# Anonymization of Datasets with Privacy, Utility and Risk Analysis

Benjamim Moreira 2020261856
José Cunha 2021223719
José Filipe 2021216675

October 20, 2024

## 1 Introduction

In a future increasingly reliant on data to drive decision-making and improve outcomes, data privacy breaches have become a major concern. One of the most significant threats is the re-identification of individuals from de-identified data, which compromises their privacy and trust in the services they use [6]. Our focus in this project will be on the healthcare sector . The main objective is to ensure the protection of individual privacy while preserving data utility for future analysis. the sensitivity of personal data makes anonymization particularly critical [1].

As part of our curricular unit, we will conduct a detailed analysis of the anonymization process applied to a healthcare dataset. This dataset will be imported into ARX, a powerful tool that will assist us in the anonymization process. The dataset attributes such as age, gender, medical conditions, and others will be classified into sensitive information, insensitive information, quasi-identifiers, and identifiers, as will be explained in more detail later in this work. This classification is crucial for applying the appropriate anonymization techniques and ensuring the balance between data privacy and utility.

Next, we will apply various privacy models such as l-Diversity, l-Diversity + t-Closeness, and k-Anonymity + b-Likeness and perform analyses to evaluate the effectiveness and optimization of the selected models. This includes assessing the risk of re-identification and measuring the balance between data utility and privacy. Previous research highlights the importance of employing robust privacy-preserving techniques in data mining to mitigate these risks.[6] Finally, we will provide a comprehensive explanation of each analysis, ensuring a clear understanding of the anonymization techniques used and their impact on safeguarding patient data.


Figure 1: Data-Driven Risk Management and Analysis.

# 2 Description and Classification of the attack, success criteria

In a hospital environment, where an insider (such as a staff member) may have partial access to patient information, such as age, gender, room number or blood type (quasi-identifiers), we would be dealing with a linkage attack. This type of attack occurs when someone with prior knowledge about an individual attempts to correlate that information with an anonymized dataset to re-identify the person [6]. For example, a hospital employee who knows this information could use those quasi-identifiers to search for the corresponding record in the dataset, potentially revealing sensitive data. To mitigate this risk, techniques like k-Anonymity, l-Diversity, and t-Closeness are applied, making it harder for the attacker to link known quasi-identifiers to specific individuals in the dataset [1]. Protecting against this kind of attack is especially relevant in hospitals, where insiders often have partial access to information about patients.

## 2.1 Definition of the goal of the anonymization process

The goal of the anonymization process is safeguarding individual privacy by transforming identifiable data into a format that prevents the re-identification of individuals while retaining the data's analytical utility. This involves techniques that obscure identifying features, such as names and specific dates, ensuring that the dataset remains valuable for research and healthcare analytics [6]. Key objectives include protecting personal information, preserving data utility for meaningful insights, mitigating risks of data breaches and building trust among patients and stakeholders. Ultimately, effective anonymization balances the need for data accessibility with the critical importance of protecting individual privacy. In the specific case of a hospital environment in which the attack comes from an insider, we will consider the anonymisation process a success if we verify that our dataset has not lost much data utility and we guarantee its protection, i.e. a prosecutor risk below 1%.

# 3 Dataset

## 3.1 Data characterization

The dataset selected contains comprehensive information on hospital-admitted patients, encompassing demographic details, medical histories, treatment records and billing data. The richness of the features provides a wide array of variables to explore, allowing for a thorough examination of various privacy models and anonymization techniques [4]. This diversity in attributes ensures that the dataset is well-suited for evaluating how different privacy-preserving methods can be applied to protect sensitive patient information while maintaining data utility [3].

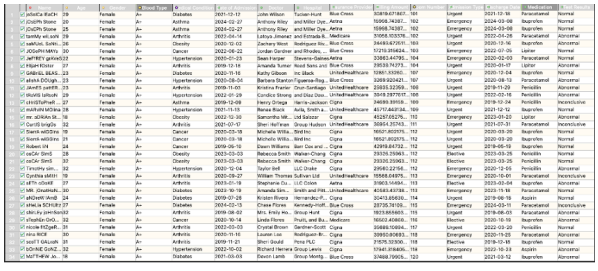


Figure 2: Dataset classification.

## 3.2 Classification of attributes

This classification is crucial for implementing effective data anonymization strategies that protect patient privacy while maintaining the utility of the data for analysis [5]. In the context of the healthcare dataset, attributes can be classified into three main categories: identifiers, quasi-identifiers, sensitive

information and insensitive information. Identifiers are attributes that can directly identify an individual. In this dataset, the patient's name serves as the primary identifier, as it uniquely links to a specific person [4].On the other hand, quasi-identifiers are attributes that, while not directly identifying, can help narrow down an individual's identity when combined with other information. In the dataset, attributes such as age, gender, blood type and room number. These can increase the risk of re-identification, especially when correlated with external datasets.

With regard to insensitive attributes, these are those that on their own or combined with other attributes do not reveal information that compromises personal information. These include elements such as: date of admission, doctor, hospital, insurance provider, billing amount, admission type, discharge date, medication and test result.

Additionally, sensitive information includes attributes that require special protection due to the potential harm or distress their disclosure could cause[6]. In this dataset the sensitive attributes is the medical condition . Revealing these details could significantly impact a patient's privacy and well-being, necessitating stringent anonymization measures.

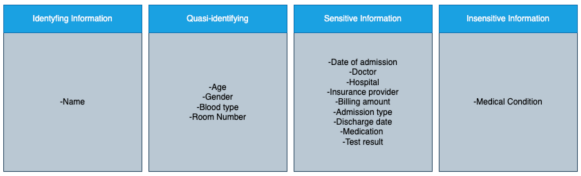



Figure 3: Dataset classification.

## 3.3 Dataset distributions

In the context of data anonymization using tools like ARX, quasi-identifiers (QIDs) and sensitive information are central concepts that affect the privacy and utility of the dataset. Understanding the distribution of a dataset is critical in anonymization because it directly impacts the effectiveness of privacy protection and the preservation of data utility. By carefully analyzing distributions, the most suitable anonymization techniques, ensure that privacy risks are minimized and maintain the value of the data for meaningful analysis.

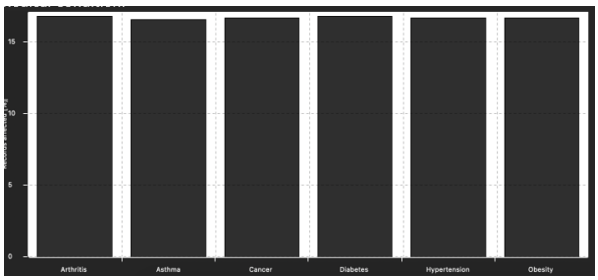



Figure 4: Medical condition distribution.

The medical condition feature showed limited variety in its attributes, though they were evenly distributed. This is important to consider, as it represents our sensitive attribute.

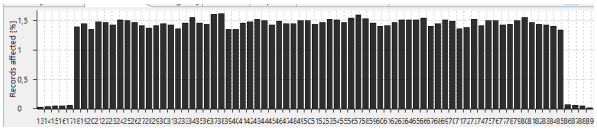



Figure 5: Age distribution.

When we analysed the distribution of the data for the age attribute, we saw a Gaussian distribution. This means that the edges, points where there is a lower rate of registrations, are more exposed to attacks and anonymisation measures need to be taken to maintain their privacy.

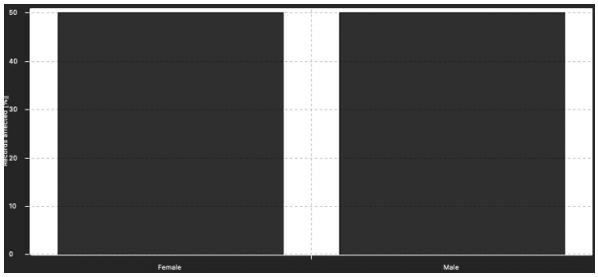



Figure 6: Gender distribution.

Analysing the distribution of the gender attribute, we conclude that the dataset has a balanced distribution.

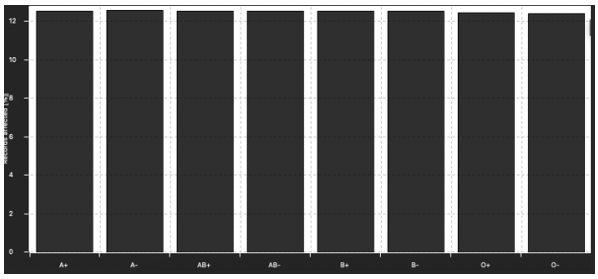



Figure 7: Blood Type distribution.

With regard to blood type, we also obtain a balanced distribution between the attributes.

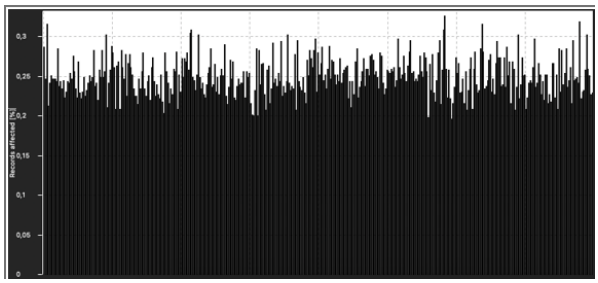



Figure 8: Room Number distribution.

The case of the room number is a very specific one, since although our dataset is balanced in this class, there are a large number of attributes causing a low percentage of appearance of each of the

possible attributes which could compromise the security of the data. It is therefore necessary to take measures to counter this security weaknesses in our data. To this end, generalised measures will have to be taken to ensure people's privacy

# 4 Privacy Models: explanation,applied models, generalisation

## 4.1 Definition of privacy and utility requirements

Privacy Requirements are centered on the protection of individuals' personal information from unauthorized access and re-identification. Dwork (2006) introduces the concept of differential privacy as a method to enhance privacy protections while allowing for data analysis.[2]

Privacy Requirements are focused on protecting individuals' personal information from unauthorized access and re-identification. Anonymization techniques, such as k-anonymity, l-diversity, and t-closeness, to ensure compliance with legal regulations and ethical standards. These techniques are aimed at mitigating privacy risks effectively. Ohm (2010) emphasizes the importance of robust anonymization methods to counter the potential for re-identification.[6]

On the other hand, utility requirements, as they focus on maintaining the usefulness of anonymized data for analysis and decision-making. To ensure that while personal information is protected, the data retains its quality and relevance. This balance allows researchers and analysts to derive meaningful insights without compromising individual privacy. Clifton et al. (2014) highlight the importance of achieving a balance between privacy and utility.[7]

## 4.2 Overview of Privacy Models

Anonymization techniques are critical for protecting sensitive information in healthcare datasets, especially when dealing with quasi-identifiers that could lead to re-identification when combined with external data. Below is an overview of the privacy models applied in this project.

### 4.2.1 l-Diversity(l=2)

This method will be applied to ensure that sensitive attributes, such as medical conditions, are sufficiently diverse within each anonymized group. Since the dataset contains only 6 medical conditions, l-Diversity ensures that each group contains at least l distinct values for these conditions, preventing all individuals in a group from sharing the same condition. This protects against homogeneity and inference attacks, where an attacker could easily deduce someone's medical condition based on the uniformity within a group.

### 4.2.2 l-Diversity (l=2)+ t-Closeness(t=0.05)

While l-Diversity guarantees diversity in sensitive attributes, it can fail if an attacker knows the global distribution of those values. Therefore, we will combine l-Diversity with t-Closeness. t-Closeness adds an extra layer of protection by ensuring that the distribution of medical conditions within each group is similar to the global distribution of the dataset. This reduces the risk of inference attacks based on external knowledge. This method is particularly useful in our case, as it prevents attackers from deducing medical conditions easily, even if they know the overall distribution of conditions in the population.

### 4.2.3 k-Anonymity(k=3)+ b-Likeness(b=0.1)

Finally, we will apply k-Anonymity combined with b-Likeness to protect both quasi-identifiers and sensitive attributes. k-Anonymity ensures that quasi-identifiers (such as age and gender) are sufficiently

generalized so that each group contains at least k indistinguishable records, preventing re-identification attacks. However, since k-Anonymity alone does not protect sensitive attributes, b-Likeness ensures that the distribution of medical conditions within each group does not deviate significantly from the global distribution, mitigating the risk of inference attacks, especially for rare conditions.

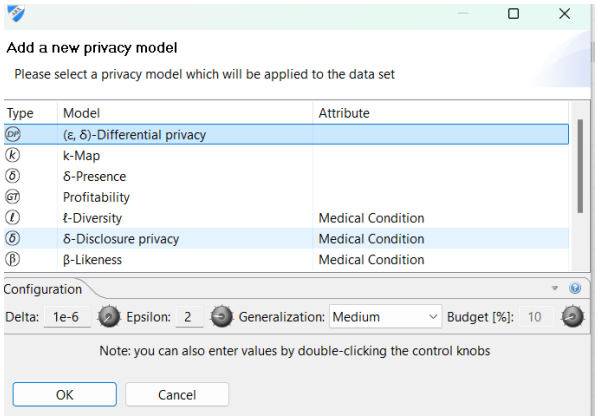



Figure 9: Privacy model settings.

## 4.3 Generalisations applied

In this section, the attributes underwent a process designed to reduce the amount of identifiable information they contain. To achieve this, hierarchies were created, creations of intervals and explicit identifiers with asterisks (*) were used to suppress certain details, ensuring that sensitive information still allowing for meaningful analysis.

In this first case, the age attribute was generalised into different intervals, making it possible to have different levels of security.

| Level-0 | Level-1 | Level-2 | Level-3 |
|---|---|---|---|
| 13 | [13, 33[ | [13, 53[ | [13, 90[ |
| 14 | [13, 33[ | [13, 53[ | [13, 90[ |
| 15 | [13, 33[ | [13, 53[ | [13, 90[ |
| 16 | [13, 33[ | [13, 53[ | [13, 90[ |
| 17 | [13, 33[ | [13, 53[ | [13, 90[ |
| 18 | [13, 33[ | [13, 53[ | [13, 90[ |
| 19 | [13, 33[ | [13, 53[ | [13, 90[ |
| 20 | [13, 33[ | [13, 53[ | [13, 90[ |
| 21 | [13, 33[ | [13, 53[ | [13, 90[ |
| 22 | [13, 33[ | [13, 53[ | [13, 90[ |
| 23 | [13, 33[ | [13, 53[ | [13, 90[ |
| 24 | [13, 33[ | [13, 53[ | [13, 90[ |
| 25 | [13, 33[ | [13, 53[ | [13, 90[ |
| 26 | [13, 33[ | [13, 53[ | [13, 90[ |
| 27 | [13, 33[ | [13, 53[ | [13, 90[ |

Figure 10: Age generalised.

In the generalisation of the second quasi identifier attribute, a generalisation was created which, if made, makes it impossible to identify the patient's gender.

Gender:

| Level-0 | Level-1 |
|---|---|
| Female | {Female, Male} |
| Male | {Female, Male} |

Figure 11: Gender generalised.

The blood type was generalised using asterisks, and 3 levels of security were created, each with a different level of data utility.

Blood Type:

| Level-0 | Level-1 | Level-2 | Level-3 |
|---|---|---|---|
| A+ | A+* | A** | *** |
| A- | A-* | A** | *** |
| AB+ | AB* | A** | *** |
| AB- | AB* | A** | *** |
| B+ | B+* | B** | *** |
| B- | B-* | B** | *** |
| O+ | O+* | O** | *** |
| O- | O-* | O** | *** |

Figure 12: Blood type generalised.

Finally, the number of rooms was generalised by creating different intervals from the first number to the last number. In this generalisation we also have different levels of privacy.

| Level-0 | Level-1 | Level-2 |
|---|---|---|
| 101 | [101, 302[ | [101, 501[ |
| 102 | [101, 302[ | [101, 501[ |
| 103 | [101, 302[ | [101, 501[ |
| 104 | [101, 302[ | [101, 501[ |
| 105 | [101, 302[ | [101, 501[ |
| 106 | [101, 302[ | [101, 501[ |
| 107 | [101, 302[ | [101, 501[ |

Figure 13: Room number generalised.

# 5 Privacy models: utility, privacy, and risk assessment

In this project, we will apply and compare three anonymization techniques—l-Diversity, l-Diversity + t-Closeness, and k-Anonymity + $\delta$-Likeness, to protect the privacy of data from a dataset containing 55,000 hospital records. First, before we start analyzing each model, we need to analyze the risks present in our dataset.

| Quasi-identifier | Distinction | Separation |
|---|---|---|
| Gender | 0.0036% | 50.00[illegible]6% |
| Blood Type | 0.01441% | 97.30139% |
| Age | 0.13874% | 98.54005% |
| Room Number | 0.72073% | 99.74955% |
| Gender, Blood Type | 0.02883% | 98.76% |
| Age, Gender | 0.27748% | 99.27018% |
| Age, Blood Type | 1.09912% | 99.81739% |
| Gender, Room Number | 1.44147% | 99.87466% |
| Blood Type, Room Number | 5.76587% | 99.96847% |
| Age, Room Number | 42.78[illegible]55% | 99.9963% |
| Age, Gender, Blood Type | 2.15499% | 99.90864% |
| Gender, Blood Type, Room Number | [illegible].52814% | 99.98407% |
| Age, Gender, Room Number | 62.42638[illegible]% | 99.99813% |
| Age, Blood Type, Room Number | 87.70248% | 99.99951% |
| Age, Gender, Blood Type, Room Number | 93.14582% | 99.99974% |

Figure 14: Quasi identifier combination table.

Analyzing the quasi-identifier combination table, we observe that individual quasi-identifiers like Gender (0.0036%) and Blood Type (0.0141%) exhibit low distinction values. However, when multiple quasi-identifiers are combined, such as Age, Gender, Blood Type and Room Number, the distinction values increase significantly, reaching as high as 93.15482

Regarding the separation metric, it remains consistently high for combinations of quasi-identifiers, with many values approaching 100%, such as Age, Gender, Blood Type, Room Number (99.99974%). This suggests that, without any privacy mechanisms in place, the dataset can be easily segmented into distinct groups, making re-identification highly feasible when multiple quasi-identifiers are used in combination. This puts the security of the individuals in our dataset at risk, we will now explore ways to prevent this.

## 5.1 l-Diversity(l=2)

This method will be applied to ensure that sensitive attributes, such as medical conditions, are sufficiently diverse within each anonymized group. Since the dataset contains only 6 medical conditions, l-Diversity ensures that each group contains at least l distinct values for these conditions, preventing all individuals in a group from sharing the same condition. This protects against homogeneity and inference attacks, where an attacker could easily deduce someone's medical condition based on the uniformity within a group.

Figure 15: l-Diversity(l=2) Dataset Transformation

We found that there was generalisation of the age and room number attributes, but not at the highest level of generalisation. This is a very positive aspect for preserving data utility for future analysis, but first we have to analyse its identification risk.

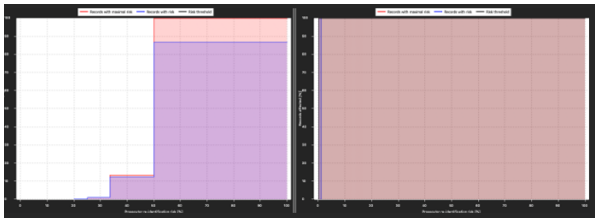


Figure 16: l-Diversity(l=2) Graphical analysis of the records at risk

On the graphic on the left before the application of l-Diversity(l=2), the records show that almost 70

percent of the data is more exposed to re-identification risks. The number of records at maximum risk increases gradually as more data is affected, indicating insufficient diversity in the sensitive attributes to prevent re-identification.

After applying l-Diversity(l=2), the results show that occured a significantly reducion in the number of records exposed to re-identification risk, even as larger portions of the data are affected. This ensures better privacy protection while maintaining data utility.

| Quasi-identifier | Distinction | Separation |
|---|---|---|
| Room Number | 0.0036% | 50.00038% |
| Gender | 0.0036% | 50.00086% |
| Age | 0.00721% | 74.12857% |
| Blood Type | 0.01441% | 87.50139% |
| Gender, Room Number | 0.00721% | 74.99995% |
| Age, Gender | 0.01441% | 87.06424% |
| Age, Room Number | 0.01441% | 87.06474% |
| Gender, Blood Type | 0.02883% | 93.75% |
| Blood Type, Room Number | 0.02883% | 93.75095% |
| Age, Blood Type | 0.05766% | 96.76658% |
| Age, Gender, Room Number | 0.02883% | 93.53226% |
| Gender, Blood Type, Room Number | 0.05766% | 96.87515% |
| Age, Gender, Blood Type | 0.11532% | 98.38777% |
| Age, Blood Type, Room Number | 0.11532% | 98.38924% |
| Age, Gender, Blood Type, Room Number | 0.23063% | 99.19128% |

Figure 17: l-Diversity(l=2) Quasi identifier table

The Distinction values are reduced in comparison to the non-privacy case. For example, Gender, Blood Type, and Room Number now have a distinction of 0.05766%, significantly lower than the original 1.52814%. Even for the most comprehensive combination of quasi-identifiers (Age, Gender, Blood Type, Room Number), the distinction is reduced to 0.23063%, compared to 93.15482% without privacy, showing that l-diversity has successfully decreased the uniqueness of records. Separation remains high but is notably lower than in the non-privacy case. For instance, Age, Gender, Blood Type, and Room Number has a separation of 99.19128%, slightly lower than the 99.99974% seen without privacy, but still high enough to preserve the dataset's utility. Other combinations, such as Room Number and Blood Type, show a similar trend, with separation values ranging from 50% to nearly 100%, albeit slightly lower than in the non-privacy case.

In conclusion, l-Diversity (l=2) reduces the Distinction metric considerably, making it more difficult to uniquely identify individuals in the dataset, which enhances privacy. While Separation is reduced, it still remains fairly high for most combinations of quasi-identifiers, indicating that the dataset retains a significant amount of utility even after the privacy method is applied. Overall, l-diversity offers a balanced trade-off, where re-identification risks are mitigated, but the ability to analyze and separate records remains mostly intact.

Analysing the risk to different types of attackers:

Figure 18: Separation and distinction table after applying privacy models.

By analysing the graph of record at risk and risk to different types of attackers, we can see that we have achieved the target we set ourselves for the success of the privacy model. To do this, we analyse the prosecutor risk, which is the risk that fits in with the attack we are simulating. We found an average prosecutor risk of 0.23063 % with a highest value of 0.3367 %. Overall we have the privacy method with a very good result for the simulated attack but also very good results for the others

## 5.2 l-Diversity (l=2)+ t-Closeness(t=0.05)

While l-Diversity guarantees diversity in sensitive attributes, it can fail if an attacker knows the global distribution of those values. Therefore, we will combine l-Diversity with t-Closeness. t-Closeness adds an extra layer of protection by ensuring that the distribution of medical conditions within each group is similar to the global distribution of the dataset. This reduces the risk of inference attacks based on external knowledge. This method is particularly useful in our case, as it prevents attackers from deducing medical conditions easily, even if they know the overall distribution of conditions in the population.

Figure 19: k-Anonymity(k=3)+ t-Closeness(t=0.05) Dataset Transformation

When analysing the privacy method applied, we found that there was a second-degree generalisation in age and blood type and a first-degree generalisation in room number. In this combination of the two methods we have already seen that for the two privacy methods to be applied there had to be a greater loss of data utility future analysis.

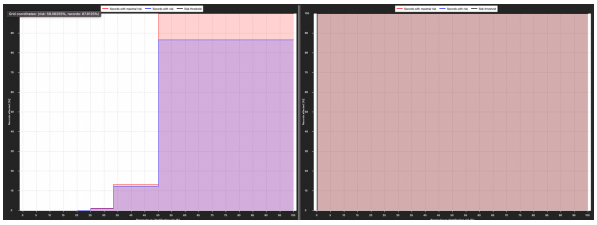
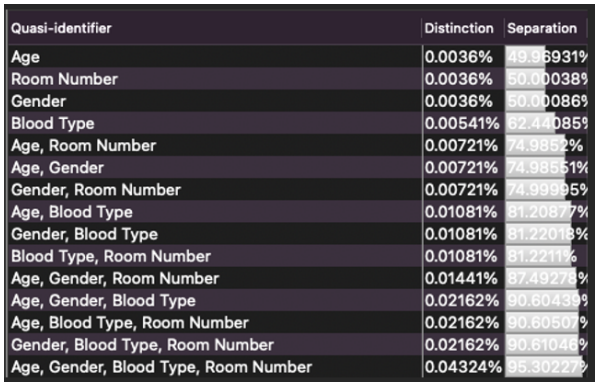


Figure 20: k-Anonymity(k=3)+ Closeness(t=0.05) Graphical analysis of the records at risk

In the graph on the left, before the application of k-Anonymity (k=3) and t-Closeness (t=0.05), a high percentage of records are exposed to re-identification risks. Approximately 80 % of the data shows a significant risk, indicating that without these anonymization techniques, the sensitive attributes are insufficiently protected, leading to increased vulnerability to re-identification.

After applying k-Anonymity (k=3) and t-Closeness (t=0.05), there is a drastic reduction in the number of records exposed to re-identification risks. Although a substantial portion of the data is still affected, the techniques significantly lower the distinction and separability of the records, contributing to more effective privacy protection. This demonstrates that even with the generalization applied, data utility is maintained while enhancing protection against inference and homogeneity attacks.



| Quasi-identifier | Distinction | Separation |
|---|---|---|
| Age | 0.0036% | 49.96931% |
| Room Number | 0.0036% | 50.000389 |
| Gender | 0.0036% | 50.000869 |
| Blood Type | 0.00541% | 62.440859 |
| Age, Room Number | 0.00721% | 74.9852% |
| Age, Gender | 0.00721% | 74.98551% |
| Gender, Room Number | 0.00721% | 74.99995% |
| Age, Blood Type | 0.01081% | 81.20877% |
| Gender, Blood Type | 0.01081% | 81.22018% |
| Blood Type, Room Number | 0.01081% | 81.2211% |
| Age, Gender, Room Number | 0.01441% | 87.49278% |
| Age, Gender, Blood Type | 0.02162% | 90.60439% |
| Age, Blood Type, Room Number | 0.02162% | 90.60507% |
| Gender, Blood Type, Room Number | 0.02162% | 90.61046% |
| Age, Gender, Blood Type, Room Number | 0.04324% | 95.30227% |

Figure 21: k-Anonymity(k=3)+ Closeness(t=0.05) Quasi identifier table.

After applying l-diversity and t-closeness, the distinction values are dramatically reduced for many of the combinations. For instance, Gender, Blood Type, and Room Number now has a distinction of only 0.0216% (compared to 1.5284% without privacy). Even the most comprehensive quasi-identifier combination, Age, Gender, Blood Type, and Room Number, shows a much lower distinction score of 0.0432% (down from 93.1542%), indicating that the privacy techniques have effectively reduced the ability to uniquely identify individuals.

The separation metric is also reduced but remains relatively high. For example, Age, Gender, Blood Type, and Room Number has a separation of 95.3027%, which is lower than the 99.9974% in the no-privacy scenario but still indicates a fairly strong level of separability between records. Some combinations, like Age and Room Number, exhibit a separation of 74.8512%, a noticeable drop from the values in the no-privacy case, reflecting how l-diversity and t-closeness reduce the dataset's ability to separate individuals into distinct groups, preserving privacy.

In conclusion, l-Diversity (l=2) and t-Closeness (t=0.05) substantially reduce both distinction and separation, making the dataset less susceptible to re-identification risks while maintaining some level of separability. In the non-privacy scenario, the dataset is much more vulnerable, as combining quasi-identifiers leads to very high distinction and separation, which could potentially lead to identification of individuals. By contrast, after applying l-diversity and t-closeness, even with multiple quasi-identifiers, the dataset becomes harder to distinguish and separate, thereby increasing privacy but still allowing for meaningful analysis of the data. This demonstrates how privacy techniques like l-diversity and t-closeness balance the trade-off between data utility and privacy protection.

Analysing the risk to different types of attackers:

Figure 22: Separation and distinction table after applying privacy models.

By examining the graph that displays the records at risk and the vulnerability to various types of attackers, we can conclude that we have reached the target we set for the success of the privacy model. To achieve this, we analyzed the prosecutor risk, which corresponds to the type of attack we are simulating. We found an average prosecutor risk of 0.04324 %, with a maximum value of 0.0615 % and we got good results for the other types of attacks too.

## 5.3 k-Anonymity(k=3)+ b-Likeness(0.1)

Finally, we will apply k-Anonymity combined with b-Likeness to protect both quasi-identifiers and sensitive attributes. k-Anonymity ensures that quasi-identifiers (such as age and gender) are sufficiently generalized so that each group contains at least k indistinguishable records, preventing re-identification attacks. However, since k-Anonymity alone does not protect sensitive attributes, b-Likeness ensures that the distribution of medical conditions within each group does not deviate significantly from the global distribution, mitigating the risk of inference attacks, especially for rare conditions.

Figure 23: k-Anonymity(k=3)+ b-Likeness(0.1) Dataset Transformation

By analysing the transformations that have been applied, we can see that there has been a generalisation (level 3) of the highest level of age, a level 2 generalisation of blood type and room number. As with the method above, we found that data was lost for future analysis.

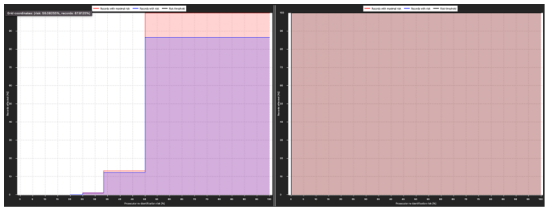


Figure 24: k-Anonymity(k=3)+ Likeness(0.1) Graphical analysis of the records at risk

In the graph on the left, before the application of k-Anonymity (k=3) and b-Likeness (b=0.1), a substantial percentage of records are at risk of re-identification. Around 75% of the data is exposed to high re-identification risks, demonstrating that without the anonymization techniques, the quasi-identifiers are not sufficiently protected, increasing the likelihood of individuals being identified.

After applying k-Anonymity (k=3) and b-Likeness (b=0.1), there is a notable reduction in the number of records exposed to re-identification risks. Even though some data is still impacted, the reduction in both distinction and separation values ensures that privacy is significantly enhanced. This method balances privacy protection and data utility, safeguarding sensitive information while still allowing for meaningful analysis.

| Quasi-identifier | Distinction | Separation |
|---|---|---|
| Age | 0.0018% | 0% |
| Room Number | 0.0036% | 50.00038% |
| Gender | 0.0036% | 50.00096% |
| Blood Type | 0.00541% | 82.44088% |
| Age, Room Number | 0.0036% | 50.00038% |
| Age, Gender | 0.0036% | 50.00096% |
| Age, Blood Type | 0.00541% | 82.44088% |
| Gender, Room Number | 0.00721% | 74.99995% |
| Gender, Blood Type | 0.01081% | 91.22018% |
| Blood Type, Room Number | 0.01081% | 91.2211% |
| Age, Gender, Room Number | 0.00721% | 74.99995% |
| Age, Gender, Blood Type | 0.01081% | 91.22018% |
| Age, Blood Type, Room Number | 0.01081% | 91.2211% |
| Gender, Blood Type, Room Number | 0.02162% | 90.61046% |
| Age, Gender, Blood Type, Room Number | 0.02162% | 90.61046% |

Figure 25: k-Anonymity(k=3)+ Likeness(0.1) Quasi identifier table

The distinction values are much lower compared to the no-privacy scenario. For example, Gender, Blood Type, and Room Number has a distinction of 0.02162%, a significant drop from the original 11.52814%. The combination of Age, Gender, Blood Type, and Room Number now has a distinction

of only 0.02162%, compared to 93.14582% without privacy, showing that k-anonymity and b-likeness successfully reduce the uniqueness of records in the dataset.

The separation values also drop significantly. For example, Gender, Blood Type, and Room Number now has a separation of 90.61046%, lower than the original 99.98407%. The combination of Age, Gender, Blood Type, and Room Number results in a separation of 90.61046%, showing that k-anonymity and b-likeness reduce the dataset's ability to separate individuals into distinct groups, while still preserving some separability.

In conclusion, k-Anonymity (k=3) combined with b-likeness (b=0.1) effectively reduces both Distinction and Separation, making it harder to identify or distinguish individuals based on quasi-identifiers. While the reduction in Separation is noticeable, especially in complex combinations of quasi-identifiers, it still remains fairly high, meaning that the dataset retains a good level of utility while significantly increasing privacy. The method provides a solid balance between protecting individuals' privacy (by reducing distinction and separation) and maintaining the usability of the dataset for analysis.

Analysing the risk to different types of attackers:

Figure 26: k-Anonymity(k=3)+ b-Likeness(0.1) - Separation and distinction table after applying privacy models.

By analysing the figure 26, we can see that we have achieved the desired success in protecting against the simulated attack. With an average prosecutor risk of 0.02162 % with its highest value being 0.02959 %. For the other types of attack, the method also proved to be very reliable.

# References


[1] Ning Cao, Ninghui Li, Tiancheng Li, and Subhash Venkatasubramanian. *k*-anonymity and *δ*-likeness. *IEEE Transactions on Knowledge and Data Engineering*, 24(1):54–67, 2011.

[2] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography (TCC)*, pages 265–284. Springer, 2006.

[3] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4):1–53, 2010.

[4] Ninghui Li, Tiancheng Li, and Subhash Venkatasubramanian. *t*-closeness: Privacy beyond k-anonymity and *ℓ*-diversity. *IEEE 23rd International Conference on Data Engineering (ICDE)*, pages 106–115, 2007.

[5] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Subhash Venkatasubramanian. *l*-diversity: Privacy beyond k-anonymity. *ACM Transactions on Database Systems (TODS)*, 42(3):1–29, 2007.

[6] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57(6):1701–1777, 2010.

[7] Latanya Sweeney. k-anonymity: A model for protecting privacy. In *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, volume 10, pages 557–570. World Scientific, 2002.