



UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

ANONYMIZATION PROCESSES IN THE BRAZILIAN
CONTEXT: INVESTIGATING PRIVACY ON OPEN DATABASES

Mariana Souza de Oliveira

Orientador

Carlos Eduardo Ribeiro de Mello

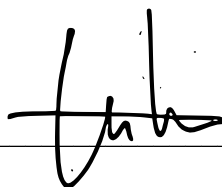
RIO DE JANEIRO, RJ – BRASIL
AGOSTO DE 2023

ANONYMIZATION PROCESSES IN THE BRAZILIAN
CONTEXT: INVESTIGATING PRIVACY ON OPEN DATABASES

Mariana Souza de Oliveira

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA
OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓSGRADUAÇÃO
EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE
JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO
ASSINADA.

Aprovado por:



Dr. Carlos Eduardo Ribeiro de Mello - UNIRIO



Dr.ª Laura de Oliveira Fernandes Moraes - UNIRIO



Dr.ª Georgina Evans - Google DeepMind

RIO DE JANEIRO, RJ – BRASIL

AGOSTO DE 2023

Catálogo informatizada pelo(a) autor(a)

O48

Oliveira, Mariana Souza de
ANONYMIZATION PROCESSES IN THE BRAZILIAN
CONTEXT: INVESTIGATING PRIVACY ON OPEN DATABASES /
Mariana Souza de Oliveira. -- Rio de Janeiro, 2023.
60 f

Orientador: Carlos Eduardo Ribeiro de Mello.
Dissertação (Mestrado) - Universidade Federal do
Estado do Rio de Janeiro, Programa de Pós-Graduação
em Informática, 2023.

1. Privacy-preserving methods. 2. Public
release. 3. Open data. 4. Risk-based anonymization
methodology. 5. Brazil. I. Mello, Carlos Eduardo
Ribeiro de, orient. II. Título.

Acknowledgements

To my family, who made this possible. To my advisor and friend, who never gave up on me and inspires me to be a better researcher. And to all my friends, for their support.

Souza de Oliveira, Mariana. **ANONYMIZATION PROCESSES IN THE BRAZILIAN CONTEXT: INVESTIGATING PRIVACY ON OPEN DATABASES**. UNIRIO, 2023. 60 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

O crescente número de iniciativas de dados abertos tem levado a uma maior disseminação de informações ao público, trazendo diversos benefícios, mas também levantando preocupações sobre a privacidade individual. No Brasil, essas questões são abordadas por meio de duas leis complementares que regulam a transparência e a privacidade na divulgação governamental de dados individuais. Por um lado, a lei de transparência conhecida como LAI (Lei de Acesso à Informação) rege o acesso da sociedade a informações públicas. Por outro lado, a lei de privacidade recentemente implementada, LGPD (Lei Geral de Proteção de Dados Pessoais), restringe a divulgação de dados individuais e impõe sanções em caso de não conformidade. Nesse contexto, para promover a transparência dos dados ao mesmo tempo em que se garante a confidencialidade, muitos órgãos governamentais brasileiros concentraram-se em identificar e remover informações pessoalmente identificáveis do conjunto de dados antes de divulgá-lo ao público. No entanto, diversos trabalhos têm debatido sobre a inadequação e ineficiência dessa técnica desde pelo menos a década de 80, a qual prova-se ineficiente para preservar a privacidade dos indivíduos. Pesquisadores brasileiros têm desenvolvido estudos para abordar esse problema, explorando a aplicação de técnicas populares de anonimização, como a k-anonimidade, ou propondo novos métodos para avaliar riscos. No entanto, esses estudos são escassos e carecem de um processo de anonimização mais sistemático e completo. Este trabalho tem como objetivo contribuir para o debate sobre privacidade em bancos de dados abertos, investigando a aplicação de uma metodologia sistemática de anonimização baseada em riscos por meio de um estudo de caso. Pretendemos discutir os fatores importantes que precisam ser considerados ao realizar um processo de anonimização, sob a ótica do contexto brasileiro. Também implementamos diferentes métodos para quantificar e mitigar riscos de privacidade e discutimos a utilidade dos dados. Os resultados indicam riscos substanciais de privacidade nos dados do estudo de caso, assim como em outros conjuntos de dados brasileiros, afetando inúmeras pessoas. Ao mesmo tempo, a metodologia estudada pode oferecer orientação na discussão da anonimização de dados ao oferecer um passo a passo adaptável e compreensível. A seleção dos métodos de preservação da privacidade, no entanto, pode ser uma tarefa desafiadora, especialmente ao considerar múltiplos riscos em um conjunto de dados grande e complexo.

Palavras-chave: Métodos de preservação da privacidade, Publicação de dados públicos, Dados abertos, Metodologia de anonimização baseada em riscos, Brasil

ABSTRACT

The growing number of open data initiatives has led to a greater dissemination of information to the public, bringing various benefits but also raising concerns about individual privacy. In Brazil, these issues are addressed through two complementary laws that regulate transparency and privacy in the governmental release of individual data. On the one hand, the transparency law known as LAI (- Lei de Acesso à Informação) governs society's access to public information. On the other hand, the recently implemented privacy law, LGPD (- Lei Geral de Proteção de Dados Pessoais) restricts the disclosure of individual data and imposes sanctions for non-compliance. In this context, to promote data transparency while guaranteeing confidentiality, many Brazilian government agencies focused on identifying and removing personally identifiable information from the data set before releasing it to the public. However, several works have shed light on the inadequacy and inefficiency of this approach since at least the 1980s, demonstrating its insufficiency in safeguarding individuals' privacy. Brazilian researchers have developed works towards this problem by exploring the application of popular anonymization techniques, such as k-anonymity, or by proposing methods to evaluate risk. These studies, nonetheless, are few and lack a more systematic and complete anonymization process. This work aims to contribute to the debate regarding privacy on open databases by investigating the application of a systematic risk-based anonymization methodology through a case study. We intent to discuss the important factors that need to be taken into consideration when performing an anonymization process under the Brazilian context. We also implement different methods to quantify and mitigate privacy risks and discuss data utility. Results indicate substantial privacy risks in the case study data and other Brazilian datasets, impacting numerous individuals. At the same time, the studied methodology can offer guidance in the discussion of data anonymization by a flexible step-by-step process. The selection of the privacy-preserving methods, however, can be a very challenging task, especially when considering multiple risks in a large and complex data set.

Keywords: Privacy-preserving methods, Public release, Open data, Risk-based anonymization methodology, Brazil.

Contents

1	INTRODUCTION.....	10
1.1	Motivation.....	10
1.2	Objectives	12
1.3	Research Questions.....	12
1.4	Contributions.....	12
1.5	Outline.....	13
2	BACKGROUND.....	15
2.1	Overview.....	15
2.2	Attribute classification.....	16
2.3	Privacy models.....	17
2.4	Types of privacy risk	18
2.5	Types of privacy metrics.....	18
2.6	Privacy-preserving techniques and utility metrics	20
2.7	Related work	21
3	METHODOLOGY.....	23
3.1	Data Collection	23
3.2	Methodology for Managing Re-Identification Risk	24
3.3	Research Design.....	27
4	RESULTS.....	31
4.1	Experimental Results of Re-Identification Attacks	31
4.2	Experimental Results of Attribute-Inference Attacks.....	40
4.3	Evaluation of Anonymization Techniques.....	46
5	DISCUSSION	54
6	CONCLUSION	57
6.1	Contributions.....	57
6.2	Challenges and Limitations.....	58
6.3	Future Works	59
	BIBLIOGRAPHY	60

List of Figures

Figure 1. Overview of Privacy-Preserving Data Analysis and Publishing.	15
Figure 2. Overall re-identification risk management process, taken from [1].	25
Figure 3. Distribution of possible values for the equivalence class.	32
Figure 5. Adversary's re-identification success by combination of quasi-identifiers.	33
Figure 6. Minimum quasi-identifying combination high mean re-identification risk.	34
Figure 7. Maximum quasi-identifying combination min. mean re-identification risk.	34
Figure 8. Histograms for the distribution of individual IRS scores.	34
Figure 9. Distribution of age (NU_IDADE_N) by values of outcome (EVOLUCAO).	37
Figure 10. Mean re-identification success by quasi-identifiers combinations with and without large number of outcomes.	37
Figure 11. Mean re-identification success for ID_UNIDADE and DT_EVOLUCA when adding another attribute.	38
Figure 12. Normalized entropy values for ID_UNIDADE by DT_NASC and ID_UNIDADE.	40
Figure 13. Distribution of sensitive values by overall data set and by equivalence class size.	42
Figure 14. Distribution of t-closeness values for each sensitive attribute.	43
Figure 15. Mutual information between sensitive and quasi-identifier attributes.	45
Figure 16. Transformations results.	49
Figure 17. Distribution of possible values for the equivalence class in the anonymized data set.	49
Figure 18. Histograms for the re-identification individual score IRS for the original and anonymized data set.	50
Figure 19. Mean re-identification success by quasi-identifier combination in the anonymized data set.	51
Figure 20. Distributions of t-closeness values for the original and anonymized data set.	52

List of Tables

Table 1. Variables chosen as quasi-identifiers in the case study.	28
Table 2. Variables chosen as sensitive attributes in the case study.	29
Table 3. Re-identification results.	32
Table 4. Outcomes by Quasi-identifier.	36
Table 5. Mean re-identification success by Quasi-identifier combinations with attributes with large number of outcomes.	38
Table 6. Attribute-inference results.	41
Table 7. Attribute-inference results for quasi-identifier combinations.	44
Table 8. Statistics regarding the transformed variables.	49
Table 9. Re-identification results for the anonymized data set.	50
Table 10. Attribute-inference results for the anonymized data set.	52

1 INTRODUCTION

1.1 Motivation

In the governmental sphere, data transparency has emerged as a crucial element, shaping the relationship between the state and citizens [17]. Open data, characterized by the unrestricted availability and accessibility of government-held information, plays a pivotal role in supporting democracy and citizenship. By making government activities and decisions visible to the public, open data empowers citizens to hold their representatives accountable and make informed decisions. This democratization of information helps build trust, increase civic engagement and incentive individuals to actively participate in public affairs.

A particular category of data collected and transparently shared by governments pertains to the public they serve [17]. By gathering information about citizens to monitor areas such as health, education and demographics, governments can enhance their capacity to allocate resources more efficiently. However, gathering and handling individuals' data comes jointly with the responsibility of protecting personal details so as to prevent any potential misuse or exploitation. Guaranteeing privacy safeguards is not only ethically important but also legally mandated in many jurisdictions. Privacy laws, as the European Union's General Data Protection Regulation (GDPR)¹, exemplify global efforts in this direction.

In Brazil, these issues are addressed through two complementary laws that regulate transparency and privacy in the governmental release of individual data. On the one hand, the transparency law, a.k.a. LAI (Access to Information Law, or *Lei de Acesso à Informação*)², enforces access to public information. On the other hand, the recently implemented privacy law, a.k.a. LGPD (General Data Protection Law, or *Lei Geral de Proteção de Dados Pessoais*)³, restricts the disclosure of individuals' data and imposes sanctions for non-compliance. Jointly, LAI and LGPD establish a comprehensive legal framework that reflects Brazil's strategy in balancing data transparency and privacy concerns.

¹ General Data Protection Regulation. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council. <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679>>

² Lei de Acesso à Informação. (2011). Law no. 12.527, Constitution of the Federal Republic of Brazil. <https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/12527.htm>

³ Lei Geral de Proteção aos Dados. (2018). Law no. 13.709, Constitution of the Federal Republic of Brazil. <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/13709.htm>

In this scenario, a common method employed by Brazilian authorities to release public data in accordance with LGPD is by means of removal of direct identifying information, such as names and government-issued unique numbers [17]. Nonetheless, despite this method prevents direct linkage between an individual and his/her corresponding database record, it is still not enough to avoid re-identification. There are certain combinations of attributes that, once cross-referenced or analyzed with external data sources, can still lead to the identification of individuals. For example, the external information about where someone lives in a certain city and which industry works could be enough to identify him/her, even if their name and address are not publicly available. In fact, Sweeney [16] was able to show in 2000 that 87% of the United States population in that census could be uniquely identified by using only a combination of the ZIP code, gender, and date of birth.

These vulnerabilities were highlighted by previous works on publicly accessible Brazilian data sets [2][18][19]. The conclusions draw attention to a high level of privacy risk, with the potential to impact a considerable portion of the population and expose sensitive information. As a result, the adoption of more sophisticated privacy-preserving methods is crucial to adequately meet privacy concerns.

At the same time, the role of ensuring access to public information while protecting personal data can prove to be rather difficult. The data anonymization process introduces various complexities, such as accurately evaluating potential malicious entities seeking to breach the dataset's confidentiality, determining the most appropriate privacy measures based on established assumptions and executing customized anonymization techniques aligned with the dataset specific context. To successfully overcome these challenges demands a deep understanding of data context, potential risks and evolving technological landscapes, all to ensure that privacy is upheld without compromising the underlying analytical value of the data.

In the Brazilian context, these concepts are notably new and require proper formalization [17]. For example, there are no accessible official guidelines on anonymization, which could offer guidance for navigating this process. This stands in contrast to other countries, like the guide provided by Canada's Information and Privacy Commissioner of Ontario. Even LGPD does not provide a clear definition of what is considered a "reasonable anonymization." Article 12 of the law states that if anonymized data can be reversed by reasonable efforts, then it is still considered personal data and is

subject to the LGPD's protections. However, the law does not specify what would be considered a reasonable effort, or which anonymization methods should be used [11].

Consequently, a systematic and adaptable guidance for conducting the anonymization process may provide great value in the Brazilian context. Addressing these questions in a practical manner, like providing insights into the key factors, challenges, and limitations observed, would illuminate the subject, and enrich the ongoing discourse on preserving privacy in Brazilian public data releases.

1.2 Objectives

The goal of this work is to investigate the application of a data anonymization methodology within the specific context of Brazilian government microdata release. Through a comprehensive case study, we seek to assess the effectiveness of the anonymization process applied to this microdata release considering how it mitigates privacy risks while maintaining an acceptable level of data utility. Thus, we intend to identify advantages and disadvantages related to the anonymization process, as well as to pinpoint the main weaknesses and strengths related to the domain application and the data set.

1.3 Research Questions

Particularly, we have the following research questions:

- **RQ1.** How applicable is the selected data anonymization methodology to the Brazilian context, and what are its advantages and limitations in ensuring data privacy in the case study data?
- **RQ2.** What are the vulnerabilities and privacy risks associated with the case study data? Which attributes contribute the most for the overall risks?
- **RQ3.** What are the challenges and trade-offs involved in balancing privacy protection and data utility in the anonymization process for the case study data? How can they be effectively addressed to optimize the anonymization outcome?

1.4 Contributions

The main contributions of this work are described as follows:

- **Systematic Anonymization Process:** to investigate a well-established and systematic anonymization process applied to a Brazilian governmental microdata through a case study. The process not only ensured effective privacy protection,

but also emphasized the preservation of data utility for publication. The applicability of the process was thoroughly discussed, highlighting its main advantages and limitations, thus providing valuable insights for future implementation in similar contexts.

- **Privacy Metrics Comparison:** to conduct a comparison of different metrics used to evaluate privacy risks. The analysis examines each metric's unique characteristics, clarifying the specific aspects they assess and their variations in measuring different privacy protection levels.
- **Attribute Impact Analysis:** to identify and analyze the impact of individual attributes on various privacy attacks. This analysis enables better-informed decisions in choosing data protection strategies to strengthen overall privacy safeguards.
- **Proposed Anonymization Solution:** to present a viable anonymization solution for the case study data, aiming to achieve a favorable balance between privacy protection and data utility. The challenges encountered during the anonymization process were discussed, contributing to the understanding of real-world implementation difficulties.
- **Review of Best Practices:** to provide a comprehensive assessment of current research and best practices in data anonymization, focusing on their relevance and applicability to the specific context of Brazilian data. The insights gained from this review enhance the discourse surrounding the implementation of anonymization strategies for privacy-preserving data publication in Brazil. Additionally, the review provides an overview of the contemporary techniques employed and discusses the findings from previous works, contributing to a deeper understanding of the state-of-the-art approaches in this domain.

1.5 Outline

In addition to this chapter, the rest of the work is organized as follows.

The **Background** section delves into fundamental concepts of data anonymization and privacy preservation data publishing research field, accompanied by a thorough review of pertinent literature concerning Brazilian microdata privacy concerns.

The **Methodology** section outlines the research design and provides detailed information on the data collection and preprocessing steps. Additionally, it presents a comprehensive description of the Methodology for Managing Re-Identification Risk, including its functioning and its application in the case study.

The **Results** section presents a comprehensive analysis of re-identification and attribute-inference risks is conducted. Furthermore, an anonymization strategy is proposed and evaluated for the case study data.

The **Discussion** section examines the findings carefully and evaluates the applicability of the studied anonymization methodology to the case study data, emphasizing the main challenges observed during the process.

The **Conclusion** section summarizes the primary contributions, discusses challenges and limitations, and proposes directions on future research.

2 BACKGROUND

In this section, we present the common ground of concepts, definitions, theories, and baselines for our comprehension of the research area and theme.

2.1 Overview

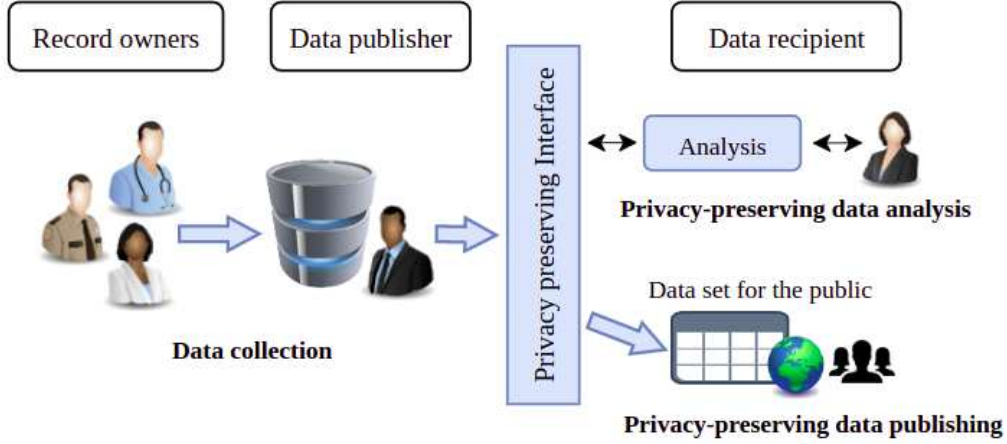


Figure 1. Overview of Privacy-Preserving Data Analysis and Publishing.

Consider the scenario in *Figure 1*. Overview of Privacy-Preserving Data Analysis and Publishing A data publisher (organizations, institutions, etc.) collects data from record owners (individuals) and makes it available to the data recipient (a third party or public domain) [8]. This availability may occur in two ways: the data publisher, via some privacy preserving interface, may allow the data recipient to interact with the data to perform queries or run analysis without directly handling the data. Alternatively, the data publisher could release a sanitized version of the collected data to the public domain. The first setting typically concerns the application context of Privacy-Preserving Data Analysis (PPDA), whereas the second setting mostly concerns Privacy-Preserving Data Publishing (PPDP) [8]. These concepts are also related to a third area called Statistical Disclosure Control (SDC), in which the information is released as statistical tables for general use or with the indirect access to data through query-based systems [7]. In this work, we consider the PPDP problem.

In practical terms, each data publishing scenario has its own assumptions and requirements, involving the data publisher, the data recipients, and the data publishing purpose [22]. We here assume that the data publisher is trustworthy and that record owners consent to share their personal information with the data publisher.

The purpose of the release is due to transparency and scientific research issues, and because of that, the data must not lose its utility, i.e., the property of a database release to satisfy the legitimate purposes of data analysts.

On the other hand, we assume that the data recipient can be malicious and untrustworthy, also called as adversary. The adversary would attempt to perform a privacy attack, which corresponds to an action or set of actions that leads to the obtention of sensitive information. Sensitive information is any kind of information considered worthy of protection, and it may be (not necessarily) defined by law. We also consider that, for conducting such attack, the adversary may access auxiliary, or external, information, i.e., any piece of information apart from the database that could provide leaks about an individual. For example, an individual's date of birth, which could have been retrieved from a social network public profile, could be used to discover sensitive information. Therefore, the private information of the individuals must be safeguarded.

The motivation for the choice of focusing on PPDP over PPDA comes from the Brazilian scenario: open data is made available, in many cases, in a non-iterative form. In other words, the data is published, with the promise of privacy, on official websites, e.g., <https://dados.gov.br/>, to anyone with an Internet access. The main concern, however, is whether such promise of privacy is accomplished.

2.2 Attribute classification

The variables (i.e., attributes) in a data set can be classified into four main categories [8]:

- **Identifiers.** These are variables that directly identify the respondent. For example, name and social security number.
- **Quasi-identifiers or key variables.** These are set of variables that an adversary can use, individually or in combination, to potentially re-identify someone. For example, while by themselves the gender, age, region and occupation variables may not reveal any respondents' identity, in combination they can uniquely identify them.
- **Sensitive variables.** These are variables that contain sensitive information on the respondent. They refer to highly critical attributes, usually protected by law and regulations. For example, religion, sexual orientation, medical conditions and political opinion.

- **Non-sensitive variables:** Those variables that contain non-sensitive information about the individual.

2.3 Privacy models

There are many methods a data publisher can apply to safeguard their published data against the aforementioned attacks, each offering distinct privacy assurances. We outline the methods employed in this study, but for a complete overview read [8].

- **K-anonymity model** [8]. A data set achieves k-anonymity when it contains k records for any given set of quasi-identifier values. Consequently, a record should be indistinguishable from a minimum of $k - 1$ other records based on these quasi-identifiers. These records that share identical quasi-identifiers constitute an equivalence class. From an attacker's perspective, the probability to successfully link his target record is never greater than $1/k$. This is the probability considers that the adversary knows a specific individual present in the data set, as well as the possible values of quasi-identifiers of the target.
- **L-diversity model** [8]. It requires that each equivalence class contains at least l different attribute values. This model serves as an early strategy to mitigate attribute disclosure and is primarily focused on safeguarding sensitive attributes. It efficiently addresses vulnerabilities linked to k-anonymity, particularly when a significant group of individuals shares an identical value for a sensitive attribute, simplifying an adversary's inference on their intended sensitive attribute. There are many variations for l-diversity, in this work we consider the traditional l distinct values for the sensitive attribute within each equivalence group.
- **T-closeness model** [8]. An equivalence class satisfies the t-closeness if the distance between the sensitive attribute distribution within the class and the corresponding distribution based on the whole table is not greater than a threshold t. A data set satisfies the t-closeness requirement, if all equivalence classes satisfy the t-closeness threshold. This model overcomes a limitation of l-diversity by accommodating cases involving skewed distributions. For example, a given equivalence class may have a high number of distinct values, say four, but these are extremely skewed, in which one possible value accounts for basically 80% of the data. In this case, an adversary could still infer the

sensitive attribute of an individual by simply looking at the most frequent value in the equivalence class.

2.4 Types of privacy risk

The literature specifies four main types of disclosure risks [4]:

- **Identity disclosure (or Re-identification risk).** It occurs when an intruder can correctly associate an individual in a privacy preserved published data set. Typically, the adversary uses outside information, which may be obtained from external databases, or its own personal knowledge. Once the linkage is successful, the intruder has access to all other information in the released data related to the specific respondent.
- **Attribute disclosure (or Attribute-inference risk).** It occurs when an adversary infers the sensitive attribute value for individuals, regardless of being successful in re-identifying them. This type of threat can be easily launched in imbalanced data sets.
- **Membership disclosure (or Membership-inference risk).** When an intruder can conclude if a certain individual is present or not in the data set. In some databases, to know if someone is there can already represent a high invasion of privacy. For example, if an intruder knows his target is in a database regarding criminal records.
- **Inferential disclosure (or Inference risk).** It occurs when the intruder can determine the value of some sensitive characteristic of an individual more accurately with the released data than it would have been possible otherwise. This risk has the potential to inflict harm on a broader group of individuals, including those whose specific information is not explicitly present in the data set. For example, how feasible is to use a machine learning model and incomplete information on individuals to infer the missing information.

In this work, we focus on identity and attribute disclosure since they represent together the main concern of many privacy regulations (identification of individuals and safeguard of sensitive information).

2.5 Types of privacy metrics

There are multiple ways to measure the risk of re-identification and attribute inference. We specifically selected the classical Prosecutor and Journalist definitions by

El Emam [1], and the Collective-target Re-identification Single database (CRS) attack to measure re-identification risk by [2]. The first definitions represent a more traditional approach in the literature, while the last is a recent definition also applied in a Brazilian data set. We detail them below.

- **Prosecutor model.** the adversary targets a specific individual who he knows is present in the data set. Generally, the probability of correct re-identification is of an individual given by $1/k$, where k is the size of the matching equivalence class in the data set (sample frequency). To obtain the overall risk, one can compute either the average risk by considering individual outcomes or opt for the maximum value calculation for a more cautious approach.
- **Journalist model.** the adversary targets a specific individual but is not sure or cannot know whether his target is in the data set. In this case, the individual risk is calculated by $1/F_k$, in which F_k represents the population frequency.
- **Collective-target Re-identification Single database (CRS) attack model.** The adversary can completely access a single database and is aware that their intended target is within it. The adversary can also gather auxiliary information on the values of a set of quasi-identifiers for every individual who holds a record in the database. The goal is to re-identify as many individuals as possible, no matter who they are, i.e., to precisely determine which record in the database corresponds to each individual of interest. This model employs a Bayesian approach, wherein the adversary's risk is initially calculated a priori, considering no access to information, and subsequently updated a posteriori, considering the use of quasi-identifier information. Furthermore, by comparing the posterior risk with the prior risk, the degradation of privacy is obtained. It calculates a deterministic risk, in which the re-identification of an individual needs to be exact, i.e., yielding only 1 record. But it also involves a probabilistic risk, which is based on the size of the equivalence class, like the Prosecutor model. It is relevant to note that the Collective-target model represents an average of individual results (Individual-target), therefore we can analyze both adversary's focus.

Regarding attribute-inference, we considered the l -diversity and t -closeness models mentioned earlier and the Collective-target Attribute-inference Single database (CAS) attack model. As we already defined the first two models in previous section, we detail below the CAS model.

- **Collective-target Attribute-inference Single database (CAS) attack model.** Similar to the CRS model, the adversary has access to a single database and is aware that their intended target is within it. He can rely on auxiliary information on the values of a set of quasi-identifiers. The goal, on the other hand, is to infer the values of an attribute considered to be sensitive for as many individuals as possible, no matter who they are. It also presents a Bayesian approach, and a deterministic and probabilistic risk measures. The CAS model assumes that the adversary's most effective strategy involves focusing on the most likely value for a sensitive attribute, which tends to be the one that occurs most often.

2.6 Privacy-preserving techniques and utility metrics

The proposed taxonomy of privacy-preserving techniques in microdata presented in [6] includes four categories: non-perturbative, perturbative, de-associative, and synthetic data. Non-perturbative techniques involve reducing detail or suppressing information, while perturbative techniques involve distorting information. Moreover, de-associative techniques aim to sever the connection between quasi-identifiers (QI) and sensitive attributes by permuting sensitive values or releasing two separate tables—one for QI attributes and another for sensitive attributes. Finally, synthetic data involves the creation of artificial data that retains the original data's properties and characteristics.

In this work, we mainly focused on non-perturbative techniques, especially generalization and suppression. We detail both below.

- **Generalization.** Refers to the replacing of specific values in a data set with more general values. For example, the age 20 might be generalized to the age range 18-25. This makes it more difficult to identify individuals in the data set, as the adversary has less specific information to work with. It can be achieved either globally (all occurrences are replaced) or locally (specific occurrences are replaced).
- **Suppression.** Refers to the removing of sensitive values from a data set altogether. For example, the income of an individual might be suppressed if it is considered to be sensitive information. This can make the data less useful, but it can also provide a higher level of privacy protection. Different levels of suppression can be used to protect privacy in data publishing, including record

suppression (entire record deleted), value suppression (specific value deleted), and cell suppression (subset of instances deleted).

2.7 Related work

Some works have previously discussed the privacy theme within Brazilian databases. Significantly, most of these studies are recent, conducted within a 2 to 3-year timeframe. This highlights the novelty of this research topic, especially in Brazil.

The work in [2] undertakes a comprehensive quantitative study that centers on the official Brazilian Educational Censuses provided by INEP, the nation's educational statistics agency. The primary contributions include rigorously formalized attack models and extensive quantitative evaluations of privacy risks on INEP's Educational Census databases, which cover a significant portion of the population. The results reveal significant vulnerabilities in INEP's current disclosure control methods, which comprises direct identifier removal and pseudo-anonymization (artificially created identification code for each record). Furthermore, the study assesses the trade-off between privacy and utility using differential privacy variants. From this analysis, the authors could conclude that differential privacy is highly dependent on the mechanism and parameter for noise addition used, but also on correlations existent in the database itself. It is also challenging to properly setting a differential privacy implementation with optimal privacy and utility.

Additionally, [17], [16], and [18] also reveals the high susceptibility of INEP's databases under background knowledge attacks, which involves using external information or knowledge to identify individuals. Especially [17] and [18] presents a k-anonymization sensitivity analysis using the ARX tool, [17] also explore l-diversity results and [18] further evaluate the impact of generalization. Most of them discusses about the inadequacy of the privacy preservation method applied to these databases. In fact, [16] also observe similar results for the data available on the Transparency Portal of the Federal Executive Power and on the Transparency Portal of the government of the state of Rio Grande do Norte.

Notably, the work in [21] analyze the privacy within a data of patients tested positive for COVID-19 in the State of Ceará, Brazil using the differential privacy definition. The authors could conclude that Laplace and Exponential mechanisms can achieve an excellent balance between privacy and utility to answer to numerical and categorical queries, respectively. This result relates, however, to a Privacy-preserving Data Analysis framework instead of the Privacy-preserving Data Publishing context.

Furthermore, it is noteworthy to mention the contributions made by [19] and [20] in the context of big data. Both studies delve into the domain of big data within this framework. The initial study concludes that effective governance practices can potentially strengthen the privacy of anonymous data inherent to Big Data contexts. It introduces a governance framework specifically aimed at addressing privacy concerns. The latter author presents a re-identification risk-based anonymization framework tailored for large-scale data analytics platforms. The focus of this work is on designing and implementing essential anonymization components, with a specific emphasis on re-identification risks and the application of k-anonymity. The outcomes of this approach demonstrate its feasibility and success, revealing an enhanced level of privacy protection while minimizing the impact on data utility.

However, these prior studies do not investigate different types of attacks with varying setups, and for the most part, they do not establish a systematic and comprehensive methodology for evaluating and addressing anonymization in a broader sense.

3 METHODOLOGY

In this chapter, the main proposal of the work is described and formally defended.

3.1 Data Collection

The case study data is a comprehensive collection of epidemiological databases that have been maintained since 2009, originally focusing on Severe Acute Respiratory Syndrome (SARS) and influenza, and later included surveillance records of COVID-19 cases, from 2020. The surveillance of SARS in Brazil is carried out by the Ministry of Health, through the Department of Surveillance in Health, since the A(H1N1)pdm09 Influenza pandemic.

The primary system used to record SARS cases and deaths is the Influenza Epidemiological Surveillance Information System (SIVEP-Gripe⁴). The dataset covers individual-level information, encompassing data from notification forms on SARS cases of Brazilian individuals nationwide. It comprehends details such as patient demographics, residence, clinical and epidemiological information, received care, laboratory results, and case outcomes. This data is updated on a weekly basis and was extracted from the DATASUS open health data portal, which is published by the Ministry of Health. The data set for the years 2020 up to 2022 is particularly relevant for evaluating and monitoring the pandemic's impact in Brazil. Ministry of Health clarifies that the databases available on the portal undergo processing that involves anonymization before publication, in compliance with the Law 13.709/2018 (LGPD). Such anonymization refers to the removal of direct identifiers.

For this case study, we chose to analyze the nationwide SARS data set focusing particularly on the data for the year 2021. This is one of the largest ever published among all SARS data sets, primarily due to the significant number of COVID-19 case records in Brazil during that year [3]. It comprises almost 2 million individuals and 166 attributes. Most of the attributes are categorical (135), some are date type (27), and only 4 are numerical (being only 1 continuous). Considering information about signs and symptoms, risk factors, laboratory results, and imaging tests as sensitive attributes, then the data set contains a total of 84 sensitive columns. Additionally, if one also considered COVID-19 vaccination information as sensitive, there would be more 12 columns.

⁴ The Anísio Teixeira National Institute of Educational Studies and Research. <<https://www.gov.br/INEP>>

Regarding data preparation, we observed a variety of missing value codes ranging from one variable to another. For example, “I” for CS_SEXO, “9” for FATOR_RISC and <NA> for EVOLUCAO. To ensure data standardization, we re-encoded them all as "<NA>". Thus, observations with only null values were filtered out. Additionally, the age variable NU_IDADE_N was transformed to represent its values in years (days and months were converted to 0 year).

Furthermore, previous research highlights the vulnerability of these data sets [11]. They argue that the abundance of attributes in the data may make it easier to potentially identify individuals when cross-referencing with information from other public databases.

3.2 Methodology for Managing Re-Identification Risk

Approaches to privacy preservation vary from basic lists of variables to be either removed or modified [5] to more broadly defined strategies, like the "cell size of five" rule [1], or yet to well-structured systematic methodologies. Though various methods exist for anonymizing data sets, following methodological guidelines can provide valuable directions to facilitate the efficient application of privacy-preserving techniques.

Notably, the work presented in [1] serves as an example of a systematic methodology, being the object of investigation of our study. The Methodology for Managing Re-identification Risk aims to evaluate and mitigate re-identification risks associated with the publication of health data. By a step-by-step process, it equips data custodians with tools and guidance to effectively reduce the likelihood of re-identification while optimizing data utility, all within acceptable and context-specific thresholds. Higher-risk disclosures, for example, will require a higher level of protection. Despite its original design for health data, this process exhibits high adaptability, making it applicable across diverse contexts, and offering a versatile and flexible framework for risk management. Different government institutions related to data privacy points to the work of [1] as a suggestion for conducting data anonymization. For example, the Singapore's Personal Data Protection Commission (PDPC) [12], the Canada's Information and Privacy Commissioner of Ontario [13], and the National Institute of Standards and Technology (NIST) in the United States [14].

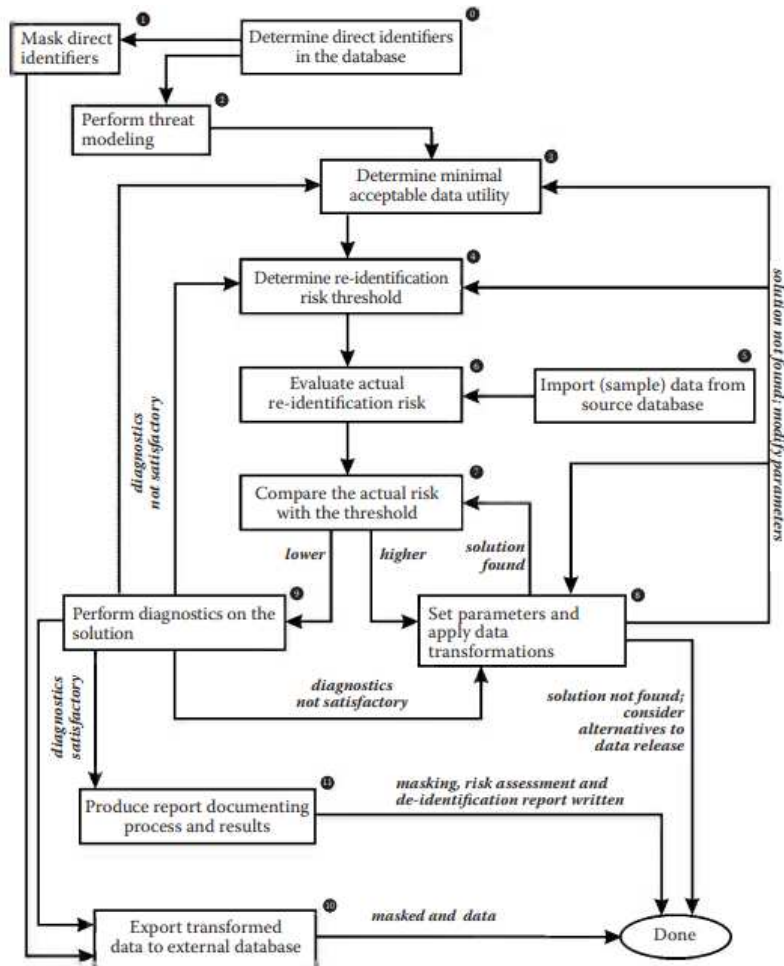


Figure 2. Overall re-identification risk management process, taken from [1].

The process for managing re-identification risk is described below and illustrated in Figure 2.

- **Step n° 0: Determine direct identifiers in the database.** An expert determines which fields in the database are direct identifiers. These fields promote the direct identification of an individual, such as name or telephone numbers. If the database has already been stripped of direct identifiers, then this step may not be needed. In this case, the step 1 is not performed, and the next step would be step 2. Otherwise, both steps 1 and 2 are executed sequentially.
- **Step n° 1: Mask direct identifiers.** The direct identifiers are either removed or replaced with pseudonyms. If the database has already been stripped of direct identifiers, then this step may not be needed. In some applications, only masking direct identifiers may be enough to prevent re-identification. For example, when releasing the data to a trusted data recipient in a controlled environment.

- **Step n° 2: Perform threat modeling.** Threat modeling consists of identifying potential adversaries and the information they might access, as well as determining the quasi-identifiers in database.
- **Step n° 3: Determine minimal acceptable data utility.** An expert examines what fields are the most appropriate given the use of the anonymized data. This step aims at imposing practical limits on how some data may be anonymized to preserve data utility.
- **Step n° 4: Determine re-identification risk threshold.** The organization determines what is an acceptable risk for releasing the dataset, based on the previous steps and regulatory requirements.
- **Step n° 5: Import (sample) data from the source database.** Importing data from the source database can be a complex task, depending on the data model. The authors explicitly included this step because it can consume significant resources and should be accounted for in the anonymization planning. For example, large datasets may require subsampling.
- **Step n° 6: Evaluate actual re-identification risk.** The actual identification risk is calculated.
- **Step n° 7: Compare the actual risk with the threshold.** Compare the actual risk, obtained in step 6, with the threshold determined in step 4.
- **Step n° 8: Set parameters and apply data transformations.** If the risk of re-identification is higher than the acceptable threshold, then privacy-preserving techniques are applied to the data, such as such as generalization and suppression. However, it may not be possible to find a solution within the specified parameters. In this case, the parameters may need to be adjusted or the threshold may need to be modified. If a solution is still not possible, then alternative data access mechanisms may need to be considered.
- **Step n° 9: Perform diagnostics on the solution.** Conduct analyses on the anonymized data to verify its adherence to assumptions and ensure adequate utility. If diagnostics are satisfactory, the data is then exported with a report. Otherwise, adjustments to re-identification parameters, risk threshold, and potentially revisiting privacy and utility thresholds may be necessary.

- **Step nº 10: Export transformed data to external dataset.** The anonymized data is exported. This step is included explicitly in the process because it can consume significant resources and must be accounted for in the anonymization planning.
- **Step nº 11: Produce report documenting process and results.** A report outlining the process and outcomes is generated and provided to the data custodian.

3.3 Research Design

To address the research questions, a case study on a Brazilian nationwide health microdata publicly available was conducted. In the study, we investigate the complete data anonymization process following the Methodology for Managing Re-Identification Risk proposed in [1]. From a real-world data set, we aim to better understand the challenges related to preserving privacy in open microdata in Brazil, while assessing the applicability of the existing techniques to the Brazilian context. We can divide the case study into three main parts, derived from [1].

Firstly, we define the privacy threats considered in the study, which covers the methodology's steps 0 to 5. We assume that the resulting protected data set will be released as a Public Use File (PUF) [6], i.e., the same data sharing approach of the case study microdata. This release model allows the public to access and use the data for multiple purposes, including for research. Therefore, due to the wide range of users who could access this data, it is challenging to evaluate accurately potential malicious adversaries who might compromise individuals' privacy [1]. In this context, it is recommended that the data custodian adopts a worst-case perspective regarding the intentions and capabilities of potential data recipients [1].

Under this assumption, we outline the adversaries considered in the case study. We assume that adversaries have complete access to a version of the dataset at a specific point in time. Their goal is supposed to either identify individuals (re-identification attack) or infer sensitive attributes (attribute-inference attack). The focus can be on a specific individual or as many as possible. They may also gather auxiliary information to accomplish the attacks, and they may or may not know whether their target(s) is(are) in the database. Given time and financial constraints, we do not expect adversaries are willing to spend money in the acquisition of other databases or to dedicate a significant amount of time to the attack, so as not to interfere with their personal life.

There are several other factors that could be considered to better evaluate a potential adversary according to the context of the data publication, for example the adversary's willingness to commit illegal acts [1]. However, for the sake of simplicity, we won't extensively consider them in this analysis, as a more comprehensive evaluation would require the involvement of subject matter experts.

To further conduct the analysis, we also define which auxiliary information an adversary may access. Due to the computational costs involved, including time and memory usage, we are limited to just a few of the dozens of attributes available in the dataset. For quasi-identifying attributes, we considered the ten attributes in Table 1. These were selected based on how easily an adversary could access that information, such as date of birth, gender, and city of residence. Similar attributes have been also considered in previous works [2]. We believe that these variables have a great potential to facilitate the re-identification of individuals in the dataset by cross-referencing information from other available data sources or utilizing prior knowledge (if the target is known). Additionally, we have chosen three attributes related to personal health to state as sensitive for posteriorly evaluating attribute-inference risks. These sensitive attributes are listed in Table 2.

Table 1. Variables chosen as quasi-identifiers in the case study.

Variable name	Description	Values
DT_NASC	Patient's birth date, format "dd/mm/YYYY"	18/09/1870 - 01/01/2022
NU_IDADE_N	Patient's age in day, month or year	0 - 99
CS_SEXO	Patient's gender	"F"- Female "M"- Male "I"- Ignored
CS_RACA	Color or race declared by the patient	1- White 2- Black 3- Yellow 4- Mixed race 5- Indigenous
ID_PAIS	Patient's country name	Example, "BRASIL" (Brazil)
SG_UF	Patient's state abbreviation	Example, "RJ" (Rio de Janeiro)
ID_MN_RESI	Patient's city of birth name	Example, "RIO DE JANEIRO"
ID_UNIDADE	Health facility name who performed the service	Example, "HOSPITAL COPA DOR" (COPA DOR Hospital)
EVOLUCAO	Case outcome	1- Recovery 2- Death 3- Death from other causes
DT_EVOLUCA	Date of hospital discharge or death	03/01/2021 - 28/02/2023

Table 2. Variables chosen as sensitive attributes in the case study.

Variable name	Description	Values
FATOR_RISC	Whether the patient presents any risk factor	1- Yes 2- No 9- Ignored
TP_AMOSTRA	Sample type collected for diagnostic testing	1- Nasopharyngeal secretion 2- Bronchoalveolar lavage 3- Post-mortem tissue 4- Other, which one? 5- Cerebrospinal fluid (CSF) 9- Ignored
RAIOX_RES	Patient's chest x-ray result	1- Normal 2- Interstitial infiltrate 3- Consolidation 4- Mixed 5- Other 6- Not performed 9- Ignored

It is worth recognizing that the selection of these quasi-identifiers and sensitive attributes is arbitrary and may vary in significance over time, and even across countries. The primary objective of this study is to portray potential real-life application scenarios, while examining the applicability of the investigated methodology. Thus, it is out of our scope classifying such attributes.

Secondly, we evaluate the actual risks of the case study data under the previously defined settings, which covers the methodology's steps 5 and 6. To evaluate quantitatively the adversary's re-identification success, we considered the classical Prosecutor and Journalist attack models [1], and the Collective-target and Individual-target Re-identification Single database (CRS and IRS, respectively) attacks [2]. For the attribute-inference attacks, we analyzed the values for the distinct 1-diversity [9], t-closeness [9], and Collective-target and Individual-target Attribute-inference Single database (CAS and IAS, respectively) attacks [2]. The Total Variational Distance (TVD) metric was adopted to calculate the t distances in the t-closeness property of each categorical sensitive attributes [7]. All the metrics were implemented in Python⁵, except the Journalist metric, which was calculated via the ARX Data Anonymization tool⁶. Also, the metrics in [2] were calculated using the Python code⁷ shared by the own authors. It is important to highlight that we did not conservatively consider the influence of missing values on the adversary's success [8]. By considering these different metrics, we aim to assess risk from multiple perspectives and to provide a comparative study of the results.

⁵ Python programming language. <<https://www.python.org/>>

⁶ ARX Data Anonymization Tool. Comprehensive open-source software for anonymizing sensitive personal data. <<https://arx.deidentifier.org/>>

⁷ BVM library. <<https://github.com/nunesgh/bvm-library>>

Finally, in the last part, we propose a possible strategy to mitigate the risks involved in the case study data, which would cover the methodology's steps 7 to 11. We evaluate, in practice, how this data could be effectively anonymized before its publication, while also guaranteeing its usefulness. We prioritized the use of straightforward and easily understandable anonymization techniques, such as generalization and suppression. After applying these techniques, we re-analyzed the privacy risk with the same metrics as before. We then discuss the results, proposing what could further be done to improve it under the light of data utility. This helps us understand the trade-off between preserving privacy and ensuring data utility.

4 RESULTS

In this section, we present the results of the case study, in which we investigate the application of the de-identification methodology proposed in [1]. It analyzes how the anonymization process might be approached in a real-world case. We first present the results for the re-identification and attribute-inference attacks on the case study data. Then, we conclude with an evaluation of traditional anonymization techniques, such as suppression and generalization.

4.1 Experimental Results of Re-Identification Attacks

The results of the quantitative analyses on the re-identification risk are organized as follows:

- *Figure 3* shows the distribution of possible values for the equivalence classes.
- Table 3 summarizes the results of the re-identification Prosecutor (minimum, average and maximum), Journalist (estimated) and CRS (deterministic, probabilistic, prior and posterior) attacks.
- *Figure 4* shows the re-identification probabilistic success for each combination of quasi-identifying attributes.
- *Figure 5* shows the minimum combination of quasi-identifying attributes to reach the highest re-identification probabilistic success and *Figure 6*, the maximum combination to reach the privacy threshold.
- *Figure 7* shows the histograms for the distribution of individual scores according to the re-identification probabilistic success for the quasi-identifying attributes in Table 1 and for some subsets of different sizes.

We start by investigating the distribution of the equivalence classes in the data set. According to the *Figure 3*, over 99% of the approximately 1.2 million patients with non-null values for quasi-identifiers attributes are sample unique, i.e., there is only one individual with the same values for the quasi-identifiers across the entire dataset. This is the worst-case scenario for re-identification, in the sense that individuals are highly exposed. An adversary interested in a target individual in this data set and aware of information about its quasi-identifiers and its record existence in the dataset can easily re-identify this individual by means of a simple query. This assertion is supported in Table

3. Here, the high re-identification metric results for the Prosecutor and CRS attack provide clear evidence of the high risk of re-identification for individuals who were notably belonging to the data set. The average Prosecutor risk, and the deterministic and probabilistic CRS attacks, are close to the maximum risk value of 100%. Furthermore, even for the Journalist configuration, in which the adversary may unknow if the target is in the database, the results are still very high.

In contrast, the highest equivalence class size reaches 3, but with an astonishing low percentage. Theoretically, the more records share the same quasi-identifiers, the more difficult it is to re-identify an individual. The records with an equivalence class size of 3 should have a re-identification risk of 33%, i.e., for each record there would have another two equal ones, being the most protected in this data set. However, these cases are very far from being representative (only few 57 individuals), and even this equivalence class size is considered low. In [1], for practical cases, the most permissive suggestion for a minimum equivalence class size is 10, which, still, would result in a re-identification success rate of 10%. Also, if we consider 5%, or even 10%, as the maximum acceptable re-identification risk (under Prosecutor and CRS assumptions), then all records would be at risk.

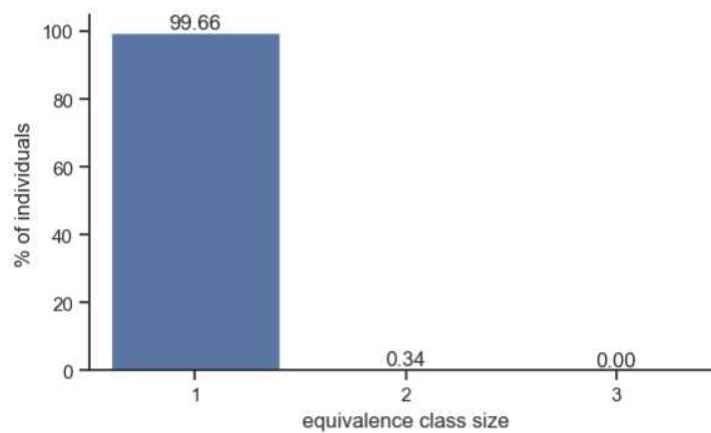


Figure 3. Distribution of possible values for the equivalence class. In the x axis, there is the size of the equivalence class, and, in y axis, there is the % of individuals present in such equivalence class size.

Table 3. Re-identification results. Results for Prosecutor (average and maximum), Journalist (estimated) and CRS (deterministic – dCR, probabilistic – pCR, prior and posterior probabilistic success) attacks.

		CRS risk			
Prosecutor risk	Estimated Journalist risk	dCR	pCR	Prior	Posterior
Min: 33%	100%	99.6%	1224562	0.00008%	99.8%
Avg: 99.8%					
Max: 100%					

One may speculate that the high privacy risk for the records is due to the number of quasi-identifying attributes. It might be assumed that having knowledge of all 10 quasi-identifiers would be necessary for a successful re-identification attack. However, the results challenge this assumption. Figure 4 shows the mean re-identification risk for each combination of quasi-identifiers, calculated considering the Prosecutor Model as defined in Section 2. Here, one should note that three combinations of only two quasi-identifiers already pose a mean re-identification success rate of 80%, or more. For example, DT_NASC and DT_UNIDADE have a mean re-identification risk of more than 90%. This means that most individuals can be re-identified by knowing these two variables. In fact, the minimum combination⁸ that achieves almost 100% of mean re-identification risk consists of only three attributes (DT_NASC, ID_UNIDADE, and DT_EVOLUCAO) as shown in Figure 5. Interestingly, knowing only these three attributes results in re-identification metrics that are strikingly similar to the results obtained when knowing all 10 quasi-identifiers, both highly concentrated at individual 100% risk, as evident in Figure 7.

In this context, an adversary who is aware of an individual's presence in the data set would likely know, or could easily discover, general information about the individual's birth date and which health facility they were hospitalized too. For example, they could access this information from social media or other external databases. This would make re-identification even easier, as the adversary would only need to know a few additional information to identify the target individual.

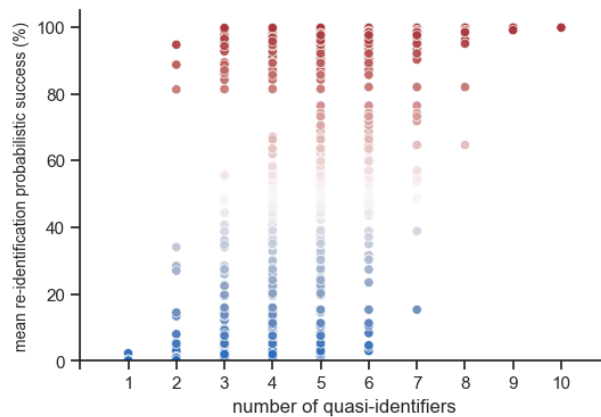


Figure 4. Adversary's re-identification success by combination of quasi-identifiers. Each dot represents the mean re-identification probabilistic success for a distinct scenario defined by the selected quasi-identifying attributes among the 1 024 (2^{10}) possibilities. For example, the combination DT_NASC and DT_UNIDADE is the first dot from top to bottom under the combinations of 2 quasi-identifiers, with the highest mean re-identification risk of more than 90%.

The horizontal axis defines the size of each combination of quasi-identifiers, while the vertical axis defines the adversary's re-identification probabilistic success.

⁸ More details on how the minimum size combination with highest mean re-identification risk was found in the Figure 5 description.

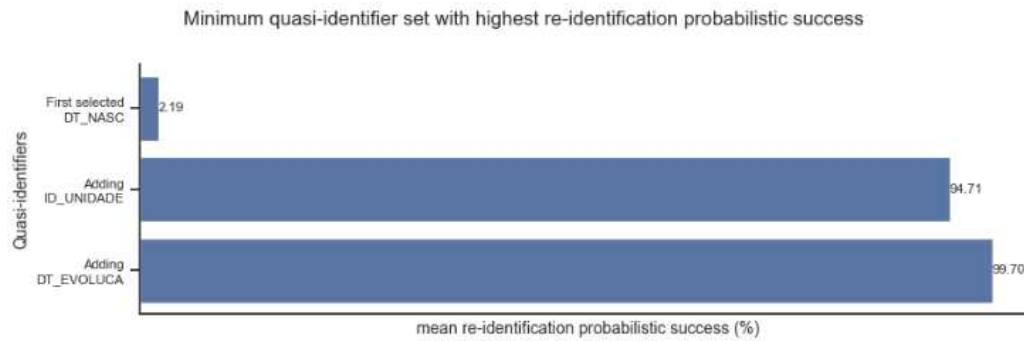


Figure 5. Minimum quasi-identifying combination high mean re-identification risk. Minimum quasi-identifying combination to achieve the highest re-identification probabilistic success. Such result was obtained by a greedy approach. The algorithm will add quasi-identifier by quasi-identifier sequentially, always choosing the one that leads to the highest mean re-identification probabilistic success. It will stop when reaches a threshold of 100%.

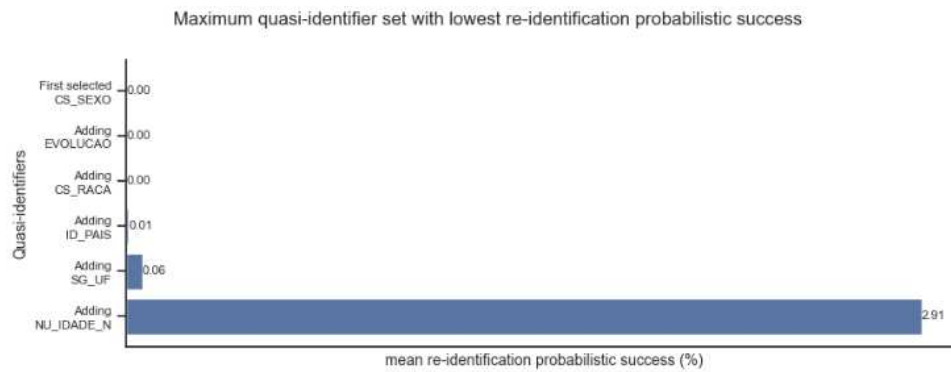


Figure 6. Maximum quasi-identifying combination minimal mean re-identification risk. Maximum quasi-identifying combination to achieve the privacy threshold of 0.05 (5%) of re-identification probabilistic success. Such result was obtained by a greedy approach. The algorithm will add quasi-identifier by quasi-identifier sequentially, always choosing the one that leads to the lowest mean re-identification probabilistic success. It will stop when reaches a threshold of 5%.

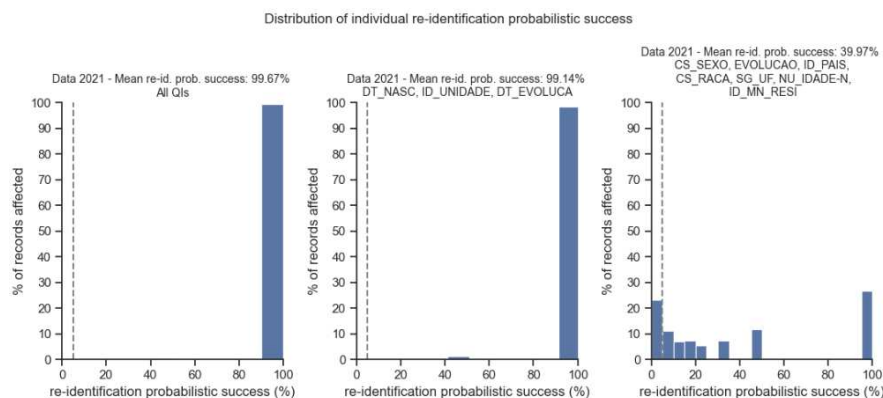


Figure 7. Histograms for the distribution of individual IRS scores. From left to right, the first histogram represents the individual re-identification probabilistic success considering the quasi-identifiers from Table N1. The second and third histograms considers the minimum and maximum combination of quasi-identifiers to achieve highest and lowest re-identification probabilistic success, respectively, as in Figure 5 and Figure 6. The dashed line corresponds to a re-identification threshold of 5%.

On the other hand, Figure 4 also shows that there are combinations of quasi-identifiers with a significantly lower mean of re-identification success rate up to size 7. For example, Figure 6 shows that the combination of CS_SEXO, EVOLUCAO, CS_RACA, ID_PAIS, SG_UF, and NU_IDADE_N is the maximum combination⁹ with a mean re-identification success rate of less than 5%. This suggests that the privacy risk for the records might be due to the specific combinations of quasi-identifiers: some combinations seem to hold more potential for re-identifying individuals than others. This evidence shows how challenging is to protect individuals in this data set.

By further analyzing these attributes, we can see that some of them have a large number of possible outcomes, as shown in Table 4. In particular, DT_NASC has the most expressive number of 38,000 possible values, followed by ID_UNIDADE (6,055), ID_MN_RESI (5,306), DT_EVOLUCA (677) and NU_IDADE_N (123). Among the 10 quasi-identifying attributes, only 4 (SG_UF, CS_RACA, EVOLUCAO, CS_SEXO) present a low number of categories, none of them unique to an individual. For all the other 6 variables, the granularity of the information is smaller, which tends to present specific categories of only one individual. For example, in the case study sample, 601 health facilities attended to only one patient, 12 municipalities have only one resident and more than 1,000 patients have birth dates that do not match with any other individual in the data set, as can be seen in Table 4.

It is also worth to highlight the attribute ID_PAIS. Despite being a variable that offers high-level information (country of residence), the context behind the studied data set makes it a potential source of disclosure. Essentially, most of the patients (>99%) are Brazilian residents and only a few (263) individuals are from another country. As the data records cases of flu-like syndrome in Brazil, it is much more likely for an individual to be a Brazilian citizen than a foreigner. In the sample, ID_PAIS register 29 distinct countries, half of them (15) refers to only one individual and most of them (24) hold less than 10 individuals. This observation suggests that foreigners in this dataset might have a higher re-identification risk if compared to Brazilian patients.

The potential re-identification risk from a specific attribute is not only related to the number of its possible outcomes, but also to their distribution. For example, an attribute with a small number of possible values that are not evenly distributed may incur a higher risk than an attribute with a large number of equally likely outcomes.

⁹ More details on how the maximum size combination with minimum mean re-identification risk was found in the Figure 6 description.

This is because the distribution of outcomes of an attribute can impact the uniqueness and distinguishability of individuals in a dataset. On top of that, the combination of attributes can also affect this outcome distribution.

Figure 8 illustrates these ideas. By observing the age variable (NU_IDADE_N) according to the patient's outcome (EVOLUCA), either recovery (Scenario 1) or death (Scenario 2), we can observe roughly the same amount of possible age values, around 120. At the same time, the corresponding age distribution for each outcome is different: Considering the patients who died, we observe higher density for the elderly and less of the younger. This is expected for flu-like syndromes, in which elderly is more likely to die. As a result, the distribution of age in Scenario 2 is more concentrated in a specific interval, which makes it more uneven. This uneven distribution may become a higher re-identification risk, being easier for an adversary to identify individuals in the data set. In fact, Scenario 1 (recovery) holds a larger equivalence class groups than Scenario 2 (death), which corroborates this hypothesis about the distribution balance.

We can relate the distribution of the attributes with the adversary's uncertainty in terms of entropy. Broadly speaking, entropy is the amount of uncertainty a random variable carries about its outcomes. The lower entropy of a random variable, the more predictable its outcomes are. In the context of the age variable, the asymmetry of the ages related to elderly incurs a lower entropy than for those individuals who recovered from COVID-19. In the example in Figure 8, we can observe that the entropy of the age distribution among patients who died is lower than among the ones who recovered. This means that an adversary would be more uncertain in Scenario 1 than in Scenario 2.

Table 4. Outcomes by Quasi-identifier.

For each quasi-identifier, the second column presents the number of possible outcomes and, next to it, it is the representativeness in overall data set in parenthesis. The third column shows the number of outcomes that contains only one record with its proportion among the distinct values of the attribute in parenthesis. The last column indicates the mean re-identification success when considering each attribute as the only information available as quasi-identifier.

Quasi-identifier	Number of outcomes	Number of outcomes with only 1 record	Mean risk
DT_NASC	37992 (2.2%)	1023 (2.7%)	2.2%
ID_UNIDADE	6055 (0.3%)	601 (9.9%)	0.3%
ID_MN_RESI	5306 (0.3%)	12 (0.22%)	0.3%
DT_EVOLUCA	677 (0%)	68 (10%)	0%
NU_IDADE_N	123 (0%)	5 (4%)	0%
ID_PAIS	29 (0%)	15 (51.7%)	0%
SG_UF	27 (0%)	0	0%
CS_RACA	5 (0%)	0	0%
EVOLUCAO	3 (0%)	0	0%
CS_SEXO	2 (0%)	0	0%

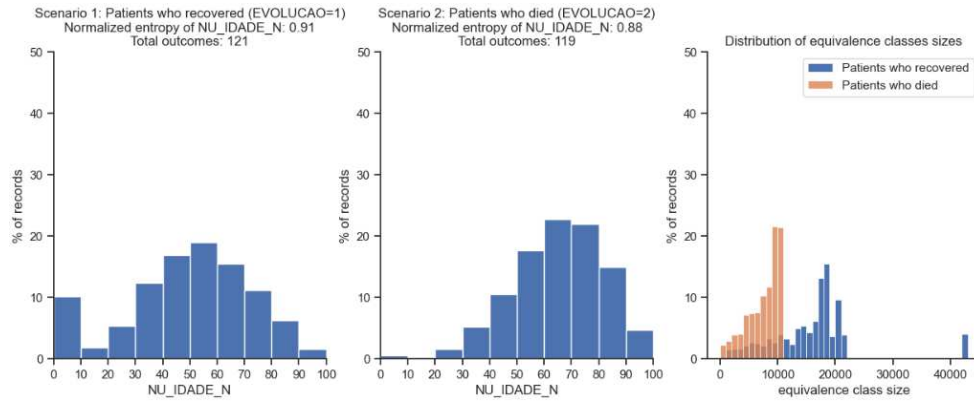


Figure 8. Distribution of age (NU_IDADE_N) by values of outcome (EVOLUCAO). Distribution of age (NU_IDADE_N) values considering two scenarios: All patients that recovered (Scenario 1, in the left considering EVOLUCAO=1) and all patients that died (Scenario 2, in the middle considering EVOLUCAO=2). The third histogram represents the distribution of equivalence class sizes among the two scenarios considering only age as quasi-identifier. The limit of y axis goes up to 50% (out of 100) to be able to better visualize the results.

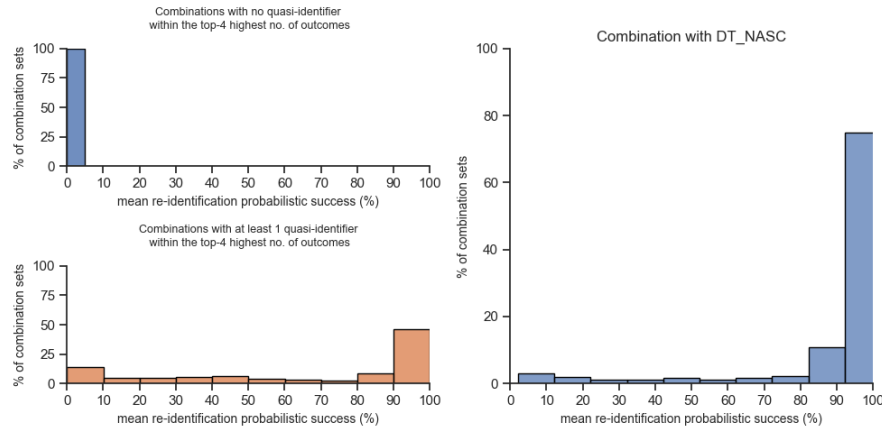


Figure 9. Mean re-identification success by quasi-identifiers combinations with and without large number of outcomes. In the top left, the histogram shows the mean re-identification success rate for combinations of quasi-identifiers without any of the top-4 attributes with the highest number of outcomes (DT_NASC, ID_UNIDADE, ID_MN_RESI, and DT_EVOLUCA). In the bottom left-hand corner, the histogram shows the mean re-identification success rate for combinations of quasi-identifiers with at least one of these attributes. In the right-hand side, the histogram of the mean re-identification success rate for all combinations of quasi-identifiers with DT_NASC.

In Figure 9, we can see that among the combinations of the quasi-identifiers that do not consider the top-4 high-outcome variables (DT_NASC, ID_UNIDADE, ID_MN_RESI, and DT_EVOLUCA), the maximum mean re-identification success observed is 3%. This means that any combination with a higher risk (above a threshold of 5%) includes at least one attribute with a significant number of outcomes. For example, in Figure 4, the three combinations of size 2 with a mean risk of 80% or more are DT_NASC and ID_UNIDADE (94.7%), DT_NASC and DT_EVOLUCA (88.7%), and DT_NASC and ID_MN_RESI (81.2%). Notably, all of them contains DT_NASC, which is the attribute with the highest number of unique values.

In fact, when combined with other attributes, DT_NASC can significantly increase the number of expected re-identifications: 87% of combinations that include DT_NASC have a mean re-identification risk of around 90%, corroborating the distribution in Figure 9.

Table 5. Mean re-identification success by Quasi-identifier combinations with attributes with large number of outcomes. Quasi-identifiers combinations of high-outcome variables and the expected re-identification probabilistic success. The sample size column indicates the total number of non-null records considering the combination.

Quasi-identifier combination	Mean re-identification risk	Sample size
DT_NASC and ID_UNIDADE	94.7%	1731072
DT_NASC and DT_EVOLUCA	88.7%	1470369
DT_NASC and ID_MN_RESI	81.3%	1730811
ID_UNIDADE and DT_EVOLUCA	34%	1471106
ID_MN_RESI and DT_EVOLUCA	27%	1470870
ID_MN_RESI and ID_UNIDADE	5%	1731979
ID_UNIDADE and ID_MN_RESI and DT_EVOLUCA	55.5%	1470870

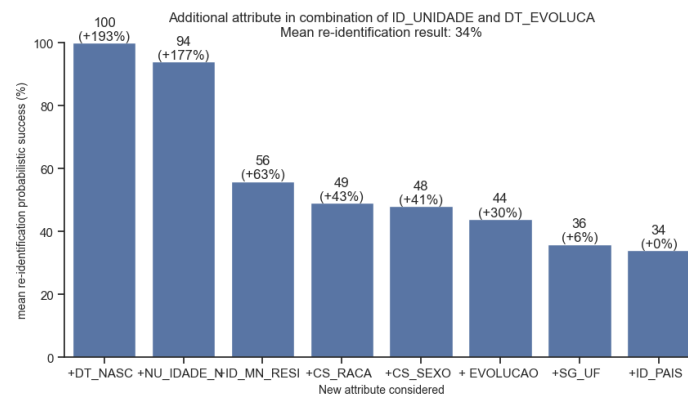


Figure 10. Mean re-identification success for ID_UNIDADE and DT_EVOLUCA when adding another attribute. Each bar corresponds to new mean re-identification success when the attribute is presence. For example, the first bar represents the results for ID_UNIDADE, DT_EVOLUCA and DT_NASC. In parenthesis is the difference between the results for ID_UNIDADE and DT_EVOLUCA and the results considering the new variable.

Conversely, when DT_NASC is not included, the other high-outcome attributes do not yield as much re-identification risk. The pair with the highest number of expected re-identifications in Table 5 is ID_UNIDADE and DT_EVOLUCA, with 34%. This is much lower than the pairs that include DT_NASC, which start from 81%. Even combining ID_UNIDADE, ID_MN_RESI and DT_EVOLUCA still yields results lower than the ones observed with DT_NASC (55.5%). It is interesting to note that given the pair ID_UNIDADE and DT_EVOLUCA, adding DT_NASC or NU_IDADE_N would both result in a substantial increase compared to the other available attributes. This can be seen in Figure 10. Although these two variables have different numbers of outcomes, they both essentially provide the same information: the individual's age. When combined with health facility and the date of case evolution, this information has a high potential

for re-identification due to rare combinations. For example, it is very unlikely to find a relevant set of people who share the exact same age, the exact same date of death, and the exact same healthcare service unit. Furthermore, it is also worth mention that the combination ID_MN_RESI and ID_UNIDADE presents a surprisingly low mean re-identification success of 5%. These two attributes are the top-2 attributes with the greatest number of outcomes. But, different from the other combinations in Table 5, their combined values still contain a relevant number of individuals that share the same attributes values. This occurs since one variable aggregates information to the other: the availability of health units varies by location, i.e., some units are only located in one municipality. In contrast, by combining ID_UNIDADE with DT_NASC, we observe a very different scenario. The health unit values do not correlate with the patient's date of birth, being more difficult to find a relevant number of individuals with sample unique for these combined values.

Figure 11 reinforces this discussion. The attribute ID_UNIDADE has a normalized entropy of 0.9, which means there is a high level of uncertainty for its outcomes. This is because ID_UNIDADE has many possible values, and these are not evenly distributed. Some health units are naturally more popular than others, due to factors such as location, capacity, and specialization.

When we join the birth date information (DT_NASC) to ID_UNIDADE, the entropy of ID_UNIDADE does not change significantly. This suggests that DT_NASC does not hold much additional information about ID_UNIDADE. However, when we add the municipality (ID_MN_RESI), we see that the entropy of ID_UNIDADE changes much more. This suggests that ID_MN_RESI is a more informative variable than DT_NASC for predicting ID_UNIDADE, i.e., both variables share common information. These findings suggest that ID_UNIDADE is more strongly dependent on ID_MN_RESI than on DT_NASC. This is likely because people are more likely to seek health units in their own municipality, regardless of their birth date.

Independent variables are more likely to have rare combinations of values because they tend to produce high entropy distributions. Thus, the number of observations required in the sample to keep privacy gets greater. On the other hand, dependent variables are related to each other, so there are fewer possible combinations of values that are rare, being of low entropy. This is less likely that observations are unprotected.

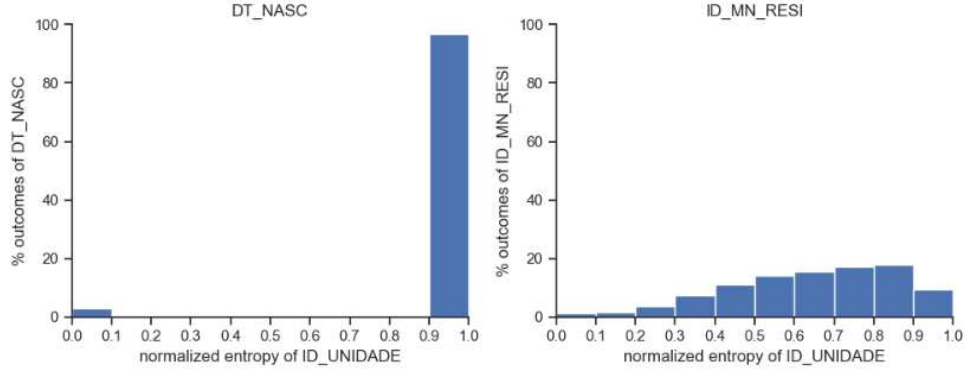


Figure 11. Normalized entropy values for ID_UNIDADE by DT_NASC and ID_UNIDADE. In the left, the distribution of entropy values of ID_UNIDADE given the outcomes of DT_NASC. For example, the entropy of ID_UNIDADE when DT_NASC equals 04/27/1998. In the right, the distribution of entropy values of ID_UNIDADE given the outcomes of ID_MN_RESI. For example, the entropy of ID_UNIDADE when ID_MN_RESI equals “RIO DE JANEIRO”.

In conclusion, we assessed this data set to have a high probability of re-identification. Our analyses suggest the volume of unique entries, the sensitivity of the data and the possibility of an adversary obtaining necessary auxiliary information are key for increasing individuals’ re-identification risk. This is also supported by the privacy metrics. Despite challenging, it is entirely possible that a motivated adversary could acquire additional data and learn sensitive information about an individual from this data set. Notably, our findings indicate that combining attributes with numerous outcomes, like DT_NASC, with a highly unbalanced distribution, like ID_PAIS, and with a degree of uncorrelatedness, like DT_NASC and ID_UNIDADE, might result in a sparse distribution that further increases the likelihood of expected re-identifications. These insights emphasize the importance of implementing robust privacy protection measures when dealing with this characteristic.

4.2 Experimental Results of Attribute-Inference Attacks

For the analysis of attribute-inference attacks, we considered the sensitive attributes listed in *Table 2*. The results are reported as follows:

- *Table 6* summarizes the results of the attribute-inference distinct l-diversity (minimum, median and maximum), t-closeness (minimum, median, and maximum) and CAS attacks (deterministic, probabilistic, prior and posterior).
- *Figure 12* illustrates the distribution of values for each sensitive attribute, considering both the overall data set and the equivalence classes.
- *Figure 13* displays the distribution of distances between the original distribution of the sensitive attribute and its distribution in an equivalence class, for each sensitive attribute.

- Table 7 shows the results of the attribute-inference distinct l-diversity (minimum, median and maximum), t-closeness (minimum, media, and maximum) and CAS (probabilistic, prior and posterior) attacks, for each sensitive attribute considering the quasi-identifiers subsets in Figure 5 and Figure 6.

Table 6. Attribute-inference results. Results of the attribute-inference distinct l-diversity (minimum l, median l, maximum l), t-closeness (minimum t, median t and maximum t) and CAS (deterministic – dCA, probabilistic – pCA, prior and posterior probabilistic success) attacks, considering the quasi-identifiers in Table N1.

Sensitive attribute	L-diversity risk	T-closeness risk	CAS risk			
			dCA	pCA	Prior	Posterior
FATOR_RISC	Min: 1 Median: 1 Max: 2	Min: 0.08 Median: 0.42 Max: 0.58	99.9%	1.79	55.8%	0.99%
TP_AMOSTRA	Min: 1 Median: 1 Max: 2	Min: 0.1 Median: 0.1 Max: 0.99	99.9%	1.12	89.2%	0.99%
RAIOX_RES	Min: 1 Median: 1 Max: 3	Min: 0.25 Median: 0.44 Max: 0.97	99.9%	1.78	56.2%	0.99%

As previously mentioned, the prevalence of sample unique records in the dataset is extremely relevant. As a result, one can reasonably expect high risk of inferring sensitive attributes. This is because once an individual is re-identified, the adversary has access to all other information available, including sensitive ones. The values presented in Table 6 confirm that. In particular, the l-diversity results show that most equivalence classes have only one distinct outcome, i.e., 1-diversity, for all the three sensitive attributes considered. This relates to the fact that the majority of this equivalence classes is of size 1 (sample uniques). Additionally, the CAS deterministic and probabilistic measures are very close to the maximum risk value of 100%. This is because these measures assume that the adversary will always choose the most frequent class in an equivalence group. Since most equivalence groups contain only one individual, the risk is very high.

However, as shown in Figure 12, even within groups of size larger than 1, a single outcome remains dominant across all sensitive attributes. For example, 76% of the equivalence classes of size 2 for FATOR_RISC have only one distinct value. This is the worst-case scenario for an attribute-inference attack. Even though the re-identification risk would be 50% in this case, the adversary could still infer the target's sensitive value because both records have the same value for the sensitive attribute.

Figure 12 additionally shows the distribution of values for the sensitive attributes. Notably, FATOR_RISC stands out due to its relatively balanced distribution, with both

of its possible values being well represented: 56% for the category “1” (Yes) and 46% for the “0” category (No). In contrast, the attributes TP_AMOSTRA and RAI0X_RES exhibit significantly imbalanced distributions. Particularly, a single TP_AMOSTRA value constitutes 89% of the records. Similarly, within the six possible values for RAI0X_RES, two of them cover approximately 80% of the individuals. These findings suggest that some sensitive attributes may incur a different overall inference risk.

For example, in an extreme imbalance distribution, such as TP_AMOSTRA, if an adversary assumed that all individuals have the most frequent value ("6"), they would be wrong for only a small percentage (10%) of individuals. This can be seen in the Prior CAS risk in Table 6, which assumes this exact strategy, i.e., the success probability of inferring a sensitive attribute relates to the frequency of the most common value of that attribute in the data set. The FATOR_RISC and RAI0X_RES present around 56% of the initial risk.

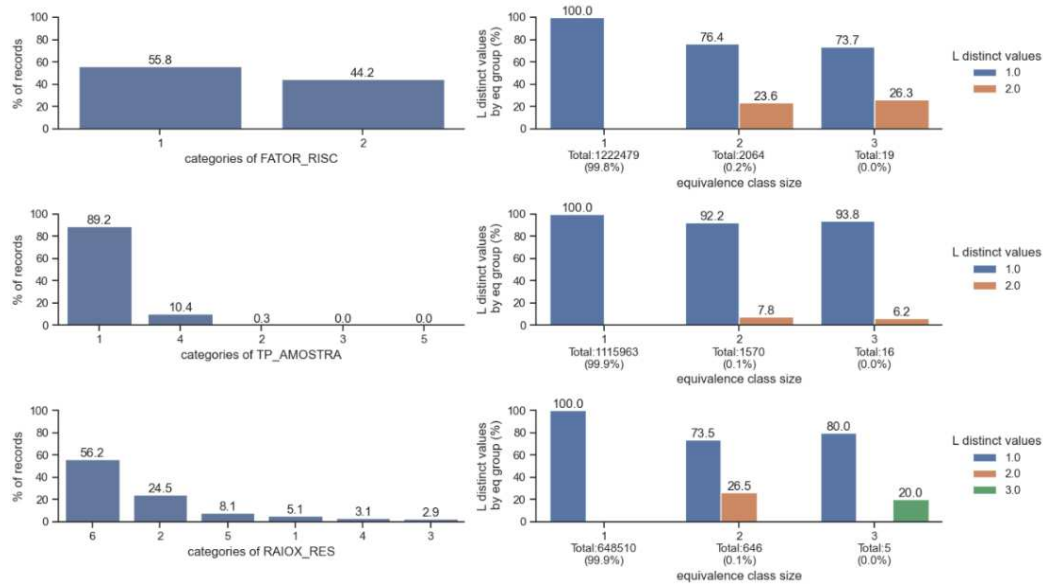


Figure 12. Distribution of sensitive values by overall data set and by equivalence class size. In the left, it can be found in each row the distribution of values for the sensitive attributes in Table N2 in the overall data set. In the right, it is shown the distribution of distinct values of a sensitive attribute by equivalence classes, considering the quasi-identifiers in Table N1. Observations with empty values for the sensitive attributes were disregarded. Also, it was only examined records that presented values for all quasi-identifiers attributes for the results in the right.

The imbalance of these distributions is reflected within the equivalence groups. For example, the class "1" of TP_AMOSTRA is the most frequent in the data set, so we expect to also observe it more frequently in the equivalence groups. This implies that, overall, the distribution of TP_AMOSTRA values inside the equivalence classes is similar to the distribution of TP_AMOSTRA values across the entire case study data set. The t-closeness results in Table 6 and Figure 13 support this observation.

In this context, we can see that most equivalence classes have 0.1-closeness regarding TP_AMOSTRA, i.e., the total variational distance between these distributions are very low (0.1).

Under the t-closeness model, this specific value would be considered optimal as it attends the proximity requirement of the distributions mentioned. However, in this case, the t value of 0.1 does not help in avoiding inference of TP_AMOSTRA. If the value "1" of TP_AMOSTRA is highly frequent in the original data, it will similarly be highly represented within equivalence classes. Consequently, this could make it easier for adversaries to infer individuals' sensitive attributes based on the dominant class (measured by the CAS risk).

On the other hand, FATOR_RISC and RAIOX_RES presents equivalence groups that differs from the global corresponding attribute distribution. For instance, taking the most representative value of "6" (Not performed) for RAIOX_RES results in a considerable deviation (0.4) from the original distribution of RAIOX_RES. Furthermore, some equivalence classes for RAIOX_RES exhibits around 0.9-closeness, meaning that their distribution is almost opposite to the primary one. One illustrative case is a sample unique with "3" (Consolidation) as RAIOX_RES value: this class has now 100% representativeness, whereas it was previously only 2.9%.

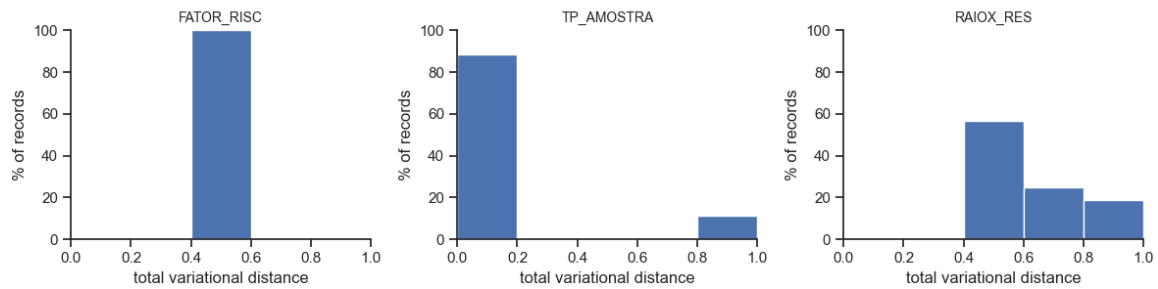


Figure 13. Distribution of t-closeness values for each sensitive attribute. Each plot holds the distribution of the distances between the distribution of a sensitive attribute in an equivalence class and the distribution of the same attribute in the whole data set, taken from the t-closeness definition.

(a) Attribute-inference results for FATOR_RISC.

			CAS risk			
Quasi-identifiers	L-diversity risk	T-closeness risk	dCA	pCA	Prior	Posterior
Min subset	Min: 1 Median: 1 Max: 2	Min: 0.07 Median: 0.42 Max: 0.58	99.5%	1.79	55.8%	99.8%
Max subset	Min: 1 Median: 2 Max: 2	Min: 0 Median: 0.26 Max: 0.58	2.1%	1.23	55.8%	69.1%

(b) Attribute-inference results for TP_AMOSTRA.

			CAS risk			
Quasi-identifiers	L-diversity risk	T-closeness risk	dCA	pCA	Prior	Posterior
Min subset	Min: 1 Median: 1 Max: 2	Min: 0.1 Median: 0.1 Max: 0.99	99.9%	1.12	89.2%	99.9%
Max subset	Min: 1 Median: 1 Max: 5	Min: 0 Median: 0.1 Max: 0.99	5.4%	1.0	89.2%	88.8%

(c) Attribute-inference results for RAIQX_RES.

Table 7. Attribute-inference results for quasi-identifier combinations. Results of the attribute-inference distinct l-diversity (minimum l, median l, maximum l), t-closeness (minimum t, median t and maximum t) and CAS (deterministic – dCA, probabilistic – pCA, prior and posterior probabilistic success) attacks, for sensitive attributes FATOR_RISC (a), TP_AMOSTRA (b) and RAIQX_RES (c) considering the subsets in Figure F4. The “min subset” represents the minimum quasi-identifying combination to achieve the highest re-identification probabilistic success (DT_NASC, ID_UNIDADE and DT_EVOLUCA) and the “Max subset” represents the maximum quasi-identifying combination to achieve the privacy threshold of 5% (CS_SEXO, EVOLUCAO, ID_PAIS, CS_RACA, SG_UF, and NU_IDADE_N).

			CAS risk			
Quasi-identifiers	L-diversity risk	T-closeness risk	dCA	pCA	Prior	Posterior
Min subset	Min: 1 Median: 1 Max: 3	Min: 0.2 Median: 0.43 Max: 0.97	99.8%	1.78	56.2%	99.9%
Max subset	Min: 1 Median: 2 Max: 6	Min: 0 Median: 0.42 Max: 0.97	2.6%	1.06	56.2%	59%

Like the previous results for re-identification risk, the subset of quasi-identifying attributes comprising only DT_NASC, ID_UNIDADE, and DT_EVOLUCA already achieves attribute-inference results comparable to the original ten quasi-identifying attributes proposed, evidenced by Table 7. This is mainly because this combination also results in a high amount of sample uniques, which implications were earlier discussed. On the other hand, the maximal set with the lowest re-identification probability portrait a better scenario for the attribute-inference measures. We can observe in Table 7 that the median and maximum values of l-diversity for FATOR_RISC and RAIQX_RES are

higher than before, implying more diverse equivalence classes regarding these attributes. Also, the CAS risk results also presents improvements: the deterministic measure is much lower for all sensitive attributes and the probabilistic (posteriori) values reduced by around 30% and 40% for FATOR_RISC and RAIOX_RES, respectively.

The TP_AMOSTRA attribute, however, shows almost no difference between the minimal and maximal sets results. This outcome relates to the fact that this attribute already poses a high inference risk due to its imbalance. In Table 7, for example, we observe that the most equivalence classes contain only one distinct value for TP_AMOSTRA (median of 1-diversity) while the Posterior CAS risk is only 0.4% away from the Prior risk of 89.2%.

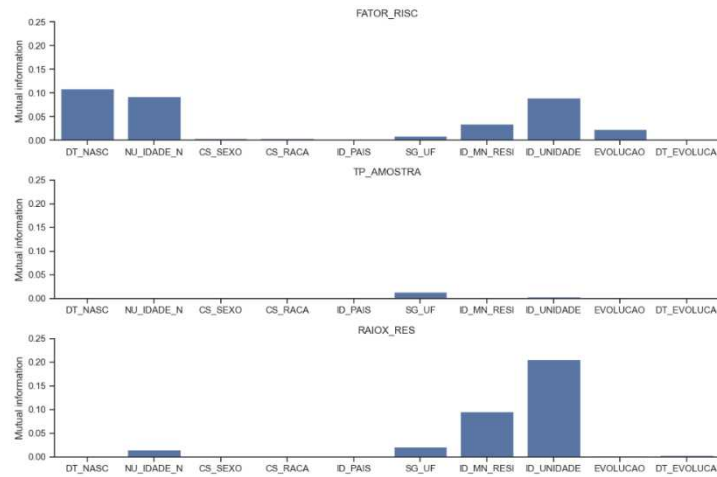


Figure 14. Mutual information between sensitive and quasi-identifier attributes. Mutual information between each sensitive attribute, one in each row, with the quasi-identifiers. The limit of y axis goes up to 0.25 (out of 1) to be able to better visualize the results.

Additionally, we can observe from Figure 14 that some of the quasi-identifying attributes are more relevant to inform about the sensitive variable than others. For example, age is more relevant to inform about an individual's health risk factor than the city they were born in or their race. This is because age is more closely related to health risk factors than the other two attributes. For example, older individuals are more likely to have certain health risk factors, like heart disease. Overall, the variables of age (NU_IDADE_N and DT_NASC) and health facility (ID_UNIDADE) are the ones that share the most information with FATOR_RISC. Regarding RAIOX_RES, ID_UNIDADE and ID_MN_RESI, these are the ones with the highest mutual information values, as health facilities may or may not offer the ideal equipment to conduct an x-ray exam. The sensitive attribute TP_AMOSTRA does not present a relevant mutual information with the quasi-identifiers.

To summarize, we conclude that the case study data set has also a high risk of sensitive attribute inference. The same factors that contribute to the high re-identification risk are also relevant to this attack. Additionally, the imbalance of sensitive attributes and the interdependence of some attributes (which implies a significant amount of shared information) further increase the inference risk. The results also suggest that addressing the re-identification risk does not necessarily address the attribute inference risk. In simpler terms, reducing the likelihood of re-identifying an individual doesn't guarantee a corresponding reduction in the likelihood of deducing their sensitive attributes. This is because the attributes and attribute combinations relevant to re-identification concerns are not necessarily the same as the ones that are relevant to the sensitive attribute inference problem.

4.3 Evaluation of Anonymization Techniques

Since the measured privacy risks are notably high, even surpassing most conservatives' thresholds, the implementation of privacy-preserving data publishing (PPDP) methods becomes essential to safeguard individual's privacy. To achieve this, a wide range of techniques and methodologies are available, each offering distinct solution paths. In this case study, we propose one viable solution by combining traditional and simple PPDP techniques to ensure adherence to the privacy thresholds without significantly compromising data utility.

In our solution, we considered two privacy models: k-anonymity and t-closeness [9]. This combination has been explored in the literature for data publishing [4]. By combining them, we expect to address re-identification and attribute-inference attacks, and thereby reducing the risks associated with privacy breaches. The k-anonymity model guarantees that each record remains indistinguishable within a group of at least k individuals, preventing the identification of specific ones. Simultaneously, the t-closeness model focuses on preserving the distribution of sensitive attributes within each group, minimizing the possibility of inferring sensitive information based on group inclusion. We aim at k value of 20, as suggested in [1] and, the t value of 0.2, as suggested in [7], to ensure a higher level of data protection.

To evaluate the effectiveness of the proposed solution, the privacy risk metrics considered previously will be reevaluated, and then the results will be discussed, considering potential improvements and how data utility might be impacted. From this discussion, the intend is to understand the balance between safeguarding individual

privacy and preserving the dataset's utility for research and decision-making purposes. By quantifying this trade-off, we hope to make informed decisions to optimize both privacy protection and data utility, ensuring the successful implementation of the PPDP methods.

Our proposed course of action is described as follows. We particularly transformed the attribute set below due to their contribution to the elevated risk:

- **Suppression of the patient's birth date (DT_NASC).** This variable can be a source of significant information when combined with others, especially with the health facility variable (ID_UNIDADE). As the age information already exists in the data set (NU_IDADE_N), we decided to remove this attribute.
- **Global re-encoding of the patient's age (NU_IDADE_N).** We re-encoded the variable NU_IDADE_N into 10-year age categories for reducing its outcomes and avoid disclosing precise details.
- **Global re-encoding of the patient's municipality of residence (ID_MN_RESI).** This variable is highly sparse, resulting in a low level of granularity concerning the patient's residence information. We then assign broader categories. The new categories represent the mesoregions (greater area granularity) of each municipality as determined by the Brazilian Institute of Geography and Statistics¹⁰. The values that couldn't be matched were defined as "NAO CADASTRADO" (not registered).
- **Global re-encoding of the health facility name (ID_UNIDADE).** Another variable with a significant number of outcomes is ID_UNIDADE, representing the name of the healthcare facility. When taken together with birth date, for example, were able to identify more than 90% of the individuals. We re-encoded it to the type of healthcare facility (hospital, ambulatory, clinic, etc.), as stated by the Brazilian National Registry of Health Establishments¹¹. The values that couldn't be matched were defined as "NAO CADASTRADO" (not registered).

¹⁰ Brazilian Territorial Division data. Instituto Brasileiro de Geografia e Estatística (IBGE).

<<https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/23701-divisao-territorial-brasileira.html>>

¹¹ DEMAS - Open Data API. Cadastro Nacional de Estabelecimentos de Saúde (CNES), Types of units.

<https://apidadosabertos.saude.gov.br/v1/#/CNES/get_cnes_tipounidades>

- **Global re-encoding of the date of hospital discharge or death (DT_EVOLUCA).** Similar to the birth date attribute, this is another date variable directly associated with an individual, yielding a considerable number of outcomes. Consequently, when combined with other data, it has the potential to reveal more information. Initially, we decided against suppressing this variable as it enables us to track the timing of patient deaths or recoveries related to the respiratory syndrome. This information holds significant relevance for researchers and decision-makers, proving to be especially crucial during the management of the COVID-19 pandemic [11]. As a result, we enhanced the granularity level of the variable DT_EVOLUCA by transforming the full date into a new format: "ww/YYYY," where "w" represents the week number of the year (with Sunday as the first day of the week) and "Y" denotes the year. With this modification, we hope to preserve the variable utility while simultaneously reducing the likelihood of expected re-identifications.

Figure 15 depicts examples of the transformations applied to the data set according to the proposed solution. Table 8 provides a summary of statistics concerning the variables' outcomes before and after the transformations. As a result, we observe a significant reduction in the total number of outcomes for all transformed variables, especially for ID_MN_RESI (-190%) and ID_UNIDADE (-198%). It is also relevant to point out that the number of classes with only one individual presented a decrease as well. For NU_IDADE_N and ID_MN_RESI this number is now down to 0, while ID_UNIDADE and DT_EVOLUCA showed a percentual reduction of 197% and 183%, respectively.

However, despite the improvement in the number of outcomes, ID_UNIDADE and DT_EVOLUCA still presents values that are unique to 4 and 3 individuals, respectively. It might be tough to create a general rule that encompass these unique cases. On the other hand, the suppression of these individuals from the data set seems to be harmless for statistical and analytics purposes.

NU_IDADE_N	ID_MN_RESI	ID_UNIDADE	DT_EVOLUCA
(30, 40] 31	METROPOLITAN REGION OF SAO PAULO SAO PAULO	GENERAL HOSPITAL SANTA CASA DE URUGUAIANA	02/2021 10/01/2021
(60, 70] 70	METROPOLITAN REGION OF SAO PAULO DIADEMA	GENERAL HOSPITAL SANTA CRUZ HOSPITAL	21/2021 25/05/2021
(10, 20] 18	NOT REGISTERED PARK WAY	EMERGENCY CARE UPA JAGUARE	08/2021 26/02/2021
(80, 90] 83	METROPOLITAN REGION OF BELÉM BARCARENA	GENERAL EMERGENCY ROOM MUNICIPAL EMERGENCY ROOM	40/2021 09/10/2021

Figure 15. Transformations results. Transformations applied to the variables NU_IDADE_N, ID_MN_RESI, ID_UNIDADE and DT_EVOLUCA. The new value is indicated in black at the top of each cell, while the previous value is shown in grey at the bottom. It is important to note that the observations are synthetic: each cell value was randomly and independently selected. This example is solely for the purpose of illustration, such combinations are not observed in the real-world.

Table 8. Statistics regarding the transformed variables. Statistics regarding the transformed variables NU_IDADE_N, ID_MN_RESI, ID_UNIDADE and DT_EVOLUCA. It lists the number of unique values, minimum class frequency, median class frequency, mean class frequency, maximum class frequency, and number of classes with frequency equals 1, in this order top to bottom. The values in the parenthesis in red represent the result in the original data set, before applying the transformation.

	NU_IDADE_N	ID_MN_RESI	ID_UNIDADE	DT_EVOLUCA
No. unique values	12 (123)	138 (5306)	30 (6055)	113 (667)
Min class freq.	1565 (1)	308 (1)	1 (1)	1 (1)
Median class freq.	112960 (11308)	6449 (56)	464 (64)	1310 (1226)
Mean class freq.	144353 (14083)	12552 (326)	57741 (286)	13018 (2172)
Max class freq.	321794 (51995)	241833 (130998)	1405868 (10236)	62560 (9858)
No. classes w/ freq. 1	0 (5)	0 (12)	4 (601)	3 (68)

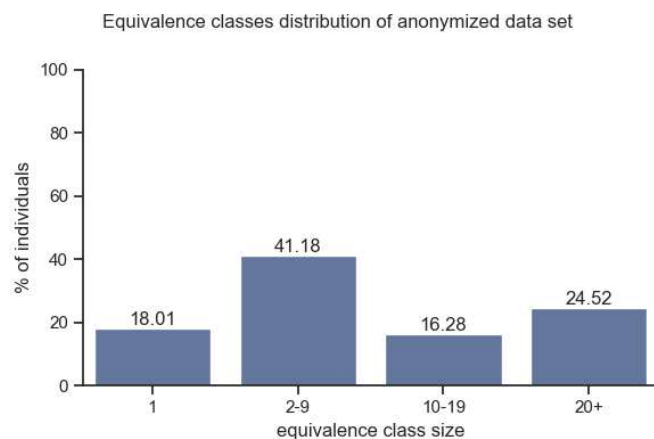


Figure 16. Distribution of possible values for the equivalence class in the anonymized data set. In the x axis, it is the equivalence class size while, in the y axis, it can be found the % of individuals number of individuals belonging to that equivalence class. The last bar represents all values from 20 onwards.

Regarding the re-identification results, there is a notable 82% of reduction from the previous 99.6% of uniquely identified individuals in the original dataset, as shown in Figure 16. In contrast to the previous findings, we observe a higher number of individuals belonging to equivalence classes with higher sizes.

A representative portion of individuals (41%) fall into an equivalence class of 10 or more, with 24.5% belonging to an equivalence class of 20 or more. The maximum number observed for an equivalence class was 311.

On the other hand, still most individuals, 59%, are within groups of less than 10. Among those, 18% are still sample uniques, which corresponds to 220,909 patients. This group presents a re-identification risk of 100%, as they can be uniquely identified within the database if the adversary is aware of their existence. Given that the dataset is intended for public release, such an outcome is not preferable. Also, 75% still do not attend the k-anonymity value of 20.

Table 9. Re-identification results for the anonymized data set. Results of the re-identification Prosecutor (average and maximum), Journalist (estimated) and CRS (deterministic – dCR, probabilistic – pCR, prior and posterior probabilistic success) attacks on SIVEP 2021, for the anonymized data set.

		CRS risk			
Prosecutor risk	Estimated Journalist risk	dCR	pCR	Prior	Posterior
Min: 0.3% Avg: 31.7% Max: 100%	100%	18%	388912	0.0001%	31.7%

The number of expected re-identifications (Average Prosecutor risk and Posterior CRS risk) is also significantly lower than the previous result of 99.8% (68% of decrease), as presented in Table 9. The distribution of individual probabilistic success is much more concentrated in lower values of re-identification (around 20%) than in the original distribution, as shown in Figure 17. However, the maximum Prosecutor risk and the estimated Journalist risk remains 100%, and again most of the individuals (75%) do not attend the re-identification threshold of 5% (k value of 20).

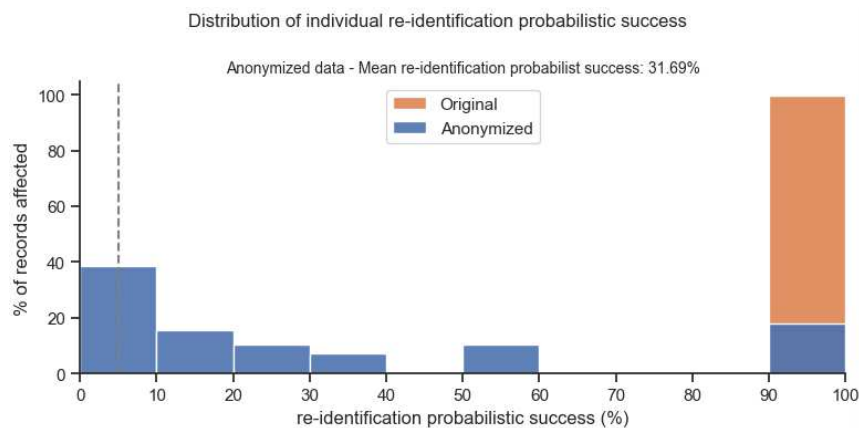


Figure 17. Histograms for the re-identification individual score IRS for the original and anonymized data set. Histogram for the distribution of individual scores according to the re-identification probabilistic success for the anonymized data set. The dashed line corresponds to the re-identification threshold of 5%. The blue bars correspond to the anonymized results while the orange bars, to the original data set results.

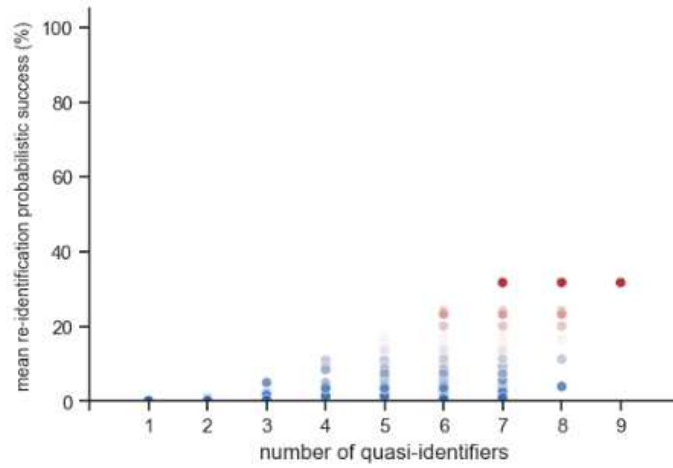


Figure 18. Mean re-identification success by quasi-identifier combination in the anonymized data set. Each dot represents a distinct scenario defined by the selected quasi-identifying attributes among the 512 (2^9) possibilities. The horizontal axis defines the size of each combination of quasi-identifiers, while the vertical axis defines the adversary's re-identification probabilistic success.

It is interesting to note that the re-identification risk associated with each possible combination of quasi-identifiers now reveal a different scenario, as shown in *Figure 18*. The risk does not increase significantly as the size of the combination increases. For instance, within the anonymized dataset, all combinations involving only two quasi-identifiers do not exceed a 0.6% risk threshold, while in the original dataset, the combination of ID_UNIDADE and DT_NASC presents a risk exceeding 90%. On the other hand, a combination of 7 attributes already achieves the same risk value as the one with all 9 attributes. Unlike previous results, it doesn't seem that one attribute, or a pair of attributes, contributes heavily for the overall risk, even within groups of 5, the risk is under 20%.

In terms of attribute-inference attack, the anonymized results remain similar to the ones observed in the original data set as presented in *Table 10*. The minimum and median values for l-diversity remains unchanged, both equal to 1. This represents that is still frequent to observe only one distinct value for the sensitive attributes among the equivalence classes. Likewise, even though there are more equivalence classes in the anonymized data with smaller values (close to 0) of t , the overall results for t -closeness exhibit no significant change, especially for TP_AMOSTRA, as shown in *Table 10* and *Figure 19*. These findings suggest that the proposed transformations might not have been as effective in preventing attribute-inference attacks as they were in mitigating re-identification attacks.

Table 10. Attribute-inference results for the anonymized data set. Results of the attribute-inference distinct l-diversity (minimum l, median l, maximum l), t-closeness (minimum t, median t and maximum t) and CAS (deterministic – dCA, probabilistic – pCA, prior and posterior probabilistic success) attacks for the anonymized data set.

Sensitive attribute	L-diversity risk	T-closeness risk	CAS risk			
			dCA	pCA	Prior	Posterior
FATOR_RISC	Min: 1 Median: 1 Max: 2	Min: 0.0004 Median: 0.41 Max: 0.58	32.6%	1.40	55.8%	78.2%
TP_AMOSTRA	Min: 1 Median: 1 Max: 4	Min: 0.003 Median: 0.1 Max: 0.99	58.1%	1.04	89.2%	92.8%
RAIOX_RES	Min: 1 Median: 1 Max: 6	Min: 0.03 Median: 0.43 Max: 0.97	36.5%	1.34	56.2%	75.4%

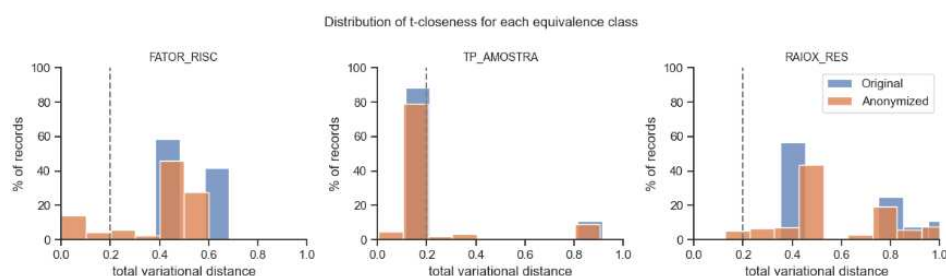


Figure 19. Distributions of t-closeness values for the original and anonymized data set. Each plot holds the distribution of the distances between the distribution of a sensitive attribute in an equivalence class and the same distribution in the whole data set, taken from the t-closeness definition. The dashed line highlights a threshold of 0.2. The blue bars correspond to the anonymized results while the orange bars, to the original data set results.

Regarding utility, the proposed solution would restrict the granularity for certain analysis. For example, it would no longer be possible to analyze the cases and deaths of acute respiratory syndrome at the municipal level, only at mesoregions level. Similarly, the ability to track daily cases and deaths related to these respiratory syndromes would not be possible, with the frequency of tracking reduced to a weekly basis. Moreover, the anonymized database would also not be able to assess the overall performance of a specific healthcare facility, such as the count of COVID cases treated or the number of hospitalizations, only by facility type. However, a broad spectrum of data analysis remains feasible. For example, many of the analyses commonly featured in Brazil's primary COVID-19 monitoring dashboards¹² could be conducted using the anonymized dataset, like the gross death rate from COVID-19 per 100,000 population by state. In these cases, the result would be the same as the original data set since the methods applied are non-perturbative.

¹² Coronavirus Brasil panel from Brazilian Ministry of Health. <<https://covid.saude.gov.br>>

However, as the current outcomes have not yet reached the intended thresholds of $k=20$ (k-anonymity) and $t=0.2$ (t-closeness), other techniques should be evaluated to lower the overall risks. In this way, one popular solution is suppression [8]. By this technique, entire rows, columns or cells are suppressed from the data set. However, in the case study scenario, this solution would significantly compromise data utility. There are many rows that still do not meet the $k=20$ threshold, and it would need to suppress a significant number of columns to effectively obtain a minimum risk of 5%, as shown in *Figure 16*. Similarly, it is also possible to generalize the data further. For example, we could transform the DT_EVOLUCA attribute to only contain the month and year, instead of the week and year. However, it would also be difficult to avoid sample uniqueness without compromising the utility of the data.

Another possible approach to protecting privacy is through perturbation [8]. Perturbation modifies or disguises records in a way that they no longer correspond to the original values. This can be done by adding noise, rounding, or swapping values. Perturbation is a more effective solution against attribute-inference attacks than generalization, as it makes it more difficult for an attacker to infer the original values of the records.

However, there are some challenges associated with perturbation [8]. One challenge is deciding which columns to perturb. If too many columns are perturbed, the data may become too noisy and lose its utility. On the other hand, if too few columns are perturbed, the data may still be vulnerable to privacy attacks. Another challenge is determining the amount of perturbation needed to achieve a satisfactory trade-off between privacy and utility. If the perturbation is too small, the data may still be vulnerable to privacy attacks. If it is too large, the data may lose its utility.

Finally, it is important to note that perturbation can sometimes result in improbable combinations of values [8]. For example, if a column is perturbed by adding noise, it is possible that a record will have a value that is outside the range of possible values for that column. This can make it difficult to use the data for some purposes.

It is important to note that the solutions discussed here only consider a small set of 10 quasi-identifiers out of the over 100 attributes in the original data. Therefore, the challenges may be even greater when considering more quasi-identifiers and sensitive variables. Additionally, if the results are still not within acceptable parameters, then the desired thresholds may need to be adjusted, as they seem to be very difficult to achieve in this data set [1].

5 DISCUSSION

The case study showed that removing direct identifiers is not enough to protect data holders' privacy, as also observed in previous works [16]. Even if only the birth date (DT_NASC) and health facility (ID_UNIDADE) attributes were released, 1.6 million patients (around 95%) would be susceptible to re-identification attacks. Similarly, under the same quasi-identifying combination, each equivalence class generally contains only one distinct sensitive value (FATOR_RISC, TP_AMOSTRA, and RAIOX_RES). This leads to a disparity in the frequency distribution of these sensitive values when compared to the entire dataset, except for TP_AMOSTRA due to its imbalance. Either way, this setting facilitates an adversary's ability to find out an individual's sensitive information: if an adversary infers the most frequent class within the groups, he would be correct most of the time.

Another key finding emphasizes the importance of better understanding the specific privacy threat under consideration to achieve a precise quantification of the associated risk. In our case study, we formulate assumptions about potential risks and adversaries, allowing us to assess privacy vulnerabilities under specific conditions. By analyzing the privacy risks using different assumptions and privacy measures, we found that some measures are closely related, while others can provide different aspects and perspectives of the risk.

For example, the l-diversity metric can be used to assess whether a sensitive attribute is sufficiently varied to prevent the dominance of a single value within a subgroup, that lead to less uncertainty. This can be important if an adversary is trying to infer sensitive information about an individual, such as their health status. However, simply assessing the distinct values within groups is not sufficient to understand the full extent of the privacy vulnerability incurred by the skewed distribution of sensitive attributes. At the same time, some metrics might yield similar results: both the Prosecutor risk (averaged and individual values) and the probabilistic re-identification definitions in [2] (collective and individual) produce comparable outcomes under similar assumptions. Therefore, it is important to understand exactly what the privacy threat is, given the data, the context, and the adversaries. When we fully understand these points, we can better select a privacy measure to effectively address the main aspects of the risk and achieve a more accurate risk quantification.

Additionally, in this context, we also highlight the importance of handling missing values. In the case study, the strategy involved excluding combinations with empty values. For example, when examining DT_NASC and ID_IDADE together, only combinations with values for both attributes were considered. But other approaches can be found in the literature [8]. Indeed, we could consider the missing value as a potential source of uncertainty for an adversary: since the intruder may not know the true attribute value, he can be less certain of his findings when there are missing values present. For example, suppose that there is only one record in the data set who was born on 27/04/1998 and assisted at the municipal hospital in Rio de Janeiro. At the same time, there are other 10 records with the exact same date but with missing values for the health facility. In this case, one may argue that there is a probability that the real target is within the records with missing values, which implies that the equivalence class size for this combination would be 11 and not only 1. It's then easy to see that, depending on the strategy taken, the measured risk obtained could differ.

Overall, the results suggest that it is important to revisit the current methods of preserving the privacy of publicly published data in Brazil. This is reinforced by other works, such as [2], [18], and [19], that have also identified open Brazilian databases exhibiting notably elevated risks. Additionally, there may be other undiscovered data sets with similar characteristics as the one in the case study that also present such risks, or even worse, are able to offer even more information when crossed. Consequently, there is a need for more advanced privacy-preserving methods and for a better understanding of how to conduct this process so as to effectively guarantee both privacy and utility.

In this context, El Emam's methodology [1] emerges as a valuable resource for informed decision-making, especially in the absence of comprehensive guidelines, as seen in the Brazilian case study. The approach integrates threat modeling, risk assessment, anonymization techniques and data utility evaluation to offer a comprehensive and adaptable solution for the challenges of preparing datasets for public release. The method application in the case study not only improved the decision-making process but also helped pinpoint the crucial factors to consider when anonymizing the dataset. This systematic approach can be very advantageous in the context of open governmental data as it may help promote transparency, build trust among stakeholders, facilitate external audits, and support accountability.

On the other hand, we observed that the methodology in [1] presents some limitations. Firstly, it does not explicitly consider the release model type, which is a crucial aspect to determine the appropriate level of privacy for a data set. For example, the public release provides the greatest availability while it demands a greater protection [6]. In this way, the release type can influence the appropriate definition of the privacy and utility thresholds as well as the selection of anonymization techniques. Additionally, the methodology primarily focuses on re-identification attacks, overlooking other potential threats. For example, an individual's sensitive attributes may be exposed through an attribute-inference attacks, even if their identity remains undisclosed. Therefore, a more comprehensive consideration of various privacy threats would further enhance the methodology effectiveness in diverse data-sharing scenarios.

Finally, we conclude that preserving privacy in microdata for public release is a very challenging task. These datasets typically hold numerous attributes, often with high dimensionality, leading to complex relationships among variables. Additionally, these attributes often have uneven distributions and may have many missing values. The challenge becomes even harder when many sensitive attributes are involved, as disclosing such data can not only violate personal privacy but also have far-reaching social consequences. Due to the complexity of these challenges, finding a one-size-fits-all, automated solution is exceptionally hard.

Then, determining the optimal version of the data set becomes an interactive process, which involves closely collaboration with subject matter experts and a careful analysis of the data relationships. In the context of public data publication, there is also the challenge of explainability: the government must be able to explain to the public the methods used to anonymize the data, and it should be understandable to everyone [2]. In this case, it is preferable to use simpler and more intuitive anonymization approaches, if possible.

6 CONCLUSION

6.1 Contributions

In this work, we have investigated the application of the methodology proposed in [1] within the specific context of a public Brazilian health microdata. The risks of re-identification and attribute-inference attacks were assessed under different assumptions and metrics. Furthermore, a potential approach for data anonymization was proposed to enhance privacy without compromising utility.

Our results revealed that El Emam's methodology [1] can offer several advantages for executing an effective anonymization process. Through its iterative process, the methodology offers a structured pathway to transform sensitive information into an anonymized format by conducting comprehensive risk assessments and implementing privacy techniques, ideally under expert guidance. This facilitates the responsible release of public data, striking a careful balance between protecting individuals' privacy and ensuring the utility of the data for analysis and research. Also, its general applicability makes it easily adaptable for use with case study data.

On the other hand, the methodology is limited to re-identification risk and don't consider other potential privacy attacks. Such a narrow focus might lead to incomplete privacy safeguards, compromised privacy objectives and an inaccurate assessment of risks associated with the data. In practical scenarios, adversaries often combine both types of attacks to compromise privacy, rendering the methodology less effective. Additionally, the methodology does not explicitly address the concept of the data release type. Understanding how the data will be released is important to properly assess the potential privacy risks associated with different dissemination methods. For example, public releases will typically require a higher level of protection. Overall, these findings help to clarify research question **RQ1**.

Regarding the research question **RQ2**, the case study data is highly susceptible to re-identification and attribute-inference attacks. In fact, an adversary with access to limited external information could, most of the time, find their target, once they know that the target is present in the database. Notably, attributes such as birth date and health facility appear to be particularly influential in this context. Moreover, the removal of direct identifiers as the only privacy-preserving method is insufficient, being consistent with existing literature.

Finally, the experimental findings of the anonymization techniques contribute to addressing the research question **RQ3**. In summary, it could be observed that the optimal solution to privacy protection comprises a set of techniques combined with a set of privacy models. This is because each model has a different focus and is better suited to mitigating a specific risk, which demands a specific set of techniques. For instance, in the examined case study, it became clear that using generalization and suppression resulted in significant improvements for mitigating the risk of re-identification, but the same degree of improvement was not observed in the context of attribute-inference attacks.

Additionally, achieving an acceptable solution that balances privacy and utility is a challenging task. Addressing both privacy risks at the same time can lead to a higher level of protection, but this may come at the cost of utility. The characteristics of the data can also add a layer of complexity. The more sensitive the data, the larger the volume of data, and the stronger the relationships between different pieces of data, the more difficult it is to protect privacy using privacy-preserving techniques.

6.2 Challenges and Limitations

We bring attention to certain drawbacks of our study.

- Our experimental results are limited to the specific data set of the case study. This data set is in a specific format, containing many categorical attributes, many of which are health-related sensitive information. A more comprehensive investigation of different data sets and data structures could provide valuable insights into the generalizability of our findings and the suitability of the overall solution for different contexts.
- In the case study, only 10 attributes were empirically selected as quasi-identifiers. However, given that the dataset encompasses more than 100 variables, a comprehensive assessment of all attributes would yield results better aligned with the complexities associated with the release of this data. Also, exploring established methods from the literature could offer valuable insights into a more systematic approach to selecting quasi-identifiers, minimizing empirical biases in the process.
- The assessment was limited to only two specific attacks. However, there exists potential value in expanding the scope of evaluation to encompass additional

attacks, like linkage attacks, which involve linking different datasets to extract more information than each dataset individually reveals.

6.3 Future Works

We also highlight some possible future work:

- Exploring an alternative set of quasi-identifiers, considering all available variables, and conducting a comprehensive analysis to better understand their influence on privacy risks could provide valuable understanding for the subsequent application of privacy-preserving techniques. It would be also interesting to evaluate whether automatic methods for selecting quasi-identifiers [10] could reduce the risk of human bias.
- The case study results explored a limited set of privacy-preserving techniques. It would be valuable to investigate a wider range of techniques and combinations to understand what would achieve the best privacy-utility trade-off. For this, a formal quantification of data utility would also be relevant.
- Considering that the SARS microdata is released annually, a practice shared with other Brazilian datasets like those published by INEP¹³, it is important to also consider the longitudinal attack [2]. In this attack, an adversary can access multiple versions of the database, each corresponding to a different point in time. This allows the adversary to combine information from different databases to increase their chances of success.
- Finally, it is recommended to apply the same approach outlined in this study to evaluate other Brazilian microdata as well. This will enable the evaluation of the extent to which these findings can be applied across various scenarios and data structures.

¹³ The Anísio Teixeira National Institute of Educational Studies and Research. <<https://www.gov.br/INEP>>

BIBLIOGRAPHY

1. EL EMAM, K. Guide to the de-identification of personal health information. CRC Press, 2013.
2. ALVIM, M. S., FERNANDES, N., MCLVER, A., MORGAN, C., NUNES, G. H. Flexible and scalable privacy assessment for very large datasets, with an application to official governmental microdata. Proceedings on Privacy Enhancing Technologies, n. 4, p. 378-399, 2022.
3. BASTOS, L. S., RANZANI, O. T., Souza, T. M. L., HAMACHER, S., BOZZA, F. A. COVID-19 hospital admissions: Brazil's first and second waves compared. The lancet respiratory medicine 9, n. 8, p. e82-e83, 2021.
4. MAJEED, A., LEE, S. Anonymization techniques for privacy preserving data publishing: A comprehensive survey. IEEE 9, p. 8512-8545, 2020.
5. SWEENEY, L., YOO, J. S., PEROVICH, L., BORONOW, K. E., BROWN, P., BRODY, J. G. Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study. Technology science, 2017.
6. CARVALHO, T., MONIZ, N., FARIA, P., ANTUNES, L. Survey on Privacy-Preserving Techniques for Microdata Publication. ACM Computing Surveys, 2023.
7. LI, N., LI, T., & VENKATASUBRAMANIAN, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In 2017 IEEE 23rd international conference on data engineering, p. 106-115, IEEE, 2006.
8. TEMPL, M., KOWARIK, A., MEINDL, B. Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro, Journal of Statistical Software, 67, n. 4, p. 1-36, 2015.
9. ZIGOMITROS, A., CASINO, F., SOLANAS, A., PATSAKIS, C. A survey on privacy properties for data publishing of relational data. IEEE Access, 8, p. 51071-51099, 2020.
10. MOTWANI, R. and XU, Y. Efficient algorithms for masking and finding quasi-identifiers. In Proceedings of the Conference on Very Large Data Bases (VLDB), p. 83-93, 2007.

11. MORAES, T. G., LEMOS, A. N. L. E., LOPES, A. K., MOURA, C., DE PEREIRA, J. R. L. Open data on the COVID-19 pandemic: anonymisation as a technical solution for transparency, privacy, and data protection. *International Data Privacy Law*, 2021.
12. Personal Data Protection Commission. Guide to basic data anonymisation techniques. Bloomberg, Singapore, 2018.
13. Information and Privacy Commissioner of Ontario. De-identification Guidelines for Structured Data., 2016.
14. GARFINKEL, S. De-identification of Personal Information. US Department of Commerce, National Institute of Standards and Technology, p. 1-46, 2015.
15. ROCHER, L., HENDRICKX, J.M. and DE MONTJOYE, Y.A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10 n. 1, p. 1-9, 2019.
16. SWEENEY, L. (2000). Simple Demographics Often Identify People Uniquely. https://kithub.cmu.edu/articles/Simple_Demographics_Often_Identify_People_Uniquely/6625769/1.
17. QUEIROZ, M. J. D., MOTTA, G. H. M. B. Transparência e preservação de privacidade em dados governamentais no Brasil: pesquisa documental e estudo de caso. Instituto de Pesquisa Econômica Aplicada (IPEA), 2017.
18. QUEIROZ, M.J. MOTTA, G.H. Privacidade e Transparência no Setor Público: Um Estudo de Caso da Publicação de Microdados do INEP. In *Anais do XV Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais*, SBC, p. 362-365, 2015.
19. SANTOS, W., SOUSA, G., PRATA, P., FERRÃO, M. E. Data Anonymization: K-anonymity Sensitivity Analysis. In *2020 15th Iberian Conference on Information Systems and Technologies, CISTI*, p. 1-6, IEEE, 2020.
20. POTIGUARA CARVALHO, A., POTIGUARA CARVALHO, F., DIAS CANEDO, E., POTIGUARA CARVALHO, P.H. Big data, anonymisation and governance to personal data protection. In *The 21st Annual International Conference on Digital Government Research*, p. 185-195, 2020.
21. SILVA, H., BASSO, T., MORAES, R., ELIA, D. and FIORE, S. A re-identification risk-based anonymization framework for data analytics platforms. In *2018 14th European Dependable Computing Conference, EDCC*, p. 101-106, IEEE, 2018.

22. MANUEL FILHO, E. B., NETO, E. R. D., MACHADO, J. C. Privacy-preserving of patients with Differential Privacy: an experimental evaluation in COVID-19 dataset. *Journal of Information and Data Management*, 12, n. 5, 2021.
23. FUNG, B. C., WANG, K., FU, A. W. C., PHILIP, S. Y. *Introduction to privacy-preserving data publishing: Concepts and techniques*. CRC Press, 2010.