UHASSELT

KNOWLEDGE IN ACTION

Faculteit Geneeskunde en Levenswetenschappen

master in systeem-en procesinnovatie in de gezondheidszorg

Masterthesis

An algorithmic solution for enhanced data sharing in open science: a systematic review and diagnostic accuracy test

Lore Menten

Scriptie ingediend tot het behalen van de graad van master in systeem-en procesinnovatie in de gezondheidszorg

PROMOTOR :

Prof. dr. ir. Liesbet PEETERS

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

2023
2024

UHASSELT

KNOWLEDGE IN ACTION

Faculteit Geneeskunde en Levenswetenschappen

master in systeem-en procesinnovatie in de gezondheidszorg

Masterthesis

An algorithmic solution for enhanced data sharing in open science: a systematic review and diagnostic accuracy test

Lore Menten

Scriptie ingediend tot het behalen van de graad van master in systeem-en procesinnovatie in de gezondheidszorg

PROMOTOR :

Prof. dr. ir. Liesbet PEETERS

Contents

List of tables

List of figures

Glossary

Term	Definition
Attribute disclosure	The attacker learns something about the individual that is not public knowledge without fully re-identifying their entry within the dataset [1].
Background knowledge attack	This attack exploits the correlation between one or more quasi-identifiers and the sensitive attribute to narrow down the range of possible values for the sensitive attribute [2].
De-identification	The processes used to remove the association between a person's identity and the data obtained from them, aiming to prevent the disclosure of their identity. The objective is to minimise the risk of connecting the data to an individual to a statistically insignificant level [1, 3].
Direct identifier (DID)	Information that can be directly linked to an individual's identity [1, 3].
Homogeneity attack	This attack occurs when all the sensitive attribute values within an equivalence class are identical. Despite the data being k-anonymised, the sensitive attribute for the equivalence class can still be predicted [2].
Identity disclosure	Full re-identification of individuals within the dataset can happen when an attacker can link a specific data item to a specific individual [1, 3].
Inferential disclosure	Occurs when information can be inferred with high confidence from statistical properties of the released data [3].
Linking attack	Attacks performed by connecting information from two datasets. This could also involve linking two de-identified datasets that originated from the same raw data to gain enough information to identify individuals that appear in both datasets. [1]
Quasi-identifier (QID)	Information that on its own cannot identify a person, but when combined it can lead to identification of an individual [1, 3].
Re-identification attack	The process of attempting to discern the identities that have been removed from de-identified data [3].
Sensitive attribute (SA)	Any piece of information related to an individual that, if exposed, could cause discrimination, harm or other negative outcomes for the individual associated with the data [4].
Similarity attack	When the sensitive attribute values within an equivalence class are different yet semantically similar, an adversary can learn important information [5].
Skewness attack	This attack occurs when the distribution of records within equivalence classes is significantly skewed. Despite meeting diversity requirements, certain equivalence classes exhibit extreme imbalances in the distribution of sensitive attribute values, leading to certain individuals having a higher or lower likelihood of possessing a particular sensitive attribute compared to the overall population [5].

Abstract

Background: High-quality real-world data (RWD) is crucial for various healthcare applications, but this data requires transformations to be shared in compliance with the General Data Protection Regulation. This regulation is known for being ambiguous, posing challenges for implementation. This thesis explored the following research questions:

- Given the already present definitions entailing quasi-identifiers and sensitive attributes, how can the identification process for these types of attributes be standardised using an algorithmic approach?
- What are the methods used to evaluate data usefulness of an anonymised dataset?

Methods: A systematic literature review was conducted using ProQuest and PubMed. English peer-reviewed publications about structured and tabular data were included, while books were excluded. Insights were integrated into the data anonymisation process, which was coded, validated, and tested using two mock datasets.

Results: Two and five publications were identified for the first and second research question, respectively. The pipeline has three stages: identification, de-identification, and quasi-identifier dimension. The identification stage contains calculating g-distinct, calculating re-identification risk, and classifying attributes. The quasi-identifier dimension stage measures k-anonymity, ℓ-diversity, t-closeness, usefulness and privacy in the de-identified dataset. Non-uniform entropy was identified as the usefulness metric. The experiment demonstrated that the pipeline is compatible with RWD.

Discussion: This thesis provides a publicly available tool for attribute identification and measuring data usefulness, contributing to the standardisation of the data anonymisation process. The findings underscore the necessity of combining methodologies, securing a robust design together with the development of open-source tools.

Introduction

Real-world data (RWD) is a term that has recently gained attention and is defined in multiple ways. In this thesis, RWD refers to health data routinely gathered from different sources within healthcare services, as opposed to data collected in experimental settings. Disease registries contribute to the accumulation of RWD through various means. This can involve directly collecting information from patients or aggregating data extracted from electronic health record systems. In some cases, a blend of both methods may be used to ensure comprehensive data collection [6].

High-quality RWD is not only essential in patient care, but also in quality improvement, safety monitoring, and research [6]. Sharing data enhances confidence and trust in research findings while also enabling reproducibility and promoting the exploration of new hypotheses. By sharing data, the efficiency of progress can be maximised by preventing unnecessary duplication and leveraging insights gained from each trial. Additionally, it satisfies the moral obligation of researchers towards participants and brings benefits to many stakeholders. As awareness regarding the importance of data sharing grows, numerous global initiatives are advocating for medical data sharing. These efforts are paving the way for open science while simultaneously safeguarding the privacy rights of patients [7].

When discussing data sharing, the General Data Protection Regulation (GDPR) inevitably enters the conversation. The GDPR stands as a cornerstone in safeguarding data privacy, forming a framework for managing, processing, and protecting personal data [8]. To share data in a GDPR-compliant manner, an original dataset must undergo transformations as shown in Figure 1. Typically, this process involves identifying attributes as either direct identifiers (DIDs), quasi-identifiers (QIDs) or sensitive attributes (SAs), applying de-identification strategies, and employing anonymisation techniques. When these steps are completed, an anonymised dataset is obtained. Lastly, a bias assessment check can be performed to assess data usefulness [9].

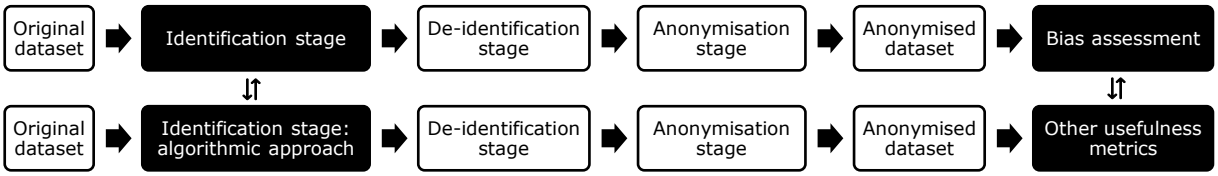


Figure 1: The data anonymisation process and contribution of this master's thesis

The research gap exists within the identification stage. The identification of DIDs is fairly easy because of their clear definition (see Glossary). For QIDs and SAs however, the definitions are more abstract (see Glossary). The GDPR is often criticised for its ambiguous language, which results in various interpretations during its implementation [10]. Organisations face significant challenges in grasping what GDPR-compliance means and figuring out how to put it into practice [11]. Some researchers even state that the GDPR has needlessly complicated the functioning of research biobanks and related data operations, without significantly enhancing privacy [12]. Because of these complexities, there is no standardised methodology to identify these attributes in a dataset, with interpretations of the GDPR guidelines being the foundation for selection. This approach is insufficient since any approach based on available definitions for QIDs may lead to re-identification [13].

This thesis aims to contribute to the standardisation of the anonymisation process in order to streamline data sharing efforts. To achieve this goal, the focus will be on two key steps of the data sharing process. Firstly, the identification stage will be executed using an algorithmic approach based on mathematical logic rather than interpretations of the GDPR guidelines. Secondly, additional data usefulness metrics will be identified. The contribution of this thesis is presented in Figure 1.

These aims result in two research questions:

- Given the already present definitions entailing quasi-identifiers and sensitive attributes, how can the identification process for these types of attributes be standardised using an algorithmic approach?
- What are the methods used to evaluate data usefulness of an anonymised dataset?

Methods

This thesis forms a bridge between healthcare and computer science research, requiring two methodologies. The decision to employ two methodologies rather than adhering to a single design was based on the need for a comprehensive literature review and validation of these literature findings. This approach was crucial to maintain a balance between a strong design and practical significance. Consequently, some sections were omitted or modified in both guidelines.

Firstly, a systematic literature review (SLR) was conducted and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline [14]. The Parsif.al tool was used to facilitate the review process [15, 16]. Since this SLR is a component of a thesis, all steps were carried out by one person unless stated otherwise.

Secondly, the findings from the SLR were implemented into the data anonymisation process in Figure 1. The pipeline was transformed into code, validated and subsequently tested on two mock datasets. To report this step, the Standards for Reporting of Diagnostic Accuracy (STARD) guideline was used [17]. The diagnosis was interpreted as the classification of attributes and the pipeline represented the diagnostic test. As data acquisition fell outside the scope of this thesis, segments relating to participants were substituted with information regarding the experimental dataset.

Literature review

Inclusion and exclusion criteria

To search for relevant publications, inclusion and exclusion criteria were applied. These criteria were formulated for study and publication details. Regarding the study details, publications were required to address structured and tabular data due to the scope of this thesis. Considering the publication details, publications had to be peer-reviewed to ensure their quality and reported in English. The only exclusion criterium used was the exclusion of books to mitigate potential accessibility issues.

Search strategy

The electronic databases ProQuest and PubMed served as information sources. Publications were gathered between December 25th of 2023 and February 11th of 2024. Because there are two research questions, two search strings were composed. To narrow down the search, field codes were applied to ensure that search terms appeared in either the title or the abstract of the publication. No Medical Subject Headings (MeSH-terms) were used to reduce the possibility of missing recent publications, since there is a delay between the time a publication is published and the moment that publication is labelled with MeSH-terms. Table 1 represents these search strings. Because of the inclusion and exclusion criteria, filters were applied to include only peer-reviewed publications written in English while excluding books.

Table 1: Search strings

	Given the already present definitions entailing quasi-identifiers and sensitive attributes, how can the identification process for these types of attributes be standardised using an algorithmic approach?	**What are the methods used to evaluate data usefulness of an anonymised dataset?**
ProQuest	TIAB(("quasi-identifier" OR "QID" OR "QI") AND ("recogni*" OR "classif*" OR "detect*" OR "discover*" OR "identif*" OR "find*" OR "solv*") AND ("algorithm") AND ("privacy"))	TIAB(("anonymi*") AND ("metric" OR "assess*" OR "evaluat*" OR "measur*") AND ("data usefulness" OR "data quality" OR "data utility"))
PubMed	((quasi-identifier[Title/Abstract]) OR (QID[Title/Abstract]) OR (QI[Title/Abstract])) AND ((recogni*[Title/Abstract]) OR (classif*[Title/Abstract]) OR (detect*[Title/Abstract]) OR (discover*[Title/Abstract]) OR (identif*[Title/Abstract]) OR (find*[Title/Abstract]) OR (solv*[Title/Abstract])) AND (algorithm[Title/Abstract]) AND (privacy[Title/Abstract])	(anonymi*[Title/Abstract]) AND ((metric[Title/Abstract]) OR (assess*[Title/Abstract]) OR (evaluat*[Title/Abstract]) OR (measur*[Title/Abstract])) AND ((data usefulness[Title/Abstract]) OR (data quality[Title/Abstract]) OR (data utility[Title/Abstract]))

Selection and management of the publications

The selection process was documented in a flowchart, visualised by Figure 2. Publications were initially screened on title and abstract to sift through the vast pool of literature available, focusing on identifying publications that directly correlated with the research objectives outlined. If the publications aligned with the research objective, the full text was read and reviewed. EndNote was used to effectively manage the publications [18].

Data items and data collection

The aims of this thesis played a pivotal role in shaping the outcomes of the SLR. The first outcome included an algorithmic approach to identify attributes. Any metrics for data usefulness served as a second outcome. Two data extraction forms were constructed to facilitate the data extraction process. Appendix 1 contains these forms.

Quality assessment

The quality of the included publications was assessed using quality assessment checklists. Because there were two research questions which targeted different kinds of publications, two separate checklists were used. In the absence of validated checklists for the type of publications in question, custom checklists were developed based on available literature and checked by the day-to-day supervisor [19]. Appendix 2 represents both checklists.

Pipeline

Another important outcome was a pipeline, which was depicted as a flowchart and would be transformed into an open-source tool if the obtained publications were sufficiently detailed. This tool was made publicly available to facilitate the use of discovered methods for attribute identification and resulting data usefulness measures. The data anonymisation process (Figure 1) was used as a blueprint for the pipeline. Explanations of the identification stage and usefulness metrics were included in the results, since these were influenced by the SLR. This section explains de-identification and anonymisation, which were not addressed in the SLR but are an essential part of the pipeline. Appendix 3 and Appendix 4 provide visual representations of respectively the de-identification strategies, and k-anonymity and ℓ-diversity.

De-identification

Suppression, masking, generalisation, and aggregation are common strategies employed to protect privacy. Suppression involves removing a value or attribute, which maximises privacy protection but often reduces data usefulness. Health data protection standards often necessitate some degree of suppression [1, 3]. Masking obscures data so that the original values cannot be readily obtained. This can involve replacing a value or a part of it with placeholders, such as asterisks or xs [3]. Generalisation enhances privacy by reducing the specificity of information, thereby decreasing its granularity. For instance, values may be represented within a range such as an age range. An important consideration is that increasing the generalisation tends to reduce data usefulness as detailed information is lost [1, 3]. Aggregation refers to the process of collecting or grouping together raw data, meaning that either statistics about the data are disclosed or values are merged. This approach allows the release of summary statistics or information about small groups within a dataset, rather than revealing the entire dataset [1].

K-anonymity

K-anonymity serves as a privacy protection model against linking attacks, thus preventing identity disclosure (see Glossary) [3, 20]. The principle is that each record is indistinguishable from at least k-1 other records for every combination of QIDs. Thus, k-anonymity offers privacy protection by ensuring that each released record relates to at least k individuals, even in cases where the records are directly linked to external information. This group of indistinguishable individuals is called an equivalence class [3, 21].

ℓ-diversity

ℓ-diversity was used because k-anonymous datasets are still vulnerable to several re-identification attacks, such as homogeneity attacks and background knowledge attacks (see Glossary). This technique helps protect against inferential disclosure by assuring diversity of SAs within an equivalence class. A table is considered ℓ-diverse if each equivalence class contains at least ℓ well-represented values for the SA [2, 3].

T-closeness

T-closeness mitigates the risks still present in the ℓ-diversity model during skewness attacks and similarity attacks (see Glossary). An equivalence class is considered to exhibit t-closeness if the distance between the distribution of SAs within this class and the distribution of it in the whole table does not exceed a threshold t. A table is deemed to have t-closeness if all equivalence classes within it exhibit t-closeness [5].

Validation

The identification stage in the state of the art example was outsourced and the original dataset is no longer accessible, making it difficult to validate the pipeline using this example [9]. However, if the papers derived from the SLR provided sufficient transparency regarding their experiments, the pipeline would be validated by using datasets from these source papers.

Experiment

Experimental dataset

The experimental dataset was a mock dataset designed to closely resemble an RWD dataset [9]. The dataset consisted of 17 attributes (also referred to as columns) and was created with 500 and 1000 rows using a mock data generator [22]. Two datasets were created to analyse the effect of dataset size on the used metrics. Appendix 5 contains a description of every attribute.

Missing data

An important practical consideration was the high likelihood of missing values, since RWD is characterised by low data quality [23]. A function was designed to handle missing values. Firstly, the missing values are filled in with the string "missing". The goal was to calculate the percentage of missing values in an attribute and subsequently drop attributes if more than 85% of the data was missing [24]. Filling missing values with a value enables the creation of an extended match within k-anonymity, allowing missing values to be matched with other missing values. This is different from a basic match, which is traditionally used in k-anonymity, where missing values do not match with other missing values, nor with any other value [25].

Results

Literature review

Study selection

The selection of the publications for both research questions was documented in flowcharts, as presented in Figure 2. This flowchart was based on the flowchart created by PRISMA [14]. Appendix 6 contains the filled out data extraction forms of all included publications.

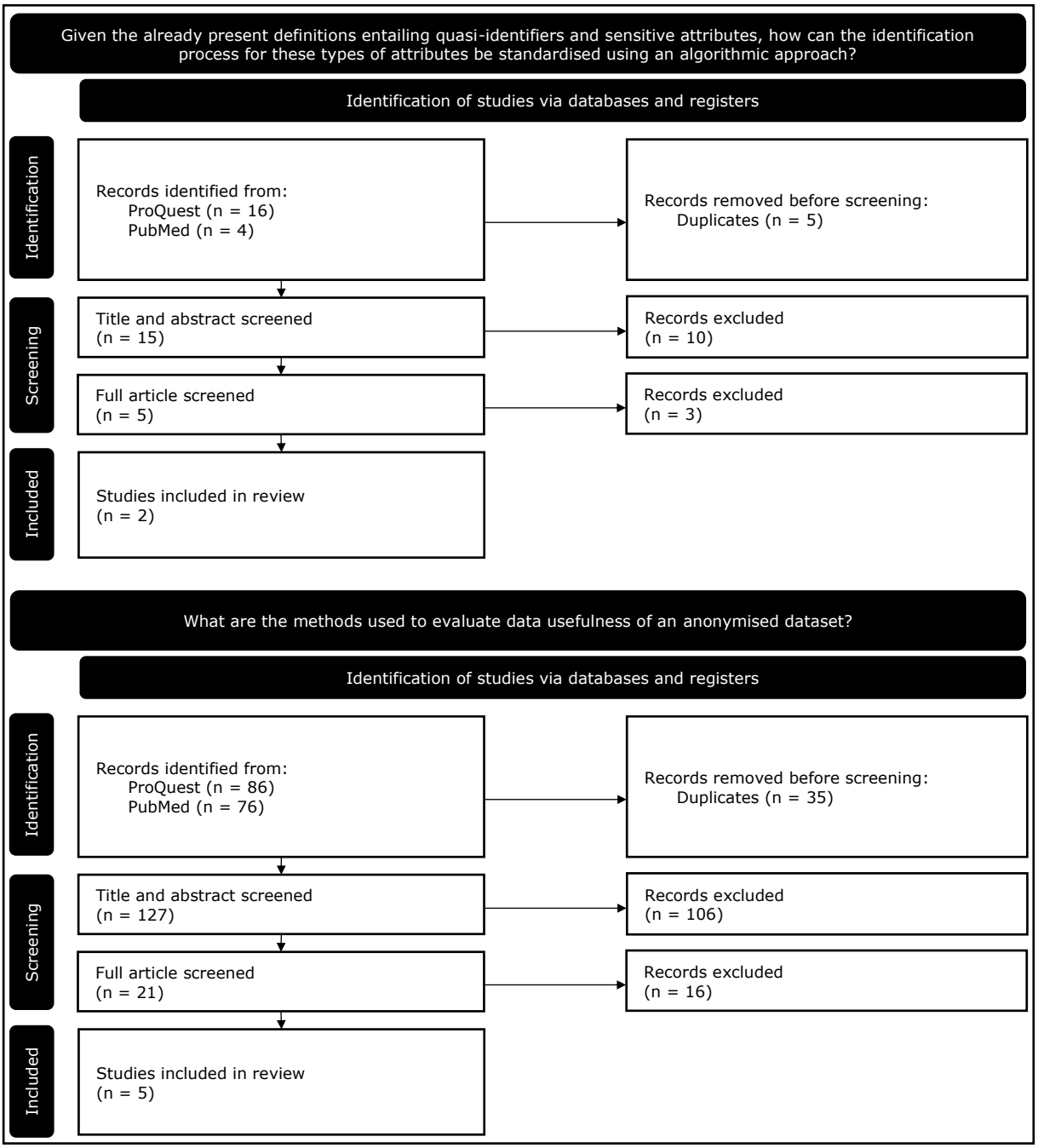


Figure 2: Flowchart of the selection process

The search strings of the first research question (see Table 1) yielded a total of 20 publications, with 16 publications from ProQuest and 4 from PubMed. Of these publications, five were duplicates. After screening the title and abstract of the remaining 15 publications, 10 publications were excluded. The full text of the five residual publications was reviewed, where three more publications were excluded, leaving two publications included in this SLR.

For the second research question (see Table 1), 86 publications were identified in ProQuest and 76 in PubMed, resulting in a total of 162 publications. Out of these publications, 35 were duplicates. Screening the titles and abstracts of these 127 publications led to the exclusion of 106 publications. The full text of the 21 remaining publications was reviewed and 16 more publications were excluded. This resulted in five publications being included in this SLR.

Quality assessment

The quality of each article was assessed using two custom quality assessment checklists (see Appendix 2). For the first research question, both articles scored five out of eight as they did not incorporate machine learning and lacked a code repository. The five articles obtained from the second research question received more divergent scores. The lowest score was zero out of five. Three articles received a score of three out of five due to not validating the methodology and the absence of a code repository. One article provided a repository, resulting in the highest quality score of four out of five. Appendix 7 shows the filled out checklists.

Pipeline

The pipeline was based on the process represented in Figure 1. Figure 3 represents a more detailed flowchart of the different steps. The structure of this pipeline was influenced by a publication retrieved for the first research question [26]. The pipeline consists of three major steps: the identification stage, the de-identification stage, and the QID dimension stage.

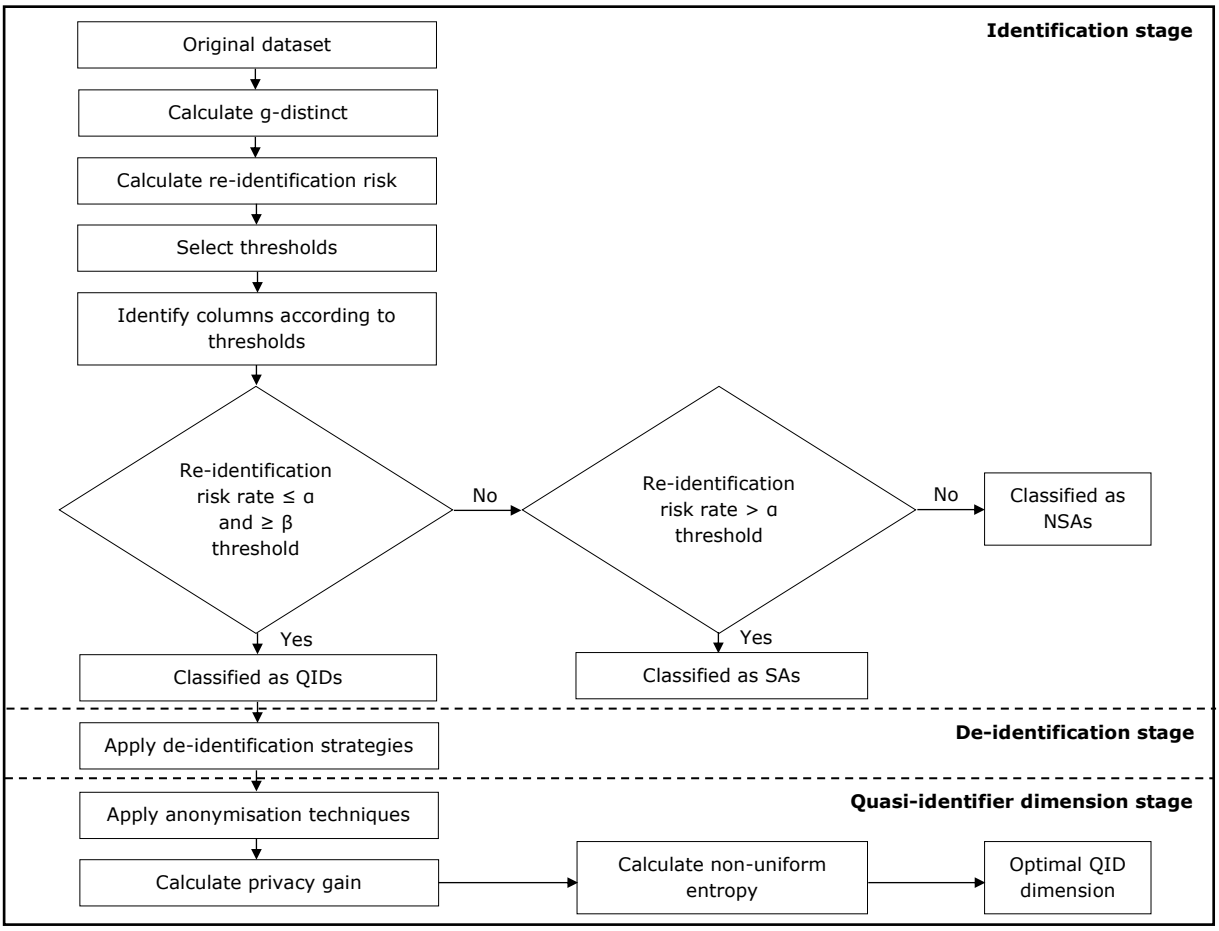


Figure 3: Flowchart of the pipeline

The code for this pipeline was made available in a GitHub repository. The identification stage and QID dimension stage are available as two separate Python files. De-identification was not included, as these strategies were discussed in the methods and were not the focus of this thesis. Additionally, the repository includes a README that contains a tutorial video, as well as the prerequisites and steps for each file. The README is available in Appendix 8. Further information about this repository is provided in the code availability section.

The identification stage

The two publications from the first research question lead to the identification of three steps for the identification stage: calculate g-distinct, calculate re-identification risk, and classify attributes according to re-identification risk thresholds [26, 27]. The calculation of the g-distinct values is based on the idea of uniqueness. Each g-distinct value represents the uniqueness of a value within an attribute. Therefore, there are as many g-distinct values in an attribute as there are unique values. Based on these g-distinct values, the re-identification risk rate of every attribute can be calculated. This risk rate is determined by the sum of all g-distinct values within an attribute. Subsequently, the α and β thresholds need to be established. These thresholds act as cut-off values to classify attributes as QIDs, SAs or non-sensitive attributes (NSAs). Attributes surpassing the α threshold are labelled as SAs. Attributes with a risk rate lower than or equal to α but higher than or equal to β are considered QIDs. The remaining attributes with a risk rate below β are classified as NSAs [26]. The classified attributes serve as the output of this stage. The code is represented in Figure 4.

```python
# Function to calculate g-distinct
def compute_g_distinct(df):
    n_cols = len(df.columns)
    n_rows = len(df)
    g_distinct_dict = {}

    for col in range(n_cols):
        cols_values_to_g = {}
        cols_values = df.iloc[:, col]
        cols_values_to_amounts = Counter(cols_values)
        for row in range(n_rows):
            value = df.iloc[row, col]
            value_count = cols_values_to_amounts[value]
            if value not in cols_values_to_g:
                cols_values_to_g[value] = 1 / value_count
        g_distinct_dict[df.columns[col]] = list(cols_values_to_g.values())

    return g_distinct_dict


# Function to calculate re-identification risk rate
def calculate_reidentification_risk(g_distinct_values):
    reidentification_risk_rates_dict = {}

    for key in g_distinct_values:
        g_distinct_attr = g_distinct_values[key]
        length = len(g_distinct_attr)
        reidentification_risk_rates_dict[key] = round(
            (np.sum(g_distinct_attr) / length) * 100, 2
        )
    return reidentification_risk_rates_dict
```

Figure 4: Code for the identification stage

The de-identification stage

Attributes identified as QIDs in the previous stage undergo de-identification in descending order of their risk rates. This implies that the QID with the highest risk rate will be de-identified first, followed by those with progressively lower risk rates, ending with the QID with the lowest risk rate. Where appropriate, SAs can also be de-identified. De-identification strategies can be found in the methods.

The quasi-identifier dimension stage

The de-identified dataset is measured in terms of k-anonymity, ℓ-diversity, t-closeness, usefulness and privacy. Based on the privacy gain and usefulness metric, an optimal selection of QIDs is suggested as represented in Figure 5. This selection process also takes into consideration a k-anonymity level of at least two. This means that options with k-anonymity of less than two are not considered as a potential best QID dimension. A publication resulting from the first research question served as a blueprint for the selection of optimal QIDs [26]. The privacy measurement was also adopted from this study, since the second research question only focussed on metrics for usefulness.

```python
# Function to calculate the optimal QID dimension
def find_optimal_qid_dimension(qid_dimension_list):
    filtered_list = list(filter(filter_k_anonymity, qid_dimension_list))
    max_record = filtered_list[0]

    for item in filtered_list:
        current_pg = item["pg"]
        current_nue = item["inverse_nue"]
        current_k = item["k_anonymity_after"]
        if (
            current_pg >= max_record["pg"]
            and current_nue >= max_record["inverse_nue"]
            and current_k >= 2
        ):
            current_difference = abs(current_pg - current_nue)
            min_difference = abs(max_record["pg"] - max_record["inverse_nue"])
            if current_difference < min_difference:
                max_record = item

    return max_record["qid_dimension"]
```

Figure 5: Code for the quasi-identifier dimension stage

Several usefulness metrics were discovered in the five included publications from the second research question [28, 29, 30, 31, 32]. Metrics considered to be used were non-uniform entropy (NUE), utility criterion and clustering. NUE received the best results for general purpose usage and was the best documented approach [29, 32]. Hence, this model was chosen as the usefulness measure. NUE is calculated by analysing the frequencies of attribute values in the de-identified dataset and comparing them to those in the original dataset. This means that NUE quantifies information loss. To represent the opposite effect, an inverse NUE was incorporated into the pipeline. Figure 6 represents the code for NUE.

```python
# Function to compute non-uniform entropy
def compute_non_uniform_entropy(df_original, df_after, selected_quasi_identifiers):
    non_uniform_entropy = 0

    for column in selected_quasi_identifiers:
        if column in df_original.columns and column in df_after.columns:
            original_values = df_original[column]
            after_values = df_after[column]
            original_values_to_amount = Counter(original_values)
            after_values_to_amount = Counter(after_values)
            for index in range(len(original_values)):
                original_value = original_values[index]
                after_value = after_values[index]
                original_amount = original_values_to_amount[original_value]
                after_amount = after_values_to_amount[after_value]
                if original_amount > 0 and after_amount > 0:
                    ratio = original_amount / after_amount
                    non_uniform_entropy -= np.log(ratio)

    return non_uniform_entropy
```

Figure 6: Code for non-uniform entropy

Validation

The absence of a code repository in the source paper necessitated the development of code from scratch, with the exception of the de-identification and anonymisation strategies outlined in the methods. The results of the experiment were reported in sufficient detail to use as a validation for the code [26]. Nevertheless, relying solely on pseudocode and term definitions when developing the code can lead to inconsistencies between this pipeline and the one described in the source paper. When the created pipeline was tested with one dataset from the source paper, the re-identification risk rates were similar. However, the classification when using the same thresholds was different. Reflections on these results are represented in the discussion.

Experiment

The experiment consists of two steps, namely the identification stage and the QID dimension stage of the discovered pipeline.

The output of the identification stage consists of re-identification risks for each attribute, together with a classification of these attributes according to the re-identification risk thresholds. For the first dataset which contained 500 rows, the α threshold was set to 25% and the β threshold to 1%. As for the second dataset with 1000 rows, α was set to 10% and β to 1%. Appendix 9 demonstrates the classification for these datasets. The attribute "covid19_self_isolation" was excluded because the missing values surpassed the predefined threshold of 85%, reaching 88% in the first dataset and 91.8% in the second dataset. The attribute "secret_name" was also suppressed, since this was a DID.

The QID dimension stage was executed after de-identification of QIDs and SAs in both datasets. For the first dataset, the optimal QID dimension was five, indicating that all QIDs needed to be de-identified. This was primarily because k-anonymity for the other dimensions was below two. For this dimension, k-anonymity was four and ℓ-diversity was two for both SAs. The t-closeness value was 0.74. K-anonymity before anonymisation was one, resulting in a privacy gain of three. NUE was 69.26% and the inverse NUE was 30.74%.

The optimal QID dimension in the second dataset was three, implying that all three QIDs should undergo de-identification. K-anonymity was six when two QIDs were de-identified, satisfying the minimum requirement of two to ensure that no person remains unique. This resulted in a privacy gain of five. ℓ-diversity varied across QIDs, with "bmi" and "ms_diagnosis_date" having a value of three and "edss" having a value of two. The t-closeness value was 0.61. NUE was 53.61% and the inverse NUE was 46.39%. After de-identifying the third QID, k-anonymity significantly improved from 1 to 110, resulting in a privacy gain of 109. ℓ-diversity for "bmi", "ms_diagnosis_date" and "edss" was three, six and two respectively, while t-closeness decreased to 0.32. NUE was 69.05%, with an inverse NUE of 30.95%.

Discussion

This master's thesis addresses the following research questions using an SLR and experiment:

- Given the already present definitions entailing quasi-identifiers and sensitive attributes, how can the identification process for these types of attributes be standardised using an algorithmic approach?
- What are the methods used to evaluate data usefulness of an anonymised dataset?

In the SLR, a method to identify attributes based on re-identification risk was found, alongside NUE as a well-supported usefulness metric. Re-identification risk seems just one of countless methods created to identify attributes, with every method claiming superiority over the others by minimising information loss or supposedly being more efficient [33, 34, 35, 36]. However, a common drawback among these methods is the lack of publicly available code, exemplifying that these publications fail to contribute to the open science culture. While most methods provide pseudocode, this significantly affects the usability of these methods as they still have to be translated to the desired programming language. In this thesis, the results were not fully identical to those in the source paper during the validation of the code. However, this source paper had methodological flaws. An example is the β threshold being set to zero, meaning that NSAs would have to be negative. This is not possible as risk rates are strictly positive. Another example is the superficial explanation of some steps of the pipeline, like the transformation of the re-identification risk rate to a percentage [26].

As for usefulness metrics, there are also many options to choose from. This became apparent during the SLR, where various metrics were identified from which one was implemented due to time and knowledge constraints. Unlike attribute identification methods, usefulness metrics are well-documented and publicly available. An example of existing open-source software is ARX [37]. It supports a variety of privacy models, data transformation models, and utility and risk analysis techniques. They provide code to compute re-identification risks for various attacker models and numerous data quality models. However, all these methods are stored in individual files in their repository, resulting in a lack of cohesion compared to the pipeline developed in this thesis. A major weakness of this software is that the code is only available in Java. Python is the programming language of choice for data scientists and developers in data analysis and numerical computations [38, 39]. Consequently, the software is less usable in these specific fields.

The most apparent observation in the experiment is that as dataset size increases, the re-identification risk decreases. This is logical, as values lose their uniqueness when they are present in a larger number of records. Consequently, smaller datasets inherently carry a greater re-identification risk, necessitating stronger de-identification strategies to satisfy required anonymity levels. This leads to a decrease in overall utility when weighed against the privacy gain [40]. In both datasets, the NUE was similar for the best QID dimensions, with 69.26% in the 500 row dataset and 69.05% in the 1000 row dataset. However, the privacy gain in the 500 row dataset was merely 3, whereas the 1000 row dataset achieved a privacy gain of 109. Therefore, even though the utility remains comparable between the two datasets, the privacy gain differs significantly.

Strengths

A major strength of this thesis is its methodology. The use of an SLR results in the incorporation of components that contribute to the quality of the methodology. Examples include the use of a predefined search string across preselected databases and quality assessments of included publications. This systematic search for literature also provides a solid scientific foundation for the pipeline. The use of an experiment further enhances the significance of the findings within the practical domain. Additionally, the reporting quality is elevated by the transparent documentation of the decisions made.

In the SLR itself, the inclusion of solely peer-reviewed publications enhances the quality of the results. The majority of the publications scored well on the quality assessment, which further improves the quality of this thesis. Most publications lost points due to a lack of methodology validation and the absence of a code repository.

The greatest asset of this thesis is its contribution to the open science culture. The code is available on GitHub, complete with comprehensive documentation and mock datasets. The inclusion of mock datasets that mimic real privacy risks without compromising actual privacy provides a safe way to develop skills in handling RWD. This significantly enhances the educational value of this thesis. The use of realistic datasets also ensures that the algorithm can effectively handle low-quality data often found in such datasets [23]. Moreover, the code can be applied to any dataset, making it an invaluable tool for people working with data anonymisation. The QID dimension stage is designed to accept datasets and calculate necessary parameters for comparing them. With some adjustments, it could compare two datasets where the same QID is de-identified differently, allowing users to determine the most effective de-identification strategy. The code was also developed with user-friendliness in mind, utilising user input prompts instead of requiring manual code changes.

Limitations

For the SLR, an important limitation is the limited amount of publications retrieved by the search strings. However, this was to be expected considering the relatively recent emergence of privacy regulations like the GDPR. Regardless, it is essential to reflect on whether this is a result of flaws in the search strings or in the article selection process. Another flaw is that almost all steps were executed by a single individual, which is also to be expected in the context of a thesis. Ideally, an SLR is executed by at least two researchers.

As for the pipeline, the arbitrary selection of the α and β thresholds poses a major challenge to the objectivity of the results. Especially the idea that SAs have a higher re-identification risk than QIDs should be interpreted with caution. This assumption has not been thoroughly tested, making methods based on this assumption less credible. The fact that there is no consensus about where these thresholds should be proves again that the definitions of QIDs and SAs are not clear enough.

The use of two reporting guidelines complicates the assessment of this thesis in terms of reporting quality. However, it is important to note that this thesis is written to obtain a master's degree in healthcare engineering, a field that is not completely a healthcare discipline nor an engineering field of study. This thesis in particular forms a bridge between healthcare and computer science research. This makes the application of guidelines from the EQUATOR network difficult, since this network primarily addresses transparency in the reporting of health research.

Recommendations

Further research is necessary to develop more objective methods for selecting re-identification risk thresholds. Regardless, the risk rates can be used to guide the order in which attributes should be de-identified. While measures of usefulness have been extensively documented, greater emphasis should be placed on creating comprehensive pipelines that integrate all steps of the anonymisation process, rather than solely focusing on generating code for individual usefulness measures.

Additionally, open-source tools should not be considered a nice-to-have but a necessity for publications in this field of study. The publication of code for new methods should be encouraged, since this promotes reproducibility, enhances usability, and contributes to the open science culture. Specific research in healthcare informatics should be supported, with a focus on combining methodologies to ensure a robust design together with the development of open-source tools.

Conclusion

The results of this thesis are promising, demonstrating that objective ways to identify attributes exist and various data usefulness metrics are available. The drawback is not the scarcity of methods but rather the absence of open-source tools to apply them effectively. In this regard, this thesis made a major contribution by providing a tool while also using robust methodology to design this tool. Since the code can be applied to various datasets and customised with minimal adjustments, it provides flexibility for users dealing with different types of data. This adaptability makes it a versatile tool for numerous applications.

Code availability

The pipeline was developed using Python and split up into two files. The first file contains the identification stage, while the second file contains the QID dimension stage. The Python libraries used are pandas 2.1.4, NumPy 1.26.2 and pyCANON 1.0.1.post2. The datasets from the experiment are provided in a CSV format. Users can access the GitHub repository at https://github.com/LoreMenten/Attribute-Identification-And-Utility-Metrics-Pipeline.

References

1. Krehling L. De-Identification Guideline. Canada: Western University; 2020.

2. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data. 2007;1(1):3–es.

3. Garfinkel SL. De-Identification of Personal Information. National Institute of Standards and Technology; 2015.

4. What Is Sensitive Data? Available from: https://www.paloaltonetworks.com/cyberpedia/sensitive-data. [Accessed 1st June 2024].

5. Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. 2007 IEEE 23rd International Conference on Data Engineering. 2007:106-15.

6. Peeters LM, Parciak T, Kalra D, Moreau Y, Kasilingam E, van Galen P, et al. Multiple Sclerosis Data Alliance - A global multi-stakeholder collaboration to scale-up real world data research. Multiple Sclerosis and Related Disorders. 2021;47:102634.

7. Hulsen T. Sharing Is Caring-Data Sharing Initiatives in Healthcare. International Journal of Environmental Research and Public Health. 2020;17(9).

8. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), (2016).

9. Khan H, Geys L, Baneke P, Comi G, Peeters LM. Patient level dataset to study the effect of COVID-19 in people with Multiple Sclerosis. Scientific Data. 2024;11(1):149.

10. Lindqvist J. New challenges to personal data processing agreements: is the GDPR fit to deal with contract, accountability and liability in a world of the Internet of Things? International Journal of Law and Information Technology. 2017;26(1):45-63.

11. Sirur S, Nurse JRC, Webb W. Are We There Yet? Understanding the Challenges Faced in Complying with the General Data Protection Regulation (GDPR). Proceedings of the 2nd International Workshop on Multimedia Privacy and Security; Toronto, Canada: Association for Computing Machinery; 2018. p. 88–95.

12. Peloquin D, DiMaio M, Bierer B, Barnes M. Disruptive and avoidable: GDPR challenges to secondary research uses of data. European Journal of Human Genetics. 2020;28(6):697-705.

13. Bettini C, Wang XS, Jajodia S. The Role of Quasi-identifiers in k-Anonymity Revisited. ArXiv. 2006;abs/cs/0611035.

14. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews.

15. Simple Complex. Parsif.al; 2021. Available from: https://parsif.al/. [Accessed 13th January 2024].

16. Carrera-Rivera A, Ochoa W, Larrinaga F, Lasa G. How-to conduct a systematic literature review: A quick guide for computer science research. MethodsX. 2022;9:101895.

17. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies.

18. The EndNote Team. EndNote. EndNote 20 ed. Philadelphia, PA: Clarivate; 2013.

19. Yang L, Zhang H, Shen H, Huang X, Zhou X, Rong G, et al. Quality Assessment in Systematic Literature Reviews: A Software Engineering Perspective. Information and Software Technology. 2021;130:106397.

20. Sweeney L. k-anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002;10(5):557–70.

21. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002;10(5):571–88.

22. Pirmani A. GDSI-Mock-DataGenerator; 2023. Available from: https://github.com/MS-DATA-ALLIANCE/GDSI-Mock-DataGenerator. [Accessed 9th May 2024].

23. Grimberg F, Asprion PM, Schneider B, Miho E, Babrak L, Habbabeh A. The Real-World Data Challenges Radar: A Review on the Challenges and Risks regarding the Use of Real-World Data. Digital Biomarkers. 2021;5(2):148-57.

24. Durgapal A. Data Preprocessing — Handling Missing Values in a dataset; 2023. Available from: https://medium.com/@ayushmandurgapal/data-preprocessing-handling-missing-values-in-a-dataset-5140f77d2a47. [Accessed 1st June 2024].

25. Ciglic M, Eder J, Koncilia C. Anonymization of Data Sets with NULL Values. In: Hameurlain A, Küng J, Wagner R, Decker H, Lhotska L, Link S, editors. Transactions on Large-Scale Data- and Knowledge-Centered Systems XXIV: Special Issue on Database- and Expert-Systems Applications. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016. p. 193-220.

26. Mansour HO, Siraj MM, Ghaleb FA, Saeed F, Alkhammash EH, Maarof MA. Quasi-Identifier Recognition Algorithm for Privacy Preservation of Cloud Data Based on Risk Reidentification. Wireless Communications & Mobile Computing (Online). 2021;2021.

27. Jadhav PS, Borkar GM. Quasi-identifier recognition with echo chamber optimization-based anonymization for privacy preservation of cloud storage. Concurrency and Computation. 2024;36(2).

28. Albright JJ. Privacy Protection in Social Science Research: Possibilities and Impossibilities. Political Science & Politics. 2011;44(4):777-82.

29. Eicher J, Kuhn KA, Prasser F. An Experimental Comparison of Quality Models for Health Data De-Identification. Studies in Health Technology and Informatics. 2017;245:704-8.

30. Ferrão ME, Prata P, Fazendeiro P. Utility-driven assessment of anonymized data via clustering. Scientific Data. 2022;9.

31. Loukides G, Gkoulalas-Divanis A. Utility-preserving transaction data anonymization with low information loss. Expert Systems with Applications. 2012;39(10):9764-77.

32. Prasser F, Bild R, Kuhn KA. A Generic Method for Assessing the Quality of De-Identified Health Data. Studies in Health Technology and Informatics. 2016;228:312-6.

33. Motwani R, Xu Y, editors. Efficient Algorithms for Masking and Finding Quasi-Identifiers2007.

34. Podlesny NJ. Quasi-identifier discovery to prevent privacy violating inferences in large high dimensional datasets. Potsdam, Germany: University of Potsdam; 2023.

35. Yan Y, Wang W, Hao X, Zhang L. Finding Quasi-identifiers for K-Anonymity Model by the Set of Cut-vertex. Engineering Letters. 2018;26(1).

36. Wafo Soh C, Njilla LL, Kwiat KK, Kamhoua CA. Learning quasi-identifiers for privacy-preserving exchanges: a rough set theory approach. Granular Computing. 2020;5(1):71-84.

37. ARX - Data Anonymization Tool | A comprehensive software for privacy-preserving microdata publishing; 2024. Available from: https://arx.deidentifier.org/. [Accessed 25th May 2024].

38. Khoirom S, Sonia M, Laikhuram B, Laishram J, Davidson Singh T. Comparative Analysis of Python and Java for Beginners International Research Journal of Engineering and Technology. 2020;7(8).

39. Nagpal A, Gabrani G, editors. Python for Data Analytics, Scientific and Technical Applications. 2019 Amity International Conference on Artificial Intelligence (AICAI); 2019.

40. Lee H, Kim S, Kim JW, Chung YD. Utility-preserving anonymization for health data publishing. BMC Medical Informatics and Decision Making. 2017;17(1):104.

41. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology. 1983;33(11):1444-52.

42. Multiple Sclerosis Trust. Expanded Disability Status Scale (EDSS); 2020. Available from: https://mstrust.org.uk/a-z/expanded-disability-status-scale-edss. [Accessed 6th May 2024].

Appendices

Appendix 1: Data extraction forms

Given the already present definitions entailing quasi-identifiers and sensitive attributes, how can the identification process for these types of attributes be standardised using an algorithmic approach?	**What are the methods used to evaluate data usefulness of an anonymised dataset?**
Title	Title
Authors	Authors
Publication year	Publication year
Objectives	Objectives
Methodology	Methodology
Performance measures	Key takeaways
Key takeaways	

Appendix 2: Quality assessment checklists

Given the already present definitions entailing quasi-identifiers and sensitive attributes, how can the identification process for these types of attributes be standardised using an algorithmic approach?	**What are the methods used to evaluate data usefulness of an anonymised dataset?**
Was there a clear description of the aims and purposes of the research?	Was there a clear description of the aims and purposes of the research?
Was the algorithm clearly described (e.g. flowchart, pseudocode, ...)	Was the experimental dataset described?
Was the experimental dataset described?	Were any metrics used to validate the methodology?
Were any metrics used to validate the methodology?	Were the metrics clearly described?
Was the quality of the anonymised data assessed?	Is there a repository of the code?
Was the quality assessment done using simple statistical methods or machine learning?	
Were there any hyperparameters that were finetuned?	
Is there a repository of the code?	

Appendix 3: De-identification strategies

Suppression

Name	Postal code	Age	Sex	Diagnosis
Emma	3500	25	F	Gastric flu
Bob	3510	32	M	Flu
Tommy	3520	36	M	COVID
Michael	3530	45	M	Gastric flu
Sara	3540	23	F	Flu
Ziggy	3550	43	M	COVID

Original data

Name	Postal code	Age	Sex	Diagnosis
	3500	25	F	Gastric flu
	3510	32	M	Flu
	3520	36	M	COVID
	3530	45	M	Gastric flu
	3540	23	F	Flu
	3550	43	M	COVID

After suppression

Masking

Name	Postal code	Age	Sex	Diagnosis
Emma	3500	25	F	Gastric flu
Bob	3510	32	M	Flu
Tommy	3520	36	M	COVID
Michael	3530	45	M	Gastric flu
Sara	3540	23	F	Flu
Ziggy	3550	43	M	COVID

Original data

Name	Postal code	Age	Sex	Diagnosis
xxxx	3500	25	F	Gastric flu
xxxx	3510	32	M	Flu
xxxx	3520	36	M	COVID
xxxx	3530	45	M	Gastric flu
xxxx	3540	23	F	Flu
xxxx	3550	43	M	COVID

After masking

Generalisation

Name	Postal code	Age	Sex	Diagnosis
Emma	3500	25	F	Gastric flu
Bob	3510	32	M	Flu
Tommy	3520	36	M	COVID
Michael	3530	45	M	Gastric flu
Sara	3540	23	F	Flu
Ziggy	3550	43	M	COVID

Original data

Name	Postal code	Age	Sex	Diagnosis
Emma	3500	20-29	F	Gastric flu
Bob	3510	30-39	M	Flu
Tommy	3520	30-39	M	COVID
Michael	3530	40-49	M	Gastric flu
Sara	3540	20-29	F	Flu
Ziggy	3550	40-49	M	COVID

After generalisation

Aggregation

Name	Postal code	Age	Sex	Diagnosis
Emma	3500	25	F	Gastric flu
Bob	3510	32	M	Flu
Tommy	3520	36	M	COVID
Michael	3530	45	M	Gastric flu
Sara	3540	23	F	Flu
Ziggy	3550	43	M	COVID

Original data

Name	Postal code	Age	Sex	Diagnosis
Emma	3500	25	F	Digestive
Bob	3510	32	M	Respiratory
Tommy	3520	36	M	Respiratory
Michael	3530	45	M	Digestive
Sara	3540	23	F	Respiratory
Ziggy	3550	43	M	Respiratory

After aggregation

Appendix 4: Anonymisation techniques

k-anonymity

Direct identifier	Quasi-identifier	Quasi-identifier	Quasi-identifier	Sensitive attribute
Name	**Postal code**	**Age**	**Sex**	**Diagnosis**
Emma	3500	25	F	Gastric flu
Bob	3510	32	M	Flu
Tommy	3520	36	M	COVID
Michael	3530	45	M	Gastric flu
Sara	3540	23	F	Flu
Ziggy	3550	43	M	COVID

Original data

Direct identifier	Quasi-identifier	Quasi-identifier	Quasi-identifier	Sensitive attribute
Name	**Postal code**	**Age**	**Sex**	**Diagnosis**
	35xx	20-29	F	Gastric flu
	35xx	30-39	M	Flu
	35xx	30-39	M	COVID
	35xx	40-49	M	Gastric flu
	35xx	20-29	F	Flu
	35xx	40-49	M	COVID

k = 2 anonymised data

ℓ-diversity

Direct identifier	Quasi-identifier	Quasi-identifier	Quasi-identifier	Sensitive attribute
Name	**Postal code**	**Age**	**Sex**	**Diagnosis**
Emma	3500	25	F	Gastric flu
Bob	3510	32	M	Flu
Tommy	3520	36	M	COVID
Michael	3530	45	M	Gastric flu
Sara	3540	23	F	Flu
Ziggy	3550	43	M	COVID

Original data

Direct identifier	Quasi-identifier	Quasi-identifier	Quasi-identifier	Sensitive attribute
Name	**Postal code**	**Age**	**Sex**	**Diagnosis**
	35xx	20-29	F	Gastric flu
	35xx	30-39	M	Flu
	35xx	30-39	M	COVID
	35xx	40-49	M	Gastric flu
	35xx	20-29	F	Flu
	35xx	40-49	M	COVID

ℓ = 2 anonymised data

Appendix 5: Description of the experimental dataset

Column name	Description
secret_name	Indicates the unique identifier for the record. The beginning letters, namely "P_" or "C_", indicate whether outcomes are patient-reported or clinician-reported, respectively. This column accepts data of the type "object".
report_source	Represents the source from which the data is collected. This column accepts data of the type "object". - "clinicians" - "patients"
sex	Shows the biological sex of the patient. This column accepts data of the type "object". - "male" - "female"
age	Contains the ages of the patients. This column accepts data of the "integer" type.
edss	Indicates the score on the Expanded Disability Status Scale (EDSS) for a patient. The EDSS has a range of 0 to 10, where higher scores indicate higher levels of disability. The scoring is determined through an examination conducted by a neurologist. EDSS steps 1.0 to 4.5 refer to people with MS who can walk without any assistance and are evaluated based on impairment in 8 functional systems. EDSS steps 5.0 to 9.5 are defined by the impairment to walking [41, 42]. This column accepts data of the type "float".
bmi	Represents the body mass index (BMI) of the patient. This column accepts data of the type "float".
covid19_admission_hospital	Contains the hospital admission status of the patient as a result of COVID-19. This column accepts data of the type "object". - "yes" - "no"
covid19_confirmed_case	Represents whether the patient had a confirmed COVID-19 diagnosis. This column accepts data of the type "object". - "yes" - "no"
covid19_diagnosis	Shows the perceived COVID-19 diagnosis of the patient. This column accepts data of the type "object". - "not_suspected" - "suspected"
covid19_symptoms	Represents the symptoms of patients. This column accepts data of the type "object". - "no" - "congestion" - "pneumonia" - "sore_throat" - "shortness_breath" - "fever" - "fatigue" - "pain" - "chills"
covid19_icu_stay	Indicates whether the patient was admitted to the intensive care unit (ICU) of the hospital as a result of COVID-19. This column accepts data of the type "object". - "yes" - "no"
covid19_outcome_recovered	Shows whether a patient recovered from their COVID-19 infection. This column accepts data of the type "object". - "yes" - "no" - "not_applicable"
covid19_self_isolation	Indicates whether the patient self-isolated. This column accepts data of the type "object". - "yes" - "no"
covid19_ventilation	Indicates whether a patient was ventilated during their hospital stay. This column accepts data of the type "object". - "yes" - "no"

Column name	Description
comorbidities	Contains the comorbidities of a patient. This column accepts data of the type "object". - "no" - "chronic_liver_disease" - "immunodeficiency" - "hypertension" - "cardiovascular_disease" - "diabetes" - "lung_disease" - "chronic_kidney_disease" - "other" - "malignancy"
ms_type	Indicates the type of MS. This column accepts data of the type "object". - "RRMS" - "CIS" - "PPMS" - "not_sure" - "SPMS"
ms_diagnosis_date	Represents the year in which the patient received their diagnosis for MS. This column accepts data of the "integer" type.

Appendix 6: Filled out data extraction forms

Given the already present definitions entailing quasi-identifiers and sensitive attributes, how can the identification process for these types of attributes be standardised using an algorithmic approach?

Title	Quasi-identifier recognition algorithm for privacy preservation of cloud data based on risk reidentification [26]	Quasi-identifier recognition with echo chamber optimization-based anonymization for privacy preservation of cloud storage [27]
Authors	Mansour HO, Siraj MM, Ghaleb FA, Saeed F, Alkhammash EH, Maarof MA	Jadhav PS, Borkar GM
Publication year	2021	2024
Objectives	- to overcome the identity disclosure resulting from QID linking - to reduce the leakage of privacy by proposing a QID recognition algorithm based on risk rate reidentification	- to identify the quasi-attributes based on clustering - to maintain privacy preservation in the cloud based on the echo chamber optimization as well as the optimized k-anonymisation process
Methodology	- data preprocessing - compute risk rate for all attributes - select classification thresholds - classify the attributes as quasi-identifiers, sensitive attributes, and non-sensitive attributes - determine the actual dimension of QIDs that should be used in an anonymisation operation that will achieve optimum case	- data preprocessing - compute risk rate - select classification thresholds - classify the dataset attributes into quasi-identifiers, sensitive attributes, and non-sensitive attributes - echo chamber optimisation
Performance measures	- privacy gain - non-uniform entropy	- average equivalent class size metric - discernibility metric - normalised certainty penalty
Key takeaways	- accurate identification of QIDs is an important issue for the success and validity methods of privacy-preserving outsourced data that seek to avoid privacy leakage caused by QID linking - the proposed identification algorithm has better performance and is more perfect in terms of privacy provided against data utility when compared with other works	- the developed optimized clustering-based algorithm with the privacy preservation model extensively minimizes the leakage of private information and the utilisation of data is well-maintained compared with other existing algorithms

What are the methods used to evaluate data usefulness of an anonymised dataset?

Title	Privacy protection in social science research: possibilities and impossibilities [28]	An experimental comparison of quality models for health data de-identification [29]	Utility-driven assessment of anonymized data via clustering [30]	Utility-preserving transaction data anonymization with low information loss [31]	A generic method for assessing the quality of de-identified health data [32]
Authors	Albright JJ	Eicher J, Kuhn KA, Prasser F	Ferrão ME, Prata P, Fazendeiro P	Loukides G, Gkoulalas-Divanis A	Prasser F, Bild R, Kuhn KA
Publication year	2011	2017	2022	2012	2016
Objectives	- contribute to an understanding of the technical issues involved with SDC	Answer the following questions: - How do common models for measuring data quality influence the way in which datasets are transformed? - If different models are used, how are the obtained results related to each other? - How well is de-identified data, obtained by using different quality models, suited for real-world applications?	- proposal to adjust the utility model to the research question in the applied field of study as complementary to data utility quantified by standard metrics, no matter the substantive applied field of study - provide insight into the differences between anonymised and original datasets and debate its relevance for research purposes	- propose a novel approach for anonymising data in a way that satisfies data publishers' utility requirements and incurs low information loss	- development of a generic variant to non-uniform entropy which can be used to assess the information loss induced by transforming data with arbitrary combinations of full-domain generalisation, local recoding and record or value suppression
Methodology	- introduction of the field of SDC by defining key terms, describing how researchers quantify risk, identifying options to minimise risk, and outlining how these decisions affect the usefulness of a data file - description of the implications of SDC for political science research, namely the problems it introduces for variance estimation in complex surveys - outline where the field of SDC is headed	The used quality models: - Average Equivalence Class Size (AECS) - discernibility - precision - loss - ambiguity - Kullback-Leibler (K.-L.) divergence - non-uniform entropy	- clustering as an utility indicator	- introduction of Utility Criterion (UC), a measure that can quantify data utility under different generalisation models and be employed by effective anonymisation algorithms - development of a novel anonymisation algorithm - experimental evaluation of the approach using two datasets	- non-uniform entropy - a generic variant to non-uniform entropy
Key takeaways	- disclosure risk may be higher than researchers realise - the proactive steps data collection organisations take to minimise disclosure risk can affect the ability of the end user to accurately estimate statistical relationships	- different models are suited best for different application scenarios - the non-uniform entropy model provides the best results for general purpose usage	- when working with low dimensionality datasets, no matter the method of anonymisation, the results obtained suggest that the replacement of original data by their anonymised versions may jeopardise the proper data analysis, the data-based inferences or deductions and even the conclusions of the scientific research	- the UAR anonymisation algorithm incurs significantly lower information loss than the state-of-the-art methods	- the used method provides a unified framework in which this model can be used to assess and compare the quality of differently transformed data to find a good or even optimal solution to a given de-identification problem

Appendix 7: Filled out quality assessment checklists

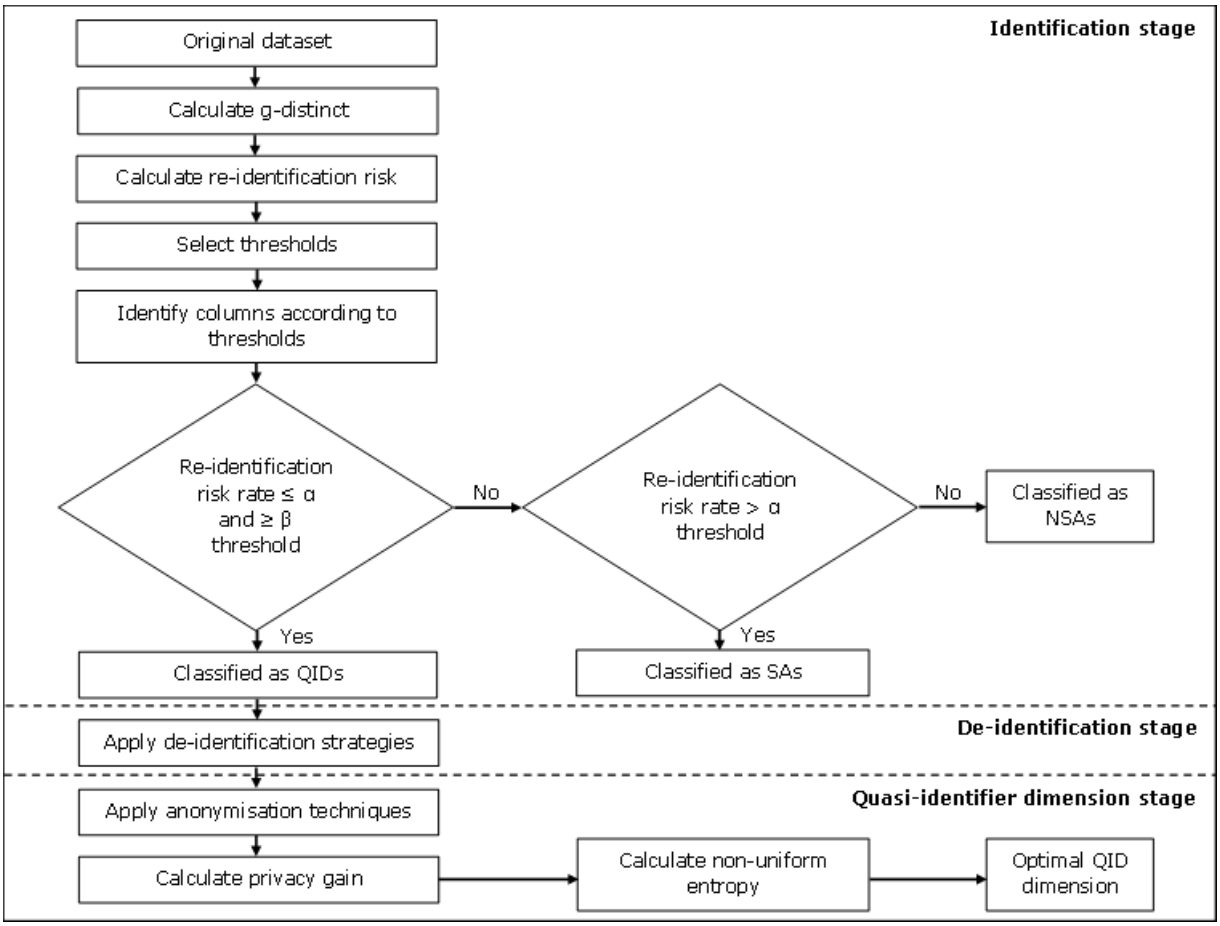
Given the already present definitions entailing quasi-identifiers and sensitive attributes, how can the identification process for these types of attributes be standardised using an algorithmic approach?		
Study	Quasi-identifier recognition algorithm for privacy preservation of cloud data based on risk reidentification [26]	Quasi-identifier recognition with echo chamber optimization-based anonymization for privacy preservation of cloud storage [27]
Was there a clear description of the aims and purposes of the research?	Yes	Yes
Was the algorithm clearly described (e.g. flowchart, pseudocode, ...)	Yes	Yes
Was the experimental dataset described?	Yes	Yes
Were any metrics used to validate the methodology?	Yes	Yes
Was the quality of the anonymised data assessed?	Yes	Yes
Was the quality assessment done using simple statistical methods or machine learning?	Statistical methods → no	Statistical methods → no
Were there any hyperparameters that were finetuned?	No	No
Is there a repository of the code?	No	No
Final score	**5/8**	**5/8**

What are the methods used to evaluate data usefulness of an anonymised dataset?					
Study	Privacy protection in social science research: possibilities and impossibilities [28]	An experimental comparison of quality models for health data de-identification [29]	Utility-driven assessment of anonymized data via clustering [30]	Utility-preserving transaction data anonymization with low information loss [31]	A generic method for assessing the quality of de-identified health data [32]
Was there a clear description of the aims and purposes of the research?	No	Yes	Yes	Yes	Yes
Was the experimental dataset described?	No	Yes	Yes	Yes	Yes
Were any metrics used to validate the methodology?	No	No	No	No	No
Were the metrics clearly described?	No	Yes	Yes	Yes	Yes
Is there a repository of the code?	No	No	Yes	No	No
Final score	**0/5**	**3/5**	**4/5**	**3/5**	**3/5**

Appendix 8: README file of the repository

Attribute Identification And Utility Metrics Pipeline

This repository contains Python scripts to identify attributes in a dataset and subsequently determine the best quasi-identifier dimension based on privacy gain and non-uniform entropy. Two original datasets together with their de-identified datasets were provided to test the scripts.
The following figure represents the flow of the pipeline. Note that the de-identification stage is not included in this repository and must be implemented using your preferred de-identification strategies.



About the datasets

Both datasets were made up using a mock data generator, which can be found **here**.
The original datasets that were used in the first script (stage_1_identification.py), are named accordingly. For the second script (stage_2_qid_dimension.py), the de-identified datasets are provided. These datasets were de-identified in order of descending risk rates and named according to the last de-identified attribute. For the 500 row dataset, an alpha threshold of 25 was used together with a beta threshold of 1. For the 1000 row dataset, these values were 10 and 1.
The original datasets contain the following columns:

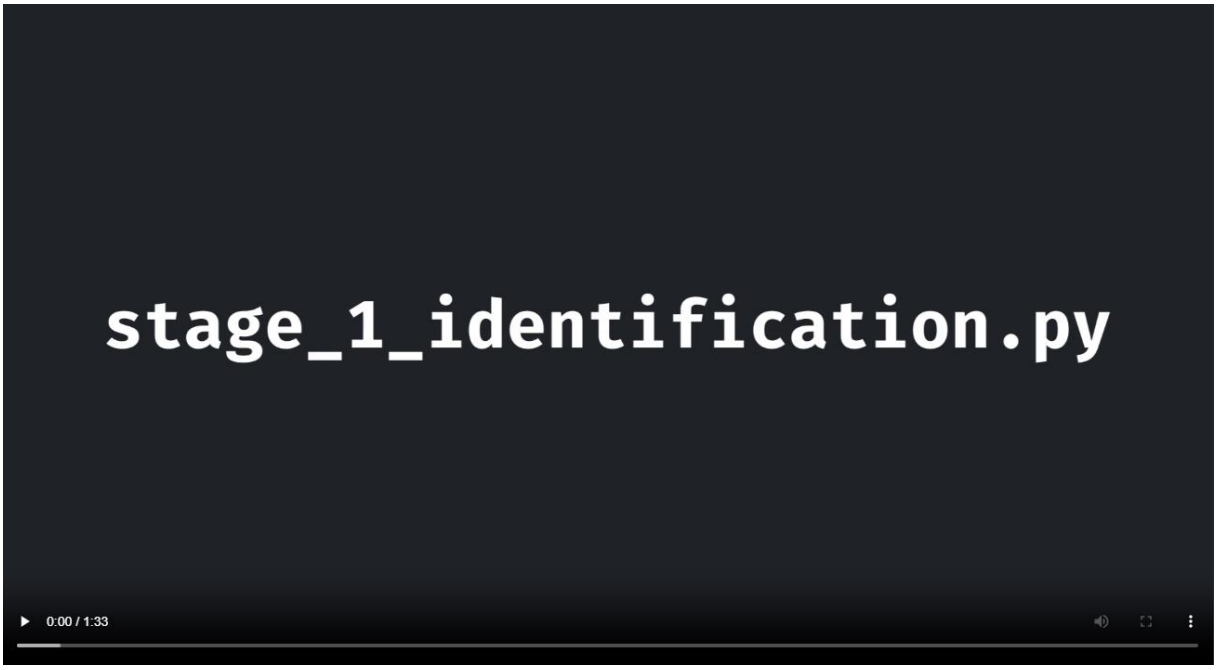
Column name	Description
secret_name	Indicates the unique identifier for the record. The beginning letters, namely “P_” or “C_”, indicate whether outcomes are patient-reported or clinician-reported, respectively. This column accepts data of the type “object”.
report_source	Represents the source from which the data is collected. This column accepts data of the type “object”. - “clinicians”

	- “patients”
sex	Shows the biological sex of the patient. This column accepts data of the type “object”. - “male” - “female”
age	Contains the ages of the patients. This column accepts data of the “integer” type.
edss	Indicates the score on the Expanded Disability Status Scale (EDSS) for a patient. The EDSS has a range of 0 to 10, where higher scores indicate higher levels of disability. The scoring is determined through an examination conducted by a neurologist. EDSS steps 1.0 to 4.5 refer to people with MS who can walk without any assistance and are evaluated based on impairment in 8 functional systems. EDSS steps 5.0 to 9.5 are defined by the impairment to walking. This column accepts data of the type “float”.
bmi	Represents the body mass index (BMI) of the patient. This column accepts data of the type “float”.
covid19_admission_hospital	Contains the hospital admission status of the patient as a result of COVID-19. This column accepts data of the type “object”. - “yes” - “no”
covid19_confirmed_case	Represents whether the patient had a confirmed COVID-19 diagnosis. This column accepts data of the type “object”. - “yes” - “no”
covid19_diagnosis	Shows the perceived COVID-19 diagnosis of the patient. This column accepts data of the type “object”. - “not_suspected” - “suspected”
covid19_symptoms	Represents the symptoms of patients. This column accepts data of the type “object”. - “no” - “congestion” - “pneumonia” - “sore_throat” - “shortness_breath” - “fever” - “fatigue” - “pain” - “chills”
covid19_icu_stay	Indicates whether the patient was admitted to the intensive care unit (ICU) of the hospital as a result of COVID-19. This column accepts data of the type “object”. - “yes” - “no”
covid19_outcome_recovered	Shows whether a patient recovered from their COVID-19 infection. This column accepts data of the type “object”. - “yes” - “no” - “not_applicable”
covid19_self_isolation	Indicates whether the patient self-isolated. This column accepts data of the type “object”. - “yes” - “no”

covid19_ventilation	Indicates whether a patient was ventilated during their hospital stay. This column accepts data of the type “object”. - “yes” - “no”
comorbidities	Contains the comorbidities of a patient. This column accepts data of the type “object”. - “no” - “chronic_liver_disease” - “immunodeficiency” - “hypertension” - “cardiovascular_disease” - “diabetes” - “lung_disease” - “chronic_kidney_disease” - “other” - “malignancy”
ms_type	Indicates the type of MS. This column accepts data of the type “object”. - “RRMS” - “CIS” - “PPMS” - “not_sure” - “SPMS”
ms_diagnosis_date	Represents the year in which the patient received their diagnosis for MS. This column accepts data of the “integer” type.

Tutorial video

Below, you can download and watch the tutorial video.



How to use stage_1_identification.py

This script is designed to assist in the process of identifying and classifying attributes in a dataset based on their re-identification risk rates. It evaluates the attributes and classifies them into 3 categories: quasi-identifiers (QIDs), sensitive attributes (SAs), and non-sensitive attributes (NSAs). This classification is based on user-defined thresholds for re-identification risk.

Prerequisites

Before running the first script (stage_1_identification.py), make sure you:

1. Have an original dataset.
2. Know which attributes are direct identifiers.

Steps

1. Download an original dataset (e.g. stage_1_df_mock_1000.csv).
2. Run the Python script. It will prompt you to enter the path of the original dataset and the attribute name of direct identifier(s).
3. The script will filter out attributes where missing values exceed 85% and display them.
4. The script calculates the re-identification risks of the remaining attributes.
5. You will be prompted to enter alpha and beta thresholds, which will be used in the identification of attributes. Attributes surpassing the alpha threshold are labelled as SAs. Attributes with a risk rate lower than or equal to alpha but higher than or equal to beta are considered QIDs. The remaining attributes with a risk rate below the beta threshold are classified as NSAs.
6. The script will classify the attributes into SAs, QIDs, and NSAs based on the selected thresholds.
7. The script gives advice on next steps to take. This includes de-identifying attributes in order of descending risk rates.

> Tip: Copy or write down the results from this stage. You will need the attribute names, their classification and the order in which you de-identified them in the next stage.

Example output with the 1000 row dataset

```
Columns excluded because of missing values:
        covid19_self_isolation: 91.8

Re-identification risk rates:
        bmi: 20.98
        ms_diagnosis_date: 13.81
        edss: 10.04
        age: 2.66
        comorbidities: 1.8
        covid19_symptoms: 1.54
        ms_type: 0.67
        covid19_ventilation: 0.52
        covid19_outcome_recovered: 0.31
        covid19_confirmed_case: 0.26
        covid19_icu_stay: 0.26
        report_source: 0.2
        sex: 0.2
        covid19_admission_hospital: 0.2
        covid19_diagnosis: 0.2

Enter the α threshold (as a float): 10.0

Enter the β threshold (as a float): 1.0

Sensitive attributes:
        bmi: 20.98
        ms_diagnosis_date: 13.81
        edss: 10.04

Quasi-identifiers:
        age: 2.66
        comorbidities: 1.8
        covid19_symptoms: 1.54

Non-sensitive attributes:
        ms_type: 0.67
        covid19_ventilation: 0.52
```

```
        covid19_outcome_recovered: 0.31
        covid19_confirmed_case: 0.26
        covid19_icu_stay: 0.26
        report_source: 0.2
        sex: 0.2
        covid19_admission_hospital: 0.2
        covid19_diagnosis: 0.2

Advice:
        De-identify the columns according to their re-identification risk rates.
        Start with the column with the highest risk rate and end with the column with the
lowest risk rate.
        Afterwards, proceed to the QID dimension stage.
```

How to use stage_2_qid_dimension.py

This script helps to evaluate the privacy-utility trade-off for de-identification processes by computing privacy metrics such as k-anonymity, l-diversity, t-closeness, privacy gain, and non-uniform entropy. It serves as a guide in determining the optimal number of QIDs to de-identify for balancing privacy with utility.

Prerequisites

Before running the second script (stage_2_qid_dimension.py), make sure you:

1. Have an original dataset where sensitive attributes are de-identified and direct identifiers and attributes where missing values exceed 85% are suppressed.
2. Have a dataset where all information is masked (values are changed to the value "x").
3. Have a new dataset for every additional attribute that is de-identified. De-identify datasets in order of descending risk rates and create a CSV file after each de-identification step. You should have as many datasets as there are QIDs. Ensure the tuples in the original dataset and the tuples in the de-identified datasets are in the same order since the algorithm compares privacy and utility based on this order.

Steps

1. Download the de-identified datasets provided or have your de-identified datasets at hand.
2. Run the script. It will prompt you to provide the following input:
    1. The path of your original dataset where sensitive attributes are de-identified and direct identifiers and attributes where missing values exceed 85% are suppressed.
    2. The path of a dataset where all information is removed.
    3. All the QIDs as a comma-separated list in order of descending re-identification risk rate.
    4. All the SAs as a comma-separated list.
3. The script will prompt you to provide the QID(s) you de-identified and the according dataset until all QIDs are processed. Again, you should enter the QIDs in order of descending re-identification risk rate.

Example output with the 1000 row dataset

```
De-identified QIDs: "age", "comorbidities", "covid19_symptoms"

        K-anonymity original:           1

        K-anonymity after:              110

        T-closeness original:           0.9750000000000001

        T-closeness after:              0.32376470588235295

                Sensitive attributes: bmi

                        L-diversity original:    1
```

```
                        L-diversity after:       3

                Sensitive attributes: ms_diagnosis_date

                        L-diversity original:    1

                        L-diversity after:       6

                Sensitive attributes: edss

                        L-diversity original:    1

                        L-diversity after:       2

        Privacy gain:                    109

        Non-uniform entropy:             4635.091083186598

        Non-uniform entropy(%):          69.05

        Inverse non-uniform entropy(%):  30.950000000000003

All QIDS were processed

Optimal QID Dimension: 3

Advice:
        You should de-identify the first 3 QIDS.
```

Appendix 9: Classification of attributes

Classification of attributes in the 500 row dataset		
Column name	**Re-identification risk (%)**	**Classification (α = 25.0, β = 1.0)**
bmi	38.50	Sensitive attribute
ms_diagnosis_date	27.65	Sensitive attribute
edss	22.58	Quasi-identifier
age	5.49	Quasi-identifier
comorbidities	3.63	Quasi-identifier
covid19_symptoms	3.12	Quasi-identifier
ms_type	1.34	Quasi-identifier
covid19_ventilation	0.96	Non-sensitive attribute
covid19_outcome_recovered	0.61	Non-sensitive attribute
covid19_icu_stay	0.53	Non-sensitive attribute
covid19_confirmed_case	0.50	Non-sensitive attribute
report_source	0.40	Non-sensitive attribute
sex	0.40	Non-sensitive attribute
covid19_admission_hospital	0.40	Non-sensitive attribute
covid19_diagnosis	0.40	Non-sensitive attribute

Classification of attributes in the 1000 row dataset		
Column name	**Re-identification risk (%)**	**Classification (α = 10.0, β = 1.0)**
bmi	20.98	Sensitive attribute
ms_diagnosis_date	13.81	Sensitive attribute
edss	10.04	Sensitive attribute
age	2.66	Quasi-identifier
comorbidities	1.80	Quasi-identifier
covid19_symptoms	1.54	Quasi-identifier
ms_type	0.67	Non-sensitive attribute
covid19_ventilation	0.52	Non-sensitive attribute
covid19_outcome_recovered	0.31	Non-sensitive attribute
covid19_icu_stay	0.26	Non-sensitive attribute
covid19_confirmed_case	0.26	Non-sensitive attribute
report_source	0.20	Non-sensitive attribute
sex	0.20	Non-sensitive attribute
covid19_admission_hospital	0.20	Non-sensitive attribute
covid19_diagnosis	0.20	Non-sensitive attribute