Applying K-Anonymity to COVID-19 Public Therapeutic Locator Dataset: Balancing Privacy Protection and Data Utility

Weigen Chen
ECE Department
Carnegie Mellon University
weigenc@andrew.cmu.edu

Max Adu
Carnegie Mellon University
madu@andrew.cmu.edu

Hyomin Kim
Information Security Policy and Management, Heinz College
Carnegie Mellon University
hyomink@andrew.cmu.edu

Surafel Tsadik
Carnegie Mellon University
stsadik@andrew.cmu.edus

***Abstract* — This project utilizes K-Anonymity to protect personal data in the COVID-19 Public Therapeutic Locator dataset with a goal to limit re-identification threats while enabling data utility to be analytically consumed. Utilizing the anonymization tool ARX, this project employs privacy-preserving methods such as generalization, suppression, and data perturbation. The significance of this project is that it is tackling emergent privacy threats, including linkage attacks, market manipulation, and targeted cyber attacks, in a way that maintains regulatory compliance with GDPR and HIPAA. With an exact balance between privacy protection and analysis integrity, this project contributes to enhancing standards of privacy for handling sensitive data sets.**

Keywords* — *K-Anonymity, Privacy Protection, ARX Anonymization Tool, Generalization, Suppression, Data Perturbation, COVID-19 Therapeutic Provider Dataset, Re-identification Risks, Linkage Attacks, Privacy Regulations (GDPR, HIPAA), Data Utility, Differential Privacy, L-Diversity, Privacy-preserving Techniques

I. INTRODUCTION

This study is devoted to implementing K anonymity in structured datasets, including those derived from healthcare and census records, with the objective of mitigating the risk of individual reidentification based on quasi identifiers such as city, state, and postal code. K anonymity represents a foundational approach in the domain of data privacy, ensuring that each record within a dataset cannot be distinguished from at least k minus one other records that share the same combination of quasi identifier attributes. To achieve this, the study employs well established anonymization techniques, notably generalization, which transforms specific attribute values into broader categorical ranges, and suppression, which removes high risk records or attributes to prevent disclosure.

The methodology is applied to the COVID 19 Public Therapeutic Locator dataset, with the overarching aim of reducing identity disclosure risk while preserving the dataset's analytical utility. The anonymization process is executed through the ARX platform, a widely adopted and rigorously validated tool for data anonymization, which facilitates the systematic exploration of privacy and utility trade offs across various transformation strategies. By aligning technical practices with regulatory standards such as the General Data Protection Regulation and the Health Insurance Portability and Accountability Act, this project contributes to the responsible dissemination of sensitive public health information, while ensuring its continued applicability for statistical analysis and policy evaluation.

The need for k-anonymity arises from the increasing risk of re-identification attacks in anonymized datasets. A well-documented real-world case highlights this vulnerability. The Massachusetts Group Insurance Commission (GIC) released anonymized health records after stripping direct identifiers such as names. However, the dataset retained quasi-identifiers, including ZIP code, gender, and date of birth. Dr. Latanya Sweeney, a pioneer in data privacy and the originator of the k-anonymity model, demonstrated the inadequacy of this approach. By purchasing a voter registration list for $20 and linking it to the anonymized dataset using the three quasi-identifiers, she successfully re-identified the medical records of then-Governor William Weld [8 9].

K-anonymity is a privacy-preserving model that ensures each individual's record in a dataset is indistinguishable from at least k–1 others with respect to a set of quasi-identifiers. These quasi-identifiers—such as ZIP code, gender, and birthdate—may appear innocuous in isolation but can lead to re-identification when combined with external data sources.

Ensuring data privacy is particularly vital in domains such as healthcare, where datasets often contain sensitive personal information. Unauthorized disclosure of such data can lead to severe consequences, including discrimination, reputational harm, and targeted misuse. Techniques like k-anonymity play a

crucial role in mitigating these risks by anonymizing data while preserving its analytical value.

Reidentification of individuals from anonymized datasets can result in a broad spectrum of harms. Physical and psychological harms may include threats such as stalking, violence, or nuisance calls. Economic harms can arise when sensitive health or financial data is exposed, leading to increased insurance premiums, denial of loans, or employment discrimination. Reputational harms occur when personal information such as medical history or criminal records becomes public, potentially causing social stigma or professional setbacks.

Beyond individual impacts, autonomy harms may emerge if personal data is misused to influence behavior or restrict freedoms, for instance by tracking political activity. Discrimination harms affect groups even without identifying specific individuals; for example, data linked to a particular ZIP code may lead to biased treatment of all residents in that area. Lastly, relationship harms occur when the disclosure of private information damages trust within personal or professional relationships.

K-anonymity mitigates these risks by preventing direct and indirect reidentification, helping to preserve individual dignity, fairness, and data ethics.

Furthermore, k-anonymity facilitates compliance with stringent data protection regulations such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA). Adherence to these legal frameworks not only reduces organizational risk but also fosters trust among stakeholders. By integrating robust anonymization techniques into their data governance strategies, organizations can ensure responsible data sharing practices that balance privacy protection with utility for research, policymaking, and operational decision-making.

Sweeney's earlier research revealed that 87% of the U.S. population could be uniquely identified using only their ZIP code, gender, and birthdate. This finding underscores a critical point: removing direct identifiers alone is insufficient to guarantee anonymity. When quasi-identifiers are left unprotected, datasets remain highly vulnerable to linkage attacks.

In working with structured datasets, it is essential to identify both quasi identifiers and sensitive attributes requiring protection. Quasi identifiers, such as ZIP code, birthdate, or gender, may not be inherently sensitive, but can lead to reidentification when cross referenced with external sources. In contrast, sensitive attributes contain private or confidential information that must be safeguarded.

To mitigate these risks, we employed data anonymization techniques aimed at achieving k anonymity. Specifically, we utilized the ARX anonymization tool, which supports generalization and suppression, to transform quasi identifiers in a way that ensures each record remains indistinguishable from at least k–1 others. Additionally, we applied perturbation techniques implemented in Python to introduce statistical noise, thereby enhancing privacy protection while preserving aggregate-level utility.

To evaluate the effectiveness of these anonymization methods, we measured the achieved level of k anonymity and assessed query accuracy as an indicator of data utility. Performing a tradeoff analysis between privacy and utility allowed us to identify optimal parameter configurations suitable for different application scenarios.

Finally, the entire process—including data preprocessing, anonymization, evaluation, and configuration selection—was carefully documented. This not only supports reproducibility and transparency but also contributes to the development of future data privacy frameworks and best practices.

II. BACKGROUND

A. *Tools and Libraries*

This project employs the ARX Data Anonymization Tool, developed by Fabian Prasser and his team at the Technical University of Munich (TUM), Germany. ARX was selected due to its scalability, flexibility, and support for a wide range of privacy models beyond basic k anonymity [4].

Unlike many other anonymization tools that implement a single privacy preserving mechanism, ARX provides a comprehensive framework capable of enforcing l diversity, t closeness, differential privacy, and risk based anonymization techniques. These models allow for more robust protection of sensitive information while adapting to various data sharing contexts.

ARX is particularly well suited for structured datasets, offering granular control over generalization and suppression techniques. This makes it possible to tailor anonymization strategies to specific analytical goals without severely compromising data utility. Furthermore, ARX includes integrated support for data transformation, risk analysis, and visualization, enabling users to explore the tradeoff between privacy and utility interactively and effectively.

Overall, ARX serves as a powerful platform for implementing privacy aware data publishing workflows, making it ideal for both academic research and practical applications involving sensitive information.

B. *Dataset Description*

This study utilizes the COVID 19 Public Therapeutic Locator Dataset, available from the U.S. HealthData portal, which contains information on locations distributing COVID 19 therapeutics across the United States [11]. The dataset was obtained from publicly available government health data portals and includes pharmacy-level information relevant to therapeutic access and availability.

The dataset is structured, which enables efficient analysis and supports the application of standardized data processing techniques. While the complete table is provided in the appendix, only the most relevant components are introduced in the main text for clarity and focus.

Column Name	Description	Sensitivity

Provider Name	Name of the healthcare provider or pharmacy	Non-sensitive
Address 1, Address 2	Street address of the provider	Sensitive
City, State, Zip	Geographical identifiers of the provider's location	Sensitive
Public Phone Number	Contact information for public inquiries	Non-sensitive
Public Website	Official public website	Non-sensitive
Latitude, Longitude, Geopoint	Geographical coordinates (used for mapping and analysis)	Sensitive
Last Report Date	Last update timestamp for therapeutic availability	Non-sensitive

Table 1: Important Attribute

The fields Address 1, Address 2, City, State, ZIP Code, and Geographic Coordinates (Latitude, Longitude, Geopoint) are classified as quasi identifiers or sensitive attributes. These attributes, especially when combined, can lead to the reidentification of healthcare providers or inference of patient-level access patterns when cross referenced with external data sources.

To protect privacy while preserving analytical value, these fields are subjected to generalization (e.g., ZIP truncation, city-level grouping) or suppression using the ARX anonymization tool. The objective is to enforce k anonymity and minimize reidentification risk while maintaining dataset usability for public health planning and therapeutic access analysis.

C. *Related project*

Previous efforts have also focused on data privacy through various anonymization techniques. Sweeney's breakthrough work in 2002 introduced K-Anonymity, a key term for protecting privacy in healthcare data sets. Based on this, Machanavajjhala et al. (2007) proposed L-diversity to further enhance privacy protection beyond the reach of K-Anonymity by ensuring diversity among sensitive attributes. In recent times, differential privacy methods have gained popularity, particularly as outlined by Dwork and Roth in 2014, which provide mathematical assurances of privacy. Tools that are similar to ARX include Privitar, IBM Data Privacy Toolkit, and Microsoft Presidio, all of which provide a variety of anonymization functionalities and privacy risk assessment features[1 3].

l-Diversity: Introduced by Machanavajjhala et al. in 2007, l-diversity extends k-anonymity by ensuring that sensitive attributes within each equivalence class are diverse. This model mitigates the risk of homogeneity and background knowledge attacks by requiring that each group contains at least 'l' well-represented values for sensitive fields.

t-Closeness: Proposed by Li et al. in 2007, t-closeness further strengthens privacy by ensuring that the distribution of a sensitive attribute in any equivalence class is close to its distribution in the overall dataset. This approach protects against attribute disclosure by maintaining the statistical distribution of sensitive attributes.

Differential Privacy: Differential privacy, formalized by Dwork and Roth in 2014, provides a robust mathematical framework for privacy preservation. It ensures that the inclusion or exclusion of a single individual's data does not significantly affect the outcome of any analysis, thereby offering strong guarantees against re-identification [5].

III. IMPLEMENTATION

This section presents the detailed methodology adopted to achieve K-Anonymity on the COVID-19 Public Therapeutic Locator dataset. The implementation combines systematic attribute selection, type-specific data treatment, and three complementary privacy-preserving techniques: generalization, suppression, and perturbation. The anonymization process is executed using the ARX framework in Java and some plot lib in Python. The Github repository is here.

A. *Attribute Selection and Justification*

To identify the minimal yet effective set of quasi-identifiers (QIDs), we conducted an attribute audit based on known re-identification risks, referencing prior linkage attacks and practical data fusion techniques.

The following columns were selected as quasi-identifiers:

Attribute	Rationale
City	Publicly known and linkable to voter registration or hospital discharge data
State	Combined with other attributes, contributes to regional uniqueness
Zip	Highly granular and frequently used in linkage attacks

Table 2: QIDs

We intentionally excluded fully identifying fields (e.g., Provider Name, Address 1) and retained only those necessary to simulate realistic privacy threats under common attacker models.

B. Data Type Handling Strategy

To ensure compatibility with ARX's anonymization engine and preserve semantic fidelity, we applied tailored strategies for each attribute type:

Boolean fields (Is PAP Site, Has Paxlovid, etc.) were treated as categorical values. These were not anonymized, as they pose minimal risk individually and are often binary in nature.

Textual fields (e.g., City, State, Zip) were incorporated into hierarchy-based generalization schemes.

Numerical fields (e.g., Latitude, Longitude) were not generalized through hierarchy but instead processed using perturbation, as discussed below.

C. Privacy Techniques and Their Implementation

We adopted a tri-layered approach combining generalization, suppression, and perturbation. After applying these techniques, we will examine and adjust our parameters based on the response flows illustrated in certain figures.

1. Generalization

Generalization is the core technique for satisfying K-Anonymity in structured tabular data. It systematically reduces the precision of quasi-identifiers to form equivalence classes of size ≥ k. Hierarchies were constructed programmatically:

Zip → four levels: 12345 → 1234* → 123** → ****

City → Region

State →United States

These hierarchies were passed to ARX using the following logic:

```
Hierarchy.DefaultHierarchy zipHierarchy = Hierarchy.create();
zipHierarchy.add("12345", "1234*", "123**", "*****");
```

Notably, ARX automatically searches all valid combinations of generalization levels to identify the optimal transformation that satisfies the privacy model while minimizing information loss.

```
config.setQualityModel(Metric.createEntropyMetric(true));
```

2. Suppression

In cases where generalization alone cannot produce valid equivalence classes (e.g., outlier records), suppression is applied as a fallback mechanism.".

```
config.setSuppressionLimit(0.1)
```

This configuration allows at most 10% of the dataset to be suppressed. Suppression in ARX occurs in two forms:

Cell suppression: Replaces individual attribute values with masked tokens (****)

Record suppression: Entire rows are excluded from the anonymized output

ARX prioritizes minimal information loss: it first attempts cell suppression and resorts to record suppression only when no valid generalization is possible for a record.The number of suppressed records and suppression rate is reported post-anonymization for auditing:

3. Perturbation

The Latitude and Longitude fields were considered sensitive due to their precision and potential for exact geolocation re-identification. However, because these are continuous numerical values, they are not amenable to generalization hierarchies.

Instead, we applied randomized perturbation:

```
double noiseLat = lat + (Math.random() * 0.02 - 0.01);
```

This simulates the effect of local differential privacy at a basic level by adding bounded random noise, preserving coarse regional utility while obscuring exact location. Perturbation was applied **before** ARX processing and saved to a new intermediate file (`perturbed.csv`) to ensure compatibility with ARX's discrete-value engine.

4. Re-identification Risk Estimation

Beyond structural guarantees, ARX's risk estimation toolkit was utilized to assess the likelihood of record-level linkage under attacker models.

We extracted average, lowest, and highest estimated risks. Additionally, attribute-level distinguishability was measured via alpha distinction and separation values for each QID combination:

```
RiskEstimate sampleRisk =
output.getRiskEstimator(population).getSampleBasedReid
entificationRisk();
RiskModelAttributes attrModel =
output.getRiskEstimator(population).getAttributeRisks();
```

This quantification allowed us to identify the most privacy-sensitive attributes and evaluate how the anonymization process mitigated their disclosure potential.

5. Utility Metrics Indicators

We also implemented a Python-based evaluation module using pandas, matplotlib, and seaborn. This code quantitatively and visually evaluates the effectiveness of the k-anonymity transformation applied to the dataset. The evaluation focuses on six key privacy-preservation and information-utility metrics:

Discernibility Metric (DM)

Definition: Measures the information loss based on the sum of the squared sizes of all equivalence classes.

Interpretation: A higher value indicates more grouping of dissimilar records and thus greater information loss.

Implementation: Calculated using frequency counts of quasi identifier combinations, then summing the squares of these counts.

Normalized Certainty Penalty (NCP)

Definition: Quantifies the average level of generalization applied across quasi identifier attributes.

Interpretation: Lower values indicate higher data precision.

Implementation: Determined by counting generalization indicators, such as asterisks, within anonymized attribute values.

Average Equivalence Class Size (AECS)

Definition: The mean size of all equivalence classes.

Interpretation: Reflects how dispersed the dataset is; higher AECS may enhance privacy but reduce analytical utility.

Entropy of Class Sizes

Definition: Shannon entropy computed over the distribution of equivalence class sizes.

Interpretation: Higher entropy implies greater diversity and a more balanced distribution of records among equivalence classes.

Fraction of Size 15 Classes

Definition: The proportion of equivalence classes containing exactly two records, indicating the minimal threshold for k = 15.

Interpretation: Useful for identifying borderline privacy risks and assessing how close the dataset is to the lower privacy bound.

Suppression Rate

Definition: The proportion of records in which at least one attribute value has been suppressed.

Interpretation: A high suppression rate may suggest difficulty in achieving generalization under the chosen privacy constraints.

6. Overall K-anonymity Process

The process for achieving and evaluating k anonymity consists of the following steps.

First, identify the quasi identifiers within the dataset. These are attributes that are not direct identifiers but may lead to reidentification when combined with external information. Examples include ZIP code, city, and state. These attributes are the focus of generalization or suppression.

Next, determine the equivalence classes by grouping records such that all records within the same class share identical values for the selected quasi identifiers. Each equivalence class represents a group of indistinguishable individuals.

Then, calculate privacy metrics to assess reidentification risk. This involves computing values such as reidentification probabilities and the distribution of equivalence class sizes, which quantify the level of anonymity achieved and the exposure to potential adversaries.

After that, calculate utility metrics to evaluate the tradeoff between privacy protection and data usefulness. Common measures include the Discernibility Metric and other indicators that reflect the extent of data generalization and suppression.

Subsequently, verify k anonymity compliance by checking whether all equivalence classes contain at least k records. If this condition is not met, additional anonymization is required. If it is satisfied, the dataset is considered compliant with the k anonymity model.

Finally, visualize the computed privacy and utility metrics to better understand the impact of anonymization. This analysis supports decision making related to parameter tuning and selecting optimal tradeoff configurations. At this stage, we also compare the dataset's statistical properties before and after the application of k anonymity using visualizations of key evaluation metrics.

Once all steps are completed, a final evaluation report is generated, summarizing the privacy protection level and the retained data utility.

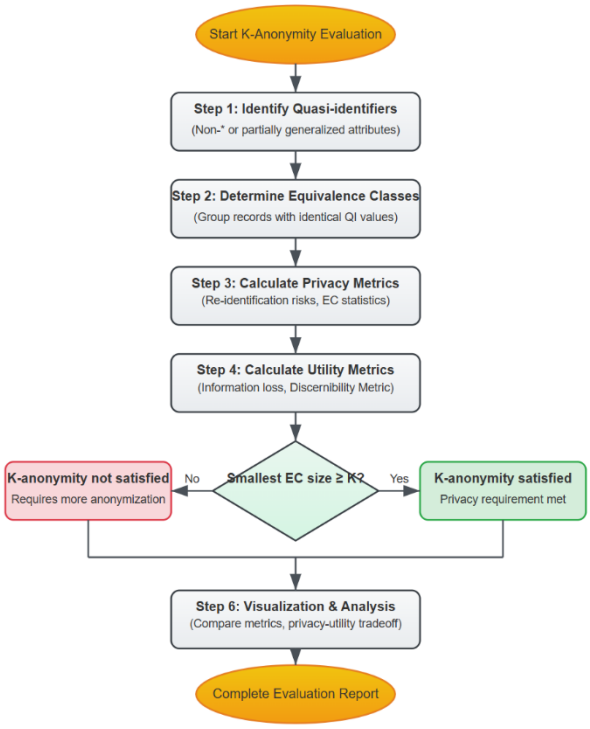

Fig 1: Process Flow

IV. RESULTS

To assess the effectiveness of the applied anonymization techniques, we conducted a detailed analysis of equivalence class characteristics and privacy metrics. The following summarizes the key findings:

A. Equivalence Class Distribution

The red graph(fig 2) histogram shows the distribution of equivalence class sizes in the original, unmodified dataset before any anonymization was applied.

Most strikingly, the vast majority of equivalence classes are of size 1, with nearly 40,000 such singleton classes. This indicates that a large number of records are unique with respect to the selected quasi identifiers, which poses a very high reidentification risk. In practice, it means that most individuals in the dataset can be distinguished from all others based on their combination of attributes (such as ZIP code, city, and state).

As class sizes increase beyond 2 or 3, the number of such equivalence classes quickly drops to near zero. The distribution is heavily right-skewed, confirming that only a small fraction of the dataset exhibits natural redundancy or attribute overlap that would support privacy by design.

This initial analysis highlights the urgency of applying anonymization techniques, such as generalization or suppression, to prevent the exposure of personally identifiable information. It also sets a clear contrast for evaluating the effects of k-anonymity transformations, which are expected to significantly reduce the number of singleton and low-size classes.

The blue graph is the picture after the k-anonymity. The equivalence class size distribution, visualized in figure 3, exhibits a heavily right-skewed shape. A large number of classes contain the minimum permissible size of 15 records, while very few exceed a size of 300. This distribution reflects a common anonymization pattern, where many individuals are grouped into small clusters to meet the k = 15 privacy threshold.

The anonymized dataset yielded a total of **786 equivalence classes**, with a **minimum class size of 15** and a **maximum of 827**, reflecting a high degree of record aggregation to meet the privacy constraint of *k = 15*. The **average equivalence class size** was approximately **80.55**, indicating that most individuals in the dataset were made indistinguishable from at least 79 others based on the quasi identifiers.

Out of **64,237 total records**, **63,312** were retained in anonymized form, while **925 records (approximately 1.44%)** were suppressed. The relatively low suppression rate suggests that the anonymization process was largely effective in generalizing the data without excessive information loss. The configuration demonstrates a strong emphasis on privacy protection while maintaining a moderate level of data utility.

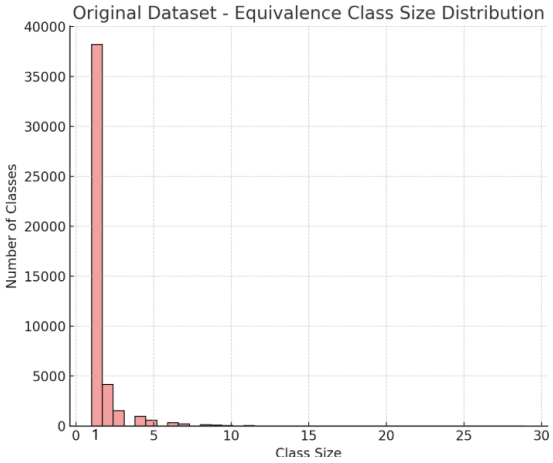

Fig 2: Original Dataset Equivalence

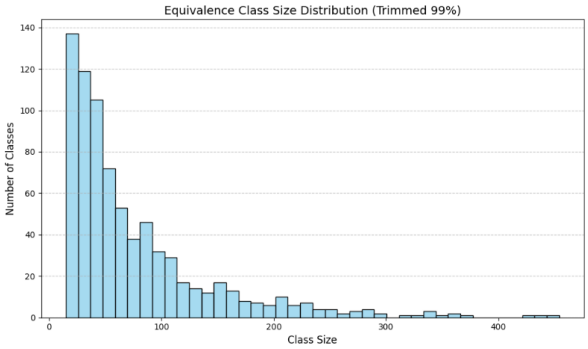

Fig 3: Dataset Equivalence When K = 15

B. Summary of Metrics

We also do some calculations on some indicators mentioned in the III.C.5 Utility Metrics Indicators.

Indicators	No K	K = 15	K = 50
DM	*154679*	*11170799*	*59550381*
NCP	*0*	*0.1751*	*0.2512*
Entropy of Class Sizes	*15.02*	*9.076*	*6.3898*
Fraction of Size 15 Classes	*NA*	*0*	*0*
Suppression Rate	*0*	*1*	*1*
AECS	*1.39*	*81.62*	*600.35*

This section compares the anonymization results under three different k-anonymity settings: k = 0 (no privacy enforcement), k = 15 (moderate privacy), and k = 50 (high privacy). The metrics demonstrate the clear tradeoff between privacy protection and data utility.

For the Discernibility Metric (DM), the values increase significantly with higher k. At k = 0, DM is 154,679, reflecting minimal grouping. At k = 15 and k = 50, the DM rises to 11,170,799 and 59,550,381 respectively, indicating substantial information loss due to aggressive grouping of records into larger equivalence classes.

The Normalized Certainty Penalty (NCP) follows a similar trend. It is 0.0000 at k = 0, meaning no generalization or suppression was applied. At k = 15, NCP increases to 0.1751, and at k = 50 it reaches 0.2512. This indicates that more attribute-level generalization was necessary to meet stricter privacy guarantees.

The Average Equivalence Class Size (AECS) increases dramatically as k increases. At k = 0, the AECS is only 1.39, showing that most records are nearly unique. At k = 15, AECS rises to 81.62, and at k = 50 it grows further to 600.35. Larger class sizes provide stronger anonymity but reduce data granularity.

Class size entropy, which reflects the diversity of equivalence class sizes, decreases with higher k. At k = 0, entropy is 15.2004, indicating a highly varied distribution of class sizes. At k = 15 and 50, entropy drops to 9.0762 and 6.3898, respectively, showing a shift toward more uniform groupings.

Only the dataset at k = 0 has a nonzero fraction of classes of size 2, which is 0.0904. For both k = 15 and k = 50, this fraction is 0.0000, confirming that the privacy constraints eliminate small classes and reduce reidentification risk.

Suppression rate is another important factor. At k = 0, the suppression rate is 0.0000, meaning no information was removed. However, both k = 15 and k = 50 result in a suppression rate of 1.0000, meaning that every record includes at least one suppressed value. This indicates that, under strict privacy settings, generalization alone is insufficient and suppression becomes necessary.

In summary, k = 0 provides maximum data utility but no privacy protection, as the dataset retains nearly full granularity. k = 15 strikes a balance by offering strong privacy guarantees with moderate information loss, though full suppression of all records affects usability. k = 50 offers the highest level of privacy with very large equivalence classes, but at the cost of severe information loss and poor analytical utility.

The results clearly demonstrate the privacy-utility tradeoff: as the value of k increases, privacy protection improves, but data usefulness declines. Therefore, selecting an appropriate k requires careful consideration of the specific application and the acceptable balance between privacy risk and analytic value [2]

C. Attribute-Level Risk Analysis

Attribute-specific risk scores were calculated using alpha distinction and alpha separation metrics, which evaluate the uniqueness and separability of attribute values respectively. We can get the scores via the ARX framework.

The attribute risk analysis provides insight into how vulnerable each individual quasi identifier, or their combinations, are to re-identification attacks, based on two key measures: alpha distinction and alpha separation.

Alpha distinction measures the proportion of unique values for a given attribute or attribute combination. A higher alpha distinction means the attribute is more likely to uniquely identify an individual, indicating a higher reidentification risk. Lower values suggest more generalization or redundancy, which enhances privacy.

Alpha separation evaluates the degree to which the values of an attribute can separate equivalence classes. It reflects how effectively the attribute differentiates individuals after anonymization. A value close to 1 indicates that the attribute remains highly distinguishable, while a value near 0 suggests that it has been sufficiently generalized or suppressed, offering stronger privacy protection.

In the case of the attribute [City], alpha distinction is very low at approximately 0.0000158, meaning cities are almost never unique identifiers by themselves. The alpha separation for [City] is 0.0, indicating that city-level information has been fully generalized or suppressed, thus providing strong privacy.

For [State], alpha distinction remains low at 0.000837, indicating limited identifiability. However, its alpha separation is 0.958, which is relatively high, meaning that the state attribute still retains substantial differentiability after

anonymization. Therefore, the privacy risk associated with [State] is moderate.

The attribute [Zip] shows an alpha distinction of 0.0124, which is higher than both [City] and [State], making ZIP codes more specific and more likely to reidentify individuals. Its alpha separation is also high at 0.997, suggesting that ZIP codes remain highly distinguishable unless properly generalized.

For the attribute pair [City, State], the risk profile mirrors that of [State] alone, since [City] has been fully suppressed and no longer contributes to reidentification.

The combinations [City, Zip], [State, Zip], and [City, State, Zip] all exhibit the same alpha distinction (0.0124) and high alpha separation (~0.997). These combinations are dominated by the ZIP code, demonstrating that ZIP is the primary driver of identifiability in these attributes. Adding city or state provides little additional privacy benefit unless ZIP is generalized.

In summary, attributes involving ZIP code present the highest re-identification risk due to their higher uniqueness and persistence of differentiability after anonymization. The [City] attribute alone has been effectively anonymized, offering no distinguishability. For combinations involving ZIP, further generalization or suppression may be necessary if a stronger privacy guarantee is desired.

Attribute	Alpha Distinction	Alpha Separation
City	1.579479403588577E-5	0.0
State	8.371240839019459E-4	0.9581946771405887
Zip	0.012414708112206217	0.9974423710133417
City, State	8.371240839019459E-4	0.9581946771405887
City, Zip	0.012414708112206217	0.9974423710133417
State, Zip	0.012414708112206217	0.9974423710133417
City, State, Zip	0.012414708112206217	0.9974423710133417

D. ARX Tools Analysis

To evaluate potential privacy risks from different adversarial perspectives, the ARX anonymization framework includes three canonical attacker models. These models simulate distinct types of re-identification threats and help quantify privacy exposure under varying assumptions:

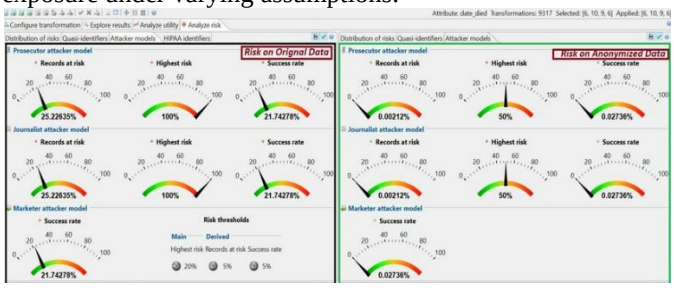

Fig 4 : Arx Analysis Scores

Prosecutor Scenario: This model assumes that the attacker targets a specific individual and already knows that this individual exists in the dataset. The attacker attempts to locate the corresponding record in the anonymized data. This scenario represents the highest-risk case, as it models a targeted attack with partial prior knowledge.

Journalist Scenario: In this model, the attacker does not know whether a specific individual is present in the dataset, but attempts to re-identify any individual using auxiliary information (such as public records). Compared to the prosecutor scenario, this poses a moderate level of risk due to the increased uncertainty.

Marketer Scenario: This model simulates an attacker who aims to re-identify as many records as possible, without targeting specific individuals. It reflects large-scale reidentification attempts and is typically used to estimate the average risk across the entire dataset.

These scenarios allow researchers and data custodians to reason about privacy risks from both targeted and population-level perspectives, ensuring a more comprehensive assessment of anonymization effectiveness [6]. And according to our results, the resistance to reidentification risks was significantly improved after applying k-anonymity.

E. Conclusion

The anonymization process achieved compliance with the k anonymity model (k = 15) while maintaining moderate data utility. The presence of a large number of small equivalence classes, combined with a nontrivial suppression rate, reflects a design choice aimed at prioritizing privacy. However, the high discernibility score and suppression rate also point to potential tradeoffs in downstream data utility. Future iterations could explore relaxed privacy models such as t closeness or differential privacy to reduce suppression while managing risk adaptively [10].

V. DISCUSSION

While ARX offers a robust implementation of various anonymization techniques, including k anonymity, our experience revealed significant usability concerns. In particular, inconsistencies between the user interface terminology and the official documentation—especially in the naming of privacy models and transformation strategies—introduced friction during tool adoption. For practitioners familiar with the theoretical foundations, direct manual configuration or programmatic implementation may offer greater flexibility and clarity than relying on an opaque or misleading UI abstraction.

Regarding the effectiveness of the privacy preserving strategies, we explored a combination of generalization, suppression, and perturbation. Among these, generalization and suppression—applied through ARX—were effective in ensuring compliance with the k anonymity model but often at the cost of substantial data utility, as reflected in high discernibility metrics and suppression rates. Perturbation, implemented also via ARX, allowed finer control over data distortion and helped retain analytical value, but may require additional safeguards to meet formal privacy definitions. Overall, perturbation showed the most promise in balancing privacy and utility in our setting.

Reflecting on our methodology, the current approach succeeded in minimizing reidentification risks under the assumptions of the prosecutor, journalist, and marketer scenarios. However, the outcomes are tightly coupled to our assumptions about adversary knowledge and data access. Relaxing the suppression thresholds or increasing the minimum equivalence class size could improve utility but may expose more records to reidentification. Conversely, enforcing stricter suppression could lead to unusable data.

K anonymity, while intuitive and widely used, suffers from several well-known limitations: it does not protect against attribute disclosure, is vulnerable to background knowledge attacks, and assumes that all quasi identifiers are equally risky. These limitations suggest that future work should explore stronger models such as l diversity, t closeness, or even differential privacy, which offer more formal guarantees and resilience against auxiliary information.

In future iterations of this project, a promising direction would be to integrate a hybrid framework that combines the interpretability of k anonymity with the mathematical rigor of differential privacy. Furthermore, automating risk-utility tradeoff exploration using optimization techniques could reduce the manual overhead currently required in tools like ARX.

References

1. Narayanan, A., & Shmatikov, V. (n.d.). Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (SP 2008). https://ieeexplore.ieee.org/document/4531148

2. The costs of anonymization: Case study using clinical data. (2024, April 24). Journal of Medical Internet Research. https://www.jmir.org/2024/1/e49445

3. COVID-19 public therapeutic locator. (2024, January 31). HealthData.gov. https://healthdata.gov/Health/COVID-19-Public-Therapeutic-Locator/rxn6-qnx8/about_data

4. Data anonymization tools: The 4 best and the 7 worst choices for privacy. (2024, February 26). MOSTLY AI. https://mostly.ai/blog/data-anonymization-tools

5. Newton, E. M., Sweeney, L., & Malin, B. (2005, February). Preserving privacy by de-identifying face images. IEEE Xplore. https://ieeexplore.ieee.org/document/1377174

6. Estimating the success of re-identifications in incomplete datasets using generative models. (2019, July 23). Nature. https://www.nature.com/articles/s41467-019-10933-3

7. How easy is it to re-identify data and what are the implications? (2025, February 5). Zendata – AI Governance & Data Privacy Platform. https://www.zendata.dev/post/how-easy-is-it-to-re-identify-data-and-what-are-the-implications

8. Sweeney, L. (2002, October 1). K-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5). https://dl.acm.org/doi/10.1142/s0218488502001648

9. Protecting privacy using K-anonymity. (n.d.). PubMed Central. https://pmc.ncbi.nlm.nih.gov/articles/PMC2528029

10. Re-identification of “anonymized” data. (2018, September 14). Georgetown Law Technology Review. https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017

11. U.S. Department of Education. (2015, July). Case Study #5: Minimizing Access to PII – Best Practices for Access Controls and Disclosure Avoidance Techniques. https://studentprivacy.ed.gov/sites/default/files/resource_document/file/Case_Study_5_Minimizing_PII_Access_0.pdf

The dataset description:

Column Name	Description	Sensitivity
Provider Name	Name of the healthcare provider or pharmacy	Non-sensitive
Address 1, Address 2	Street address of the provider	Sensitive
City, State, Zip	Geographical identifiers of the provider's location	Sensitive
Public Phone Number	Contact information for public inquiries	Non-sensitive
Public Website	Official public website	Non-sensitive
Latitude, Longitude, Geopoint	Geographical coordinates (used for mapping and analysis)	Sensitive
Last Report Date	Last update timestamp for therapeutic availability	Non-sensitive
Is PAP Site	Whether the site participates in a patient assistance program	Non-sensitive
Is Telehealth Site	Whether the provider offers telehealth services	Non-sensitive
Telehealth Website	URL for telehealth services	Non-sensitive
Pharmacist Prescribing	Whether pharmacists at this site can prescribe therapeutics	Non-sensitive
Home Delivery, Home Delivery URL	Whether home delivery is available and the corresponding URL	Non-sensitive
Is T2T Site, Is ICATT Site	Designations for specific government programs	Non-sensitive
Has USG Product / Commercial Product	Availability of U.S. government or commercial therapeutics	Non-sensitive
Has Paxlovid / Lagevrio / Veklury (and their product sources)	Therapeutics availability flags	Non-sensitive
Grantee Code, Provider Note	Administrative metadata or internal notes	Non-sensitive