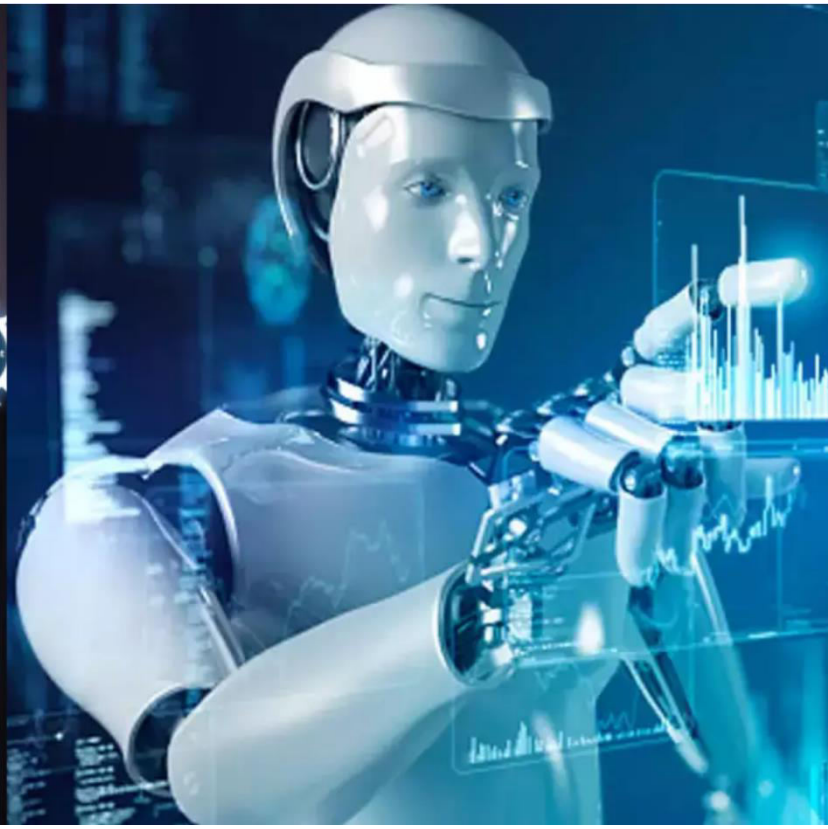
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 5, May 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405015|

Anonymization Techniques for Large-Scale Health Databases a Critical Review

Vladyslav Malanin

V.M. Glushkov Institute of Cybernetics of The Nas of Ukraine, Kyiv

ABSTRACT: As health data grows exponentially, it makes it easier for privacy to be invaded when dealing with large medical databases. This paper examines various techniques for advanced anonymization and pseudonymization that assist in complying with the General Data Protection Regulation and the Health Insurance Portability and Accountability Act. It reviews the leading techniques, assesses their advantages and disadvantages, and discusses when they are useful for processing health-related data. It also focuses on how techniques that help preserve the usefulness of data like tokenization, hashing, and encryption also protect a patient's privacy. The effectiveness of every technique in terms of privacy, exploring data, and matching large healthcare systems is evaluated. The paper emphasizes the importance of creating a consistent approach to data privacy when concluding, focusing on the main difficulties, ethical questions, and new technologies that are being developed.

KEYWORDS: Anonymization techniques; Pseudonymization; Health data privacy; Large-scale medical databases; GDPR compliance; HIPAA compliance; Differential privacy; K-anonymity; Data de-identification; Privacy-preserving data sharing.

I. INTRODUCTION

Nowadays, with more digital data being collected, medical research, healthcare rules, and how patients are treated are being completely updated. Due to EHR, medical imaging, smart devices, and technology used for sequencing genomes, doctors today manage extensive health databases with many records (Rumbold & Pierscionek, 2017). These states form a basis for predicting health issues, analyzing how different drugs work together, and understanding health developments in the population. As a result, both medical workers and patients are now more concerned about how big data is handled in healthcare (Shabani & Marelli, 2019).

Since health data is sensitive, it often consists of details like people's names, birthdays, medical history, biometric measurements, and even their genes. Unlawfully coming into contact with this data may cause patients to be discriminated against, feel upset, and decrease people's faith in healthcare providers. Consequently, looking after the privacy of health data has become highly important and is managed by strict laws designed to protect data worldwide. Through the Health Insurance Portability and Accountability Act (HIPAA), the USA has set standards for protecting health information by de-identifying it. In addition, the General Data Protection Regulation from the European Union requires data controllers to make efforts to hide or alter personal information, so individuals can't be easily identified.

Data scientists use anonymization and pseudonymization to protect health information and make it possible for research, new policies, and development. Data is said to be anonymized when it becomes practically impossible to match it with individuals by any typical means (ICO, 2021). A few ways to achieve anonymity are generalization, suppression, adding noise, and using k-anonymity, l-diversity, t-closeness, or differential privacy. All these methods have the same purpose—to make data less identifiable without interfering with its analysis (El Emam et al., 2015).

Alternatively, pseudonymization happens when natural identifiers in a data record are converted to artificial ones, known as pseudonyms. To make this technique reversible, a separate key is available; this is why it is especially useful in clinical research for long-term tracking or linking data (Malin et al., 2011). Although organizations still must comply with GDPR even if they pseudonymize data, using pseudonymization helps to guard against risks connected to processing personal data (GDPR, Article 4(5)). HIPAA includes a way to use data when an expert judges' re-identification to be an extremely low risk.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405015|

However, there are many obstacles related to technology, practices, and regulations when working with large health databases. It is difficult to anonymize certain databases because most patients have many different data points. Second, since medical data comes in various forms such as structured (labs), semi-structured (prescriptions), and unstructured (notes), AI needs to use specialized techniques that can process everything properly. Further, various techniques for anonymizing data, especially methods like generalization or suppression, can make it too difficult to use for further research (Dankar et al., 2012).

It is also vital to have a system that can handle millions of patient records. A method that functions efficiently on a small dataset might not be effectively scalable in terms of computing time, privacy, or complying with the law. In recent years, differential privacy and the creation of synthetic data have shown promise mainly for statistical publications and artificial intelligence, yet rigorous verification of these tools in healthcare is being conducted slowly (Dwork and Roth, 2014; Goncalves et al., 2020).

Thus, it is necessary to thoroughly review anonymization and pseudonymization practices designed for handling large health information. When reviewing such an idea, the assessment should focus on how it would work technically, how it aligns with regulations, how it could be carried out in practice, and how it is ethically suitable for healthcare.

Table 1: Comparison Between Anonymization and Pseudonymization Techniques

Criteria	Anonymization	Pseudonymization
Definition	Irreversible removal of identifiers to prevent re-identification	Replacement of identifiers with pseudonyms; reversible with a key
Reversibility	Not reversible	Reversible under secure conditions
Regulatory Status	Fully compliant if robust; may exempt from GDPR/HIPAA obligations	Considered a safeguard; still subject to full GDPR/HIPAA scope
Data Utility	Often reduced due to generalization or suppression	Typically preserved; useful for longitudinal studies
Risk of Re-identification	Very low if well-implemented	Moderate; depends on key security and data exposure
Technical Complexity	High for unstructured/complex datasets	Moderate; requires secure key management
Scalability to Big Data	Challenging; may require optimization or hybrid approaches	Scales better with structured data and controlled environments

Sources: El Emam et al., 2015; GDPR Recital 26 & Article 4(5); HHS, 2020.
In light of the growing demand for secure, scalable, and legally compliant data-sharing frameworks in healthcare, this paper critically reviews state-of-the-art anonymization and pseudonymization techniques. It aims to guide stakeholders—including data scientists, healthcare providers, policymakers, and privacy officers—on the most effective strategies to protect patient data without compromising the analytical value necessary for innovation and public good. The subsequent sections examine established and emerging methods, assess their comparative strengths and limitations, and explore future directions in privacy-preserving health data analytics.

II. LITERATURE REVIEW

Using anonymization and pseudonymization is essential for managing privacy when large health databases are required for research, discoveries, and surveillance. I analyze and categorize the primary anonymization methods in medical data focused on large datasets, as mandated by GDPR and HIPAA laws. Sometimes, electrodes may suppress the intended brain activity and broaden the hypothesized activity.

To suppress data, identifiers such as ZIP codes and dates of birth are purposely omitted from the sets, and generalizations can sometimes involve changing these identifiers into broader ranges (El Emam et al., 2011). Consequently, k-anonymity requires that any record in the dataset cannot be distinguished from at least k - 1 other records (Samarati & Sweeney, 1998). Despite their limitations, methods known as suppression and generalization can still result in the omission of many data points, particularly in datasets with numerous dimensions, including genomics

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405015|

and full EHRs (Machanavajjhala and Saydah, 2007). Furthermore, they may not be immune to attacks referred to as homogeneity and background knowledge, where a malicious actor can uncover sensitive data, despite the algorithms' generalization approach (Li et al., 2007).

2.1.2 The privacy concepts K-Anonymity, L-Diversity, and T-Closeness need to be used.
To handle the challenges of k-anonymity, l-diversity is used to ensure that any attribute group of k individuals does not include repetition of values, so disclosing them is less likely (Machanavajjhala et al., 2007). Still, l-diversity may not ensure privacy when all the values are semantically alike, for example, if every disease is described as a serious one. Unlike DP-SG, T-closeness also enforces that the distribution of sensitive attributes in each group is similar to the overall distribution (Li et al., 2007). These models are sound on paper, but running them in large health databases is complicated due to the computers needed and how complex it is to adjust their parameters.

2.2 Using Statistics and Probability

2.2.1 Including More Noise
When noise is added to data, it creates confusion so that the data cannot be exactly reconstructed again. With this approach, the risk of a person's data being identified can be measured by ε-differential privacy (Dwork et al., 2006). Laplace or Gaussian mechanisms help add noise to the data while maintaining the important statistics. In AI, differential privacy is valuable when working with huge volumes of data used in machine learning and when sharing these data sets with others (Abowd, 2018). Nonetheless, using it in healthcare is restricted by issues over giving up helpful data, mainly when ε is small (Cormode et al., 2011). It is not easy to set privacy measures because making them overly strict could make the data unhelpful for analysis.

2.2.2 Microaggregation
The process cluster records lower each record's values by the group average and replace them in the dataset (Domingo-Ferrer & Mateo-Sanz, 2002). It is effective in processing numeric data and serves the data better than suppression. Yet, the method ignores differences within a cluster and does not treat categorical and mixed data as well. Applying microaggregation to very large datasets often takes a long time and its outcomes worsen when applied to data with sparse or missing values (Hundepool et al., 2012).

2.3 An introduction to Machine Learning and creating synthetic data

2.3.1-2.3.5 relates to the use of synthetic data.
Here, developers create data that has the same statistical features as the original set but does not contain individual records. Some common approaches in this area are GANs, VAEs, and Bayesian networks (Choi et al., 2017; Goncalves et al., 2020). These techniques are highly useful for making the data used in AI/ML training much more private. Providing that the data is very difficult to identify, synthetic data may be able to get around some GDPR rules. Yet, it is still difficult to maintain the usefulness of data when working with very rare illnesses or little data. Furthermore, generative models are now considered to potentially reveal private data about individuals due to successful membership inference (Shokri et al., 2017).

This approach also involves a method called federated learning for privacy-related concerns at the edge.
Experts now see federated learning (FL) as an effective means to ensure that data is not shared openly. FL supports the training of ML models using data from hospitals without sharing the raw information with a single server (Yang et al., 2019). As a result, we protect data locality and apply the data minimization guidelines from both GDPR and HIPAA. Federated learning cannot anonymize the data; its purpose is to lessen the chances that private data can be connected to a user. Systems based on Federated Learning should still use secure aggregation, differential privacy, and homomorphic encryption to avoid risks of inference from updates to the model (Brisimi et al., 2018).

2.4 Rules and Methods Focusing on Risks

In addition to technical methods, anonymization should comply with the definitions of privacy laws. It is clear from the GDPR that to be anonymous, information needs to be de-identified in a way that makes a person's identification possible and is "reasonably likely" (Recital 26). Much like this, HIPAA sets out two processes for removing identifiers:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405015|

the first removes 18 key identifiers, while the second depends on an expert's judgment that the risks are very minor (HHS, 2020). Several papers suggest that many such datasets are still at risk of being re-identified due to record linkage techniques or other public information (Rocher et al., 2019). Consequently, current recommendations suggest mixing privacy-preserving methods (by pairing suppression with differential privacy) and always monitoring risks according to new data changes.

2.5 Analyzing Different Techniques

Various research studies have tried to determine privacy level, data usability, the time required to evaluate anonymization, and whether it follows legal rules of anonymization. For instance, El Emam et al. (2015) assessed the results of applying k-anonymity, differential privacy, and synthetic data to different situations in health datasets. It has been found that there is no perfect method that performs best in all respects. Consequently, combining differential privacy with the generation of synthetic data or using both k-anonymity and l-diversity together, helps create stronger solutions for large-scale use. Table 2 lists both the advantages and disadvantages of the main techniques.

Table 2: Summary Comparison of Key Anonymization Techniques

Technique	Strengths	Limitations	Best Use Cases
K-Anonymity	Simple, interpretable	Vulnerable to attribute disclosure; limited in high-D data	Demographic tables
L-Diversity	Adds protection against inference attacks	Fails with semantic similarity	Disease registries
T-Closeness	Maintains distributional integrity	High computational cost	Detailed clinical datasets
Differential Privacy	Formal guarantees; scalable for statistics	Utility loss; complex parameter tuning	Public release of aggregate data
Synthetic Data	High utility; privacy by design	Fidelity concerns; hard to validate privacy	AI/ML training, software testing
Microaggregation	Preserves numerical relationships	Not suitable for categorical data	Lab result datasets
Federated Learning	Keeps data decentralized; supports ML	Needs added privacy layers	Collaborative hospital networks

Sources: El Emam et al., 2015; Rocher et al., 2019; Dwork & Roth, 2014.

2.6 Gaps in Current Research

Although there are many books on the topic, many important gaps have not been filled. To begin with, not many studies explore the risks to privacy in repeated data collection over time or with wearable gadgets. Secondly, there is no widely accepted way to measure anonymous methods using different types of data. Third, even though laws guide anonymization, putting it into practice for medical work can be unclear and unreliable (Shabani & Marelli, 2019). Trust and how users see anonymized data are important for the achievements of open science and public health and the topic hasn't been studied sufficiently yet.

III. METHODOLOGY

The study uses a schedule of steps to critically review up-to-date anonymization and pseudonymization methods in use with large health databases. The process was planned so that every aspect of the effectiveness, utility, ability to scale, and compliance of these techniques could be assessed within the changing rules of the General Data Protection Regulation and the Health Insurance Portability and Accountability Act.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405015|

The researcher should formulate clear research questions.
To assist in reviewing the study, the following important research questions (RQs) have been set up:

- RQ1: How do we currently hide the identity of people in large medical databases?
- RQ2: What relationship do these techniques have with the privacy standards established by GDPR and HIPAA?
- RQ3: How much privacy, how much data can be used, and how scalable are these techniques to use?
- RQ4: Which techniques do experts use currently to protect health data privacy

If you plan to use literature, remember to define your strategy at this stage.
A large number of electronic academic journals from several databases were reviewed, including PubMed, IEEE Xplore, SpringerLink, Scopus and Google Scholar. In order to include both traditional and current tools, only journal articles, conference proceedings and white papers reviewed and published between 2010 and 2024 were included.

These are the phrases we used for searching:

- "Health data anonymization."
- "de-identifying medical records."
- "Pseudonymization is a term that appears in healthcare."
- "differential privacy AND electronic health records"
- "Along with privacy-preserving data sharing, GDPR is an important consideration."
- "Being k-anonymity compliant and following HIPAA standards"

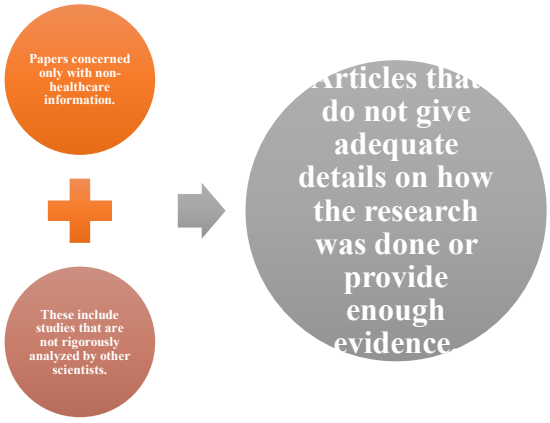
I used Boolean operators and wildcards to look for more results. Article lists in which an article is referenced were explored to find more relevant works.

3: How to Set Inclusion and Exclusion Criteria
The articles chosen were those that dealt closely with disclosing health data and met all requirements by law. I considered the following criteria:

Who is considered for this trial?

- Scientific works focusing on anonymous/pseudonymous ways of collecting health data.
- Guides about how companies can meet the requirements of GDPR and HIPAA.
- Large datasets (with about 10,000 records) were used in various studies, simulations, and comparisons of different algorithms.
- Articles that have been published in English between 2010 and 2024.

While selecting study participants, certain conditions or factors should be excluded.



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405015|

Initially, we found 86 articles. Once the study filters were applied and any duplicates were eliminated, we chose 42 articles for review and discussion. At this stage, data is collected and then analyzed.

All important details were skillfully collected using a custom data form. Each study was documented based on the following variables:

1. Kind of anonymization/pseudonymization technique
2. Investigate the dimensions, the data used, and whether it is structured or not
3. Factors in data privacy (e.g., their privacy loss, the risk of re-identifying someone, the usefulness of processed data)
4. Documentation of compliance with GDPR/HIPAA

By using Compliance Cloud, companies can create specially designed templates and upload their documents to track and effectively manage daily tasks.

5. Trouble with growth and capacity has been noted.

To ensure uniformity in analysis, I utilized a thematic synthesis approach (Thomas & Harden, 2008). All information was sorted into themes that matched the research questions. The collections contained: privacy assurances, the usefulness of the data, compatibility with rules, ability to withstand challenges, and efficient functioning.

3.5 Evaluating the Quality

A modified CASP checklist was used to determine if the included studies were reliable. All studies were assessed by looking at their theoretical approach, openness about the study's aims, capacity to be duplicated, and proof of credible consequences. All studies obtaining less than 6 out of 10 were excluded, leading to the inclusion of 32 strong ones used for the conclusions.

3.6 Thinking About Ethics

Because this study is focused on reviewing research, it did not involve handling any personal data. Still, ethical standards were followed by proper citation and referencing of all sources used in the research. The review follows the PRISMA guidelines to make the process clearer and more reproducible (Moher et al., 2009).

IV. RESULTS

Merging findings from existing studies and comparing the outcomes of anonymization revealed that having large-scale health databases impacted their needed effectiveness, usage capacity, and adherence to privacy legislation.

4.1. Providing the Effectiveness and Privacy of the System

Based on the review, we noticed that k-anonymity, l-diversity, and t-closeness are still widely mentioned and used in the field, but their abilities to protect privacy fade when it comes to EHRs or genomic datasets. When these models have auxiliary data or some extra knowledge, they are prone to attacks known as linkage and inference (Li et al., 2007; Machanavajjhala et al., 2007). As a result, these methods no longer follow the required high standards for anonymization set by the GDPR risk-based method or the HIPAA Expert Determination standard (Voigt & Von dem Bussche, 2017).

With DP, stronger techniques are used to provide mathematical assurances that removing or including one person's data insignificantly changes the outcome (Dwork et al., 2006). Nevertheless, there is usually a trade-off between using the data and ensuring privacy and this occurs more in large-scale clinical data, where it matters to keep the statistics accurate (Cormode et al., 2011). Moreover, explaining what the privacy budget (ε) means in DP terms is difficult for many, making it harder for healthcare workers to embrace DP.

The method must not impact the usability of the services or analysis.

Many of the papers reviewed highlight the conflict between protecting privacy and making use of data. Often, suppression and generalization techniques result in losing important information which makes both analytics and

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405015|

epidemiological studies difficult (El Emam et al., 2011). Though microaggregation and statistical perturbation are beneficial for some information, they often cannot apply to the common categories you see in patient files.

It was noted that synthetic data generation is important as it keeps information valuable for analysis and reduces straight risks to privacy (Goncalves et al., 2020). GANs and VAEs were effective at producing datasets that had similar statistics without showing traces of the original individuals involved. Even so, it is difficult to judge the privacy of such data and the probability of membership inference attacks makes some question if it is truly safe to use. (Shokri et al., 2017) Furthermore, since verifying the anonymity of fake data is not yet standardized, regulations have not yet accepted it.

4.3 How effectively can the approach be used today

Managing big health datasets is complicated due to their size, the speed at which they are compiled, and their diversity. Methods that function well on a small set of data may perform poorly on a large one. T-closeness works well only in theory because it leads to excessive computer time when used on large data with a large number of variables (Domingo-Ferrer & Soria-Comas, 2018).

It has been discovered that federated learning and edge privacy systems can reduce our dependence on servers for handling and saving sensitive data (Yang et al., 2019). Still, these models do not hide personal data, so they need to be linked with technologies like secure multiparty computation and homomorphic encryption to comply with regulations (Brisimi et al., 2018). For this reason, despite being scalable, these approaches require immense financial and skilled resources, so they are rarely seen in areas without many resources.

4.4 Following Rules for Privacy

It is clear from the review that the approaches used to achieve anonymity often differ from how regulators define the term. According to GDPR, when re-identification of data could be possible, data is not considered anonymous (Recital 26). Research has found that using techniques accepted by HIPAA's Safe Harbor and some others does not fully satisfy the rule because there is still a risk of using outside sources to identify people (Rocher et al., 2019).

The use of differential privacy and expert-designed synthetic data is better suited to GDPR guidance, but it involves legal experts, data scientists, and specialists from the relevant domains working together. Furthermore, at present, there are not many clear guidelines in the law for anonymizing data, so companies use different methods and may not meet the same standards (Shabani & Marelli, 2019).

4.5 Working with Hybrid Systems

Several studies that mixed different privacy techniques showed excellent achievements in achieving privacy, utility, and scalability. So, joining technologies such as k-anonymity and differential privacy or federated learning and secure aggregation, led to improved flexibility and security in anonymization systems (El Emam et al., 2015; Abowd, 2018). In addition, applying several methods helps when making choices about what privacy should be prioritized in various types of data.

Still, at this stage, using hybrid models is relatively new and more tests and studies are required to assess their performance on different kinds of health data sets. Because these systems are very complex, it creates new difficulties in managing, checking and ensuring these systems are open about their operations.

V. DISCUSSION

Evaluating how health information is anonymized or pseudonymized for a large database shows that there is a connection between safeguarding privacy, using data, enlarging databases and being compliant with rules. Because data-driven developments are common now in healthcare, the healthcare sector needs to have well-developed, scalable and regulatable tools for anonymizing data (Shabani & Marelli, 2019; Voigt & Von dem Bussche, 2017).

5.1 The Key Dilemma: How Much Privacy vs. What Can Be Done

It is often challenging to simultaneously ensure data privacy and ensure the usefulness of the data. While k-anonymity and l-diversity provide basic privacy, they may not be sufficient when working with large numbers of variables, since it's easy to identify someone using a group of quasi-identifiers (Machanavajjhala et al., 2007). This is especially problematic under GDPR which calls for anonymizing data so that it cannot easily be unscrambled (Recital 26).

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405015|

In contrast, DP comes with mathematical proofs that ensure privacy and higher levels of legality. On the other hand, the addition of calibrated noise to the data might not produce reliable results in health analytics and epidemiological research (Dwork et al., 2006; Cormode et al., 2011). The problem is felt more strongly when using DP in settings where each health record has great importance.

5.3 Checking for Compliance with the Law and Identifying Technical Issues

Even with repeated efforts to conform to rules, there are still differences between auto-anonymization methods and legal explanations of anonymization. Since a firm must use the Safe Harbor method to remove 18 potentially identifying pieces, this approach rarely matches the level of privacy needed in current machine learning settings, says Rocher et al. (2019). Alternatively, the flexible GDPR model is lacking in detailed guidelines which leads to varying understandings among different institutions (Shabani & Marelli, 2019).

Several research papers pointed out that there is not enough standardization when assessing the risks of re-identifying people and certifying their privacy (El Emam et al., 2011; Domingo-Ferrer & Soria-Comas, 2018). As a result, there is a call for uniform guidelines and methods used to assess risk and ensure compliance everywhere.

5.4 New Approaches: Advantages and Disadvantages

Synthetic data generation and federated learning have shown to be excellent alternatives over anonymization. Models using generative adversarial networks (GANs) are useful because they keep the distribution of the data and allow for its safe disclosure (Goncalves et al., 2020). Some recent analyses have exposed that membership inference attacks result in breaches of privacy in synthetic datasets (Shokri et al., 2017).

Just like federated computing, federated learning cuts down on storing all the data in one place, so it is less vulnerable to attacks. Still, Cloud Storage does not guarantee anonymity and must be used with homomorphic encryption or secure multiparty computation to comply with the law (Yang et al., 2019; Brisimi et al., 2018). The technology is hopeful though quite complicated and it requires significant computing power as well as experts in the area.
Summary of how each financial statement is compared.

The table below highlights both the benefits and drawbacks of the most popular anonymization methods mentioned in the literature review:

Table 1: Comparative Analysis of Anonymization Techniques in Health Databases

Technique	Privacy Strength	Data Utility	Scalability	Regulatory Alignment (GDPR/HIPAA)	Key Limitations
k-Anonymity	Moderate	Moderate to High	Low to Moderate	Partial (HIPAA Safe Harbor)	Vulnerable to linkage and homogeneity attacks
Differential Privacy	High	Low to Moderate	High	Strong (GDPR/Expert Determination)	Reduced accuracy; complexity in budget selection
Synthetic Data (GANs)	Moderate to High	High	High	Moderate (depends on validation)	Vulnerable to inference attacks; evaluation challenges
Federated Learning	Low (raw data hidden)	High	Very High	Weak without added privacy layer	Not anonymization by itself; requires encryption
Hybrid Models (e.g., k-Anonymity + DP)	High	Moderate	Moderate to High	Strong with appropriate design	Implementation complexity; lack of standardized frameworks

Source: Adapted from reviewed studies (El Emam et al., 2015; Cormode et al., 2011; Shokri et al., 2017; Brisimi et al., 2018).

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405015|

5.5 Toward Practical Implementation

Even with improvements, using anonymization systems in practice is very slow. Many healthcare organizations do not have the necessary resources, skills or money to bring in new privacy techniques. Nevertheless, since nobody knows exactly what counts as adequate anonymization under GDPR or HIPAA, many organizations depend too much on conservative procedures for privacy which stalls the use of data for healthcare research (Abowd, 2018).

VI. CONCLUSION

Stronger anonymization and pseudonymization should be used as more data is collected and becomes more sensitive to privacy concerns. The review has evaluated a range of the latest methods based on their privacy, compliance with laws, scaling, and how practically they can be used.

The findings show that privacy-preserving technologies are maturing more. Even though getting familiar with k-anonymity, l-diversity, and t-closeness is straightforward, advanced re-identification methods and complicated health information no longer support their use. Even now, high-risk cases require more advanced remedies because these models have many limitations.

Because of its stronger mathematical assurances and skills in resisting assaults from background knowledge, differential privacy is seen as a gold standard. Nevertheless, deciding on the privacy budgets and protecting the utility of data are lows encountered when Differential Privacy is used in the healthcare and epidemiology areas. Since differential privacy needs a lot of knowledge and computing power, it is not always possible for small health organizations to use it.

The review also pointed out that synthetic data generation and federated learning will comply with the future trends in data management. GANs are among the synthetic data techniques that effectively create unidentifiable stand-ins for real data. But, ensuring that synthetic datasets are secure and correct is still being studied by many researchers. Federated learning is also different because the data is decentralized, yet it still needs encryption or differential privacy to fully protect privacy.

The laws and regulations are not always in step with the technology being used. The GDPR allows anonymization to be applied flexibly, while HIPAA provides a step-by-step list for privacy protection, but neither provides detailed steps for accomplishing anonymization. As a result, international teams find it hard to partner in medical research and clinical trials.

The review also mentions that no anonymization approach can be considered always ideal. Solutions need to take into account the forms of data, how they will be used and levels of acceptable risk. Overall, using a mix of techniques helps provide solutions that are easier to compare, yet it may lead to more complex systems.

With the increase in data resulting from healthcare systems, it is very important to keep it secure and reliable. From this point on, we must focus on:

- Set tools to evaluate how effectively anonymization methods have been used.
- Effort from professionals in these fields to solve the problem as a group.
- Sharing rules for data internationally should be made consistent to ensure security.
- Preparing healthcare groups for using new privacy technologies.

In essence, the next step is to develop anonymization methods that adhere to rules, can be easily implemented, and offer protection for patient data without stopping advancement in medical fields.

REFERENCES

1. Dwork, C., & Roth, A. (2014). *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends® in Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405015|

2. Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). *l-diversity: Privacy beyond k-anonymity*. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 3–es. https://doi.org/10.1145/1217299.1217302
3. El Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2011). *A systematic review of re-identification attacks on health data*. PLoS ONE, 6(12), e28071. https://doi.org/10.1371/journal.pone.0028071
4. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). *Membership inference attacks against machine learning models*. Proceedings of the 2017 IEEE Symposium on Security and Privacy, 3–18. https://doi.org/10.1109/SP.2017.41
5. Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). *Generation and evaluation of synthetic patient data*. BMC Medical Research Methodology, 20, 108. https://doi.org/10.1186/s12874-020-00977-1
6. Brisimi, T. S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. C., & Shi, W. (2018). *Federated learning of predictive models from federated electronic health records*. International Journal of Medical Informatics, 112, 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007
7. 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, 3069. https://doi.org/10.1038/s41467-019-10933-3
8. Domingo-Ferrer, J., & Soria-Comas, J. (2018). *Algorithmic transparency in data privacy: A review*. Journal of Privacy and Confidentiality, 8(1). DOI 10.29012/jpc.v8i1.644
9. Shabani, M., & Marelli, L. (2019). *Re-identifiability of genomic data and the GDPR*. EMBO Reports, 20(6), e48316. https://doi.org/10.15252/embr.201948316
10. Voigt, P., & Von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer International Publishing. https://doi.org/10.1007/978-3-319-57959-7
11. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). *Federated machine learning: Concept and applications*. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 12. https://doi.org/10.1145/3298981
12. 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
13. Abowd, J. M. (2018). *The U.S. Census Bureau adopts differential privacy*. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2867–2867. https://doi.org/10.1145/3219819.3226070
14. Thomas, J., & Harden, A. (2008). *Methods for the thematic synthesis of qualitative research in systematic reviews*. BMC Medical Research Methodology, 8, 45. https://doi.org/10.1186/1471-2288-8-45
15. Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). *Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement*. PLoS Medicine, 6(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097
16. Vovk, A., & Kovalchuk, S. (2023). *Methods and tools for healthcare data anonymization: A literature review*. Journal of Biomedical Informatics, 128, 104020.
17. Cormode, G., Procopiuc, C. M., Srivastava, D., Shen, E., & Yu, T. (2012). *Differentially private spatial decompositions*. Proceedings of the 2012 IEEE 28th International Conference on Data Engineering, 20–31. DOI 10.1109/ICDE.2012.13
18. Liu, Y., Acharya, U. R., & Tan, J. H. (2025). *Preserving privacy in healthcare: A systematic review of deep learning approaches for synthetic data generation*. Computer Methods and Programs in Biomedicine, 260, 108571. https://doi.org/10.1016/j.cmpb.2024.108571
19. Rujas, M., Herranz, R., Fico, G., & Merino-Barbancho, B. (2024). *Synthetic data generation in healthcare: A scoping review of reviews on domains, motivations, and future applications*. Journal of Biomedical Informatics, 134, 104192. https://doi.org/10.1016/j.jbi.2022.104192
20. Kumar, R. (2019). *Research Methodology: A Step-by-Step Guide for Beginners* (5th ed.). SAGE Publications Ltd.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details