universität koblenz

weiter:denken

COMPARATIVE ANALYSIS OF PRIVACY PRESERVING TECHNIQUES IN HEALTHCARE DATA

BY

ABHISHEK ABHISHEK

(222100422)

MASTER'S THESIS

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Web and Data Science

First supervisor:	Prof. Dr Andreas Mauthe Institut fur Wirtschafts- und Verwaltungsinformatik
Second supervisor:	Christopher Latz

Koblenz, 23 August 2025

Statutory declaration

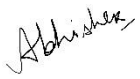
I hereby confirm that this thesis was written independently and that I have not used any resources other than those indicated—in particular, no internet sources not listed in the bibliography—and that I have not previously submitted this thesis to any other examination procedure. The paper submission is identical to the submitted electronic version.

Do I agree to have the work placed in the library? ☒ Yes ☐ No

Do I agree to the publication of this work on the Internet? ☒ Yes ☐ No

Koblenz, 23rd August 2025

(Place and Date)



(Signature)

ABSTRACT

In today’s data-driven healthcare landscape, patient information plays a vital role since it supports clinical decisions, advances medical research throughout, and guides health policies. Safeguarding patient privacy is now a meaningful difficulty too. Due to medical records' highly sensitive nature, this protection can be difficult. In order to tackle this issue, researchers have developed various approaches which preserve privacy, and they designed all of them in order to balance data's usefulness and safeguard individual confidentiality.

This thesis focuses on a systematic and practical comparison of four well known anonymization approaches—k-anonymity, l-diversity, t-closeness, and differential privacy - using a Diabetes 130-US Hospitals (1999–2008) dataset obtained from the UCI Machine Learning Repository. Building on systematic literature review, the study will first identify existing knowledge gaps by examining advanced implementations of these techniques in healthcare.

The practical component will involve applying each technique to the dataset, followed by evaluating three critical parameters for determining the feasibility and effectiveness of privacy-preserving methods in real-world clinical environments. First, algorithmic execution time will be measured to display computational overhead. Second, data utility will be assessed through query accuracy method, measuring how closely the anonymized data answers typical queries. Third, re-identification risk assessment will be conducted using the ARX anonymization tool, an all-in-one open-source software that implements multiple privacy criteria and provides methods for analyzing re-identification risks.

By comparing performance across these parameters, the research aims to highlight the strengths, weaknesses, and practical trade-offs of each technique. This complete evaluation will offer actionable insights for healthcare practitioners, data scientists, and policymakers tasked with protecting patient privacy while leveraging data for improved healthcare outcomes. In the end, these findings aim to push forward the conversation around privacy-conscious data analysis in healthcare. They can also provide practical guidance on choosing the right anonymization methods depending on different clinical needs and regulatory requirements.

The importance of this study is in offering practical, evidence-backed direction to healthcare organizations as they manage the challenges of data privacy, all while ensuring that medical research and clinical decisions can still benefit from valuable data insights. As healthcare data breaches continue to pose significant threats to patient confidentiality, this research aims at creating durable privacy protections frameworks that can be implemented in diverse healthcare settings.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my first supervisor, Prof. Dr. Andreas Mauthe, and my second supervisor, Christopher Latz, for their invaluable guidance and support throughout the development of this thesis.

I am incredibly grateful to Christopher Latz for his dedicated involvement at every step of the process - from helping me brainstorm and narrow down my ideas into a focused topic, assisting in the formulation of my thesis proposal, and solving my queries in regular meetings. His encouragement, timely feedback, and patience in addressing my doubts helped me remain focused and motivated throughout this journey.

I would also like to thank my family and friends for their unwavering encouragement and support during the challenging phases of my research.

Finally, my sincere thanks go to the University of Koblenz for providing the academic resources and an enriching environment that enabled me to successfully complete this research.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

API — Application Programming Interface

ARX — Anonymization Tool for Data Privacy (ARX**)**

CSV — Comma-Separated Values

EHR — Electronic Health Record

GDPR — General Data Protection Regulation

HIPAA — Health Insurance Portability and Accountability Act

ICD-9 — International Classification of Diseases, Ninth Revision

ML — Machine Learning

DP — Differential Privacy

ARE — Average Relative Error

MRE — Mean Relative Error

CV — Co-efficient of Variation

QI — Quasi Identifier

1 INTRODUCTION

The exponential growth of digitized medical records and the widespread adoption of electronic health record (EHR) systems have transformed the way healthcare data is managed, shared, and analysed [1][2]. This transformation has created extraordinary opportunities for clinical research, population health analytics, and evidence-based decision-making. However, the sensitive nature of healthcare data, combined with the increasing sophistication of re-identification techniques, poses significant privacy challenges. Simply removing direct identifiers is no longer sufficient, as quasi-identifiers—such as age, gender, and ZIP code—can be linked to external sources to reveal personal identities [3][4][6].

These challenges highlight a critical tension: healthcare data must be protected to maintain patient confidentiality and comply with regulations such as HIPAA and GDPR [6][7][8], yet it must also retain enough analytical value to remain useful for research, policy-making, and innovation. Over the past two decades, a variety of privacy-preserving models—such as k-anonymity, ℓ-diversity, t-closeness, and differential privacy—have been proposed to address this tension. While their theoretical foundations are well-established, their practical performance, scalability, and impact on data utility in real-world healthcare contexts remain less understood when they are put side by side and compared.

This thesis focuses on bridging this gap by conducting a systematic, practical comparison of these four privacy-preserving techniques using a large, real-world healthcare dataset. By assessing execution performance, data utility, and re-identification risk, the study aims to provide actionable insights for selecting suitable anonymization approaches in healthcare data publishing scenarios where both privacy protection and data usefulness are critical.

1.1 Motivation

Despite decades of progress in privacy-preserving techniques, a critical gap remains in understanding how these methods perform side by side when applied to large, complex healthcare datasets under realistic operational conditions [9]. Most prior studies have evaluated methods in isolation, focused on simplified datasets, or examined only a subset of relevant performance dimensions. This leaves healthcare practitioners with little practical evidence to guide the selection of anonymization strategies that balance privacy, data utility, and computational efficiency.

Healthcare data presents a particularly demanding test case. It contains rich combinations of quasi-identifiers and sensitive attributes that make re-identification risk a constant concern, while also serving as a vital resource for clinical research, public health, and policy-making. Excessive anonymization can severely degrade data utility, undermining the very purpose of data sharing, whereas insufficient anonymization can compromise patient privacy and violate

regulatory standards. These competing priorities demand a careful and evidence-based approach to privacy protection [10].

The demand for this kind of approach becomes even more evident in real-world operations. In hospitals and research environments, datasets can include millions of records and often need constant updating. This makes it essential for anonymization techniques to be not only reliable but also efficient to run. Even methods that look strong in theory may prove unusable in practice if they require too much time or computing power. Yet, performance and scalability considerations have received limited attention in academic research, creating a disconnect between theoretical models and their practical implementation in healthcare environments [10].

This issue is especially pressing because it sits at the heart of modern healthcare. Without strong and scalable privacy protections, the ability to share data for research and innovation is limited, and, just as importantly, patient trust is put at risk. By directly confronting the real-world constraints faced by hospital IT departments, clinical researchers, and policymakers, privacy-preserving data analytics becomes not just an academic exercise but a practical necessity for improving healthcare outcomes.

Addressing this challenge is essential for enabling safe and effective healthcare data sharing. By conducting a direct, comparative evaluation of k-anonymity, ℓ-diversity, t-closeness, and differential privacy on a real-world healthcare dataset, this study aims to provide practitioners with actionable insights into how these methods perform not only in protecting privacy, but also in maintaining analytical value and meeting computational constraints. Providing evidence-based guidance is essential to make sure anonymization practices uphold the ethical duty to protect patient privacy while also meeting the broader need to advance medical research and knowledge.

1.2 Research Objectives

This thesis seeks to provide actionable, evidence-based guidance for protecting patient privacy while maintaining the utility of healthcare data. To achieve this, the study performs a rigorous comparative analysis of four leading anonymization models—k-anonymity, ℓ-diversity, t-closeness, and differential privacy—using a large, authentic healthcare dataset, the Diabetes 130-US Hospitals (1999–2008) dataset [73].

Three core objectives structure this investigation:

- **Computational Efficiency and Scalability:** Evaluating how each privacy technique performs with realistic, large-scale datasets, including measurement of execution times and scalability as data volume increases. The goal is to determine which methods are practical for operational healthcare environments, where new data must be anonymized regularly and quickly. By systematically testing each technique across varying dataset sizes (e.g., 25,000, 50,000, 75,000, 100,000 records), we will characterize how computational requirements scale with data volume and identify potential bottlenecks that could limit practical deployment [11].

- **Data Utility Preservation: Systematically assessing how anonymization methods affect the** analytical usefulness of healthcare data—whether the protected data remains accurate and reliable for common research queries and clinical analytics [12]. This ensures privacy protection does not accidently undermine the very goals of data sharing [36]. High data utility indicates that researchers and practitioners can continue deriving meaningful, actionable insights from protected data, ensuring that privacy protection does not unintentionally undermine the research objectives that motivate data sharing [38].
- **Privacy Risk Mitigation**: Quantifying each technique's ability to reduce re-identification risk, under various attacker models and scenarios. This includes formal privacy metrics, practical assessment, and an assessment of any remaining vulnerabilities, making sure the selected methods live up to strict standards of confidentiality. This objective involves quantifying re-identification risk under various attack scenarios, utilizing established privacy metrics, and conducting controlled experiments to assess remaining privacy vulnerabilities [13]. The evaluation ensures comparison of methods based on their fundamental purpose: lowering the chances that individual patients could be traced or identified from the shared datasets.

In pursuit of these objectives, the thesis addresses three interconnected research questions that collectively ensure holistic evaluation across all relevant dimensions:

Research Question 1 (Computational Performance): What are the algorithmic runtimes of each anonymization technique when applied to healthcare datasets of varying sizes and complexity?

This question systematically compares methods regarding computational efficiency: processing time required to produce anonymized datasets and behavior as dataset size increases. Through controlled experiments using different data subset sizes (ranging from 25,000 to 100,000+ records), we will characterize how execution time grows with data volume and identify whether any method becomes computationally prohibitive for large-scale applications. This question directly addresses the practical feasibility of each approach in real-world deployment scenarios where time face operational constraints.

Research Question 2 (Data Utility): How effectively do k-anonymity, ℓ-diversity, t-closeness, and differential privacy preserve the analytical value and research utility of healthcare data?

This question explores whether analyses carried out on anonymized healthcare data produce findings that are consistent with those derived from the original, non-anonymized datasets. The evaluation encompasses multiple analytical scenarios like simple and complex queries involving multiple attributes and conditions. We will assess utility through metrics such as query accuracy (measuring differences between results from original versus anonymized data). This research question aims to identify which techniques best maintain data utility under various analytical requirements and determine the conditions under which utility preservation is maximized [14].

Research Question 3 (Privacy Protection): How effectively does each anonymization method reduce re-identification risk and protect patient privacy, under different attacker models?

This question investigates the core privacy protection capabilities of each technique through both theoretically (for differential privacy) and practical assessment. The evaluation involves conducting controlled re-identification experiments, calculating formal privacy risk metrics, and assessing vulnerability to different attacker models (Prosecutor Model, Journalist Model, Marketer Model). For differential privacy, this includes assessing the relationship between privacy parameter (ε) values and actual privacy protection achieved. The goal is to compare privacy protection levels across methods and verify whether each technique meets acceptable privacy standards while identifying any residual vulnerabilities that could compromise patient confidentiality [13].

These research questions are designed to provide complementary perspectives that, when considered together, enable comprehensive evaluation of each privacy-preserving technique. Rather than optimizing any single dimension, the analysis explicitly recognizes that practical anonymization success requires achieving acceptable performance across all three areas simultaneously. The core assumption behind this study is that no single technique will outperform all others in every situation. Rather, each method is expected to come with its own trade-offs and be best suited to specific contexts. By systematically answering these research questions, the thesis will generate actionable insights that transform the general question "which privacy model is best?" into elegant, evidence-based guidance about which model proves most suitable under specific operational conditions, analytical objectives, and organizational priorities.

1.3 Thesis Structure

The remainder of this thesis is organized into six more comprehensive chapters, each building upon previous findings to develop a complete understanding of privacy-preserving techniques in healthcare data applications.

Chapter 2 – Literature Review surveys foundational and contemporary research in healthcare data anonymization, detailing the theoretical basis of k-anonymity, ℓ-diversity, t-closeness, and differential privacy. It critically examines prior comparative studies, highlights methodological limitations, discusses computational complexity and privacy risk assessment gaps, and identifies the research contributions of this work. **Chapter 3 – Theoretical Foundations** defines the formal privacy models, explains data utility and relative error metrics, outlines re-identification risk measurement methodologies, and presents the capabilities of the ARX anonymization tool used in this study. **Chapter 4 – Methodology** details the experimental environment, tools, and implementation for each privacy model, including parameterization, generalization and suppression strategies, compliance verification, query-based utility evaluation, and re-identification risk analysis. It also explains the batch processing of different dataset sizes and parameter values. **Chapter 5 – Dataset and Pre-processing** describes the Diabetes 130-US Hospitals dataset, its structure and attributes, pre-processing steps, and the

selection of quasi-identifiers and sensitive attributes. **Chapter 6 – Results and Discussion** presents execution performance, data utility outcomes, and re-identification risk results for all techniques, accompanied by comparative analyses. The discussion interprets the results in light of the research questions, explores privacy–utility–performance trade-offs, and relates findings to prior work. **Chapter 7 – Conclusion, Limitations and Future Work** synthesizes the main findings, provides recommendations for selecting privacy-preserving methods under specific operational conditions, and proposes future research directions, including hybrid approaches and evaluation on diverse healthcare datasets.

2 LITERATURE REVIEW

2.1 Privacy-Preserving Techniques in Healthcare

The field of privacy-preserving data publishing has advanced considerably, largely in response to growing awareness of the risks of re-identifying individuals within datasets that were thought to be anonymized. Healthcare data presents unique challenges due to its inherent sensitivity, complex correlational structures, and regulatory requirements that vary across jurisdictions [15]. This section critically examines four foundational approaches that have shaped the field: k-anonymity, ℓ-diversity, t-closeness, and differential privacy, analysing not merely their theoretical contributions but also their practical limitations and real-world applicability in healthcare contexts.

2.1.1 k-Anonymity: Foundational Concepts and Critical Assessment

Introduction of k-anonymity in [3] a paradigm shift from ad hoc identifier removal toward systematic, measurable privacy protection [3]. Her seminal demonstration involving Massachusetts voter records and hospital discharge data revealed that about 87% of people in United States could be uniquely identified using just three details: their 5-digit ZIP code, gender, and date of birth [3]. This discovery overturned long-held assumptions about data anonymity and led to the development of k-anonymity as the first formal privacy model offering measurable protections.

The k-anonymity principle requires that, a dataset is said to provide k-anonymity when the information for any individual cannot be distinguished from at least k-1 other people whose data is also included in the same release [3]. While conceptually straightforward, the practical implementation involves complex optimization problems that balance information loss against privacy protection [17][16].

However, critical analysis reveals fundamental limitations that compromise k-anonymity's effectiveness in healthcare applications. [18] showed that the basic k-anonymity model tends to be overly cautious when applied to the journalist re-identification scenario. This results in significant information loss, which becomes even more severe when working with small sampling fractions. This finding has profound implications for healthcare datasets, which often contain rare conditions or treatments that create natural sparsity.

More fundamentally, k-anonymity's vulnerability to homogeneity attacks represents a critical flaw in healthcare contexts where patients with similar demographics may share common diagnoses or treatments [19]. Recent systematic reviews have confirmed that no approaches have yet been developed for protection against membership disclosure attacks on diagnosis codes [20], pointing to ongoing weaknesses that remain even after more than twenty years of research.

The computational complexity of achieving optimal k-anonymity presents additional challenges for large-scale healthcare implementations. A research even demonstrated that

while their Optimal Lattice Anonymization (OLA) algorithm improved upon existing approaches, it still required significant computational resources that may be prohibitive for routine clinical data sharing [8]. This challenge of scalability is especially pressing when it comes to the real-time data processing demands of today's electronic health record systems.

2.1.2 ℓ-Diversity: Addressing Homogeneity While Introducing New Vulnerabilities

The introduction of ℓ-diversity in 2007 represented a principled response to k-anonymity's homogeneity vulnerability [19]. Under this approach, an equivalence class is considered to have ℓ-diversity when it includes at least ℓ well-represented values for the sensitive attribute [19]. This requirement ensures that even successful group identification cannot lead to definitive inference about sensitive attributes.

The concept of "well-represented" values proves problematic in practice. As observed in another study, knowing that values are different is not enough if those values are all equally undesirable or very similar [21]. In healthcare, this similarity attack could apply to related diagnoses (e.g., different types of diabetes) or treatment categories that patients might consider equally sensitive.

Recent systematic reviews of healthcare anonymization reveal that ℓ-diversity is "utilized mainly for demographic data, often in combination with another algorithm" [20], suggesting recognition among practitioners that ℓ-diversity alone provides insufficient protection. This observation aligns with theoretical critiques highlighting the model's vulnerability to skewness attacks where adversaries exploit knowledge of global attribute distributions [22].

2.1.3 t-Closeness: Statistical Sophistication at Computational Cost

In 2007 introduction of t-closeness represented the culmination of group-based anonymization approaches, addressing distributional vulnerabilities through rigorous statistical constraints [21]. The model specifies that the distribution of a sensitive attribute within a given class should not differ from its overall distribution in the dataset by more than a set threshold t [21], typically measured using Earth Mover's Distance or similar metrics.

The strength of t-closeness comes from the way it directly addresses distributional attacks. By ensuring that the distribution of attributes within each group stays close to the overall dataset—within a tolerance level of t—the model makes it harder for attackers to exploit unusual patterns to draw inferences. However, this theoretical strength comes at significant practical costs that limit its healthcare applicability.

Computational complexity represents t-closeness's most significant limitation. Unlike k-anonymity or ℓ-diversity, achieving t-closeness requires complex distributional calculations across potentially numerous sensitive attributes. Research on parallel anonymization algorithms confirms that t-closeness may impose a heavy computational burden, particularly "when processing the very large datasets commonly found in medical contexts" [23].

Furthermore, t-closeness's distributional constraints can lead to excessive information loss in naturally skewed healthcare data [21]. Clinical datasets often contain rare conditions, procedures, or outcomes that deviate significantly from population norms. Enforcing

distributional similarity may require substantial generalization or suppression that eliminates precisely the outlier cases that are often most clinically significant.

Recent comparative studies reveal that t-closeness achieves "strong protection against attribute disclosure but came with higher computation time and complexity" [24]. This trade-off becomes particularly problematic in healthcare environments requiring frequent data updates or real-time processing capabilities.

2.1.4 Differential Privacy: Theoretical Rigor Meets Practical Challenges

Introduction of differential privacy in [25] marked a fundamental departure from group-based approaches toward probabilistic guarantees that are robust against attackers with arbitrary background knowledge. The model guarantees that whether a single person's data is included or not in the dataset will not have a meaningful impact on the algorithm's output [25], providing mathematically provable protection regardless of adversary capabilities.

The theoretical strengths of differential privacy have made it as the "gold standard" for privacy protection across multiple domains [26]. Unlike syntactic approaches that rely on assumptions about adversary knowledge, differential privacy provides semantic guarantees that hold even against arbitrarily powerful attackers with unlimited external information [27].

However, applying differential privacy techniques to healthcare data reveals significant challenges that have limited its adoption despite theoretical advantages. One review pointed out several key limitations of the model and its proposed mechanisms, with particular concern about the abstract nature of the privacy parameter epsilon. and its implications for determining the level of privacy protection patients would receive [28].

The epsilon parameter selection problem represents differential privacy's most significant practical challenge in healthcare applications. As noted in recent surveys, choosing appropriate ε values requires balancing competing demands for privacy and analytical utility in ways that are often non-intuitive to healthcare practitioners [29]. Healthcare stakeholders typically lack the mathematical sophistication to interpret epsilon values in terms of concrete privacy risks, creating barriers to informed decision-making about acceptable privacy-utility trade-offs.

Moreover, healthcare applications often require complex query patterns that challenge classical differential privacy implementations. A recent systematic review on the use of differential privacy in healthcare noted that, although these algorithms can stop adversaries from recovering private details from the original medical data but practical applications face significant challenges related to data dimensionality, correlation structures, and interpretability requirements [29].

The composition problem further complicates differential privacy, where multiple analyses may be conducted on the same dataset over time. Each query consumes privacy budget, potentially exhausting available privacy protection and necessitating complex privacy accounting that may be impractical for many healthcare organizations [30].

2.1.5 Critical Synthesis: Beyond Individual Model Limitations

A fundamental issue across all models is the lack of standardized evaluation metrics that would enable meaningful comparison. Recent comprehensive surveys point out that there is still no agreed-upon standard for consistently evaluating both the privacy and the usefulness of synthetic data, which continues to limit its wider adoption. [31]. This evaluation gap becomes particularly problematic when healthcare organizations must select appropriate privacy techniques without clear guidance on relative merits.

The regulatory compliance dimension adds another layer of complexity. Privacy rules are interpreted differently across jurisdictions. For instance, in the United States, biomedical data is often shared in a "de-identified" form for research, following the specific criteria set out in the Safe Harbor method, while "in the European Union anonymization is more challenging to apply in practice, due to ambiguities in the legal definition" [32]. This regulatory fragmentation complicates the selection of appropriate privacy techniques for multi-jurisdictional healthcare collaborations.

2.2 Comparative Studies and Critical Analysis of Existing Research

Research comparing privacy-preserving techniques shows a fragmented field, often limited by narrow evaluation methods, weak methodologies, and little focus on real-world implementation challenges. Although many studies explore individual privacy models, comprehensive comparative analyses under controlled conditions are still quite rare—especially in the context of healthcare.

2.2.1 Foundational Comparative Work and Its Limitations

Early comparative efforts established important precedents but were constrained by limited computational resources and simplified datasets. An earlier work provided comprehensive theoretical frameworks for understanding privacy-preserving data publishing, categorizing approaches and analysing their theoretical properties [10].

Another comparative study, though ground-breaking in its systematic approach, relied heavily on synthetic datasets that lack the clinical complexity of authentic health records—they contain no multiple diagnoses, medication histories, laboratory results, or longitudinal visit patterns that characterize real medical data [33]. This limitation has continued to appear in much of the later research, raising important concerns about how well the findings can be applied to healthcare settings.

2.2.2 Contemporary Comparative Analyses: Progress and Persistent Gaps

More recent comparative efforts have attempted to address some limitations of earlier work while introducing new insights about privacy-utility trade-offs. However, critical analysis reveals constant methodological issues that limit the value of existing comparative research.

One study—though among the few to systematically compare all four major privacy models—also highlights the typical limitations seen in comparative privacy research [24]. The evaluation, conducted on the UCI Adult dataset, found that t-closeness provided strong protection against attribute disclosure but came with higher computation time and complexity,

whereas differential privacy offered the strongest theoretical privacy guarantee but at the cost of added noise which can reduce the accuracy of data analysis [24].

While these findings provide valuable insights, critical examination reveals several methodological limitations that compromise their broader applicability. First, the reliance on the Adult dataset—containing basic demographic and income information—fails to capture the complexity, correlation structures, and sensitivity patterns of healthcare data. Medical datasets often have high dimensionality, time-based dependencies, and complex hierarchical coding systems, all of which significantly change the balance between privacy and data utility [34].

Second, the evaluation metrics employed in many comparative studies fail to capture the multifaceted nature of data utility in healthcare contexts. Measures of information loss don't always capture how anonymization affects real-world uses such as clinical decision-making, epidemiological studies, or predictive modeling—key applications that drive healthcare data sharing [35].

2.2.3 Computational Complexity: The Overlooked Dimension

A critical gap in existing comparative research concerns computational performance and scalability analysis [37]. While theoretical complexity analyses are common, practical evaluation of execution times, memory usage, and scalability characteristics receives insufficient attention. This omission is particularly problematic for healthcare applications where datasets may contain millions of records with dozens of attributes [37].

Recent work on parallel anonymization algorithms has begun to address this gap. The P4 algorithm showed that running it in parallel with 12 threads strikes an effective balance—cutting computation time by 59.2% to 88.2% (a speedup of 2.45 to 8.46) while only slightly reducing data utility, with losses ranging from 0.54% to 14.4% [23]. However, such detailed performance analyses remain rare in the broader comparative literature.

2.2.4 Privacy Risk Assessment: Beyond Theoretical Guarantees

Existing comparative studies often rely heavily on theoretical privacy guarantees rather than practical risk assessment under realistic attack scenarios. This emphasis on theoretical analysis, while mathematically rigorous, may not reflect actual privacy risks in operational healthcare environments [7].

The attack model assumptions underlying different privacy techniques represent another critical gap in comparative evaluation. K-anonymity-based approaches typically assume limited adversary knowledge, while differential privacy provides guarantees against arbitrarily powerful attackers [39]. Comparative evaluations rarely acknowledge these fundamental differences in threat models, leading to potentially misleading conclusions about relative privacy protection [39].

2.2.5 Emerging Challenges: Cross-Border Data Sharing and Regulatory Compliance

Recent systematic reviews have identified emerging challenges that further complicate comparative evaluation of privacy techniques. A quantitative analysis of anonymized biomedical data usage found that "cross-border sharing was rare (10.5% of studies)" and that

"differences between countries with comparable regulations underscore the need for global standards" [32]. This finding has important implications for comparative privacy research. Privacy techniques that perform well under one regulatory framework may be inadequate or excessive under different jurisdictions.

2.3 Evaluation Metrics: A Critical Framework for Comprehensive Assessment

Evaluating privacy-preserving techniques requires frameworks that can capture the multidimensional trade-offs between privacy, utility, and performance. However, most current evaluation methods rely on narrow metrics that often fall short, as they fail to reflect the full complexity of real-world deployment—especially in healthcare, where the stakes are particularly high.

2.3.1 Data Utility: Beyond Simple Accuracy Measures

Traditional approaches to utility evaluation have relied heavily on simple metrics that may not reflect the true value of data for healthcare applications [14]. These approaches, while mathematically convenient, often fail to capture the complex ways that healthcare data supports clinical decision-making, research, and population health monitoring.

The concept of preservation of statistical relationships represents a more sophisticated approach to utility assessment. Healthcare data derive their value not only from individual data points but from complex patterns, correlations, and predictive relationships that enable clinical insights [40]. Some anonymization methods may seem effective when judged by simple metrics because they preserve overall statistics. However, if they disrupt important correlations within the data, they can end up being practically useless for many healthcare applications.

2.3.2 Computational Overhead: Scalability in Real-World Context

The computational requirements of anonymization techniques represent a crucial but often under examined dimension of practical deployment. Healthcare organizations typically operate under resource constraints and require timely data processing which may rule out the use of privacy protection methods that are too computationally demanding [13].

Execution Time Analysis must consider not only the absolute processing times but scaling characteristics as dataset size and complexity increase as well. Recent work has demonstrated that achieving minimal information loss via generalization/suppression in biomedical data came with significant computational cost [41]. This finding has important implications for routine clinical data sharing where processing delays may compromise care coordination or research timelines.

Memory utilization patterns represent another critical but under examined aspect of computational overhead. Healthcare datasets may contain millions of records with hundreds of attributes, requiring anonymization algorithms that can operate within the memory constraints of typical healthcare IT infrastructure. Algorithms that perform well on small datasets may become impractical when scaled to institutional data volumes [11].

The parallelizability of anonymization algorithms represents an increasingly important consideration as healthcare organizations seek to leverage distributed computing resources. Recent work on parallel anonymization has demonstrated significant speedup potential, but with trade-offs in terms of data utility that must be carefully evaluated [23].

2.3.3 Privacy Risk Quantification: Moving Beyond Theoretical Guarantees

Existing approaches to privacy evaluation often rely heavily on theoretical guarantees that may not reflect actual privacy risks in operational environments. This emphasis on theoretical analysis, while mathematically rigorous, may provide false confidence about privacy protection while overlooking practical vulnerabilities [42].

Assessing re-identification risk empirically calls for advanced methods that can mimic realistic attack scenarios, rather than depending only on theoretical limits. Recent work utilizing advanced risk assessment tools has revealed significant gaps between theoretical privacy guarantees and practical protection levels [43].

Modelling attack scenarios needs to take into account the entire range of possible privacy threats, rather than concentrating only on a few specific types. Healthcare data face diverse privacy risks including identity disclosure, attribute inference, and membership detection attacks that may require different protective strategies [44]. Comprehensive privacy evaluation must assess vulnerability across multiple attack scenarios to provide realistic risk profiles.

2.4 Research Contributions and Novel Methodological Approach

Building on a critical review of the existing literature, this research tackles the identified gaps with a comprehensive, multi-dimensional evaluation framework that moves beyond the limits of earlier comparative studies and offers more practical and realistic insights for healthcare privacy practitioners.

2.4.1 Addressing Methodological Limitations in Existing Research

The biggest strength of this work is that it tackles the persistent gap between theoretical privacy analysis and practical healthcare deployment requirements. Previous comparative studies have typically operated under unrealistic assumptions—simplified datasets, limited evaluation metrics, and artificial constraints that fail to capture real-world complexity [45].

This research employs a realistic healthcare dataset that exhibits the complexity, correlational structures, and sensitivity patterns characteristic of actual clinical data. The Diabetes 130-US Hospitals dataset provides an authentic testbed with multiple diagnoses, treatment histories, and administrative complexities that previous studies using generic benchmarks have failed to capture [46].

The comprehensive evaluation framework developed for this research moves beyond the narrow metrics employed in previous studies to capture the multifaceted nature of privacy-utility-performance trade-offs. Rather than relying solely on theoretical privacy guarantees or simple accuracy measures, the evaluation incorporates empirical risk assessment and query-specific utility analysis.

2.4.2 Contribution to Healthcare Privacy Practice

The ultimate contribution of this research lies in bridging the persistent gap between academic privacy research and practical healthcare implementation.

The evidence-based decision support framework developed in this study offers healthcare stakeholders clear guidance on choosing privacy techniques that match their specific needs, constraints, and risk tolerance. Instead of promoting one-size-fits-all solutions, the research highlights which approaches work best under different operational conditions.

The practical deployment insights generated through comprehensive evaluation under realistic conditions provide valuable guidance for healthcare organizations seeking to implement privacy-preserving data sharing. By identifying scalability bottlenecks, parameter sensitivities, and performance trade-offs, this research enables more informed decisions.

2.4.3 Advancing the Field Through Systematic Comparison

By offering a systematic comparison of the four leading privacy methods in real-world healthcare scenarios, this research makes an important contribution to privacy-preserving data publishing. While previous work has typically examined techniques in isolation or under highly constrained conditions, this research provides a unified evaluation framework that enables direct comparison of approaches under equivalent conditions [47].

The standardized evaluation protocol developed for this research provides a methodological framework that can be adapted for future comparative studies. By establishing rigorous evaluation procedures and comprehensive metrics, this work contributes to the development of standardized approaches to privacy technique evaluation that have been notably absent from the field.

In conclusion, this research addresses critical gaps in existing privacy evaluation literature while introducing methodological innovations that advance the field toward more practical, evidence-based approaches to healthcare privacy protection. The comprehensive evaluation framework provides unprecedented insights into privacy-utility-performance trade-offs while offering actionable guidance for healthcare privacy practitioners facing real-world deployment challenges.

3 THEORETICAL FOUNDATIONS

3.1 Models of Privacy and Anonymization

Privacy-preserving data anonymization relies on formal models that define how individual records are protected against re-identification and sensitive attribute disclosure. Four influential models in this context are k-anonymity, ℓ-diversity, t-closeness, and differential privacy. Each model is designed to handle different threat scenarios and assumptions about what an attacker might know, reflecting how privacy-preserving techniques have evolved—from simple suppression of identifiers to more advanced methods that provide probabilistic guarantees [3][47].

k-Anonymity:

The k-anonymity model requires that every record in a released dataset looks the same as at least k–1 other records when it comes to the quasi-identifier attributes [3]. Quasi-identifiers (QIs) are combinations of details like birthdate, gender, and ZIP code that may not identify someone on their own but can be linked with outside data to re-identify individuals [48]. Formally, a dataset is considered k-anonymous if each unique combination of quasi-identifier values shows up in at least k different records within the dataset [3]. This ensures each individual is hidden within a group of size k, known as an equivalence class [17].
This model prevents exact identity disclosure because an attacker cannot narrow down a record's identity to fewer than k possibilities. For instance, if k = 5, an attacker who knows a person's quasi-identifier values—like age, gender, and ZIP code—would find at least five records in the anonymized dataset with the same combination, making it unclear which record, if any, belongs to the target individual [3]. k-Anonymity protects against identity disclosure under the prosecutor model—where the attacker is assumed to know the target is in the dataset—by guaranteeing a maximum re-identification probability of 1/k for each record [49].

However, as shown by [19], k-anonymity alone does not protect against attribute disclosure attacks. If an equivalence class has little variation in its sensitive attributes, an adversary may still infer important information about an individual—even without being able to pinpoint their exact record [19]. If all individuals in an equivalence class share the same sensitive attribute value, an attacker could still learn that sensitive information despite the k-anonymity guarantee. This limitation, known as the homogeneity attack, motivated the development of stronger privacy models [19].

ℓ-Diversity:

To address weaknesses of k-anonymity in preventing attribute disclosure, ℓ-diversity was introduced by [19]. The ℓ-diversity principle states that each equivalence class must contain at least ℓ well-represented values for every sensitive attribute [19]. This means that even if an attacker isolates an equivalence class of size k, they will observe at least ℓ different values for the sensitive attribute, preventing certainty about that information.

There are several ways to define what counts as "well-represented." These include distinct ℓ-diversity, which requires at least ℓ different sensitive values; entropy ℓ-diversity, which ensures the entropy of sensitive values stays above a certain threshold; and recursive (c,ℓ)-diversity. [19]. By ensuring intra-group diversity of sensitive values, ℓ-diversity defends against both background knowledge attacks and homogeneity attacks that can compromise k-anonymity [50]. For instance, if all individuals in a k-anonymous group have the same medical diagnosis, an attacker knowing someone belongs to that group would immediately infer that diagnosis. ℓ-Diversity diminish this by requiring a mixture of diagnoses or other sensitive values in each group [19].

However, ℓ-diversity faces its own vulnerabilities. The model can still be compromised if the overall dataset's sensitive attribute distribution is heavily skewed [21]. When one sensitive value dominates the dataset, ensuring ℓ distinct values in every group may not prevent an attacker from making probabilistic inferences based on the global distribution. This limitation led to the development of t-closeness [21].

t-Closeness:

The t-closeness model, introduced by [21], enhances protection against attribute disclosure by looking at how closely the distribution of sensitive attributes in each equivalence class matches their distribution in the overall dataset. Formally, an equivalence class is said to satisfy t-closeness if the distance between these two distributions does not exceed a set threshold t [21]. The distance is typically measured using statistical metrics such as Earth Mover's Distance (EMD) or Kullback-Leibler divergence [51]. Earth Mover's Distance, also known as the Wasserstein metric, represents the smallest amount of effort needed to turn one probability distribution into another. [51]. By enforcing distributional similarity within tolerance t, t-closeness ensures that even if an attacker knows an individual belongs to a particular equivalence class, they learn minimal additional information beyond what is available from the overall population distribution [52].
For example, if 10% of all patients in the dataset have a certain disease, then in any anonymized group, the fraction of patients with that disease will be within t of 10%. This prevents both skewness attacks (where sensitive values in a group are heavily skewed compared to the population) and similarity attacks (where different sensitive values are semantically similar) [21]. t-Closeness provides stricter privacy guarantees than ℓ-diversity, though often at the cost of greater data generalization and reduced utility [41].

Differential Privacy:

Differential privacy, introduced by [25], represents a paradigmatic departure from the previous syntactic approaches by providing semantic privacy guarantees [25]. Formally, A randomized algorithm A is said to provide ε-differential privacy if, for any two datasets D_1 and D_2 that differ by only a single record and all $S \subseteq Range(A): Pr[A(D_1) \in S] \leq exp(\varepsilon) \times Pr[A(D_2) \in S]$" [25]. In practical terms, this guarantees that the presence or absence of any single individual in the dataset has negligible impact on the algorithm's output, up to a small parameter ε (epsilon)

[26]. Smaller ε values provide stronger privacy protection, with ε→0 meaning outputs remain virtually indistinguishable whether or not any particular individual is included [53].
Differential privacy is usually implemented by adding precisely calibrated random noise to query outputs or by applying randomized response methods [54]. Unlike syntactic models that rely on assumptions about adversary knowledge, differential privacy provides provable protection against arbitrarily knowledgeable attackers [27].

This model offers a unified framework that guards against membership, identity, and attribute disclosure by mathematically limiting how much information about individuals can leak [14]. However, achieving differential privacy often requires adding substantial noise, which can significantly impact data accuracy and utility [55]. This represents a fundamental trade-off: stronger privacy guarantees necessitate greater accuracy loss [56].

3.2 Data Utility and Relative Error

A key factor in any anonymization method is its effect on data utility—that is, how valuable the anonymized data remains for meaningful analysis. Data utility refers to the preservation of meaningful information and statistical properties in the anonymized dataset so that it can still support valid conclusions, analyses, and machine learning models [47]. As observed by [47], There is an unavoidable trade-off between privacy and utility: achieving stronger privacy usually means introducing more distortion into the data (generalization, suppression, noise addition), which can degrade the accuracy or fidelity of the data" [47].

A key goal in privacy-preserving data publishing is therefore to minimize information loss while achieving required privacy guarantees [33]. Various metrics have been proposed to quantify information loss or, conversely, remaining data utility. These include simple measures like the number of suppressed values, more complex entropy-based information loss metrics, and application-specific measures such as classification accuracy on anonymized data [14].

For our purposes, we focus on measuring utility in terms of how accurately the anonymized data can answer analytical queries compared to the original data. One common way to quantify this accuracy loss is through relative error on query results or statistical aggregates. Relative error measures the difference between a value calculated on the anonymized data and the true value from the original data, relative to the true value:

$$\textbf{Relative Error} = |v_{anon} - v_{orig}| / |v_{orig}| \times 100\%$$

where v_orig is the true value and v_anon is the value obtained from anonymized data [57]. This yields a percentage indicating deviation from the original result. For example, if a query count is 200 in original data and 180 in anonymized data, the relative error would be |(180-200)/200| × 100% = 10%.

Utility Classification Framework:

To interpret the magnitude of relative error in qualitative terms, we define three broad tiers of utility based on relative error percentages, following established practices in measurement accuracy and statistical reporting contexts:

- **Good Utility:** We classify relative error below 5 % as Good—indicating a small difference between anonymized and original answers. This threshold is informed by prior healthcare anonymization work that introduces and applies Matching Relative Error (MRE) and Average Relative Error (ARE) to quantify data utility. [58] and [59] define MRE and interpret very low values as reflecting minimal distortion between anonymized and original counts, while [60] use ARE/MRE to evaluate whether anonymization preserves the accuracy of routine descriptive or surveillance queries. Although none of these studies specify a fixed 5 % cut-off, their consistent treatment of small MRE/ARE values as desirable supports our classification of < 5 % as a "small difference" zone with high analytical utility.
- **Moderate Utility:** We define relative errors between 5 % and 15 % as indicating Moderate utility—noticeable deviation from the original data that remains usable for exploratory analyses, feasibility assessments, or monitoring tasks. This threshold is informed by empirical findings from [61], who evaluated the SHARE system's privacy-preserving release of healthcare data and reported relative errors ranging from near 0 % to approximately 30 % on longitudinal pattern queries [61]. Although they did not prescribe a specific "Moderate" range, their results show that analyses remained informative even with double-digit errors, supporting our placement of 5–15 % in a zone of reduced but still acceptable analytical value outside of high-stakes inference.
- **Poor Utility:** We classify relative error greater than 15 % as Poor—indicating a level of distortion that can materially affect the validity and interpretability of analytical results. [61] report empirical relative error values ranging from near 0 % to as high as 30 % in differentially private releases of health data. While they do not define a fixed acceptability threshold, their findings show that double-digit relative errors can substantially alter counts and rates, supporting our conservative demarcation of > 15 % as a point where analytical reliability is at risk. Although these thresholds are grounded in measurement science and official statistics practices. For instance, Statistics Canada considers estimates with CV (coefficient of variation) which is another statistical measure, above 16.5% as reliable for publication for some sources but with warning to the users [62].

3.3 Re-identification Risk Metrics

Alongside utility, re-identification risk represents the other critical dimension in evaluating anonymization methods. This describes the risk that an attacker could re-identify someone in an anonymized dataset by matching it with external information or by taking advantage of unique attribute patterns. [63]. Different metrics exist to measure re-identification risk, corresponding to different attacker models and assumptions [5].

Attacker Models:

As noted by [49], it is important to define who the attacker is and what they know [49]. Three common attacker models in privacy risk analysis include:

- **Prosecutor Model:** In this case, the attacker focuses on a particular individual and already knows that this person's record is included in the dataset. They also possess background information (quasi-identifiers) about that person [64]. The risk here is the probability that the attacker can single out the target's record.
- **Journalist Model:** Here, the attacker is focused on a particular individual but is unsure whether that person's information is actually part of the dataset [64]. If the person is not in the dataset, any match represents false identification. This model typically yields lower risk estimates than the prosecutor model because uncertainty about dataset membership dilutes the attack probability [65].
- **Marketer Model:** In this scenario, the attacker's goal is to re-identify as many records as possible, instead of focusing on specific individuals. [64]. This model emphasizes average risk across the dataset and is relevant for attackers seeking to build comprehensive profiles for commercial purposes [66].

These attacker scenarios influence risk computation. The prosecutor model typically examines worst-case risk (highest probability of re-identification for targeted records), whereas the marketer model focuses on expected proportion of records that could be re-identified on average [49].

Record Uniqueness and Population-Level Risk:

A fundamental driver of re-identification risk is the presence of unique or rare combinations of attributes [4]. If an individual's quasi-identifier combination is unique in the dataset (sample unique), and especially if that combination is also unique in the broader population (population unique), that record faces high re-identification risk [5].
Population uniqueness represents a critical concept in risk assessment. As explained by [43], population uniqueness is often used as a measure of re-identification risk because it estimates how many sample uniques likely correspond to population uniques [43]. Statistical models such as the Poisson log-linear model, Zayatz estimator, can predict population uniqueness from sample data [67].

Recent research has demonstrated that population uniqueness remains problematically high even in large datasets. [68] found that "99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes. This finding challenges assumptions that dataset size alone provides adequate privacy protection.

Risk Thresholds and Acceptable Levels:

Organizations often define thresholds for acceptable re-identification risk. The concept of "very small risk" appears in regulations like HIPAA, which requires expert determination that re-identification risk is "very small" [69]. However, as noted in recent analyses, "HIPAA does not define the level of risk of re-identification other than to say it should be 'very small'" [69].
Recent research on population-scale datasets suggests that maximum risk thresholds around 9-10% may be appropriate, with 95th percentile risk measures providing more robust assessment than simple maximum values [70].

3.4 Overview of ARX Tool Capabilities

To conduct our anonymization and risk analysis, we make use of ARX, an all-in-one open-source tool designed for data anonymization. ARX is well-suited for our study because it supports a wide range of privacy models, transformation techniques, and provides built-in facilities to analyse both data utility and re-identification risk [34].

Supported Privacy Models:
While differential privacy typically is not achieved through generalization, ARX includes functionality for evaluating differential privacy guarantees under specific assumptions [71].
The tool's flexibility allows users to specify any combination of privacy models simultaneously. For example, one can require data to satisfy both k-anonymity with k=5 and 2-diversity concurrently [34]. ARX will search for transformations meeting all specified criteria, providing significant versatility for complex privacy requirements.

Utility Analysis Capabilities:
ARX includes comprehensive utility analysis features enabling comparison between original and anonymized data. The tool computes various data quality metrics including discernibility metrics, average equivalence class sizes, non-uniform entropy, and domain-specific utility measures such as classification accuracy on anonymized data [34][40].
ARX's graphical interface provides immediate feedback on information loss associated with any transformation, displaying information loss scores according to chosen metrics as users navigate the solution space [34].

Risk Analysis Module:
Perhaps ARX's most distinguished feature is its comprehensive risk analysis module, implementing the metrics described in Section 3.3. ARX calculates re-identification risk distributions under prosecutor, journalist, and marketer models, reporting metrics including highest risk, lowest risk, average risk, and fraction of records exceeding specified risk thresholds [71][72].
The tool implements population uniqueness estimation methods when supplied with population statistics or external datasets [71]. ARX also includes automatic identification of potential quasi-identifiers by analysing which attributes make records unique or rare, and built-in checks for HIPAA Safe Harbor identifiers [71]. For healthcare applications, this automated analysis helps ensure comprehensive identification of privacy-sensitive attributes.

Computational Performance and Scalability
ARX addresses the computational challenges posed by privacy-preserving data transformation through highly optimized algorithms. Performance evaluations demonstrate that ARX can "evaluate 12,960 transformations in about 150 seconds" on datasets containing 1.2 million records with eight quasi-identifiers each [34].
The framework implements space-time trade-offs allowing memory consumption reduction when required, and supports parallel processing capabilities for improved performance on large datasets [34]. Recent versions include enhanced algorithms that achieve speedups of "up to a factor of 955 compared to baseline execution time" for monotonic privacy criteria [34].

Workflow and Integration:
ARX follows a systematic workflow: (1) Configuration—import data, set attribute roles, choose privacy models and parameters; (2) Execution—search for anonymized dataset satisfying criteria; (3) Exploration—examine alternative solutions with different utility-privacy trade-offs; (4) Analysis—evaluate chosen solution using utility and risk analysis perspectives [34].
The tool provides both comprehensive graphical interface for interactive exploration and Java API for programmatic access [34]. This dual interface approach enables both exploratory analysis through the GUI and automated processing for production deployments.

By leveraging ARX's capabilities, our experimental comparison ensures that differences in outcomes are attributable to the privacy techniques themselves rather than implementation inconsistencies. ARX's adherence to well-documented models and metrics provides credibility to our evaluation methodology, while its comprehensive feature set enables fair comparison across all four privacy-preserving approaches under standardized conditions.

4 METHODOLOGY

4.1 Experimental Environment and Tools

This study's experimental implementation was conducted using the Python programming language, leveraging its comprehensive ecosystem of data manipulation, statistical analysis, and visualization libraries. Python was chosen for its wide adoption in data science research, strong community support, reproducibility, and compatibility with open-source tools essential for privacy-preserving transformations.

4.1.1 Programming Language

All pre-processing steps, anonymization technique implementations, query evaluations, and performance tracking scripts were developed in Python 3. Python's flexibility in handling heterogeneous data, combined with its capabilities for algorithm development, made it well-suited for this research—from raw data handling to advanced experimental analysis.

4.1.2 Libraries and Packages Used

The following Python libraries were employed extensively throughout the implementation phase:

- **pandas** – For structured data manipulation, filtering, aggregation, and CSV output generation.
- **numpy** – For numerical computations and efficient array.
- **matplotlib** – For visualizations, including execution time charts and data utility plots.
- **typing** – For type hinting and function annotations, improving code readability and maintainability.
- **logging** – To systematically capture execution logs, including time measurements and process summaries.
- **time** – For precise performance measurement during anonymization executions.
- **IPython.display** – For enhanced visualization and result presentation within the development environment.

Together, these libraries enabled efficient pre-processing of large datasets, execution of anonymization algorithms, systematic collection of performance metrics, and clear presentation of experimental results.

4.1.3 Hardware and System Configuration

All implementations and experiments were executed on a consumer-grade personal computer with the following specifications:

- **Device Name:** DESKTOP-SSU1OPF

- **Processor:** Intel® Core™ i5-5200U CPU @ 2.20 GHz (dual-core)
- **RAM:** 8 GB
- **Storage:** 932 GB (ST1000LM024 HN-M101MBB) and 238 GB SSD (256 GB EVM SSD)
- **Graphics:** NVIDIA GeForce 940M (2 GB dedicated VRAM) and Intel® HD Graphics 5500 (128 MB shared memory)
- **Operating System:** Windows 10 (64-bit, x64-based processor)

This configuration was sufficient for all dataset sizes used in the experiments (25k to 100k+ rows), eliminating the need for distributed or cloud-based computation. The execution times reported in later sections reflect performance on this hardware and are therefore indicative of what is achievable on standard consumer machines without specialized accelerators.

4.1.4 ARX Tool Setup

For re-identification risk evaluation, the **ARX Data Anonymization Tool**—an established open-source platform for privacy-preserving data publishing—was employed. ARX provides a Java-based environment supporting **k-anonymity**, **ℓ-diversity**, **t-closeness**, and **differential privacy**, along with a robust set of risk analysis utilities, including the *prosecutor* and *journalist* attacker models.

In this study, anonymized datasets generated using Python scripts were imported into ARX for risk assessment. The tool's graphical interface facilitated interactive evaluation, enabling the computation of re-identification risks and the capture of visual evidence (screenshots) for qualitative comparison. ARX's advanced modelling capabilities significantly reduced the complexity that would otherwise arise if these analyses were implemented directly in Python.

4.2 Dataset Preparation

The dataset used in this study is the *Diabetes 130-US Hospitals for Years 1999–2008* dataset, obtained from the UCI Machine Learning Repository [73]. This dataset was selected because it provides a large-scale, real-world clinical resource comprising over 100,000 inpatient encounters, with diverse demographic, diagnostic, and outcome variables [46]. Its richness and complexity make it highly representative of modern electronic health records (EHRs) while also posing non-trivial privacy challenges due to the presence of multiple quasi-identifiers.

All pre-processing and preparation steps are detailed comprehensively in **Section 5** of this thesis. In summary, the dataset underwent extensive cleaning and transformation, including:

- Removal of direct identifiers and attributes with excessive missing values or limited analytical relevance (e.g., weight, payer_code, medical_specialty).
- Standardization of categorical values, such as consolidating ICD-9 codes into clinically meaningful groups, mapping missing race values to "Unknown," and retaining rare gender categories to preserve distributional characteristics.

- Designation of quasi-identifiers (QIs)—including age, gender, and race—and specification of the primary diagnosis group as the sensitive attribute.
- Stratified, nested sampling to create four dataset sizes (25k, 50k, 75k, and 100k rows), maintaining consistent class distribution for the readmission variable.

This structured preparation ensured data integrity, facilitated fair comparison across anonymization techniques, and allowed scalability testing under controlled conditions.

4.3 k-Anonymity Implementation

The objective of this stage is to implement k-anonymity on the Diabetes 130-US Hospital dataset to ensure that each quasi-identifier (QI) combination appears in at least *k* records. This guarantees that no individual can be uniquely identified based on QIs such as age, race, and gender. The implementation follows the formal definition, which defines k-anonymity as a property of data releases that prevents identity disclosure by grouping records into equivalence classes of size $\geq k$ [3].

The anonymization process was operationalized through the following sequential steps:

1. **Quasi-Identifier (QI) Selection:**

    - Age, race, and gender were designated as QIs based on their high re-identification potential, consistent with Sweeney's foundational work.
    - To preserve maximum analytical value, race and gender were retained at their original granularity, while age was generalized hierarchically.

2. **Generalization Hierarchy for Age:**

    - A fixed, auditable hierarchy was applied:

        - Level 0: original 10-year age bands (most specific).
        - Level 1: broader bands [0–30), [30–60), [60–100).
        - Level 2: [0–50), [50–100).
        - Level 3: [0–100).

    - Level 1 generalization was identified as sufficient to achieve k-anonymity (for k = 2, 5, 10) while minimizing information loss.

3. **Equivalence Class Construction and Compliance Verification:**

    - Records were grouped by QIs, and equivalence classes were checked to ensure each contained $\geq k$ records.

4. **Suppression of Outliers:**

- Records falling into classes with size < *k* were suppressed to enforce compliance. Suppression levels were kept as low as possible to minimize data distortion.

5. **Post-Processing:**

   - Datasets were re-verified for compliance, ensuring no equivalence class violated k-anonymity.
   - Random shuffling of rows was performed prior to export to mitigate positional inference risks.

6. **Parameterization and Batch Execution:**

   - The process was executed for k = {2, 5, 10} across four dataset sizes (25k, 50k, 75k, and 100k records).

7. **Metrics Logging:**

   - Logged metrics included execution time, suppression count, retained records, equivalence-class sizes, and age diversity reduction and was saved in a CSV named *k_anonymity_performance_summary.csv*.

k-Anonymity was selected as the starting point of this comparative study because it represents the foundational model of privacy-preserving data publishing [3]. Its intuitive grouping mechanism directly addresses the re-identification risk posed by quasi-identifiers while allowing relatively straightforward implementation and evaluation. For healthcare data specifically, k-anonymity provides a transparent way to balance privacy and utility: the hierarchical age generalization minimizes analytical distortion, while suppression ensures compliance without excessive loss of records. Although more advanced models exist, using k-anonymity as a baseline establishes a clear reference point for comparing trade-offs in utility, runtime, and privacy risk across all methods evaluated in this thesis.

4.4 ℓ-Diversity Implementation

The objective of this stage is to implement next method which is l-diversity. The ℓ-diversity model strengthens k-anonymity by ensuring that within every equivalence class defined by quasi-identifiers (QIs), there are at least ℓ well-represented values for the sensitive attribute(s). This prevents attribute disclosure even when k-anonymity is satisfied but sensitive values are highly homogeneous. In this work, we implement ℓ-diversity using the distinct-values criterion, which requires that each equivalence class contain at least ℓ distinct values of the sensitive attributes [19].

The anonymization process was operationalized through the following sequential steps:

1. **Define QIs and Sensitive Attributes:**

    - Age (generalized), race, and gender are designated as quasi-identifiers. Primary and secondary diagnoses (diagnosis_1 and diagnosis_2) are treated as sensitive attributes, aligning with the model's focus on protecting medically meaningful attributes.

2. **Generalize Age:**

    - Age is coarsened into three bins: [0–30], [30–60], and [60–100]. This structured generalization follows the generalize-then-test workflow recommended by [19], reducing QI granularity while retaining analytical relevance.

3. **Form Equivalence Classes:**

    - Records are grouped by the QI set {Age*, Race, Gender} to produce q*-blocks. These equivalence classes are the basis for diversity evaluation.

4. **Check ℓ-Diversity:**

    - For each equivalence class, the number of distinct values in diagnosis_1 and diagnosis_2 is computed. A class is retained only if both attributes meet the required diversity threshold ($\ell \in \{2, 3, 4\}$). This instantiates the "well-represented" condition in the simplest and most tractable form: distinct-value counts.

5. **Enforce with Class-Level Suppression:**

    - If an equivalence class fails the diversity check, all its records are suppressed. This class-level suppression enforces ℓ-diversity decisively without resorting to extensive cell suppression or over-generalization. Rows are then shuffled prior to release to mitigate positional inference.

6. **Parameterization and Batch Runs:**

    - The process is repeated for ℓ values {2, 3, 4} across datasets of size 25k, 50k, 75k, and 100k. Anonymized outputs are produced for each configuration.

7. **Metrics Logging:**

    - For each run, we record execution time, number of records dropped, percentage retention, and per-attribute diversity statistics and was saved in a CSV named *l_diversity_dropping_performance_summary.csv*. These provide a quantitative basis for evaluating privacy–utility trade-offs.

The distinct-values criterion was selected because it is computationally straightforward and better suited to healthcare datasets, which often contain skewed or imbalanced distributions of diagnoses. Entropy- or recursive-based definitions, though stronger, can excessively restrict data utility in such contexts [19]. Similarly, class-level suppression was preferred over pure generalization or cell suppression because our healthcare data exhibited strong skew especially in age, making it infeasible to achieve ℓ-diversity with generalization alone or in some cases even over generalizing. Suppressing entire violating classes strikes a balance between enforceability and preservation of analytical value [19]. Finally, although the original definition suggests handling multiple sensitive attributes by incorporating them into the QI, we pragmatically enforced per-attribute ℓ-diversity within the same QI classes. This relaxation reduces information loss while maintaining compliance with the model's intent, and its implications are addressed in our evaluation.

4.5 t-Closeness Implementation

The implementation of t-closeness in this study follows the definition: an equivalence class is said to have *t-closeness* if the distance between the distribution of a sensitive attribute within the class and the global distribution of that attribute in the dataset does not exceed a predefined threshold *t* [21]. The chosen distance metric is the Earth Mover's Distance (EMD), which quantifies the "effort" required to transform one probability distribution into another. By enforcing this constraint, the model prevents attribute disclosure that arises when certain sensitive values are overly concentrated within small equivalence classes, a weakness not fully addressed by k-anonymity or ℓ-diversity [21].

The enforcement of t-closeness was carried out through a structured sequence of steps:

1. **Quasi-Identifiers (QIs) and Sensitive Attribute Selection:**

    - Equivalence classes were formed using the quasi-identifiers {Age*, Race}, where Age* denotes a potentially generalized version of age. The sensitive attribute considered was *diagnoses_1*. This selection reflects the need to balance comparability across runs while avoiding excessive fragmentation of the dataset.

2. **Computation of the Global Distribution:**

    - The overall distribution of *diagnoses_1* in the full dataset was computed and served as the reference distribution. Each equivalence class was then compared to this global distribution using EMD.

3. **Minimal Generalization of Age:**

- Age was generalized into three semantic bins: 0–30, 30–60, and >60. This limited generalization was sufficient to reduce fragmentation without causing unnecessary information loss.

4. **Calculation of EMD for Each Class:**

    - For each equivalence class, the EMD was calculated between the class distribution and the global distribution. Classes whose EMD exceeded the threshold *t* were marked as violating.

5. **Suppression of Non-Compliant Classes:**

    - Rather than relying solely on coarser generalization, violating classes were suppressed (row-dropping). This approach ensured that only the problematic classes were removed, thereby preserving the overall utility of the dataset while still enforcing the t-closeness constraint.

6. **Parameterization and Batch Runs:**

    - The implementation was evaluated across datasets of size 25k, 50k, 75k, and 100k, with *t* values set at {0.4, 0.7, 1.0}. Each (dataset, *t*) configuration produced a corresponding anonymized dataset and associated logs.

7. **Metric Logging:**

    - For each run, execution time, number of rows suppressed, final record count, maximum EMD after enforcement, and the level of generalization applied were recorded and was saved in a CSV named *t_closeness_performance_summary.csv*. These metrics provided the basis for evaluating the privacy–utility–performance trade-offs.

Several design choices were made deliberately in order to remain faithful to the t-closeness framework while ensuring practicality on healthcare data. The original definition normalizes EMD for ordered domains into the [0,1] range using a 1/(m−1) factor [21]. In this implementation, normalization was omitted, and *t* thresholds (0.4, 0.7, 1.0) were calibrated to the unnormalized scale. This ensured consistency within our experimental framework without altering the fundamental enforcement rule (distance ≤ *t*). While the original model suggests equal-distance or taxonomy-based ground distances for categorical attributes, this study used a tractable one-dimensional CDF-based EMD as a surrogate. This ensured reproducibility and computational efficiency while avoiding subjective modelling of complex disease hierarchies. The trade-off was documented as a limitation, with the possibility of future extension to taxonomy-aware ground distances. Restricting QIs to {Age*, Race} was necessary to prevent over-fragmentation, which would otherwise lead to excessive suppression. Pure generalization strategies were found insufficient under skewed clinical distributions, as many classes

continued to exhibit high EMD values. By contrast, class-level suppression removed only non-compliant groups, thereby aligning with fact that t-closeness can be integrated with suppression in practice [21].

Overall, this implementation preserves the theoretical rigor of the t-closeness model while adapting it to the realities of healthcare data, where skewness and the need for analytical utility necessitate practical trade-offs.

4.6 Differential Privacy Implementation

In this thesis *ε-differential privacy* has been applied in the **interactive query model**, rather than by publishing a sanitized dataset. In this setting, each analytic query is answered directly on the original data, and calibrated noise is injected at response time to satisfy ε-differential privacy [25]. This guarantees that the presence or absence of any single individual has a negligible impact on outputs, even under arbitrary auxiliary information and repeated querying. The method is particularly well-suited for workloads where releasing a static anonymized microdata table is either infeasible or too utility-degrading [25].

The implementation follows the canonical Laplace mechanism with workload-aware ε budgeting:

1. **Privacy Model and Adjacency:**
    - Neighboring datasets are defined under *replacement adjacency*, meaning one individual may move between groups.
    - Under this definition, histogram or crosstab queries have an ℓ_1 sensitivity of at most 2 (one cell increases by 1, another decreases by 1).

2. **Mechanism and Calibration:**
    - Each query vector is perturbed by adding Laplace noise scaled to $\Delta f/\varepsilon q$ where Δf is the query's global sensitivity and εq is its allocated privacy budget.
    - Sensitivity is computed once for the whole vector rather than per cell, following [25] histogram treatment.

3. **Budgeting Strategy:**
    - Global ε values of {2.0, 5.0, 8.0} are tested (strict → relaxed privacy).
    - Each global ε is divided across the query workload using weighted allocation: high-priority or noise-sensitive queries (e.g., means, readmission rates) are assigned larger shares of ε, while low-priority or robust queries (e.g., counts) receive smaller shares.

4. **Pre-Committed Bounds and Minimum Group Sizes:**

- To ensure data-independent sensitivity, numeric attributes are clipped to predefined ranges, and categorical groups are merged to enforce a minimum group size n_min
- Sensitivities are then instantiated as:
    - Counts/crosstabs: $\Delta f = 2$
    - Proportions: $\Delta f = 2 / n\_min$
    - Means: $\Delta f = 2R / n\_min$, where R is the predefined numeric range.

5. **Application to Queries:**

    - The mechanism is applied to both *simple queries* and *complex queries*.
    - Age is post-processed into three bins (0–30, 30–60, >60) to align with other anonymization techniques; this post-processing does not consume ε.
    - Query semantics and definitions are described in detail in the subsequent *Query-Based Utility Evaluation* section.

6. **Execution Across Datasets and ε Levels:**

    - The pipeline is executed on all four dataset sizes (25k, 50k, 75k, 100k rows) under $\varepsilon \in \{2.0, 5.0, 8.0\}$.
    - For each configuration, runtime, noise calibration, and relative error are logged, with outputs stored as CSVs named *dp_timing_simple_summary.csv* and *dp_timing_complex_summary.csv* rather than anonymized datasets.

The design choices in this implementation were guided both by the theoretical foundations of differential privacy as defined by [25] and by the practical requirements of healthcare analytics. The interactive query model was adopted rather than static dataset release, since the original framework explicitly introduces differential privacy in this setting and guarantees ε-privacy for each query, even under composition. The Laplace mechanism was selected as it is the canonical instantiation proposed in [25] for real-valued query outputs, with calibration determined by the global ℓ_1 sensitivity of the query function. In line with the definition of neighboring datasets differing in at most one individual, we enforced bounded sensitivity by applying clipping to numeric attributes and merging sparse categorical groups — a practical instantiation of the principle that sensitivity must be fixed independently of the data.

Beyond the original formulation, two pragmatic extensions were introduced to tailor the model to the dataset and workload. First, we adopted a replacement-style interpretation of adjacency, which operationalizes the requirement that only one individual's data can change between neighbouring datasets. Second, since ε accumulates across queries through composition, we employed weighted ε allocation to prioritize high-value or noise-sensitive queries while assigning smaller budgets to more robust queries. These extensions are not explicitly prescribed by the [26] but remain consistent with its theoretical guarantees, allowing us to balance privacy and utility in a workload-aware manner.

Overall, this implementation balances strict theoretical privacy guarantees with practical utility considerations, producing privatized analytic outputs aligned with both the DP literature and the empirical needs of healthcare data analysis.

4.7 Query-Based Utility Evaluation

The purpose of this stage is to evaluate the utility of data after applying each privacy-preserving technique—**k-anonymity**, **ℓ-diversity**, **t-closeness**, and **differential privacy**. Utility is assessed by executing a consistent set of analytical queries on both original and anonymized outputs (or noisy answers in the DP case) and then measuring relative error against the ground truth. This provides a quantitative and comparable basis for understanding how each method balances privacy with analytic value.

The evaluation process follows a structured series of steps:

1. **Dataset Slices and Parameter Grids:**

    - Dataset slices: 25k, 50k, 75k, and full dataset.
    - Method parameters: $k \in \{2, 5, 10\}$; $\ell \in \{2, 3, 4\}$; $t \in \{0.4, 0.7, 1.0\}$; $\varepsilon \in \{2.0, 5.0, 8.0\}$.

2. **Simple Queries:**

    - Designed to measure foundational characteristics of the dataset.
    - Five queries applied across all methods:

        - Q1: Age-3bin counts (0–30, 30–60, >60).
        - Q2: Race counts.
        - Q3: Gender × Admission Type counts.
        - Q4: Average medications by Age-3bin.
        - Q5: Readmission (<30 days) percentage by race.

    - Implemented identically across methods, differing only by input dataset.

3. **Complex Queries:**

    - Simulate multi-attribute analytics, reflecting higher-order real-world tasks.
    - Five queries applied across all methods:

        - Q1: Avg. lab procedures by Age-3bin × Admission Type.
        - Q2: Readmission% (<30 days) by Diagnosis1 × Gender.
        - Q3: Avg. medications by Diagnosis2 × Gender.
        - Q4: Avg. medications by Race × Admission Type.
        - Q5: Avg. lab procedures by Diagnosis2 × Admission Type.

4. **Execution Workflow:**

    - For **k-anonymity, ℓ-diversity, and t-closeness**:

        - Queries run on original datasets → true aggregates.
        - Queries run on anonymized datasets → anonymized aggregates.
        - Relative errors computed between the two.

    - For **differential privacy**:

        - No anonymized dataset is released.
        - Queries executed directly on original data.
        - Laplacian noise added to answers according to ε allocation.
        - Same relative error calculation applied to noisy vs. true answers.

5. **Relative Error Computation:**

    - Formula:

    Relative Error = |v_anon - v_orig| / |v_orig| × 100%

    - Where v_anon is anonymized/noisy output, v_orig is the true value.
    - Errors classified into tiers: Good (<5%), Moderate (5–15%), Poor (>15%).

6. **Outputs and Aggregation:**

    - CSV summaries generated for each method and query set:

        - k-anonymity: *k_anonymity_simple_query_utility_summary.csv*, *k_anonymity_complex_query_utility_summary.csv*.
        - ℓ-diversity: *l_diversity_simple_query_utility_summary.csv*, *l_diversity_complex_query_utility_summary.csv*.
        - t-closeness: *t_closeness_simple_query_utility_summary.csv*, *t_closeness_complex_query_utility_summary.csv*.
        - DP: *dp_utility_simple_summary.csv*, *dp_utility_complex_summary.csv*.

    - Each file records relative error, tier classification, and aggregation across parameter settings.

We decided to go with the following approach due to the following reasons:

- **Comparability across methods**: Using the same query set and dataset slicing ensures fair "apples-to-apples" comparisons across anonymization frameworks.
- **Respecting model interfaces**: Table-release models (k, ℓ, t) are evaluated on anonymized datasets, while DP is evaluated on noisy query answers, consistent with each technique's intended design.

- **Balanced query design**: Simple queries capture essential demographic/statistical aggregates, while complex queries mimic multi-dimensional analyses common in healthcare applications.
- **Error classification for interpretability**: The tier system (Good, Moderate, Poor) translates numeric errors into qualitative judgments, supporting clearer comparative insights.

4.8 Implementation of Re-Identification Risk Evaluation

The goal of this stage is to quantitatively assess the likelihood of re-identification in both original and anonymized datasets, thereby evaluating the degree of privacy protection each method provides. Re-identification risk evaluation was carried out using the **ARX Data Anonymization Tool**, which is a widely recognized platform supporting standardized attacker models and risk metrics [34]. This evaluation was applied to datasets anonymized under **k-anonymity**, **ℓ-diversity**, and **t-closeness**. Datasets generated under **differential privacy** were excluded from ARX-based analysis, since DP does not release modified microdata but instead provides formal, mechanism-level guarantees.

A consistent procedure was followed across all release-based anonymization techniques:

1. **Dataset Upload:**
    - Each dataset (original and anonymized with varying parameter values) was imported individually into ARX.
    - DP outputs were excluded as they do not yield anonymized microdata.
2. **Quasi-Identifier and Sensitive Attribute Selection:**
    - Quasi-identifiers (QIs) and sensitive attributes were defined consistently with earlier methodological sections to ensure risk estimates aligned with realistic attacker knowledge.
3. **Regional Configuration:**
    - The region was set to the **United States**, aligning ARX's population models with the dataset's context for realistic risk estimates.
4. **Risk Analysis Execution:**
    - ARX's dedicated Risk Analysis module was activated.
    - Built-in attacker models quantified re-identification risks automatically.
5. **Result Capture:**

- For each dataset and parameter setting, ARX's outputs (risk metrics) were recorded and documented for comparative analysis across techniques.

Risk Models Applied:

Three canonical attacker models built into ARX were used:

- **Prosecutor Model** – Attacker knows the target is in the dataset.
- **Journalist Model** – Attacker is uncertain whether the target is included.
- **Marketer Model** – Evaluates likelihood of successful record re-identification.

Metrics Reported:

- **Records at Risk (%)** – Proportion of records with risk above the threshold.
- **Highest Risk (%)** – Highest risk of a single record.
- **Success Rate (%)** – Proportion of records that can be re-identified on average.

Risk Thresholds Applied:

- **Highest Risk:** 20%
- **Records at Risk:** 5%
- **Success Rate:** 5%
  Values below these thresholds indicate acceptable protection.

It should be noted that, unlike the anonymization techniques discussed earlier, ε-differential privacy does not generate a static anonymized dataset suitable for release. Instead, it operates in the interactive query model, where calibrated noise is added to query responses at runtime [25][26].Consequently, the outputs of differential privacy are privatized query results rather than microdata tables, which precludes evaluation through ARX's re-identification risk analysis that assumes record-level datasets.

5 DATASET AND PRE-PROCESSING

5.1 Dataset Description and Source

This research utilizes the Diabetes 130-US Hospitals (1999–2008) dataset, a widely cited real-world clinical dataset available from the UCI Machine Learning Repository and originally prepared for a large-scale study on diabetes readmissions over a ten-year span across 130 U.S. hospitals[73]. The dataset comprises 101,766 distinct inpatient encounter records for diabetic patients, each containing approximately 50 attributes spanning demographics (e.g., age, race, gender), admission and hospital details, laboratory results, primary and secondary ICD-9 diagnoses, medication data (including specific diabetes drugs with dosage change flags), treatment and outcome indicators, and record-specific variables such as length of stay[46][73]. The primary research objective for which these data were collected was to predict early hospital readmission (within 30 days), which serves as a proxy measure of care quality [46].

De-identification and Privacy Considerations:

The dataset was partially de-identified prior to public release: direct identifiers (such as names and precise dates) were omitted, and certain quasi-identifiers were generalized[46]. In line with HIPAA Safe Harbor guidelines, patient age is reported in 10-year bands (e.g., "[50–60)") and all ages over 90 are grouped, a standard approach to reduce re-identification risk in health datasets[74]. Unique surrogate codes replace direct IDs, and the dataset omits explicit personal identifiers. Nevertheless, combinations of remaining quasi-identifiers—such as age group, race, and gender—retain latent risk for re-identification when linked with external databases[49]. This realistic mix of privacy challenge and analytic richness makes the dataset a valuable testbed for comparative anonymization studies.

5.2 Data Pre-processing Workflow

Before applying any privacy-preserving transformations, the data underwent a standardized and transparent pre-processing pipeline, structured as follows:

1. Removal of Direct Identifiers:
The "encounter_id" and "patient_nbr" columns, which serve solely as visit and patient tracking codes, were immediately dropped to guarantee the absence of direct identifiers in the analysis-ready data.

2. Handling Missing Data:
We systematically evaluated all features for missing, placeholder, or ambiguous values:
Weight: The "weight" attribute was missing in 97% of cases and was dropped.

Payer code: "payer_code" had ~52% missingness and was also removed to prevent distortion.

Medical specialty: "medical_specialty" was missing for ~53% of encounters and exhibited high cardinality; it was dropped to increase data robustness.

Race and Gender: The "race" column's missing values (encoded as "?") were mapped to a new "Unknown" category to preserve those records without spurious imputation. The rare "Unknown/Invalid" gender values (~3 rows) were retained as a third category ("Unknown").

This policy ensures that only features with analytic and privacy relevance remain, and that no sample bias is introduced by aggressive record removal.

3. Pruning Redundant or Constant Columns:
Medication columns taken from the raw data and found to be all- or nearly-all "No" (such as "examide", "citoglipton", or rare combinations) were dropped. The third diagnosis code ("diag_3") was also excluded due to frequent missingness or limited clinical relevance.

4. Consolidating and Encoding Categorical Variables:
Diagnosis Codes: ICD-9 diagnosis codes ("diag_1" and "diag_2") were mapped to nine broad clinical categories (e.g., Circulatory, Respiratory, Diabetes, Neoplasms, etc.) using established rules [46]. This consolidation allows meaningful grouping for both anonymization and analysis and reduces dimensionality.

Medications: Only metformin, insulin, two combination drugs (metformin-rosiglitazone and metformin-pioglitazone), and the diabetesMed flag were retained, based on prevalence and analytic relevance; less frequently prescribed drugs were removed.

Admission Type: Admission type codes were mapped to human-readable values (e.g., "Emergency", "Elective"), dropping codes with no analytic value or high overlap.

Other Features: All mapped/encoded categorical values maintained a clear and finite set of modalities, suitable for downstream ML and privacy tools.

5. Feature summary and Final Set:
At the conclusion of these steps, the prepared data retained the following features: race, gender, age group, admission type, diagnoses_1, diagnoses_2, counts of lab procedures and medications, four medication columns, and the outcome "readmitted". All missing or ambiguous categorical values are encoded as "Unknown" or "Missing" categories, and no rows contain placeholder characters.

5.3 Defining Quasi-Identifiers and Sensitive Attributes

Drawing on privacy best practices, we explicitly designate age, gender, and race as quasi-identifiers (QIs)—attributes plausibly available to an adversary and routinely implicated in re-identification risks. The primary and secondary diagnosis ("diagnoses_1" and "diagnoses_2") is the sensitive attribute; it directly reflects the patient's main clinical condition, aligning with regulatory interpretations of "protected health information".

5.4 Stratified and Nested Sampling for Experimental Scaling

To evaluate algorithmic scalability and privacy–utility trade-offs, we constructed three more nested, stratified subsamples of the healthcare dataset: 25k (24,951 records), 50k (49,942 records), 75k (74,942 records) from the full dataset (101,742 records) which we received after pre-processing. Stratification was applied over the quasi-identifiers age, race, and gender to preserve the original population distributions and mitigate sampling bias. The 4 datasets after pre-processing and stratification were named as follows:

- 25k dataset: *diabetic_data_25k.csv*
- 50k dataset: *diabetic_data_50k.csv*
- 75k dataset: *diabetic_data_75k.csv*
- 100k dataset: *diabetic_data_final.csv*

The nested design ensures that each larger subset fully contains all records from the smaller ones, thereby enabling stepwise scalability analysis under consistent demographic composition. This structure isolates computational performance differences to sample size alone, without introducing confounding from distributional variation.

Across all subsets, the same feature set and pre-processing pipeline were maintained. This uniform design guarantees fair comparison of privacy-preserving methods: observed differences in privacy loss and computational efficiency can be attributed to the techniques themselves rather than inconsistencies in data composition.

6 RESULTS AND DISCUSSION

6.1 Execution Performance

Evaluating the execution performance of privacy-preserving techniques is crucial for understanding their computational feasibility, especially when handling healthcare datasets of varying sizes. In this study, execution performance was measured in terms of **total processing time** required to apply each anonymization method (k-anonymity, ℓ-diversity, t-closeness, and differential privacy) on datasets containing **25k, 50k, 75k, and 100k records**. For each method, three parameter configurations were considered, representing **small**, **moderate**, and **large** privacy guarantees.

Execution time provides insight into the scalability of these methods, highlighting trade-offs between privacy strength and processing efficiency. It should be noted that the visualisations in this section are created using the CSVs we created in metrics logging step in methodology section. Also the information about the number of records/rows dropped after implementation of each section can be found in Appendices.

6.1.1 K-Anonymity

For k-anonymity, execution time was measured for **k = 2 (small)**, **k = 5 (moderate)**, and **k = 10 (large)** across the four dataset sizes.

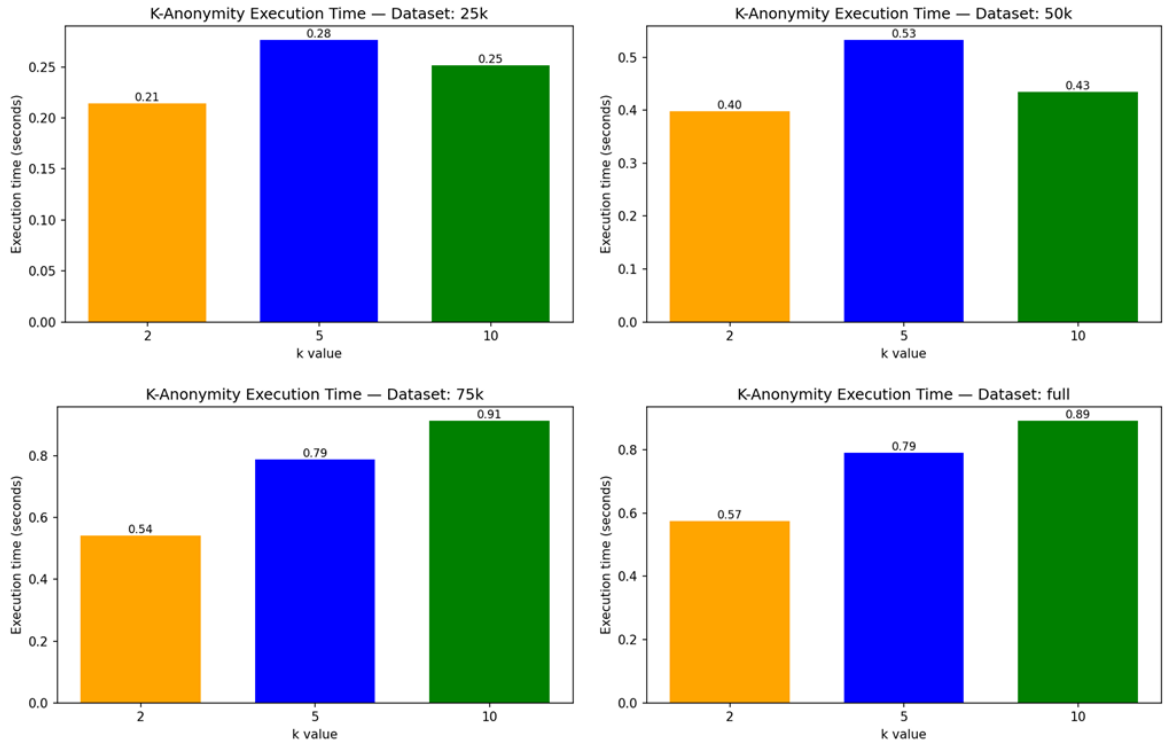


Figure 6.1 *Execution time for k-anonymity (for all dataset sizes)*

Figure 6.1 above shows execution time for k values of 2, 5, and 10 on a 25k-record dataset. The runtime starts at 0.21 seconds for k=2, increases to 0.28 seconds for k=5, and then slightly decreases to 0.25 seconds for k=10. The rise from k=2 to k=5 is expected, as more stringent anonymity requires additional generalization and suppression to ensure each equivalence class meets the minimum size. Interestingly, the small drop at k=10 occurs because the suppression step removes a larger portion of violating classes early in the process, reducing the dataset size for subsequent operations. This effect outweighs the extra cost of checking stricter k at this smaller scale. All runtimes are well below 0.3 seconds, showing that k-anonymity is computationally lightweight at this dataset size.

For 50k records, execution time increases from 0.40 seconds at k=2 to 0.53 seconds at k=5, before dropping to 0.43 seconds at k=10. The jump between k=2 and k=5 (+0.13 seconds) reflects the extra grouping and verification needed for larger equivalence classes. However, as in the 25k case, k=10's early and extensive suppression of small groups reduces the table size, which leads to lower runtime than k=5. This non-monotonic pattern demonstrates that suppression can sometimes offset the computational demands of stricter k values. Even with 50k rows, processing time stays under 0.6 seconds, confirming efficient scaling up to this dataset size.

The 75k dataset shows a clear monotonic increase: 0.54 seconds (k=2), 0.79 seconds (k=5), and 0.91 seconds (k=10). At this scale, fewer equivalence classes are small enough to be removed by suppression; instead, achieving higher k primarily requires more generalization, which increases processing time consistently. The runtime jump from k=2 to k=5 is the largest (+0.25 seconds), indicating that initial increases in k have the most significant computational impact. The added +0.12 seconds from k=5 to k=10 still shows that tighter privacy constraints require more work. This steady growth highlights that at larger sizes, stricter k values directly translate to higher computational costs.

With the full dataset, execution time rises from 0.57 seconds (k=2) to 0.79 seconds (k=5) and 0.89 seconds (k=10). The proportional increases are smaller than in the 75k dataset, suggesting diminishing marginal cost as dataset size and generalization levels increase. Once a certain degree of generalization is applied for k=5, the transition to k=10 requires comparatively fewer additional changes, keeping the added runtime low. Across all k values, the execution time remains under 0.9 seconds, confirming that k-anonymity remains feasible for datasets of this scale on standard hardware.

6.1.2 ℓ-Diversity

For ℓ-diversity, execution time was measured for **ℓ = 2 (small)**, **ℓ = 3 (moderate)**, and **ℓ = 4 (large)**.

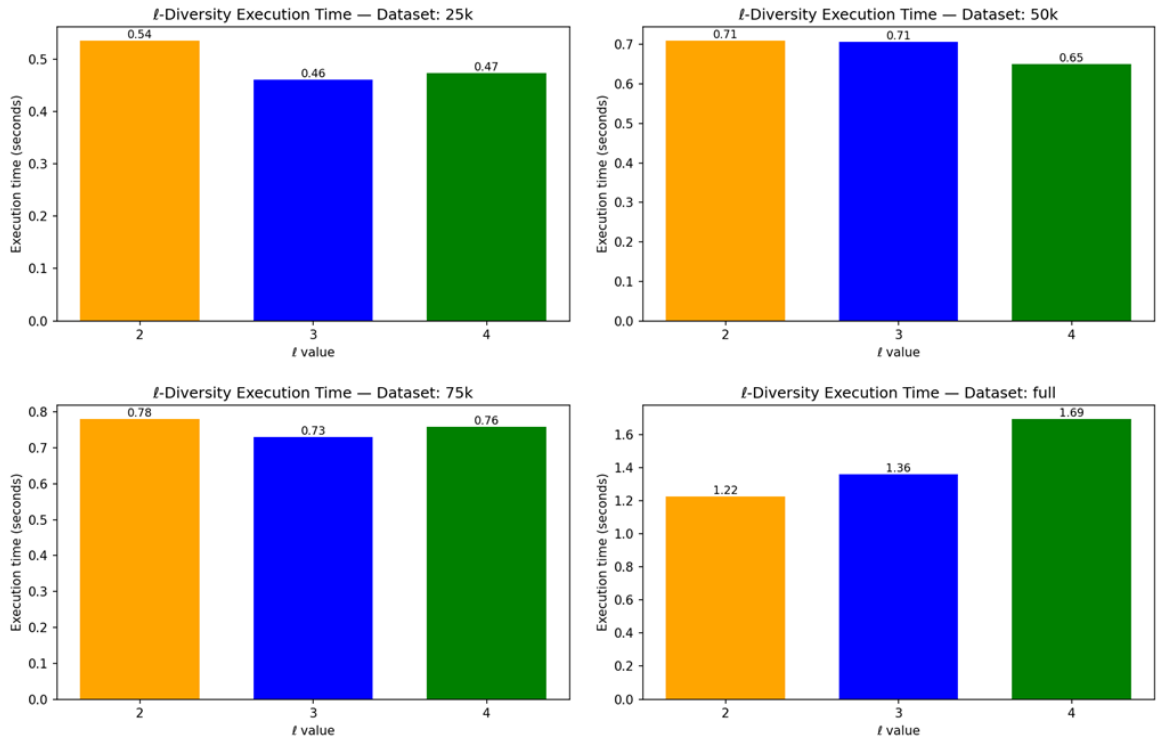


Figure 6.2 *Execution time for l-diversity (for all dataset sizes)*

As shown in figure 6.2 above, for the 25k record dataset, execution times for ℓ-diversity remain under one second across all ℓ values. Specifically, ℓ=2 requires the highest execution time (0.54 seconds), while ℓ=3 and ℓ=4 take slightly less, 0.46 and 0.47 seconds, respectively. This indicates that for smaller datasets, increasing ℓ beyond 2 does not substantially increase computational overhead. In fact, execution time decreases slightly, suggesting that at smaller scales, the grouping process for higher ℓ values may be computationally manageable. The sub-second performance shows that ℓ-diversity is efficient for datasets of this size, making it practical for real-time or frequent anonymization tasks.

With 50k records, execution time rises moderately compared to the 25k dataset. At ℓ=2, execution time is 0.71 seconds, identical to ℓ=3 (0.71 seconds), while ℓ=4 is slightly lower at 0.65 seconds. This pattern suggests that scaling up the dataset doubles the data size but only increases execution time by ~0.2 seconds. Interestingly, the lowest execution time is observed at ℓ=4, which may reflect dataset-specific efficiencies in grouping diverse sensitive values. Overall, performance remains strong: even at 50k records, the execution remains well below one second, indicating good scalability for medium-scale healthcare data.

At 75k records, execution times show a more noticeable increase but remain under one second. ℓ=2 takes 0.78 seconds, while ℓ=3 drops slightly to 0.73 seconds, and ℓ=4 rises again to 0.76 seconds. This non-linear variation across ℓ values highlights that execution cost is not strictly tied to increasing ℓ; rather, it depends on how quasi-identifiers and sensitive attributes distribute across the dataset. Still, the execution times are consistent, with a maximum difference of only 0.05 seconds between ℓ values. The scalability pattern shows that ℓ-diversity continues to handle datasets up to 75k efficiently without prohibitive time costs.

For the full dataset of 100k records execution time increases more significantly. At ℓ=2, execution requires 1.22 seconds, rising to 1.36 seconds at ℓ=3 and peaking at 1.69 seconds for ℓ=4. This demonstrates a clear upward trend: as both dataset size and ℓ value increase, execution times scale upward, reflecting the growing complexity of ensuring diversity within equivalence classes. The increase beyond one second signals that for large-scale healthcare datasets, higher ℓ thresholds may impose noticeable computational costs. Nevertheless, even at 100k records and ℓ=4, the maximum execution time (1.69 seconds) remains relatively modest compared to the dataset's scale, suggesting that ℓ-diversity remains computationally feasible for anonymizing large healthcare datasets.

6.1.3 t-Closeness

For t-closeness, execution time was measured for **t = 0.4 (small)**, **t = 0.7 (moderate)**, and **t = 1.0 (large)**.

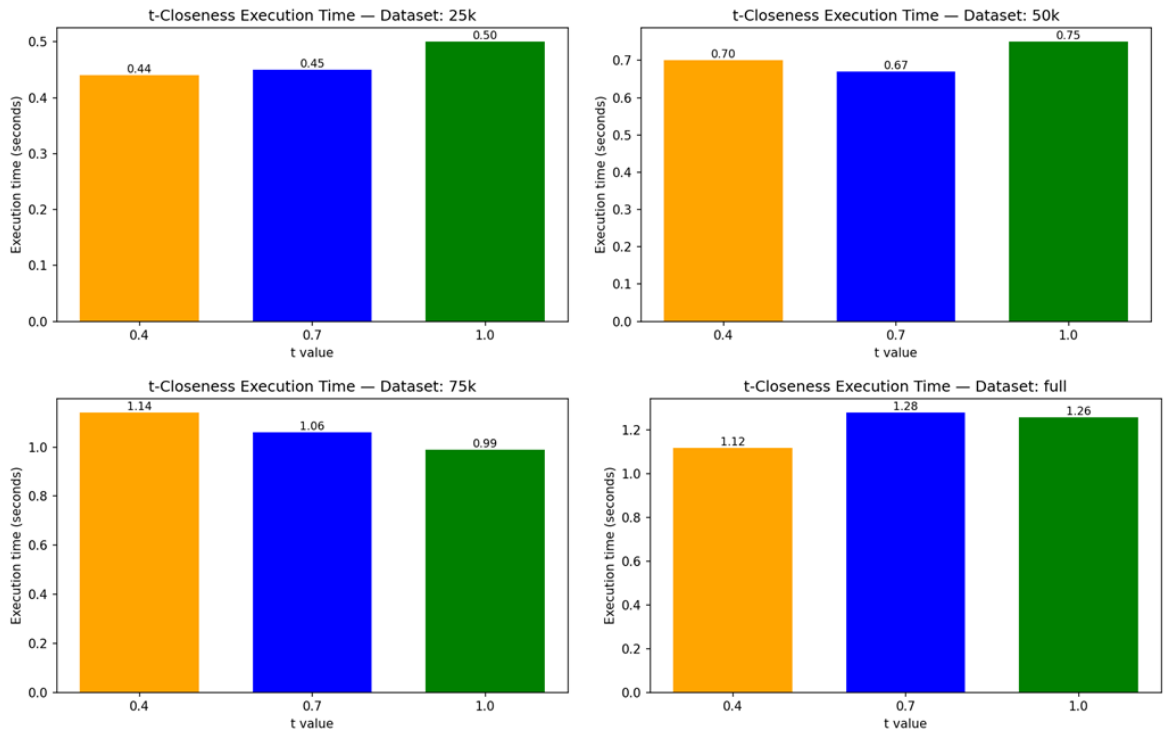


Figure 6.3 *Execution time for t-closeness (for all dataset sizes)*

Figure 6.3 presents execution times for applying t-closeness on a dataset containing 25k records. The execution times are 0.44 seconds for t = 0.4, 0.45 seconds for t = 0.7, and 0.50 seconds for t = 1.0. The results show that execution time remains relatively stable for smaller datasets, with only a marginal increase (0.06 seconds) as t grows from 0.4 to 1.0. This stability suggests that at smaller scales, the impact of stricter t-closeness thresholds on computational cost is minimal, making the method feasible for real-time or small-scale healthcare data anonymization tasks.

For the 50k record dataset, the times recorded are 0.70 seconds (t = 0.4), 0.67 seconds (t = 0.7), and 0.75 seconds (t = 1.0). Here, we observe a small fluctuation: execution is slightly faster at t = 0.7 compared to t = 0.4, but the time rises again at t = 1.0. This pattern highlights that

execution times for moderate dataset sizes are not strictly monotonic with respect to t values, likely due to variations in group distributions during partitioning. Nonetheless, all three t values keep execution times under 0.8 seconds, maintaining good efficiency.

Results for the 75k record dataset, the recorded execution times are 1.14 seconds (t = 0.4), 1.06 seconds (t = 0.7), and 0.99 seconds (t = 1.0). Interestingly, the trend here reverses compared to smaller datasets: the execution time decreases as t increases. For example, execution time reduces by 0.15 seconds when moving from t = 0.4 to t = 1.0. This indicates that with larger datasets, looser privacy constraints (higher t values) may reduce the computational effort needed for group balancing. Hence, t-closeness shows greater sensitivity to dataset size than to the exact t parameter itself.

For the full dataset of 100k records, execution times are 1.12 seconds (t = 0.4), 1.28 seconds (t = 0.7), and 1.26 seconds (t = 1.0). In contrast to the 75k dataset, execution time increases at higher t values here, peaking at 1.28 seconds for t = 0.7. The variation across t values (0.16 seconds difference between the lowest and highest times) reflects the complexity of enforcing different closeness constraints in larger datasets, where group balancing and distribution matching become more computationally intensive. Even so, execution times remain reasonable, staying below 1.3 seconds.

6.1.4 Differential Privacy

For differential privacy, execution time was measured for ε = 8.0 (small privacy), ε = 5.0 (moderate privacy), and ε = 2.0 (high privacy). The total execution time was obtained by summing the processing times for simple and complex queries under each configuration.

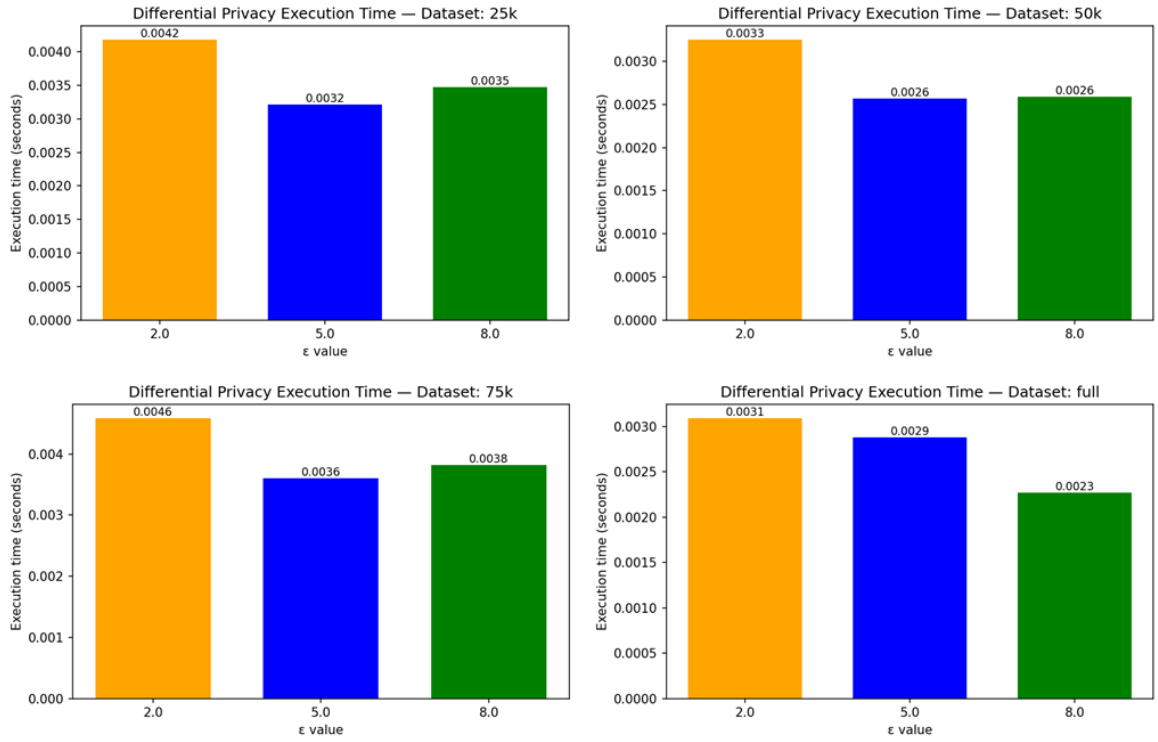


Figure 6.4 *Execution time for Differential Privacy (25k dataset)*

The figure 6.4 above illustrates the execution time of the Differential Privacy mechanism on a dataset of 25k records with varying values of the privacy parameter ε (epsilon). When ε is set to 2.0, the execution time reaches 0.0042 seconds, the highest among the three settings. For ε = 5.0, the execution time decreases significantly to 0.0032 seconds, while for ε = 8.0, it slightly increases again to 0.0035 seconds. This trend suggests that smaller values of ε (indicating stronger privacy guarantees) are associated with higher computational overhead, as the mechanism must add more noise to the data. Conversely, larger ε values (weaker privacy) reduce the time complexity slightly.

For a dataset of 50k records, at ε = 2.0, the execution time is 0.0033 seconds. This decreases to 0.0026 seconds for both ε = 5.0 and ε = 8.0, which remain identical. Compared to the 25k dataset, the execution times are slightly lower, possibly due to optimization effects or variance in how noise addition scales with intermediate dataset sizes. The results again confirm that stricter privacy guarantees (smaller ε) come with marginally higher execution costs, although the difference across ε values is not large. Importantly, the execution remains under 0.004 seconds, showing consistent efficiency.

For the 75k record dataset, execution time increases compared to smaller datasets. At ε = 2.0, the mechanism requires 0.0046 seconds, which is the highest observed so far across all datasets. When ε is increased to 5.0, the time decreases to 0.0036 seconds, while for ε = 8.0, it slightly rises again to 0.0038 seconds. These results align with the trend observed earlier: stronger privacy demands (smaller ε) incur higher costs, whereas weaker privacy allows for faster execution. The absolute execution times, however, remain extremely small (all below 0.005 seconds), reaffirming the efficiency of the implementation even as dataset size grows.

For the full 100k dataset, at ε = 2.0, the execution time is 0.0031 seconds, which decreases to 0.0029 seconds for ε = 5.0, and further down to 0.0023 seconds for ε = 8.0. Interestingly, just like the previous datasets, the execution time is maximum at the smallest ε. Across all dataset sizes, the effect of ε on runtime becomes more apparent in a monotonic manner. Importantly, even at full scale, execution times are extremely low — all well under 0.004 seconds — demonstrating that Differential Privacy can be applied to large datasets without imposing significant runtime costs.

6.1.5 Brief Comparative Analysis

When comparing all four techniques:

- Dataset-level methods (k-anonymity, ℓ-diversity, t-closeness) showed increases in execution time with both dataset size and stricter privacy parameters.
- Differential privacy exhibited more stable execution times across dataset sizes, as it processes queries rather than transforming entire datasets.
- t-Closeness consistently required the most time for large datasets, while the fastest was the fastest under all parameters.

This analysis emphasizes the trade-off between privacy level and computational efficiency, which is critical for real-world healthcare applications where timely data access is essential.

6.2 Data Utility

In this study, utility was measured through the execution of both simple queries and complex queries. The accuracy of query results was evaluated using Relative Error (%), where lower values indicate higher fidelity to the original dataset, and Utility, which quantifies the retained usefulness of anonymized data.

6.2.1 K-Anonymity

Tables 6.1 and 6.2 present the utility results for K-Anonymity across all dataset sizes for simple and complex queries, respectively.

For **simple queries** as shown in table 6.1 below, the pattern is similar, though slightly higher errors are observed compared to complex queries.

- In the **25k dataset**, relative errors increase slightly with larger k. For example, Q1 has **0.60% error at k=5** and **1.26% at k=10**, and Q5 has **0.79% at k=5** and **1.27% at k=10**. While these values are higher than complex queries, they remain comfortably within the **"Good" utility threshold (<5%)**.
- The **50k dataset** follows the same trend, with most values between **0.00% and 0.56%**. The highest occurs at Q5, k=10, with **0.56%**.
- For the **75k dataset**, errors remain very low, generally under **0.25%**, such as **0.13% (Q1 at k=10)** and **0.25% (Q5 at k=10)**.
- In the **full dataset**, results are nearly identical, with maximum errors like **0.26%** (Q4 and Q5 at k=10).

The trend shows that simple queries are slightly more sensitive to anonymization, but the overall distortion is still very low. The relative error never exceeds 1.27%, which is far below the 5% threshold. Thus, k-anonymity maintains high analytical reliability for simple, record-level queries.

Table 6.1 *Data utility results for K-Anonymity (Simple Queries) for k = 2, 5 and 10*

K-Anonymity Utility — Simple Queries

	Query	RelErr_2	Utility_2	RelErr_5	Utility_5	RelErr_10	Utility_10
25k	Q1	0.00%	Good	0.60%	Good	1.26%	Good
25k	Q2	0.00%	Good	0.05%	Good	0.12%	Good
25k	Q3	0.00%	Good	0.02%	Good	0.06%	Good
25k	Q4	0.00%	Good	0.77%	Good	1.24%	Good
25k	Q5	0.00%	Good	0.79%	Good	1.27%	Good
50k	Q1	0.00%	Good	0.16%	Good	0.49%	Good
50k	Q2	0.00%	Good	0.03%	Good	0.04%	Good
50k	Q3	0.00%	Good	0.00%	Good	0.09%	Good
50k	Q4	0.00%	Good	0.33%	Good	0.55%	Good
50k	Q5	0.00%	Good	0.33%	Good	0.56%	Good
75k	Q1	0.00%	Good	0.04%	Good	0.13%	Good
75k	Q2	0.00%	Good	0.00%	Good	0.03%	Good
75k	Q3	0.00%	Good	0.00%	Good	0.00%	Good
75k	Q4	0.00%	Good	0.07%	Good	0.25%	Good
75k	Q5	0.00%	Good	0.07%	Good	0.25%	Good
full	Q1	0.00%	Good	0.03%	Good	0.13%	Good
full	Q2	0.00%	Good	0.00%	Good	0.02%	Good
full	Q3	0.00%	Good	0.00%	Good	0.00%	Good
full	Q4	0.00%	Good	0.05%	Good	0.26%	Good
full	Q5	0.00%	Good	0.05%	Good	0.26%	Good

The table 6.2 below for **k-anonymity under complex queries** shows consistently strong utility across all datasets and k-values.

- For the **25k dataset**, the relative error (RelErr) values remain extremely low, ranging from **0.00% to 0.25%** across queries Q1–Q5. For instance, at k=2, every query has 0.00% error, while at k=10 the maximum observed error is 0.25% (Q3). All values fall well below the 5% threshold, classifying the utility as **"Good."**
- For the **50k dataset**, relative errors are again negligible, mostly **0.00%**, with a maximum of **0.99%** at k=10 for Q3. This small spike still lies within the "Good" utility range.
- For the **75k dataset**, errors remain at or near **0.00%** for most queries, with very minor variations up to **0.03%** (Q3 at k=10).
- For the **full dataset**, relative error values are virtually **0.00%** across all queries and k-values, showing almost perfect preservation of analytical accuracy.

The results indicate that k-anonymity introduces almost no analytical distortion for complex queries, regardless of dataset size or k-value. Even at higher k (which imposes stronger anonymity), the impact on query accuracy is negligible. This demonstrates that k-anonymity provides robust privacy protection with **minimal trade-offs in complex analytical tasks**.

Table 6.2 Data utility results for K-Anonymity (Complex Queries) for k = 2, 5 and 10

K-Anonymity Utility — Complex Queries

	Query	RelErr_2	Utility_2	RelErr_5	Utility_5	RelErr_10	Utility_10
25k	Q1	0.00%	Good	0.04%	Good	0.17%	Good
25k	Q2	0.00%	Good	0.02%	Good	0.05%	Good
25k	Q3	0.00%	Good	0.12%	Good	0.25%	Good
25k	Q4	0.00%	Good	0.03%	Good	0.10%	Good
25k	Q5	0.00%	Good	0.05%	Good	0.11%	Good
50k	Q1	0.00%	Good	0.01%	Good	0.07%	Good
50k	Q2	0.00%	Good	0.00%	Good	0.01%	Good
50k	Q3	0.00%	Good	0.05%	Good	0.99%	Good
50k	Q4	0.00%	Good	0.00%	Good	0.02%	Good
50k	Q5	0.00%	Good	0.01%	Good	0.04%	Good
75k	Q1	0.00%	Good	0.00%	Good	0.01%	Good
75k	Q2	0.00%	Good	0.00%	Good	0.00%	Good
75k	Q3	0.00%	Good	0.01%	Good	0.03%	Good
75k	Q4	0.00%	Good	0.00%	Good	0.00%	Good
75k	Q5	0.00%	Good	0.00%	Good	0.01%	Good
full	Q1	0.00%	Good	0.00%	Good	0.02%	Good
full	Q2	0.00%	Good	0.00%	Good	0.00%	Good
full	Q3	0.00%	Good	0.00%	Good	0.04%	Good
full	Q4	0.00%	Good	0.00%	Good	0.01%	Good
full	Q5	0.00%	Good	0.00%	Good	0.01%	Good

6.2.2 L-Diversity

Tables 6.3 and 6.4 show the performance of L-Diversity for simple and complex queries, respectively.

For **simple queries** as shown in table 6.3 below, ℓ-diversity again introduces slightly higher error rates compared to complex queries, but all remain within the "Good" range.

- In the **25k dataset**, relative errors are **0.00–1.27%**. Notably, Q5 records **1.27% error at ℓ=3 and ℓ=4**, the highest in this dataset.
- In the **50k dataset**, values range from **0.00% to 0.88%**, such as Q1 at **0.88% (ℓ=4)** and Q4 at **0.33% (ℓ=3)**.
- For the **75k dataset**, errors remain modest, mostly under **0.40%**, such as **0.13% (Q1 at ℓ=3)** and **0.40% (Q4 at ℓ=4)**.
- In the **full dataset**, errors are very small, with the maximum at **0.44% (Q4 at ℓ=4)** and most queries at or near **0.00%**.

Simple queries under ℓ-diversity show slightly higher errors compared to complex queries, but the maximum observed error of **1.27%** is still very low. This confirms that ℓ-diversity preserves utility effectively even for granular query types, ensuring that data remains useful while providing stronger attribute-level privacy protection.

Table 6.3 *Data utility results for L-Diversity (Simple Queries) for l = 2, 3 and 4*

ℓ-Diversity Utility — Simple Queries

	Query	RelErr_2	Utility_2	RelErr_3	Utility_3	RelErr_4	Utility_4
25k	Q1	0.16%	Good	1.26%	Good	1.26%	Good
25k	Q2	0.02%	Good	0.12%	Good	0.12%	Good
25k	Q3	0.00%	Good	0.06%	Good	0.06%	Good
25k	Q4	0.33%	Good	1.24%	Good	1.24%	Good
25k	Q5	0.33%	Good	1.27%	Good	1.27%	Good
50k	Q1	0.11%	Good	0.16%	Good	0.88%	Good
50k	Q2	0.03%	Good	0.03%	Good	0.10%	Good
50k	Q3	0.00%	Good	0.00%	Good	0.11%	Good
50k	Q4	0.22%	Good	0.33%	Good	0.79%	Good
50k	Q5	0.22%	Good	0.33%	Good	0.80%	Good
75k	Q1	0.00%	Good	0.13%	Good	0.31%	Good
75k	Q2	0.00%	Good	0.03%	Good	0.06%	Good
75k	Q3	0.00%	Good	0.00%	Good	0.06%	Good
75k	Q4	0.00%	Good	0.25%	Good	0.40%	Good
75k	Q5	0.00%	Good	0.25%	Good	0.26%	Good
full	Q1	0.00%	Good	0.03%	Good	0.35%	Good
full	Q2	0.00%	Good	0.00%	Good	0.04%	Good
full	Q3	0.00%	Good	0.00%	Good	0.05%	Good
full	Q4	0.00%	Good	0.05%	Good	0.44%	Good
full	Q5	0.00%	Good	0.05%	Good	0.32%	Good

The **ℓ-diversity complex query table** as shown in table 6.4 below also indicates strong preservation of utility across all datasets and ℓ-values.

- In the **25k dataset**, errors are very small, ranging from **0.00% to 0.25%**. For example, Q1 records **0.03% error at ℓ=2** and **0.17% at ℓ=4**. Similarly, Q3 has a maximum of **0.25%** across settings.
- For the **50k dataset**, relative errors remain under **1.05%**. The highest appears in Q3 at ℓ=4 (**1.05%**), which is still classified as "Good." Most other queries maintain errors between **0.00% and 0.14%**.
- In the **75k dataset**, the results are consistent, with errors such as **0.95% (Q3 at ℓ=4)** but otherwise close to **0.00%**.
- The **full dataset** shows negligible distortion, with errors mostly around **0.00–0.10%**, except for one outlier at **0.57% (Q3 at ℓ=4)**.

Complex queries under ℓ-diversity experience slightly higher error spikes than k-anonymity, particularly at ℓ=3. However, errors remain well below the 5% "Good" threshold, ensuring reliable analytical accuracy. Thus, ℓ-diversity preserves **complex query results effectively while enhancing protection against attribute disclosure.**

__Table 6.4__ Data utility results for L-Diversity (Complex Queries) for l = 2, 3 and 4

ℓ-Diversity Utility — Complex Queries

	Query	RelErr_2	Utility_2	RelErr_3	Utility_3	RelErr_4	Utility_4
25k	Q1	0.03%	Good	0.17%	Good	0.17%	Good
25k	Q2	0.00%	Good	0.05%	Good	0.05%	Good
25k	Q3	0.06%	Good	0.25%	Good	0.25%	Good
25k	Q4	0.00%	Good	0.10%	Good	0.10%	Good
25k	Q5	0.01%	Good	0.11%	Good	0.11%	Good
50k	Q1	0.01%	Good	0.01%	Good	0.14%	Good
50k	Q2	0.00%	Good	0.00%	Good	0.08%	Good
50k	Q3	0.04%	Good	0.05%	Good	1.05%	Good
50k	Q4	0.00%	Good	0.00%	Good	0.04%	Good
50k	Q5	0.01%	Good	0.01%	Good	0.08%	Good
75k	Q1	0.00%	Good	0.01%	Good	0.10%	Good
75k	Q2	0.00%	Good	0.00%	Good	0.00%	Good
75k	Q3	0.00%	Good	0.03%	Good	0.95%	Good
75k	Q4	0.00%	Good	0.00%	Good	0.01%	Good
75k	Q5	0.00%	Good	0.01%	Good	0.01%	Good
full	Q1	0.00%	Good	0.00%	Good	0.10%	Good
full	Q2	0.00%	Good	0.00%	Good	0.00%	Good
full	Q3	0.00%	Good	0.00%	Good	0.57%	Good
full	Q4	0.00%	Good	0.00%	Good	0.01%	Good
full	Q5	0.00%	Good	0.00%	Good	0.01%	Good

6.2.3 T-Closeness

Tables 6.5 and 6.6 depict the utility performance of T-Closeness across different t-values.

With simple queries as shown in table 6.5 below, t-Closeness demonstrates much stronger utility compared to complex queries.

- **For t = 0.4**, relative errors remain high for most queries, e.g., Q1 (**35.01%**), Q2 (**33.52%**), Q3 (**17.73%**) for 25k dataset and same trend for 50k, 75k and 100k dataset as well. These lead to *Poor* utility in nearly all cases. Only Q5 occasionally achieves moderate utility (e.g., **7.12% RelErr at 25k**, rated *Moderate*).
- **For t = 0.7**, utility improves substantially, with errors dropping below **1%** in many cases. For example, Q2 at 25k has just **0.19% RelErr (Good)**, and Q4 at 50k shows **1.72% RelErr (Good)**. However, Q1 (at 50k, 75k and 100k) still suffer from **RelErr (Poor)**.
- **For t = 1.0**, almost all queries perform exceptionally well. Relative errors fall close to zero, e.g., Q2 at full dataset has **0.03% RelErr (Good)**, and Q4 has **0.45% RelErr (Good)**. Only Q3 shows higher errors occasionally, such as **12.52% RelErr at 25k (Moderate)**.

For simple queries, t-Closeness utility is highly effective at t = 0.7 and t = 1.0, providing near-accurate results with low error. In contrast, t = 0.4 consistently leads to poor performance.

Table 6.5 *Data utility results for T-Closeness (Simple Queries) for t = 0.4, 0.7 and 1.0*

t-Closeness Utility — Simple Queries

	Query	RelErr_0.4	Utility_0.4	RelErr_0.7	Utility_0.7	RelErr_1	Utility_1
25k	Q1	35.01%	Poor	3.33%	Good	0.71%	Good
25k	Q2	33.52%	Poor	0.19%	Good	0.03%	Good
25k	Q3	17.73%	Poor	16.12%	Poor	12.52%	Moderate
25k	Q4	15.31%	Poor	6.84%	Moderate	0.91%	Good
25k	Q5	7.12%	Moderate	3.98%	Good	0.94%	Good
50k	Q1	35.03%	Poor	22.46%	Poor	0.60%	Good
50k	Q2	33.52%	Poor	0.45%	Good	0.01%	Good
50k	Q3	18.00%	Poor	16.66%	Poor	12.59%	Moderate
50k	Q4	15.08%	Poor	1.72%	Good	0.69%	Good
50k	Q5	6.60%	Moderate	1.24%	Good	0.71%	Good
75k	Q1	35.62%	Poor	31.46%	Poor	0.62%	Good
75k	Q2	33.58%	Poor	0.74%	Good	0.06%	Good
75k	Q3	17.28%	Poor	13.73%	Moderate	12.57%	Moderate
75k	Q4	21.17%	Poor	1.66%	Good	0.66%	Good
75k	Q5	8.38%	Moderate	0.63%	Good	0.52%	Good
full	Q1	34.32%	Poor	32.16%	Poor	0.54%	Good
full	Q2	33.48%	Poor	0.04%	Good	0.03%	Good
full	Q3	17.81%	Poor	14.22%	Moderate	12.55%	Moderate
full	Q4	10.33%	Moderate	2.13%	Good	0.45%	Good
full	Q5	4.11%	Good	0.62%	Good	0.23%	Good

The table 6.6 below for t-Closeness with complex queries provides insights into how query accuracy is affected by varying thresholds of *t* (0.4, 0.7, 1.0).

- **For t = 0.4**, the relative error values are significantly high for most queries. For example, at 25k queries, Q1 has a RelErr of **40.97%** (utility: *Poor*), Q2 records **18.15%**, and Q3 shows **19.99%**. Even at larger datasets (full), Q1 still exhibits **42.47%** relative error, again rated *Poor*. These consistently high errors across queries and dataset sizes indicate that t = 0.4 is too restrictive, leading to substantial information loss.
- **For t = 0.7**, performance improves considerably in terms of relative error. Many queries show very low relative errors, such as **0.32% for Q4 (25k)** and **0.94% for Q5 (75k)**, with utility marked *Good*. However, queries like Q1 still have higher errors (e.g., **18.46% at 25k**, *Poor*), highlighting that improvements depend on query type.
- **For t = 1.0**, the trend further improves. Several queries achieve near-zero relative error. For instance, Q4 at all dataset sizes records less than **0.05% RelErr** with *Good* utility, and Q5 also consistently shows values below **0.1% RelErr**. However, Q1, Q2, and Q3 still show higher errors (e.g., Q1 at **13.70% for 50k**) with utilities varying between *Moderate* and *Poor*.

t-Closeness utility improves as *t* increases, with t = 1.0 striking a better balance between privacy and data usability. Q4 and Q5 queries are consistently reliable, but Q1–Q3 suffer from higher inaccuracies under stricter thresholds.

Table 6.6 *Data utility results for T-Closeness (Complex Queries) for t = 0.4, 0.7 and 1.0*

t-Closeness Utility — Complex Queries

	Query	RelErr_0.4	Utility_0.4	RelErr_0.7	Utility_0.7	RelErr_1	Utility_1
25k	Q1	40.97%	Poor	18.46%	Poor	13.68%	Moderate
25k	Q2	18.15%	Poor	16.15%	Poor	15.63%	Poor
25k	Q3	19.99%	Poor	19.06%	Poor	15.89%	Poor
25k	Q4	3.89%	Good	0.32%	Good	0.01%	Good
25k	Q5	1.94%	Good	0.59%	Good	0.06%	Good
50k	Q1	41.06%	Poor	24.39%	Poor	13.70%	Moderate
50k	Q2	18.47%	Poor	17.67%	Poor	15.39%	Poor
50k	Q3	21.37%	Poor	18.26%	Poor	15.62%	Poor
50k	Q4	3.77%	Good	2.18%	Good	0.04%	Good
50k	Q5	1.80%	Good	0.82%	Good	0.05%	Good
75k	Q1	41.03%	Poor	19.86%	Poor	13.76%	Moderate
75k	Q2	18.29%	Poor	17.57%	Poor	15.40%	Poor
75k	Q3	23.19%	Poor	15.44%	Poor	15.26%	Poor
75k	Q4	4.03%	Good	3.41%	Good	0.01%	Good
75k	Q5	1.75%	Good	0.94%	Good	0.04%	Good
full	Q1	42.47%	Poor	21.46%	Poor	12.62%	Moderate
full	Q2	17.06%	Poor	15.31%	Poor	14.72%	Moderate
full	Q3	20.49%	Poor	15.16%	Poor	14.84%	Moderate
full	Q4	3.76%	Good	3.34%	Good	0.01%	Good
full	Q5	1.13%	Good	0.80%	Good	0.04%	Good

6.2.4 Differential Privacy

Tables 6.7 and 6.8 illustrate utility results for Differential Privacy at varying ε-values.

For simple queries as shown in table 6.7 below, differential privacy performs markedly better.

- **At ε = 2**, accuracy is reasonable for many queries. For example, Q1 at 25k has just **0.33% RelErr (Good)**, and Q2 shows **2.26% RelErr (Good)**. However, some queries like Q3 (25k) still show **107.17% RelErr (Poor)** and Q5 (25k) sshow **27.79% RelErr (Poor)**, highlighting vulnerability. However, as dataset size increases, utility got better for even these queries
- **At ε = 5**, errors reduce further, and utilities are frequently rated *Good*. For instance, Q1 at 25k shows only **0.07% RelErr (Good)**, and Q5 at 50k has **2.65% RelErr (Good)**. Still, occasional queries (e.g., Q3 at 75k, **20.16% RelErr, Poor**) deviate for all dataset sizes.
- **At ε = 8**, the majority of queries deliver near-accurate results. For example, Q1 at full dataset has just **0.02% RelErr (Good)**, Q2 has **0.12% RelErr (Good)**, and Q4 has **0.21% RelErr (Good)**. Utilities in these cases are overwhelmingly *Good*, with only a few queries (e.g., Q3 at 25k, **8.17% RelErr, Moderate**) slightly lagging.

Differential privacy utility is highly effective for simple queries when $\varepsilon \geq 5$, balancing accuracy and privacy. Performance for complex queries which we will discuss next, remains weaker, indicating query sensitivity plays a key role in DP's effectiveness.

Table 6.7 *Data utility results for Differential Privacy (Simple Queries) for epsilon = 2, 5 and 8*

Differential Privacy Utility — Simple Queries

	Query	RelErr_2	Utility_2	RelErr_5	Utility_5	RelErr_8	Utility_8
25k	Q1	0.33%	Good	0.07%	Good	0.33%	Good
25k	Q2	2.26%	Good	0.88%	Good	0.71%	Good
25k	Q3	107.17%	Poor	15.36%	Poor	8.17%	Moderate
25k	Q4	2.77%	Good	0.74%	Good	0.57%	Good
25k	Q5	27.99%	Poor	10.66%	Moderate	8.98%	Moderate
50k	Q1	0.05%	Good	0.26%	Good	0.04%	Good
50k	Q2	1.30%	Good	0.67%	Good	0.21%	Good
50k	Q3	21.28%	Poor	10.43%	Moderate	7.34%	Moderate
50k	Q4	0.50%	Good	0.60%	Good	0.21%	Good
50k	Q5	7.63%	Moderate	2.65%	Good	5.17%	Moderate
75k	Q1	0.03%	Good	0.00%	Good	0.02%	Good
75k	Q2	0.73%	Good	0.63%	Good	0.12%	Good
75k	Q3	6.89%	Moderate	20.16%	Poor	11.98%	Moderate
75k	Q4	0.98%	Good	0.24%	Good	0.08%	Good
75k	Q5	5.30%	Moderate	1.13%	Good	2.12%	Good
full	Q1	0.20%	Good	0.16%	Good	0.02%	Good
full	Q2	0.22%	Good	0.41%	Good	0.12%	Good
full	Q3	7.61%	Moderate	22.04%	Poor	2.48%	Good
full	Q4	0.57%	Good	0.30%	Good	0.21%	Good
full	Q5	7.41%	Moderate	1.17%	Good	1.36%	Good

The results as shown in table 6.8 below for differential privacy with complex queries highlight the challenges of balancing noise addition with accuracy.

- **At ε = 2**, the errors are extremely high, rendering the utility largely unusable. For instance, Q4 at 25k shows a staggering **2163.67% RelErr (Poor)**, and Q2 records **1423.13% RelErr (Poor)**. Even at full dataset, Q1 has **818.93% RelErr (Poor)**.
- **At ε = 5**, the errors decrease but remain substantial. For instance, Q1 at 25k has **381.80% RelErr (Poor)**, and Q4 at 50k has **1761.67% RelErr (Poor)**. Utility ratings improve only for select cases, e.g., Q5 at 50k (RelErr **4.82%**, rated *Good*).
- **At ε = 8**, the noise is much reduced, and results improve. For example, Q5 at 25k yields **7.93% RelErr (Moderate)**, while Q5 at full achieves **1.25% RelErr (Good)**. However, queries like Q4 still suffer from very high errors (e.g., **944.61% at 25k**, *Poor*).

Differential privacy utility under complex queries is poor at lower ε values (high privacy). Higher ε (weaker privacy) improves accuracy significantly, but not uniformly across queries. Q5 is the most resilient query, consistently reaching *Good* ratings at ε = 8, while Q1–Q4 remain problematic.

Table 6.8 *Data utility results for Differential Privacy (Complex Queries) for epsilon = 2, 5 and 8*

Differential Privacy Utility — Complex Queries

	Query	RelErr_2	Utility_2	RelErr_5	Utility_5	RelErr_8	Utility_8
25k	Q1	816.58%	Poor	381.80%	Poor	204.07%	Poor
25k	Q2	1423.13%	Poor	609.10%	Poor	381.56%	Poor
25k	Q3	72.79%	Poor	24.19%	Poor	12.29%	Moderate
25k	Q4	2163.67%	Poor	1125.89%	Poor	944.61%	Poor
25k	Q5	16.98%	Poor	10.57%	Moderate	7.93%	Moderate
50k	Q1	1601.26%	Poor	465.24%	Poor	299.44%	Poor
50k	Q2	1856.52%	Poor	772.51%	Poor	391.70%	Poor
50k	Q3	30.92%	Poor	13.42%	Moderate	9.70%	Moderate
50k	Q4	3364.49%	Poor	1761.67%	Poor	869.78%	Poor
50k	Q5	10.34%	Moderate	4.82%	Good	3.68%	Good
75k	Q1	1087.70%	Poor	390.55%	Poor	282.27%	Poor
75k	Q2	1881.98%	Poor	699.80%	Poor	560.72%	Poor
75k	Q3	27.03%	Poor	6.15%	Moderate	5.81%	Moderate
75k	Q4	2640.85%	Poor	1147.83%	Poor	954.16%	Poor
75k	Q5	10.98%	Moderate	3.25%	Good	1.60%	Good
full	Q1	818.93%	Poor	545.66%	Poor	300.12%	Poor
full	Q2	1351.24%	Poor	753.05%	Poor	317.70%	Poor
full	Q3	14.62%	Moderate	4.03%	Good	5.79%	Moderate
full	Q4	2348.65%	Poor	1207.35%	Poor	615.86%	Poor
full	Q5	7.08%	Moderate	2.62%	Good	1.25%	Good

6.2.5 Brief Comparative Analysis

Although a detailed cross-method comparison is presented later, several trends are immediately observable:

1. **Effect of parameter strictness** – Across all methods, stricter privacy settings (higher k, higher ℓ, lower t, and lower ε) consistently reduced utility, with smaller datasets suffering the most.

2. **Dataset size resilience** – Larger datasets maintained better utility under stricter parameters, as the higher number of records allowed more flexibility in meeting privacy constraints.

3. **Method sensitivity** – Differential Privacy showed the largest sensitivity to parameter changes in simple queries, while K-Anonymity and L-Diversity suffered more in complex queries. T-Closeness tended to fall between the two extremes.

4. **Overall patterns** – At low privacy levels (small k, small ℓ, high t, high ε), all methods preserved utility well; however, Differential Privacy still added some unavoidable noise, whereas generalization-based methods retained near-original results for simple queries.

6.3 Re-identification Risk Evaluation

Re-identification risk evaluation is a crucial step in understanding the privacy protection offered by anonymization techniques. While execution performance and data utility focus on efficiency and accuracy, re-identification risk directly addresses the core objective of privacy preservation — ensuring that sensitive information about individuals cannot be inferred by adversaries, even when they have access to external datasets.

In this study, re-identification risks were quantified using the ARX Data Anonymization Tool, which evaluates datasets under three attacker models — Prosecutor, Journalist, and Marketer — each reflecting different levels of prior knowledge an adversary may have. Metrics such as **Records at Risk (%), Highest Risk (%), and Success Rate (%)** were recorded for the original datasets as well as for datasets anonymized using **k-anonymity, ℓ-diversity, and t-closeness**.

Differential privacy datasets were excluded from this stage because they do not produce static anonymized datasets; rather, they release query results with controlled noise, making ARX-based re-identification modelling inapplicable.

6.3.1 Results for the 25k Dataset

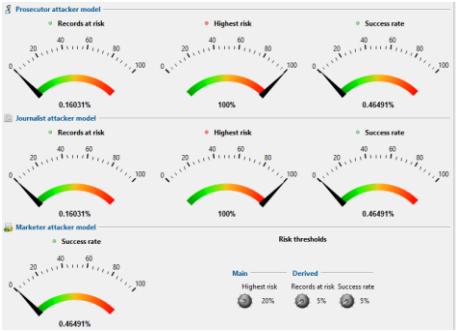


Figure 6.5 Re-identification Risk in Unanonymized 25k dataset

Figure 6.5 shows the baseline risk distribution of the 25k dataset shows high re-identification risks, with a significant portion of records being uniquely identifiable. This highlights that small-scale datasets are highly vulnerable.

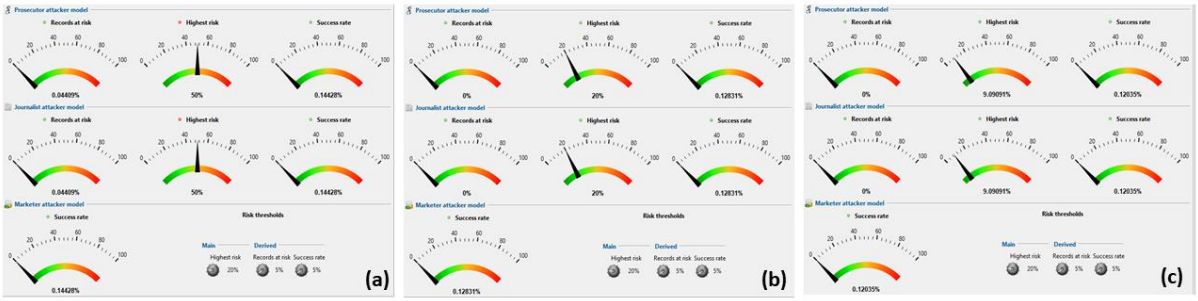


Figure 6.6 Re-identification Risk in 25k anonymized dataset at (a) k=2, (b) k=5, (c) k=10

Figure 6.6 shows that introducing k-anonymity reduces risks proportionally to the value of *k*. At k=2, risk drops but remains moderate, as groups of two records are not robust enough against re-identification. At k=5, risk decreases substantially, and at k=10, most records exhibit very low risk values, demonstrating strong identity protection.

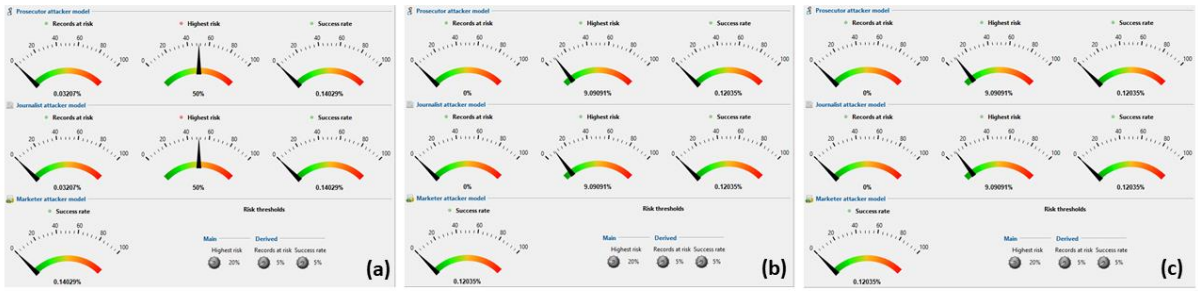


Figure 6.7 *Re-identification Risk in 25k anonymized dataset at (a) l=2, (b) l=3, (c) l=4*

Figure 6.7 shows that l-Diversity introduces protection against attribute disclosure by ensuring diversity in sensitive attributes within each group. At l=2, risks remain somewhat high, but with l=3 and l=4, the distributions shift downward, offering stronger protection. However, the improvements over k-anonymity are less pronounced in smaller datasets due to limited attribute variance.

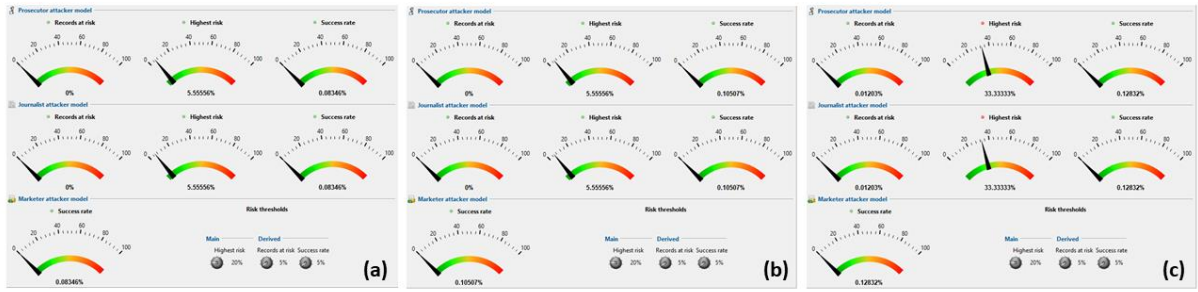


Figure 6.8 *Re-identification Risk in 25k anonymized dataset at (a) t = 0.4, (b) t=0.7, (c) t=1.0*

Figure 6.8 shows that t-Closeness further constrains sensitive attribute distributions. At t=0.4, risks are minimized, showing strong resilience against attribute disclosure. At t=0.7, risk is moderate, while at t=1.0, risks increase slightly since the constraint is relaxed. Compared to k-anonymity and l-diversity, t-closeness balances both identity and attribute disclosure.

6.3.2 Results for the 50k Dataset

The 50k dataset followed a similar pattern to the 25k dataset but with slightly different baseline risk magnitudes due to dataset size and attribute distribution changes.

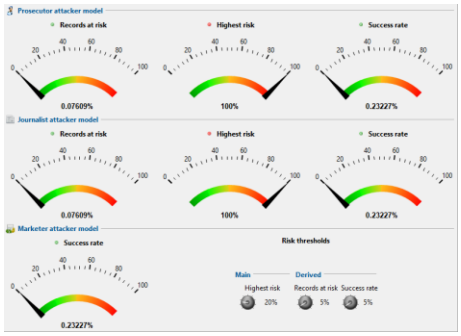


Figure 6.9 *Re-identification Risk in Unanonymized 50k dataset*

Figure 6.9 above shows that the original 50k dataset still shows high baseline re-identification risk, though slightly reduced compared to 25k due to the larger population size. Unique combinations of quasi-identifiers remain present, leaving many records at significant risk.

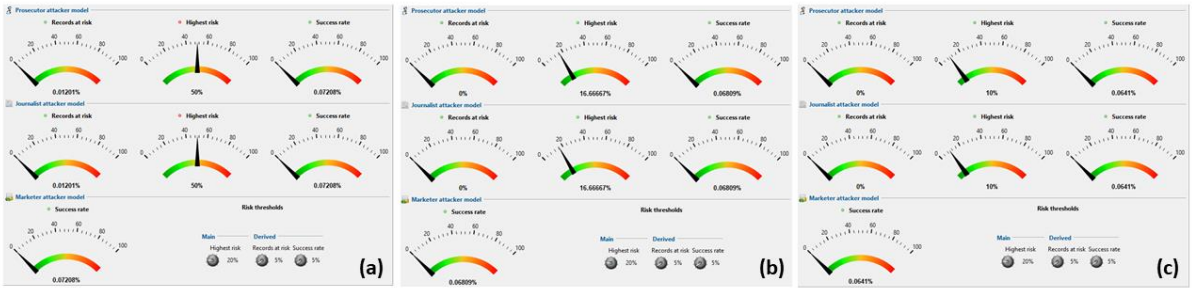


Figure 6.10 *Re-identification Risk in 50k anonymized dataset at (a) k=2, (b) k=5, (c) k=10*

Figure 6.10 above shows that at k=2, risks drop but remain noticeable. Increasing to k=5 provides strong reduction, and k=10 drives risks to very low levels, with most records nearly indistinguishable. Compared to the 25k dataset, the 50k dataset exhibits more stable anonymity, benefiting from greater population diversity.

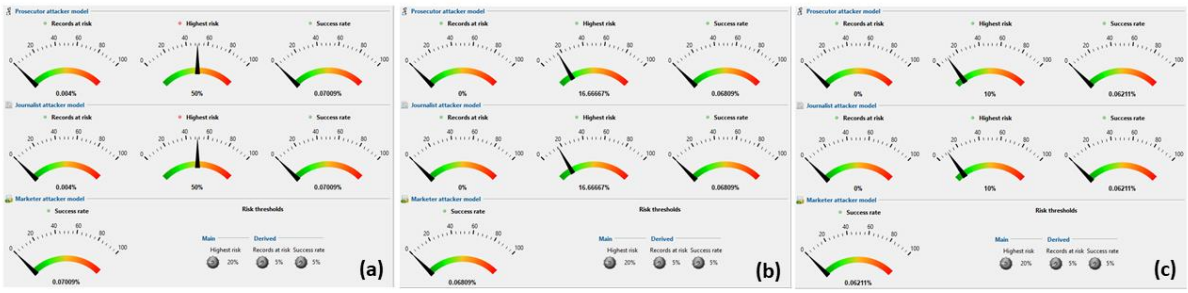


Figure 6.11 *Re-identification Risk in 50k anonymized dataset at (a) l=2, (b) l=3, (c) l=4*

Figure 6.11 shows that the l-diversity models show clear reductions in re-identification risk, especially at l=3 and l=4, where sensitive attribute disclosure is significantly minimized. The impact is stronger in 50k compared to 25k, as the larger dataset supports more diverse attribute groupings.

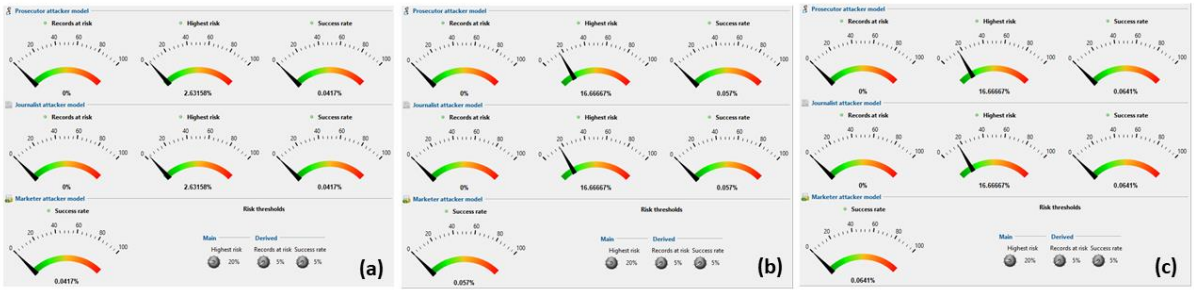


Figure 6.12 *Re-identification Risk in 50k anonymized dataset at (a) t=0.4, (b) t-0.7, (c) t=1.0*

Figure 6.12 shows that t-Closeness again demonstrates strong control over sensitive attribute risks. At t=0.4, risks are minimal; at t=0.7, moderate; and at t=1.0, slightly higher but still lower than the baseline. This shows t-closeness is particularly effective in larger datasets where sensitive distributions can be matched closely with the global distribution.

6.3.3 Results for the 75k Dataset

In the 75k dataset, anonymization effects were more pronounced due to the larger volume of quasi-identifiers and sensitive attribute combinations.

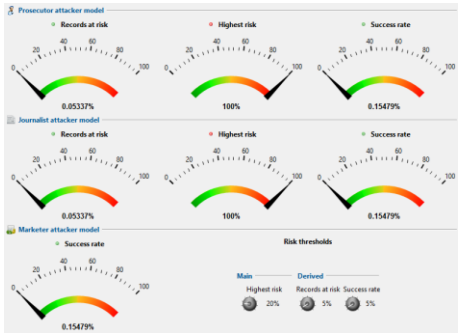


Figure 6.13 *Re-identification Risk in Unanonymized 75k dataset*

In figure 6.13 above, the baseline risk in the 75k dataset continues to be high but is reduced compared to 25k and 50k datasets due to the larger record count. Nevertheless, unique records persist, leaving identifiable risks without anonymization.

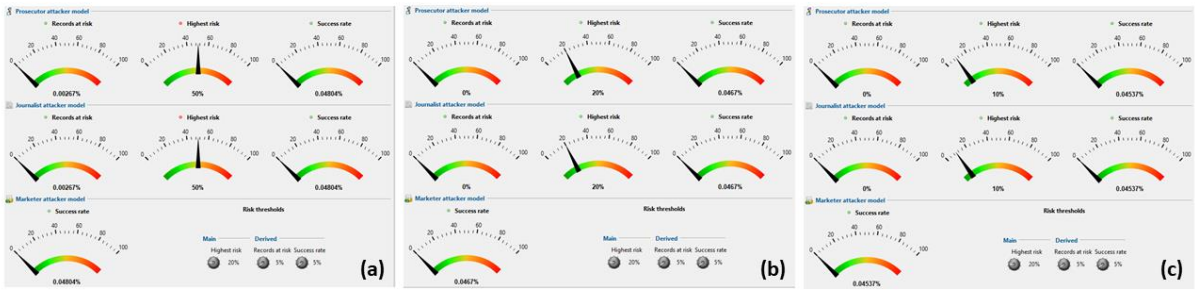


Figure 6.14 *Re-identification Risk in 75k anonymized dataset at (a) k=2, (b) k=5, (c) k=10*

In figure 6.14 above, the effect of k-anonymity is clearly visible, with significant drops in risk as *k* increases. At k=10, nearly all records show negligible re-identification risk. Compared to smaller datasets, the 75k dataset demonstrates almost similar protection, reflecting the advantage of scale.

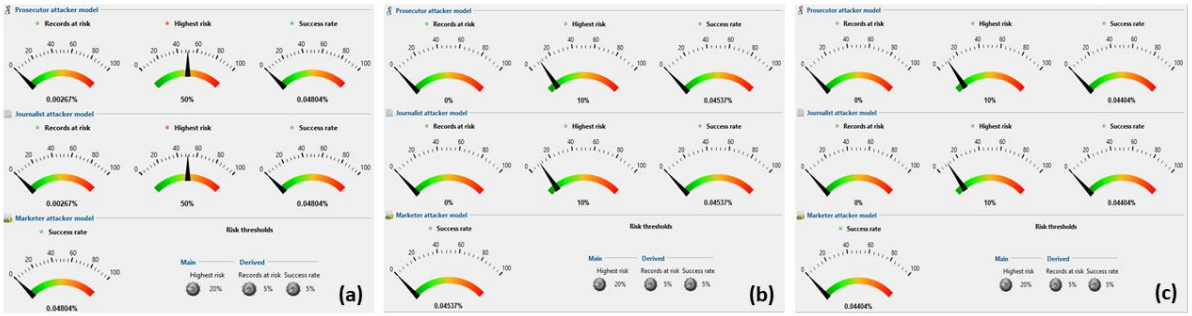


Figure 6.15 *Re-identification Risk in 75k anonymized dataset at (a) l=2, (b) l=3, (c) l=4*

In figure 6.15 above, l-Diversity provides stronger protection in the 75k dataset than in smaller datasets. At l=3 and l=4, re-identification risk is reduced substantially, with most records exhibiting very low probabilities of disclosure. This highlights the effectiveness of attribute diversity when supported by larger data volumes.

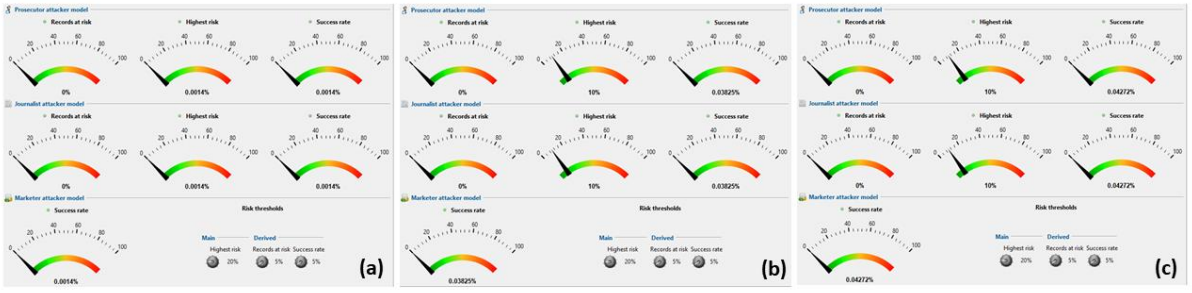


Figure 6.16 *Re-identification Risk in 75k anonymized dataset at (a) t=0.4, (b) t=0.7, (c) t=1.0*

In figure 6.16 above, t-Closeness again ensures risk reduction across all parameters. The t=0.4 model minimizes risks most effectively, while t=0.7 and t=1.0 provide moderate protection. Importantly, even with relaxed parameters, the 75k dataset maintains stronger protection than smaller datasets, showing scalability benefits.

6.3.4 Results for the Full (100k) Dataset

The full dataset demonstrated the cumulative effects of each technique at scale. Larger k and ℓ values continued to correlate with reduced re-identification risk, and t-closeness again achieved the lowest risks at *t=0.4*.

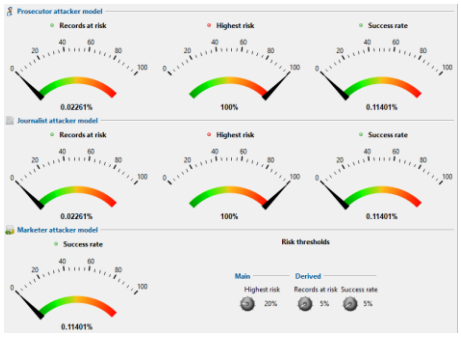


Figure 6.17 *Re-identification Risk in Unanonymized 100k dataset*

Figure 6.17 above shows that the baseline re-identification risk in the full 100k dataset, while high, is lower than that in smaller datasets due to population diversity. However, unprotected data remains highly vulnerable to identity disclosure.

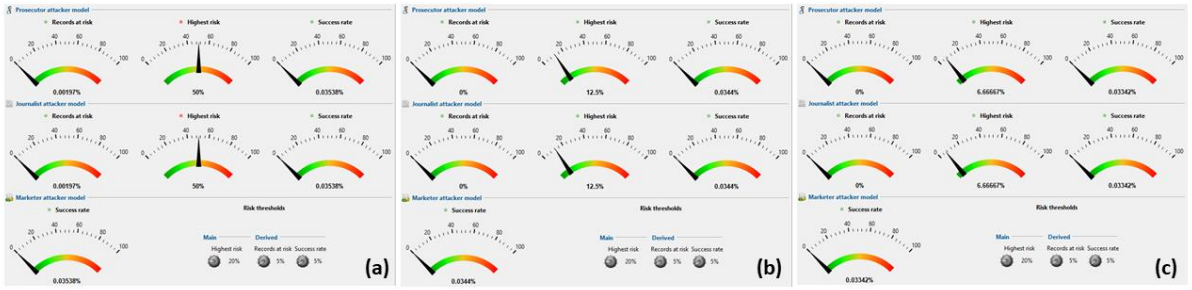


Figure 6.18 *Re-identification Risk in 100k anonymized dataset at (a) k=2, (b) k=5, (c) k=10*

In figure 6.18 above, at k=2, risks are moderately reduced; at k=5, risks drop significantly; and at k=10, risks are negligible, with most records effectively anonymized. The larger dataset size ensures more stability and consistency in achieving anonymity.

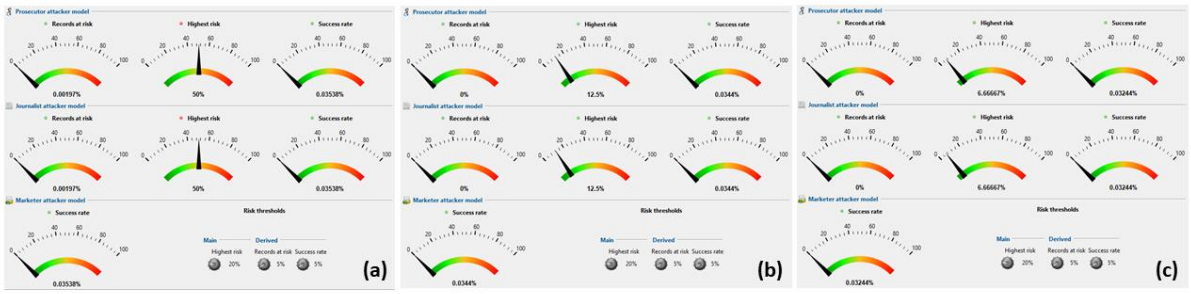


Figure 6.19 *Re-identification Risk in 100k anonymized dataset at (a) l=2, (b) l=3, (c) l=4*

In figure 6.19 above, l-Diversity demonstrates strong performance in the full dataset. At l=3 and l=4, risk distributions flatten considerably, minimizing both identity and attribute disclosure. This indicates that l-diversity achieves higher effectiveness when applied to large-scale data.

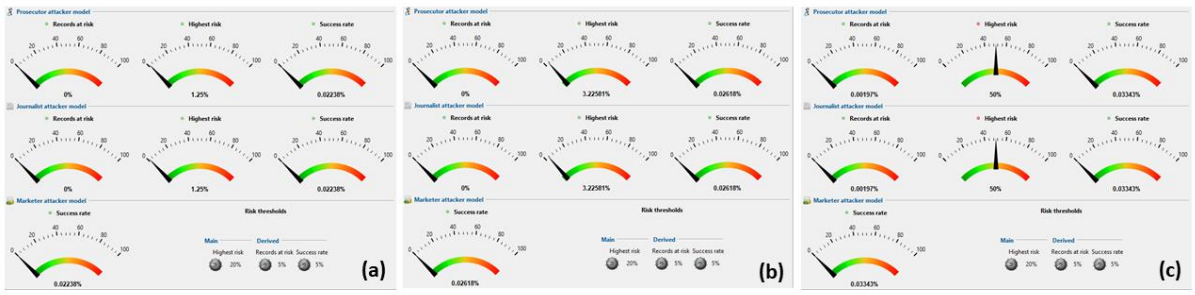


Figure 6.20 *Re-identification Risk in 100k anonymized dataset at (a) t=0.4, (b) t=0.7, (c) t=1.0*

In figure 6.20 above, t-Closeness provides the strongest attribute disclosure protection in the 100k dataset. At t=0.4, risks are nearly eliminated; at t=0.7, they remain low; and even at t=1.0, risks are substantially reduced compared to the baseline. The full dataset highlights the scalability and effectiveness of t-closeness in balancing identity and attribute protection.

6.3.5 Brief Comparative Analysis

Across all dataset sizes, several trends emerge:

- **k-anonymity**: Increasing *k* consistently lowers re-identification risk, and the degree of reduction increases as *k* becomes large.
- **ℓ-diversity**: Provides stronger protection than k-anonymity alone, especially if we keep the parameters at moderate values.
- **t-closeness**: Offers the most balanced and lowest risk levels, particularly at smaller *t* values, as it controls both identity disclosure and attribute disclosure risks.

- **Dataset size impact**: Larger datasets tend to start with slightly lower relative risk percentages due to more diverse equivalence classes but benefit similarly from anonymization techniques.

This stage confirms that all three evaluated techniques can reduce re-identification risk effectively, but their efficiency varies depending on parameter choice and dataset size. The most privacy-preserving configurations are not always the most utility-friendly — an aspect explored further in later sections.

6.4 Comparative Summary

The preceding sections presented detailed evaluations of execution performance, data utility, and re-identification risk for four privacy-preserving techniques — **k-anonymity, ℓ-diversity, t-closeness**, and **differential privacy** — across four dataset sizes (25k, 50k, 75k, and 100k) and three parameter settings (small, moderate, large). In this section, these results are synthesised to highlight overarching trends and trade-offs.

6.4.1 Execution Performance Overview

The comparative evaluation demonstrates that **k-anonymity consistently achieved fast execution times** across all dataset sizes and parameter settings, especially at lower k values. **ℓ-diversity** showed slightly higher execution times than k-anonymity, attributable to the added requirement of ensuring intra-group diversity. **t-closeness** was the slowest among the generalization-based models, as enforcing distributional similarity is computationally more intensive. For **differential privacy**, execution time remained relatively stable as well as fastest across different ε values, reflecting its query-based noise addition mechanism rather than full dataset transformation. Overall, the results confirm that differential privacy and k-anonymity is most efficient for large-scale datasets, whereas t-closeness imposes the heaviest computational cost.

6.4.2 Data Utility Overview

Utility analysis highlights that **k-anonymity at lower k values provided the best overall utility**, preserving accuracy in both simple and complex queries with relative error below the 5% “Good” threshold in most cases. As k increased, generalization reduced precision, but still keeping the utility good. **ℓ-diversity** retained acceptable utility at lower as well as higher ℓ values. **t-closeness** consistently exhibited greater utility loss compared to the other models, particularly at stricter thresholds, as distributional balancing leads to significant generalization and suppression. For **differential privacy**, utility varied widely: at higher ε values (weaker privacy), query accuracy was somewhat comparable to k-anonymity and l-diversity, while at lower ε values (stronger privacy), error rates increased sharply, especially for complex queries. Thus, the trade-off between privacy strength and analytical accuracy is most pronounced in DP.

6.4.3 Re-identification Risk Overview

Risk assessment using ARX metrics shows that all three generalization-based methods effectively reduced re-identification risks. ℓ-diversity at higher ℓ values and t-closeness at lower t values achieved the strongest reductions, producing uniformly low Highest Risk and Success Rate values across large datasets. k-anonymity, while effective, left higher residual risks compared to ℓ-diversity and t-closeness. Importantly, increasing dataset size amplified the risk reduction benefits, as larger cohorts yielded more uniform risk distributions. For differential privacy, ARX risk metrics were not applicable since no microdata was released; however, by design, DP provides strong theoretical guarantees of protection, particularly at lower ε values. Overall, t-closeness emerged as the most risk-averse techniques, while k-anonymity and l-diversity offered only moderate protection unless moderate and higher k and l values were applied.

6.4.4 Cross-Metric Trade-off Analysis

Synthesizing the results across execution time, utility, and risk reveals that no single technique dominates in all dimensions. k-anonymity (low k) provides the best balance of speed and utility but weaker privacy, making it suitable for scenarios prioritizing analytical accuracy and computational efficiency over strict confidentiality. ℓ-diversity (high ℓ) offers stronger privacy while retaining utility as well, positioning it as a balanced approach for sensitive datasets. t-closeness (low t) delivers the strongest privacy safeguards but at the cost of execution efficiency and data usability, making it best for high-stakes contexts where minimizing risk outweighs analytic needs. Differential privacy offers formal, tunable guarantees, but its effectiveness depends heavily on ε: high ε preserves utility but weakens privacy, while low ε strengthens privacy but diminishes analytical value. Hence, the most suitable method depends on institutional priorities—speed and utility (k-anonymity), privacy balance (ℓ-diversity), maximal confidentiality (t-closeness), or provable privacy in query-driven use (DP).

Table 6.9 *Cross-Metric Trade-off Analysis*

Technique	Execution Performance	Data Utility	Re-identification Risk	Suitable Contexts
k-Anonymity	★★★★★ (fast, especially at low k)	★★★★★ (at all three chosen values of k)	★★☆☆☆ (weaker at low k; improves with higher k)	Large-scale datasets needing speed and good utility; moderate privacy tolerance

ℓ-Diversity	★★★★☆ (slightly slower than k-anonymity)	★★★★★ (good at all three chosen values)	★★★★☆ (strong at higher ℓ)	Balanced privacy–utility needs in sensitive datasets
t-Closeness	★★☆☆☆ (slowest)	★★☆☆☆ (utility loss at stricter t)	★★★★★ (strongest privacy protection)	High-stakes scenarios prioritizing maximum confidentiality
Differential Privacy	★★★★★ (fastest)	Variable: ★★★★☆ at high ε; ★☆☆☆☆ at low ε	Theoretically ★★★★★	Interactive query systems requiring provable privacy

6.4.5 Summary of Findings

The overall comparison underscores the **multi-dimensional trade-offs** of privacy-preserving techniques. **k-anonymity** is the most practical where **performance and analytic fidelity** are paramount, though residual risks remain. **ℓ-diversity** provides a more robust **balance between utility and risk reduction**, particularly effective on large datasets. **t-closeness** excels in **risk minimization**, but its high computational burden and substantial utility loss limit its applicability in real-time or utility-driven contexts. **Differential privacy** stands apart by not releasing microdata and offering **theoretical privacy guarantees**, making it ideal for **interactive query systems**, but its utility strongly depends on ε. Collectively, the findings demonstrate that the "best" method is **context-dependent**: for **research requiring high accuracy and speed**, k-anonymity is preferable; for **sensitive data sharing with risk controls**, ℓ-diversity is optimal; for **maximum privacy assurance**, t-closeness is suitable; and for **provable privacy in dynamic environments**, differential privacy is the most appropriate choice. As an exploratory step, we also examined whether modifying the age brackets (e.g., 0–20, 20–40, and so forth, instead of the current choice) would lead to noticeable changes in the outcomes. The results, however, remained largely similar across these alternative groupings. We also decided not to expand the brackets further (e.g., 0–40, 40–80, 80–100), as such divisions would result in uneven intervals and provide limited analytical value for comparative purposes.

7 CONCLUSION, LIMITATIONS AND FUTURE WORK

Conclusion:

This study presented a comprehensive comparative analysis of four prominent privacy-preserving techniques—k-anonymity, ℓ-diversity, t-closeness, and differential privacy—applied to healthcare datasets of varying sizes. Through systematic experimentation, we assessed their performance across three key dimensions: **execution efficiency**, **data utility**, and **re-identification risk**. The results highlighted the trade-offs inherent to each method: k-anonymity and ℓ-diversity generally preserved higher utility but offered weaker protection under sophisticated attacks, t-closeness improved resistance to attribute disclosure at the cost of greater data loss, and differential privacy provided strong formal guarantees with noticeable utility degradation at lower ε values.

The multi-faceted evaluation underscores the importance of **context-driven selection** of privacy techniques. No single approach proved universally optimal; instead, the best choice depends on the sensitivity of the data, acceptable utility loss, computational constraints, and anticipated adversary capabilities. By combining empirical evidence with attacker-model-based risk assessment, this work offers a practical reference framework for privacy engineering in healthcare analytics.

Limitations:

While this study provides a comprehensive comparative analysis of four major privacy-preserving techniques—k-anonymity, ℓ-diversity, t-closeness, and differential privacy—applied to healthcare datasets of varying sizes, several limitations should be acknowledged:

1. **Dataset Scope and Representativeness**
   The evaluation was conducted on a specific healthcare dataset with defined quasi-identifiers and sensitive attributes. Results may not directly generalize to other healthcare datasets with different structures, attribute distributions, or privacy risks.

2. **Exclusion of Certain Risk Models**
   Re-identification risk evaluation relied solely on ARX's built-in attacker models (Prosecutor, Journalist, Marketer). Other advanced attack models—such as machine-learning-assisted re-identification or linkage attacks using external data—were not explored.

3. **Differential Privacy Constraints**
   For differential privacy, risk evaluation via ARX was not possible because anonymized datasets are not directly generated. As such, re-identification risk for differential privacy was discussed only conceptually, without empirical validation using a standardized risk analysis tool.

4. **Simplified Utility Measurement**
   Data utility evaluation relied on a set of predefined simple and complex aggregate queries. While these queries cover diverse analytical patterns, they do not represent the full range of possible statistical or machine learning tasks that could be applied to anonymized healthcare data.

5. **Computational Environment**
   All execution time measurements were performed in a single computing environment. Results might differ under different hardware specifications, parallelization strategies, or optimized implementations.

6. **Assumptions in Implementation**
   Some implementations, particularly in t-closeness and differential privacy, involved methodological simplifications (e.g., approximated EMD computation, chosen ground distances, variable ε allocations). These decisions, while justified, could influence comparability with results from studies adhering strictly to original theoretical definitions.

Future Work:

1. **Evaluation on Diverse and Multi-Source Datasets**
   Apply the comparative framework to datasets from multiple healthcare institutions, regions, and modalities (e.g., electronic health records, imaging, genomic data) to assess generalizability.
2. **Expansion of Parameter Ranges**
   Explore broader parameter ranges, especially for ε in differential privacy, t thresholds in t-closeness, and diversity values, to reveal finer-grained trade-offs.

3. **Advanced Risk Assessment Models**
   Integrate adversarial machine learning-based re-identification or linkage attacks combining anonymized datasets with external sources for more realistic privacy risk evaluations.

4. **Hybrid Privacy Models**
   Investigate combining multiple privacy techniques, such as applying differential privacy noise injection to k-anonymized datasets, to balance strengths and weaknesses.

5. **Task-Specific Utility Metrics**
   Complement aggregate query accuracy with downstream performance measures (e.g., predictive modeling, F1-score, survival analysis metrics) relevant to clinical applications.

6. **Differential Privacy Risk Quantification**
   Develop empirical or probabilistic models to estimate re-identification risk for differential privacy outputs where ARX-based tools are not applicable.

7. **Longitudinal and Dynamic Data**
   Extend the analysis to longitudinal datasets or continuously updated healthcare databases, addressing privacy preservation in real-time or evolving data contexts.

REFERENCES

[1] Charles, D., Gabriel, M., & Furukawa, M. F. (2014). Adoption of Electronic Health Record Systems among U.S. Non-federal Acute Care Hospitals: 2008-2013.

[2] Henry, J., Pylypchuk, Y., Searcy, T., & Patel, V. (n.d.). Data Brief 35: Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015.

[3] Sweeney, L. (2002). L. Sweeney. k-anonymity: a model for k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY 1. In International Journal on Uncertainty, Fuzziness and Knowledge-based Systems (Vol. 10, Issue 5).

[4] Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Sweeney.

[5] Golle, P. (2006). Revisiting the Uniqueness of Simple Demographics in the US Population.

[6] SUMMARY OF THE HIPAA PRIVACY RULE HIPAA Compliance Assistance O C R P R I V A C Y B R I E F i SUMMARY OF THE HIPAA PRIVACY RULE Contents. (n.d.). http://www.hhs.gov/ocr/hipaa.

[7] Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. Journal of the American Medical Informatics Association, 17(2), 169–177. https://doi.org/10.1136/jamia.2009.000026

[8] el Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J. P., Walker, M., Chowdhury, S., Vaillancourt, R., Roffey, T., & Bottomley, J. (2009). A Globally Optimal k-Anonymity Method for the De-Identification of Health Data. Journal of the American Medical Informatics Association, 16(5), 670–682. https://doi.org/10.1197/jamia.M3144

[9] Kifer, D., & Machanavajjhala, A. (n.d.). A Rigorous and Customizable Framework for Privacy.

[10] Fung, B. C. M., Wang, K., Fu, A. W., & Yu, P. S. (2010). Introduction to Privacy-Preserving Data Publishing Concepts and Techniques. Chapman & HallCRC Press. (n.d.).

[11] Meyerson, A., & Williams, R. (n.d.). On the Complexity of Optimal K-Anonymity.

[12] Adams, T., Birkenbihl, C., Otte, K., Ng, H. G., Rieling, J. A., Näher, A. F., Sax, U., Prasser, F., & Fröhlich, H. (2025). On the fidelity versus privacy and utility trade-off of synthetic patient data. *IScience*, *28*(5). https://doi.org/10.1016/j.isci.2025.112382

[13] el Emam, K., Dankar, F. K., Vaillancourt, R., Roffey, T., & Lysyk, M. (2009). Evaluating the risk of re-identification of patients from hospital prescription records.

Canadian Journal of Hospital Pharmacy, 62(4), 307–319. https://doi.org/10.4212/cjhp.v62i4.812

[14] LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2008). Workload-aware anonymization techniques for large-scale datasets. ACM Transactions on Database Systems, 33(3). https://doi.org/10.1145/1386118.1386123

[15] Vovk, O., Piho, G., & Ross, P. (2021). Anonymization methods of structured health care data: A literature review. In Model and Data Engineering (pp. 175-189).

[16] Bayardo, R. J., & Agrawal, R. (2005). Data privacy through optimal k-anonymization. In 21st International Conference on Data Engineering (pp. 217-228). IEEE. (n.d.).

[17] Samarati, P. (n.d.). Protecting Respondents' Identities in Microdata Release.

[18] el Emam, K., & Dankar, F. K. (2008). Protecting Privacy Using k-Anonymity. Journal of the American Medical Informatics Association, 15(5), 627–637. https://doi.org/10.1197/jamia.M2716

[19] Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (n.d.). ℓ-Diversity: Privacy Beyond k-Anonymity.

[20] Sepas, A., Bangash, A. H., Alraoui, O., el Emam, K., & El-Hussuna, A. (2022). Algorithms to anonymize structured medical and healthcare data: A systematic review. In Frontiers in Bioinformatics (Vol. 2). Frontiers Media SA. https://doi.org/10.3389/fbinf.2022.984807

[21] Li, N., Li, T., & Venkatasubramanian, S. (n.d.). t-Closeness: Privacy Beyond k-Anonymity and-Diversity.

[22] Xiao, X., & Tao, Y. (2007). M-invariance: towards privacy preserving re-publication of dynamic datasets. In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data (pp. 689-700). ACM.

[23] Halilovic, M., Meurers, T., Otte, K., & Prasser, F. (2025). Parallel privacy preservation through partitioning (P4): a scalable data anonymization algorithm for health data. BMC Medical Informatics and Decision Making, 25(1). https://doi.org/10.1186/s12911-025-02959-z

[24] P, J. A., & Thanamani, A. S. (2017). Comparison and Analysis of Anonymization Techniques for Preserving Privacy in Big Data. Advances in Computational Sciences and Technology, 10(2), 247. https://doi.org/10.37622/acst/10.2.2017.247-253

[25] Dwork, C. (n.d.). Differential Privacy.

[26] Dwork, C., & Roth, A. (2013). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–487. https://doi.org/10.1561/0400000042

[27] Nissim, K., Raskhodnikova, S., & Smith, A. (2011). Smooth Sensitivity and Sampling in Private Data Analysis *.

[28] Dankar, F. K., & Emam, K. el. (2013). Practicing Differential Privacy in Health Care: A Review. In TRANSACTIONS ON DATA PRIVACY (Vol. 5).

[29] Liu, W. K., Zhang, Y., Yang, H., & Meng, Q. (2024). A Survey on Differential Privacy for Medical Data Analysis. In Annals of Data Science (Vol. 11, Issue 2, pp. 733–747). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s40745-023-00475-3

[30] Dwork, C., Rothblum, G. N., & Vadhan, S. (2010). Boosting and differential privacy. Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS, 51–60. https://doi.org/10.1109/FOCS.2010.12

[31] Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Kulynych, B., Prasser, F., & Raisaro, J. L. (2025). A scoping review of privacy and utility metrics in medical synthetic data. Npj Digital Medicine, 8(1). https://doi.org/10.1038/s41746-024-01359-3

[32] Meurers, T., Otte, K., Abu Attieh, H., Briki, F., Despraz, J., Halilovic, M., Kaabachi, B., Milicevic, V., Müller, A., Papapostolou, G., Wirth, F. N., Raisaro, J. L., & Prasser, F. (2025). A quantitative analysis of the use of anonymization in biomedical research. Npj Digital Medicine, 8(1). https://doi.org/10.1038/s41746-025-01644-9

[33] Iyengar, V. S. (2002). Transforming data to satisfy privacy constraints. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[34] Prasser, F., Lautenschlaeger, R., Kohlmayer, F., Lautenschläger, R., & Kuhn, K. A. (2014). ARX - A Comprehensive Tool for Anonymizing Biomedical Data. https://www.researchgate.net/publication/276065720

[35] Mohammed, N., Fung, B. C. M., Hung, P. C. K., & Lee, C. K. (2009). Anonymizing healthcare data: A case study on the blood transfusion service. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1285–1293. https://doi.org/10.1145/1557019.1557157

[36] Pilgram L, Meurers T, Malin B, Schaeffner E, Eckardt K, Prasser F, GCKD Investigators The Costs of Anonymization: Case Study Using Clinical Data J Med Internet Res 2024;26:e49445 URL: https://www.jmir.org/2024/1/e49445 DOI: 10.2196/49445

[37] Ashkouti, F., & Khamforoosh, K. (2023). A distributed computing model for big data anonymization in the networks. *PLoS ONE*, *18*(4 April). https://doi.org/10.1371/journal.pone.0285212

[38] Im, E., Kim, H., Lee, H., Jiang, X., & Kim, J. H. (2024). Exploring the tradeoff between data privacy and utility with a clinical data analysis use case. *BMC Medical Informatics and Decision Making*, *24*(1). https://doi.org/10.1186/s12911-024-02545-9

[39] Ganta, S. R., Kasiviswanathan, S. P., & Smith, A. (2008). Composition Attacks and Auxiliary Information in Data Privacy. http://arxiv.org/abs/0803.0032

[40] Xiao, X., & Tao, Y. (2006). Anatomy: Simple and Effective Privacy Preservation.

[41] Kohlmayer, F., Prasser, F., & Kuhn, K. A. (2015). The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss. Journal of Biomedical Informatics, 58, 37–48. https://doi.org/10.1016/j.jbi.2015.09.007

[42] Skinner, C. J., & Elliot, M. J. (2002). A measure of disclosure risk for microdata. In J. R. Statist. Soc. B (Vol. 64).

[43] Dankar, F. K., el Emam, K., Neisa, A., & Roffey, T. (2012). Estimating the re-identification risk of clinical data sets. BMC Medical Informatics and Decision Making, 12(1). https://doi.org/10.1186/1472-6947-12-66

[44] Loukides, G., Denny, J. C., & Malin, B. (2010). The disclosure of diagnosis codes can breach research participants' privacy. Journal of the American Medical Informatics Association, 17(3), 322–327. https://doi.org/10.1136/jamia.2009.002725

[45] Ciriani, V., de Capitani Di Vimercati, S., Foresti, S., & Samarati, P. (n.d.). K-ANONYMOUS DATA MINING: A SURVEY. http://www.springerlink.com/content/gt152012204u3538/fulltext.pdf

[46] Strack, B., Deshazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. BioMed Research International, 2014. https://doi.org/10.1155/2014/781670

[47] Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys, 42(4). https://doi.org/10.1145/1749603.1749605

[48] Samarati, P., & Sweeney, L. (n.d.). Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression.

[49] Emam, K. el. (n.d.). Guide to the De-Identification of Personal Health Information.

[50] Marius Truta, T., & Vinay, B. (2006). Privacy Protection: p-Sensitive k-Anonymity Property.

[51] Rubner, Y., Tomasi, C., & Guibas, L. J. (n.d.). The Earth Mover's Distance as a Metric for Image Retrieval.

[52] Brickell, J., & Shmatikov, V. (2008). The cost of privacy: destruction of data-mining utility in anonymized data publishing. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 70-78). ACM.

[53] McSherry, F., & Talwar, K. (2008). Mechanism Design via Differential Privacy. 94–103. https://doi.org/10.1109/focs.2007.66

[54] Dwork, C., Mcsherry, F., Nissim, K., & Smith, A. (n.d.). Calibrating Noise to Sensitivity in Private Data Analysis.

[55] Sarathy, R., & Muralidhar, K. (2011). Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data. In TRANSACTIONS ON DATA PRIVACY (Vol. 4).

[56] Blum, A., Dwork, C., Mcsherry, F., & Nissim, K. (2005). Practical Privacy: The SuLQ Framework.

[57] Cormode, G., Procopiuc, C. M., Shen, E., Srivastava, D., & Yu, T. (2013). Empirical privacy and empirical utility of anonymized data. *Proceedings - International Conference on Data Engineering*, 77–82. https://doi.org/10.1109/ICDEW.2013.6547431

[58] Tamersoy, A., Loukides, G., Nergiz, M. E., Saygin, Y., & Malin, B. (2012). Anonymization of longitudinal electronic medical records. IEEE Transactions on Information Technology in Biomedicine, 16(3), 413–423. https://doi.org/10.1109/TITB.2012.2185850

[59] Loukides, G., Liagouris, J., Gkoulalas-Divanis, A., & Terrovitis, M. (2014). Disassociation for electronic health record privacy. Journal of Biomedical Informatics, 50, 46–61. https://doi.org/10.1016/j.jbi.2014.05.009

[60] Loukides, G., Gkoulalas-Divanis, A., & Malin, B. (2010). Anonymization of electronic medical records for validating genome-wide association studies. Proceedings of the National Academy of Sciences of the United States of America, 107(17), 7898–7903. https://doi.org/10.1073/pnas.0911686107

[61] Gardner, J., Xiong, L., Xiao, Y., Gao, J., Post, A. R., Jiang, X., & Ohno-Machado, L. (2013). Share: System design and case studies for statistical health information release. Journal of the American Medical Informatics Association, 20(1), 109–116. https://doi.org/10.1136/amiajnl-2012-001032

[62] Guide to the Labour Force Survey 2017. (n.d.). www.statcan.gc.ca

[63] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte, E., Spicer, N. K., de Wolf, P.-P., Domingo-Ferrer, H., Giessing, F., Spicer De Wolf, N., Nordholt, E. S., & Spicer, K. (n.d.). Statistical Disclosure Control Statistical Disclosure Control WILEY SERIES IN SURVEY METHODOLOGY.

[64] Kniola, L. (n.d.). PhUSE 2017 Plausible Adversaries in Re-Identification Risk Assessment. www.patientslikeme.com

[65] Xia, W., Liu, Y., Wan, Z., Vorobeychik, Y., Kantacioglu, M., Nyemba, S., Clayton, E. W., & Malin, B. A. (2021). Enabling realistic health data re-identification risk assessment through adversarial modeling. Journal of the American Medical Informatics Association : JAMIA, 28(4), 744–752. https://doi.org/10.1093/jamia/ocaa327

[66] De-identification 301. (n.d.). https://privacy-analytics.com/resources/white-papers/de-identification-301-understanding-your-likely-attackers/

[67] Zayatz, L. V. (n.d.). BUREAU OF THE CENSUS STATISTICAL RESEARCH DIVISION REPORT SERIES SRD Research Report Number: CENSUS/SRD/RR-91/08 ESTIMATION OF THE PERCENT OF UNIQUE POPULATION ELEMENTS ON A MICRODATA FILE USING THE SAMPLE.

[68] Rocher, L., Hendrickx, J. M., & de Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications, 10(1). https://doi.org/10.1038/s41467-019-10933-3

[69] HIPAA Journal. (2025). De-identification of protected health information: 2025 update. Retrieved from https://www.hipaajournal.com/de-identification-protected-health-information/#:~:text=The%20de%2Didentification%20of%20Protected,of%20the%20HIPAA%20Privacy%20Rule.

[70] Xia, W., Basford, M., Carroll, R., Clayton, E. W., Harris, P., Kantacioglu, M., Liu, Y., Nyemba, S., Vorobeychik, Y., Wan, Z., & Malin, B. A. (2023). Managing re-identification risks while providing access to the All of Us research program. Journal of the American Medical Informatics Association, 30(5), 907–914. https://doi.org/10.1093/jamia/ocad021

[71] Risk analysis documentation. Retrieved from https://arx.deidentifier.org/anonymization-tool/risk-analysis/

[72] Privacy models documentation. Retrieved from https://arx.deidentifier.org/overview/privacy-criteria/

[73] Clore, J., Cios, K., DeShazo, J., & Strack, B. (2014). Diabetes 130-US Hospitals for Years 1999-2008 [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5230J.

[74] Ocr. (2012). Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

APPENDICES

This appendix provides rows dropped after implementation of k-anonymity, l-diversity and t-closeness. These tables have been extracted from summary CSVs.

Table 7.1 *Number of Records dropped after k-anonymity implementation*

dataset	k_value	original_records	final_records	records_retained_pct	records_dropped
25k	2	24951	24951	100	0
25k	5	24951	24940	99.95591	11
25k	10	24951	24928	99.90782	23
50k	2	49942	49942	100	0
50k	5	49942	49936	99.98799	6
50k	10	49942	49924	99.96396	18
75k	2	74942	74942	100	0
75k	5	74942	74940	99.99733	2
75k	10	74942	74935	99.99066	7
100k	2	101742	101742	100	0
100k	5	101742	101740	99.99803	2
100k	10	101742	101732	99.99017	10

Table 7.2 *Number of Records dropped after l-diversity implementation*

dataset	l_value	original_records	final_records	records_retained_pct	records_dropped
25k	2	24951	24948	99.98798	3
25k	3	24951	24928	99.90782	23
25k	4	24951	24928	99.90782	23
50k	2	49942	49938	99.99199	4
50k	3	49942	49936	99.98799	6
50k	4	49942	49910	99.93593	32
75k	2	74942	74942	100	0
75k	3	74942	74935	99.99066	7
75k	4	74942	74925	99.97732	17
100k	2	101742	101742	100	0
100k	3	101742	101740	99.99803	2
100k	4	101742	101716	99.97445	26

Table 7.3 *Number of Records dropped after t-closeness implementation*

dataset	t_value	original_records	final_records	records_retained_pct	records_dropped
25k	0.4	24951	23963	96.04024	988
25k	0.7	24951	24745	99.17438	206
25k	1	24951	24938	99.9479	13
50k	0.4	49942	47959	96.02939	1983
50k	0.7	49942	49122	98.3581	820
50k	1	49942	49920	99.95595	22
75k	0.4	74942	71555	95.4805	3387
75k	0.7	74942	73211	97.69021	1731
75k	1	74942	74908	99.95463	34
100k	0.4	101742	98321	96.63757	3421
100k	0.7	101742	99322	97.62143	2420
100k	1	101742	101701	99.9597	41