records, we demonstrate that reconstruction attacks can achieve 96.7% accuracy with 200 queries when no defenses are employed. Our experimental analysis reveals that Gaussian noise provides the strongest protection with the best privacy-utility trade-off, transitioning at $\sigma = 2$. Rounding offers moderate protection at $R = 5$, while subsampling proves completely ineffective, failing to prevent attacks at any parameter setting. Additionally, we provide a Bayesian interpretation of Membership Inference Attacks, highlighting the critical importance of minimizing false positive rates for reliable inference.

Contents

1 Introduction

Data-driven healthcare systems have created new opportunities for medical research and personalized treatment. However, these systems also pose significant privacy risks. Even when sensitive attributes are not directly disclosed, adversaries can exploit statistical queries to reconstruct private information about individuals.

This report examines reconstruction attacks, which are a class of privacy attacks where an adversary uses multiple statistical queries to infer sensitive attributes of individuals in a dataset. We focus on a healthcare scenario where an attacker attempts to reconstruct medical test results (normal vs. abnormal) using only aggregate queries over public demographic attributes.

1.1 Research Questions

Our investigation addresses three primary questions:

1. **Attack Feasibility:** How accurately can an adversary reconstruct sensitive binary attributes using subset-sum queries?
2. **Defense Effectiveness:** Which defense mechanisms (rounding, noise addition, or subsampling) effectively prevent reconstruction while maintaining data utility?
3. **Privacy-Utility Trade-offs:** What are the optimal parameter settings that balance privacy protection with query accuracy?

Additionally, we analyze Membership Inference Attacks (MIAs) from a Bayesian perspective to understand the role of false positive rates in privacy breaches.

1.2 Contributions

This work makes the following contributions:

- Empirical evaluation of reconstruction attacks across 100 parameter settings per defense mechanism
- Identification of precise transition points where defenses become effective
- Comparative analysis revealing Gaussian noise as the optimal defense strategy
- Demonstration that subsampling alone is insufficient for privacy protection
- Bayesian framework for understanding membership inference attack reliability

2 Background and Related Work

2.1 Reconstruction Attacks

Reconstruction attacks exploit the fact that multiple statistical queries can leak information about individual records. When an adversary can issue many queries, they can construct a system of linear equations and solve for individual sensitive values. The feasibility of such attacks depends on:

- The number of queries allowed
- The diversity of query predicates
- The presence of noise or other defenses

2.2 Defense Mechanisms

Three primary defense strategies have been proposed:

1. **Output Perturbation:** Adding noise to query results (e.g., Laplace or Gaussian noise)
2. **Rounding:** Discretizing query outputs to reduce information leakage
3. **Input Perturbation:** Modifying the underlying dataset through sampling or generalization

Our work evaluates representative mechanisms from each category in a controlled experimental setting.

3 Methodology

3.1 Dataset Description

We utilize a synthetic healthcare dataset containing 100 patient records. Each record consists of four public attributes and one sensitive binary attribute:

Public Attributes (known to the adversary):

- `age`: Patient age in years (range: 0–100)
- `sex`: Binary gender indicator (0 = male, 1 = female)
- `blood`: Blood type encoded as integers 0–7 representing A+, A-, B+, B-, AB+, AB-, O+, O-
- `admission`: Hospital admission type (0 = elective, 1 = urgent, 2 = emergency)

Sensitive Attribute (reconstruction target):

- `result`: Medical test outcome (0 = normal, 1 = abnormal)

The dataset contains 65 normal results (65%) and 35 abnormal results (35%), yielding a majority baseline accuracy of 0.650. This baseline serves as our threshold for attack success: reconstruction accuracy below 65% indicates effective defense.

3.1.1 Data and Code Sources

The synthetic healthcare dataset used in this study was obtained from the OpenDP CS208 course repository[1]. The initial implementation framework was adapted from the course's problem set starter code[2], which was extended to include the defense mechanisms, experimental framework, and evaluation metrics presented in this report.

The complete implementation, including all experimental code, data files, and generated visualizations, is available in the project repository[3].

[1] https://github.com/opendp/cs208/blob/main/spring2025/data/fake_healthcare_dataset_sample100.csv
[2] https://github.com/opendp/cs208/blob/main/spring2025/homeworks/ps2/ps2_starter.py
[3] https://github.com/btranTFT/CS5510-Homework-1

3.2 Attack Model

The adversary has access to a query interface that computes subset-sum queries. For any boolean predicate q defined over the public attributes, the interface returns:

$$\text{Answer}(q) = \sum_{i:q(x_i)=\text{True}} \text{result}_i \tag{1}$$

where x_i represents the public attributes of patient i.

3.2.1 Attack Strategy

Our reconstruction attack proceeds as follows:

1. **Query Generation:** Generate 200 random boolean predicates over the public attributes (corresponding to $2n$ queries for $n = 100$ patients)
2. **Query Execution:** Submit each predicate to the query interface and collect responses
3. **System Construction:** Formulate a system of linear equations $Ax = b$ where:
   - A is a 200×100 binary matrix (each row represents which patients satisfy a predicate)
   - x is the unknown 100-dimensional sensitive attribute vector
   - b is the 200-dimensional vector of query responses
4. **Optimization:** Solve the overdetermined system using least-squares optimization
5. **Thresholding:** Round the continuous solution to binary values using threshold 0.5

The implementation uses least-squares regression to solve the optimization problem.

3.3 Defense Mechanisms

We evaluate three defense mechanisms, each parameterized to allow fine-grained analysis of privacy-utility trade-offs.

3.3.1 Defense 1: Rounding (Parameter R)

Each query answer is rounded to the nearest multiple of R:

$$\text{Answer}_{\text{defended}} = R \cdot \left\lfloor \frac{\text{Answer}_{\text{exact}}}{R} + 0.5 \right\rfloor \tag{2}$$

Conclusion: Rounding reduces the precision of query answers, making it harder to distinguish between similar queries. Larger R provides stronger privacy but introduces more distortion.

3.3.2 Defense 2: Gaussian Noise (Parameter σ)

Random noise sampled from $\mathcal{N}(0, \sigma^2)$ is added to each answer:

$$\text{Answer}_{\text{defended}} = \text{Answer}_{\text{exact}} + \mathcal{N}(0, \sigma^2) \tag{3}$$

Conclusion: Noise obscures the exact query result, introducing uncertainty that propagates through the reconstruction process. Larger σ provides stronger privacy but reduces answer accuracy.

3.3.3 Defense 3: Subsampling (Parameter t)

A random subset of t patients is sampled, the query is computed on this subset, and the result is scaled:

$$\text{Answer}_{\text{defended}} = \frac{n}{t} \cdot \sum_{i \in S_t} \text{result}_i \tag{4}$$

where S_t is a uniformly random subset of size t.

Conclusion: Subsampling introduces variance by computing on incomplete data. Smaller t provides stronger privacy but increases variance dramatically.

3.4 Experimental Design

3.4.1 Parameter Ranges

To evaluate each defense, we test the full range of integer parameters from 1 to 100:

- **Rounding:** $R \in \{1, 2, 3, \ldots, 100\}$
- **Gaussian Noise:** $\sigma \in \{1, 2, 3, \ldots, 100\}$
- **Subsampling:** $t \in \{1, 2, 3, \ldots, 100\}$

3.4.2 Evaluation Protocol

For each defense and parameter value:

1. Run 10 independent trials with different random seeds
2. Generate 200 fresh random predicates per trial
3. Execute the reconstruction attack
4. Compute evaluation metrics

3.4.3 Evaluation Metrics

We measure both utility (query accuracy) and privacy (reconstruction difficulty):

- **Root Mean Squared Error (RMSE):** Measures the average distortion between defended and exact query answers:

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{j=1}^{m} (\text{Answer}^{(j)}_{\text{defended}} - \text{Answer}^{(j)}_{\text{exact}})^2} \tag{5}$$

  where $m = 200$ is the number of queries.

- **Success Rate:** Fraction of correctly reconstructed sensitive values:

$$\text{Success Rate} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[\text{predicted}_i = \text{actual}_i] \tag{6}$$

- **Transition Point:** The smallest parameter value where success rate falls below the majority baseline (0.650)

4 Results

4.1 Baseline Attack Performance

Without any defenses, the reconstruction attack achieves great accuracy:

- **Success Rate:** 96.7% (97 out of 100 sensitive values correctly reconstructed)
- **Exact Reconstruction:** Achieved in multiple trials
- **RMSE:** 0.00 (queries return exact values)

This demonstrates the severe privacy risk posed by unrestricted statistical queries, even when only aggregate information is released.

4.2 Defense 1: Rounding

Table 1 summarizes the performance of the rounding defense across key parameter values.

Table 1: Rounding Defense Performance

Parameter R	**RMSE**	**Success Rate**	**Status**
1	0.00	0.967	Attack succeeds
2	0.65	0.733	Attack succeeds
3	0.76	0.688	Attack succeeds
4	1.10	0.666	Attack succeeds
5	**1.20**	**0.606**	**Attack fails**
6	1.56	0.584	Attack fails
8	2.16	0.570	Attack fails
10	2.62	0.585	Attack fails

Key Findings:

- **Transition Point:** $R = 5$ is the threshold where reconstruction success drops to 60.6%, falling below the majority baseline
- **Sharp Transition:** Success rate drops from 66.6% at $R = 4$ to 60.6% at $R = 5$, indicating a relatively sharp phase transition
- **Utility Cost:** At the transition point, RMSE is only 1.20, representing minimal distortion
- **Discrete Behavior:** Both RMSE and success rate exhibit step-like patterns due to the discrete nature of rounding

Interpretation: Rounding provides moderate privacy protection with predictable utility degradation. The defense becomes effective at relatively small parameter values, making it practical for scenarios requiring deterministic query answers.

4.3 Defense 2: Gaussian Noise

Table 2 presents the performance of the Gaussian noise defense.

Table 2: Gaussian Noise Defense Performance

Parameter σ	RMSE	Success Rate	Status
1	1.02	0.653	Attack succeeds
2	**1.99**	**0.598**	**Attack fails**
3	3.03	0.545	Attack fails
4	4.06	0.534	Attack fails
5	5.02	0.524	Attack fails
6	5.88	0.519	Attack fails
8	8.11	0.519	Attack fails
10	10.34	0.526	Attack fails

Key Findings:

- **Transition Point:** $\sigma = 2$ provides effective protection, with success dropping to 59.8%
- **Sharpest Transition:** The defense exhibits the most extreme phase transition among all three mechanisms
- **Excellent Trade-off:** At $\sigma = 2$, RMSE is only 1.99 while providing strong privacy protection
- **Linear RMSE Growth:** RMSE increases approximately linearly with σ, making utility costs predictable
- **Diminishing Returns:** Success rate plateaus around 51% for large σ, suggesting a lower limit

Interpretation: Gaussian noise emerges as the most effective defense, offering the best privacy-utility trade-off. Even minimal noise ($\sigma = 2$) significantly degrades attack performance while maintaining reasonable query accuracy.

4.4 Defense 3: Subsampling

Table 3 shows the performance of the subsampling defense.

Table 3: Subsampling Defense Performance

Parameter t	RMSE	Success Rate	Status
1	22.42	0.655	Attack succeeds
5	8.65	0.658	Attack succeeds
10	6.90	0.681	Attack succeeds
20	4.85	0.716	Attack succeeds
50	2.05	0.801	Attack succeeds
75	1.10	0.883	Attack succeeds
90	0.67	0.929	Attack succeeds
95	0.44	0.951	Attack succeeds
100	**0.00**	**0.962**	**No defense**

Key Findings:

- **No Privacy Protection:** Subsampling fails to prevent the attack across the entire parameter range. All success rates remain at or above the 65% baseline
- **Success Rate Increases with t:** As t increases, success rate rises from 65.5% at $t = 1$ to 96.2% at $t = 100$
- **Extreme Utility Cost at Small t:** Small t values produce massive RMSE (22.42 at $t = 1$), making the data nearly unusable
- **Complete Failure as Defense:** Even at the most extreme setting ($t = 1$), reconstruction success remains above baseline, indicating the defense is completely ineffective

Interpretation: Subsampling is ineffective as a privacy defense in this setting. Even with extreme subsampling ($t = 1$), the attack succeeds at 65.5%, above the 65% majority baseline. As t increases, utility improves but privacy protection remains absent. Unlike noise and rounding, which create barriers to reconstruction, subsampling only introduces variance that does not prevent the attack from succeeding.

4.5 Comparative Analysis

The following sections compare all three defense mechanisms based on their effectiveness, privacy-utility trade-offs, and practical applicability.

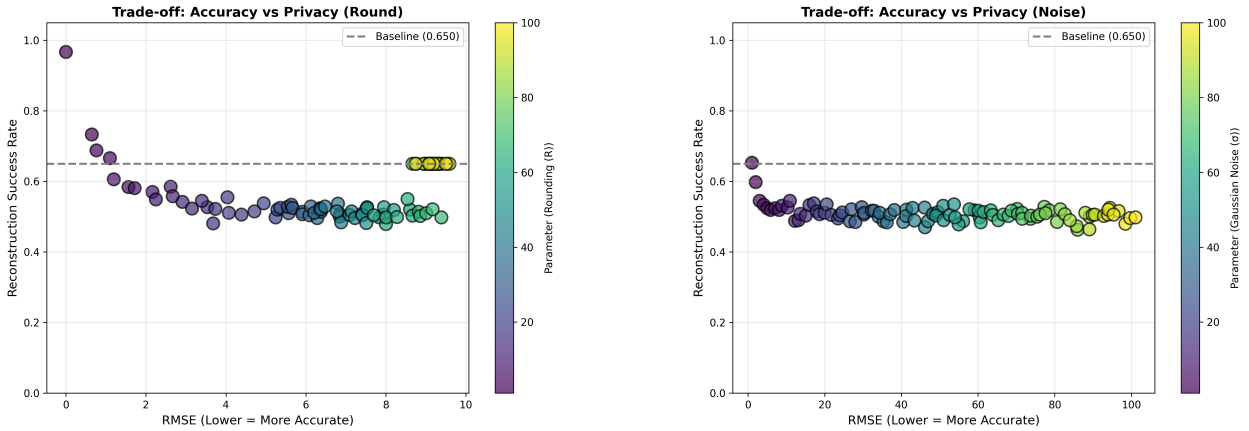
4.5.1 Defense Effectiveness Ranking

1. **Gaussian Noise** ($\sigma^* = 2$): Strongest protection with best trade-off. Success drops to 59.8%, RMSE = 1.99
2. **Rounding** ($R^* = 5$): Moderate protection with excellent utility. Success drops to 60.6%, RMSE = 1.20
3. **Subsampling**: Completely ineffective. All parameter values fail to prevent the attack (success $\geq$ 65% baseline across entire range from $t = 1$ to $t = 100$)

4.5.2 Privacy-Utility Trade-offs

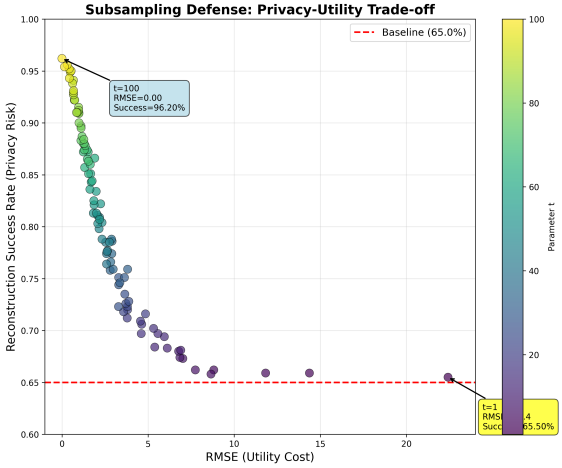
- **Rounding:** Offers the best utility (lowest RMSE = 1.20) at the transition point, but provides weaker privacy than noise
- **Gaussian Noise:** Provides the strongest privacy with acceptable utility cost (RMSE = 1.99)
- **Subsampling:** Provides no privacy protection at any parameter value. Even at extreme settings ($t = 1$, RMSE=22.42), attack success remains above baseline

Figure 1 illustrates the privacy-utility trade-offs for all three defense mechanisms, showing the relationship between reconstruction success rate (privacy) and RMSE (utility) across different parameter values.



(a) Rounding

(b) Gaussian Noise



(c) Subsampling

Figure 1: Privacy-Utility Trade-offs: Comparison of all three defense mechanisms showing success rate vs. RMSE. Lower success rates and lower RMSE values are desirable (bottom-left corner). Gaussian noise provides the best trade-off, while subsampling fails to provide meaningful privacy protection.

4.5.3 Practical Recommendations

Based on our empirical findings:

- **Recommended:** Use Gaussian noise with $\sigma \geq 2$ for effective privacy protection with minimal utility loss (RMSE ≈ 2)
- **Alternative:** Use rounding with $R \geq 5$ when deterministic answers are required, offering slightly better utility (RMSE ≈ 1.2)
- **Avoid:** Do not use subsampling as a standalone privacy defense. It provides no protection against reconstruction attacks at any parameter setting

5 Bayesian Interpretation of Membership Inference Attacks

Beyond reconstruction attacks, we analyze Membership Inference Attacks (MIAs) through a Bayesian lens to understand when such attacks provide reliable evidence of membership.

5.1 Problem

Consider an attacker who believes that an individual Alice is in a dataset with prior probability p. We express this belief as prior odds:

$$O_{\text{prior}} = \frac{p}{1-p} \tag{7}$$

An MIA algorithm examines the dataset and outputs either "In" or "Out". The algorithm is characterized by:

- **True Positive Rate (TPR):** $P(\text{MIA says "In"} \mid \text{Alice actually in})$
- **False Positive Rate (FPR):** $P(\text{MIA says "In"} \mid \text{Alice actually out})$

Question: If the MIA returns "In", what are the posterior odds O_{post} that Alice is actually in the dataset?

5.2 Derivation

Using Bayes' theorem, we compute the posterior odds:

$$O_{\text{post}} = \frac{P(\text{In} \mid \text{"In"})}{P(\text{Out} \mid \text{"In"})} \tag{8}$$

$$= \frac{P(\text{"In"} \mid \text{In}) \cdot P(\text{In})}{P(\text{"In"} \mid \text{Out}) \cdot P(\text{Out})} \tag{9}$$

$$= \frac{\text{TPR} \cdot p}{\text{FPR} \cdot (1-p)} \tag{10}$$

$$= \frac{\text{TPR}}{\text{FPR}} \cdot \frac{p}{1-p} \tag{11}$$

Therefore:

$$\boxed{O_{\text{post}} = \frac{\text{TPR}}{\text{FPR}} \cdot O_{\text{prior}}} \tag{12}$$

The ratio TPR/FPR is called the **likelihood ratio** and quantifies how much evidence the MIA provides.

5.3 Significance of Small False Positive Rate

The likelihood ratio reveals why minimizing FPR is critical for reliable membership inference, even when TPR is high.

5.3.1 Example 1: Large FPR (Weak Evidence)

Consider an MIA with perfect detection but poor specificity:

- TPR = 1.0 (detects all members)
- FPR = 0.5 (50% false positive rate)
- Likelihood ratio = $1.0/0.5 = 2$

A positive result only *doubles* our prior belief. If we initially believed Alice had a 10% chance of membership ($O_{\text{prior}} = 0.111$), the posterior odds become $O_{\text{post}} = 0.222$, corresponding to only 18% probability, which is unlikely.

5.3.2 Example 2: Small FPR (Strong Evidence)

Now consider an MIA with excellent specificity:

- TPR = 1.0 (detects all members)
- FPR = 0.01 (1% false positive rate)
- Likelihood ratio = $1.0/0.01 = 100$

A positive result multiplies our belief by 100! With the same 10% prior, the posterior odds become $O_{\text{post}} = 11.1$, corresponding to 91% probability, which is a very strong evidence of membership.

5.3.3 Conclusion

A small FPR is important because:

1. It makes the likelihood ratio large, providing strong evidence when the MIA claims membership
2. With large FPR, many false positives occur, making positive results unreliable
3. High TPR alone is insufficient, because specificity (low FPR) determines the evidential value
4. This explains why differential privacy mechanisms focus on limiting distinguishability (related to FPR control)

6 Discussion

6.1 Implications for Privacy Protection

Our results demonstrate that reconstruction attacks pose a serious threat to privacy in systems that release aggregate statistics. Even with 200 queries on a dataset of 100 individuals, an adversary can reconstruct 96.7% of sensitive values without any defenses.

The effectiveness of defenses varies:

- **Gaussian noise** provides strong protection with minimal utility loss, making it a good choice for most applications
- **Rounding** offers an alternative with slightly better utility but weaker privacy
- **Subsampling** is ineffective as a standalone defense in this setting and should not be relied upon for privacy protection

6.2 Limitations

Our study has several limitations:

1. **Dataset Size:** We evaluate on a small dataset (100 records). Larger datasets may exhibit different behavior, though reconstruction attacks typically become *easier* with more queries.
2. **Binary Attributes:** We focus on binary sensitive attributes. Continuous or categorical attributes may require different attack strategies.
3. **Query Budget:** We fix the number of queries at 200 ($2n$). Real-world scenarios may involve different query budgets.
4. **Single Defense:** We evaluate defenses in isolation. Combining multiple defenses (e.g., noise + rounding) may provide stronger protection.
5. **Optimization Method:** Our attack uses least-squares optimization. Other attack algorithms might achieve higher success rates.

6.3 Future Work

Several directions warrant further investigation:

- **Adaptive Attacks:** Develop attacks that adapt query strategies based on intermediate results
- **Combined Defenses:** Evaluate combinations of multiple defense mechanisms
- **Differential Privacy:** Compare our defenses to formal differential privacy mechanisms
- **Real-World Data:** Validate findings on actual healthcare datasets
- **Utility Metrics:** Develop application-specific utility metrics beyond RMSE
- **Composition:** Analyze how privacy degrades over multiple query sessions

7 Conclusion

This report presents an empirical evaluation of reconstruction attacks and privacy defenses on healthcare data. Our key findings are:

1. Reconstruction attacks are highly effective, achieving 96.7% accuracy with 200 queries on undefended data
2. Gaussian noise provides the strongest privacy protection, with a sharp transition at $\sigma = 2$ and excellent privacy-utility trade-off (RMSE = 1.99)
3. Rounding offers moderate protection at $R = 5$ with better utility (RMSE = 1.20) but weaker privacy than noise
4. Subsampling is ineffective as a privacy defense in this setting, failing to prevent attacks across all parameter values from $t = 1$ to $t = 100$
5. From a Bayesian perspective, small false positive rates are critical for reliable membership inference, regardless of true positive rates

These results show the importance of carefully designed privacy defenses in systems that release statistical information. Organizations handling sensitive data should implement strong protections, particularly Gaussian noise or rounding, to prevent reconstruction attacks while maintaining data utility for legitimate analyses.

The tension between privacy and utility is unavoidable. However, our findings show that with appropriate defenses and parameter choices, it is possible to achieve meaningful privacy protection with acceptable utility costs. As data-driven systems become more common in healthcare and other sensitive domains, such privacy-preserving mechanisms will be important for maintaining public trust and protecting individual rights.

Tools and Software

This project utilized the following tools and software packages:

Implementation and Analysis:

- **Python 3.x**: Primary programming language for implementation
- **NumPy**: Numerical computing, matrix operations, and least-squares optimization
- **Pandas**: Data manipulation and analysis

Visualization:

- **Matplotlib**: Generation of all plots and figures including RMSE plots, success rate curves, and privacy-utility trade-off visualizations

Report Preparation:

- **LaTeX**: Document typesetting and formatting
- **AI Assistance**: Artificial intelligence tools were used to assist with grammar checking, sentence structure refinement, and formatting consistency throughout this report. All technical content, analysis, and conclusions are the author's own work.

Acknowledgments

This work was completed as part of CS5510 Data Privacy and Security. The implementation uses standard Python scientific computing packages (NumPy, Pandas, Matplotlib).