Privacy-preserving of patients with Differential Privacy: an experimental evaluation in COVID-19 dataset

Manuel E. B. Filho, Eduardo R. Duarte Neto, Javam C. Machado

Computer Science Department
Universidade Federal do Ceará, Brazil
{edvar.filho,eduardo.rodrigues,javam.machado}@lsbd.ufc.br

Abstract. The pandemic of the new coronavirus (COVID-19) has brought new challenges to health systems in almost every corner of the world, many of them overburdened. The data analysis has given support in the fight against the coronavirus. Through this analysis, government authorities, together with health care providers, adopted effective strategies. Yet, those strategies can not be careless of privacy concerns. The individuals' privacy is a right of each citizen. Privacy techniques guarantee the analysis of health data without exposing individuals' private information. However, a balance between data privacy and utility is essential for a good analysis of the data. This work will demonstrate that it is possible to guarantee the privacy of infected patients and maintain the utility of the data, allowing a sound analysis on them, from the visualization of the application of differentially private mechanisms on queries in the data of patients tested in the State of Ceará - Brazil.

Categories and Subject Descriptors: H.2 [**Database Management**]: Miscellaneous; H.3 [**Authorization, Privacy, and Security**]: Miscellaneous

Keywords: COVID-19, differentially private publication, data analysis

1. INTRODUCTION

The new Coronavirus pandemic (COVID-19) has brought new challenges to health systems around the world. The coronavirus has demonstrated its ability to spread worldwide within weeks or even days, overwhelming current public health and medical care capacities. Therefore, health systems have generated much data concerning patients infected with the coronavirus. Data analysis is an important task to provide information to the government authorities and health care providers to adopt effective strategies to combat the pandemic. Consequently, health organizations have integrated the collected health data about covid patients, which is essential to improve collaborative researches [Haas et al. 2011]. The various units that make up the health system, and are responsible for patient care, collect data from their patients. This information is then loaded into an integrated database and used for analysis. Many Brazilian institutions have performed this fundamental role of integrating health data of patients affected by the coronavirus. The state of Ceará has played that role, assisting, through data analysis, the planning of effective strategies to fight the virus. The analysis includes, among other things, identifying which regions of the city are most affected, the relationship between the most severe cases. Also, understand why people with comorbidities are less resistant to the virus.

Although data analysis is an essential technique, this procedure must consider patients' privacy. Health records often contain sensitive patient information like names, CPFs, addresses, and illness onset. The improper access to this patient information has generated a discussion on how to guarantee the analysis of this data privately [Machado et al. 2019]. The combination of data published by various means allows identifying individuals on the dataset, leaking private information about the individual.

This research was partially supported by Capes - Brazil under grant number 88887.494811/2020-00 and LSBD/UFC.

For instance, consider a patient tested positive for COVID-19. In his record, one finds only the street name, like “17th street”, but not the house number. In addition, the record shows that the patient has AIDS, which is considered sensitive information. If there is no other person with COVID on that street, an adversary with some background information about the area could identify the individual and leak the private information.

Several techniques of privacy have already been proposed [Samarati 2001; Erlingsson et al. 2014; Dwork et al. 2006], among them, Differential Privacy (DP). A differential privacy mechanism is responsible for adding randomness to the query answer, and it is essential to ensure privacy. The mechanism choice and the parameter adjustments are necessary to balance the privacy level and the information utility. Those tasks are heavily dependent on the query. The higher the privacy level required, the more distant the data from real ones, decreasing its utility. Conversely, the lower the level of privacy, the closer the data will be to the actual data, and therefore, the more useful for analysis. However, in this case, the individuals are more exposed.

Efficient and fast data sharing is the basis for public health action in a pandemic scenario. As a result, it is necessary a great effort to increase the availability of COVID-19 data. However, the privacy of the individuals who provided their data to build up the sharing dataset must also be guaranteed. In this work, after applying differential privacy, we investigate the utility achieved in noisy health data sets. We measure the utility by the difference between the correct answer and the noise response to numerical and categorical queries performed on the data set. The goal is to achieve a balance between guaranteed privacy and preserved utility.

This work is an extended version of the work presented at SBBD 2020 [Bento Filho et al. 2020], including the following new experiments and new contributions:

—We use a real dataset to carry out experiments that show the individuals’ privacy preservation.

—We apply the Laplace mechanism for numerical queries using two or more attributes of the dataset.

—We apply the exponential mechanism for categorical queries using the COVID-19 dataset in Ceará.

—We adjust the privacy budget for different queries, where we analyze the balance between privacy and achieved utility, finding the one with the best trade-off.

We organize the rest of the paper as follows. Related work is presented in Section 2. In Section 3, we present the background knowledge necessary to understand the techniques proposed in this work. We describe the experiments which indicate how much privacy we have achieved as we adjust the privacy budget consumed by the differential privacy mechanisms, in Section 4. Section 5 presents our final considerations.

2. RELATED WORK

This section covers recent work on privacy and COVID-19 data publishing. [Fahey and Hino 2020] discuss the trade-off between individual’s privacy and data utility in COVID-19 data release. The paper considers the dilemma of using data first or privacy-first in some applications. In the first case, there is a significant impact on citizens’ privacy; in contrast, researchers and epidemiologists are highly valued. In the second case, the applications aim to protect citizens’ privacy at the cost of limited access to COVID-19 patient information. In [Lenert and McSwain 2020], we have the study showing the main changes proposed to privacy regulations, with the goal of reducing the impact of the flow of information while preserving the privacy of individuals. [Kuhn et al. 2021] analyzes some applications that track individuals and their interactions and claim to comply with GDPR, anonymity and other forms of privacy. None of these approaches perfectly protect the identity of individuals from malicious third parties.

Bringing an approach that uses syntactic models, the work [Lee et al. 2021] applies k-anonymity and l-diversity to publish quasi-identifier attributes. Using a data set with records that show the frequency for a group of individuals, the approach applies a classification process over a set of variables. The strategy excludes all personal identifiers present in the information, applying the k-anonymity and l-diversity techniques to the data to be published.

In [Reiter 2019], we have differential privacy in publishing nationwide data, showing the benefits and limitations of using differential privacy in federal data. One of the limitations we want to highlight is configuring the privacy budget to use it in the mechanisms. Some federal entities use values of ϵ equal to 0.1 or higher, such as 9. One of the most promising alternatives to accomplish this problem is to analyze a finite set of diverse values of budgets and use the one that guarantees the best balance between privacy and utility.

In [Wang et al. 2016], we have a comparison of randomized response [Chaudhuri and Mukerjee 1988] with the Laplace mechanism [Dwork et al. 2006], presenting an extension of randomized response to work with multiple polychotomous attributes. Based on an empirical evaluation, the paper shows that randomized response surpasses the Laplace mechanism. However, the queries applied to evaluate the mechanisms are categorical.

The work [Aktay et al. 2020] presents the process of aggregation and anonymity of Google's COVID-19 Community Mobility Reports, with the objective that no personal data is inferred from the analysis of the publication metrics. In this work a set of anonymous metrics is generated from user data and against a baseline. If these anonymous metrics meet reliability standards, they will be published along with the baseline, which will be private. Otherwise, the metrics will not be published. The work fixes the values of the privacy budget and, based on the composition theorem [McSherry 2009], which is used in the calculations of the metrics, the privacy budget can be high, ensuring better utility, but causing relaxation on the level of privacy. In addition, each user's contribution to the dataset is restricted, as although a user has multiple locations in their history, only 4 locations are randomly chosen for each day.

With a new approach and to guarantee the privacy of individuals, a blockchain-based medical support platform is presented in [Yu et al. 2021]. It provides data sharing while preserving the privacy of information owners. In the platform, hospitals and institutions are nodes in the chain. Doctors feed the chain with COVID-19 records to ensure privacy, allowing authenticated researchers and other doctors to access the records. In addition, research results may also be published in the chain.

Our work aims to present an experimental approach for publishing information on COVID-19 patients in the State of Ceará following numerical and categorical queries on these data. After a pre-processing stage of data cleaning, we apply the most indicated mechanisms in the literature to the type of query performed. We investigate several values for privacy budget, showing the impact of privacy on data utility. We propose the best budget values to guarantee that malicious third parties do not have access to sensitive information about patients.

3. THEORETICAL FOUNDATION

Vast quantities of individual information are currently collected and analyzed by a broad spectrum of health organizations. Although these data are essential for analysis, health organizations must collect the data with solid privacy premises. Publications without special care regarding the individuals' privacy can cause the exposure of sensitive information about the patients, such as the presence or absence of diseases. This information, combined with other published data, can increase the risk of privacy breaches.

Different strategies have been proposed to preserve the privacy of individuals and protect them from possible attacks. Encryption, tokenization, and anonymization are some of these more well-known strategies. The most common is to anonymize the data, where the original data goes through a transformation process. Many techniques have been proposed, with different assumptions and guarantees of privacy. One of the biggest challenges in data privacy is to keep the data utility even after the data pass through a process of anonymization. Thus, a balance between privacy and utility is necessary, or the data may become useless for analysis, losing its purpose.

The anonymization model modifies the original data. The new modified data does not resemble the original data but maintains the original semantics. Generalization, suppression, and perturbation are the most used techniques for anonymizing and publishing data [Fung et al. 2010a]. These techniques differ according to the loss of information and the preservation of privacy proposed by each one.

The syntactic privacy model applies an anonymization process to the data set to ensure that the modified dataset will reach some syntactic condition. The k-anonymity [Sweeney 2002] privacy model is the best known in the field of data anonymization. The anonymization process takes place on semi-identifying attributes, which, although they do not explicitly identify individuals, can be used in combination with other information for this purpose. Acting on the indistinguishability principle, the k-anonymity guarantees that for each combination of semi-identifiers, there are at least k records in the published dataset, setting up an equivalence class. This guarantee ensures that each record can not be re-identified with a probability greater than $\frac{1}{k}$. In this model, the value of k will define the level of privacy and, at the same time, acts directly on the loss of information. Although k-anonymity is a prevalent technique, it is ineffective against various types of attacks. Therefore, based on the same principle of indistinguishability, the l-diversity [Machanavajjhala et al. 2007], t-closeness [Li et al. 2007], and the δ-presence [Nergiz et al. 2007] were also proposed.

3.1 Differential Privacy and Mechanisms

Differential privacy is a popular privacy preservation technique in data analysis that provides solid privacy guarantees [Dwork et al. 2006; McSherry and Talwar 2007; Dwork 2008; Lecuyer et al. 2019]. Unlike the syntactic models, the differential privacy model acts on the queries' response to a dataset. Thus, the anonymization process will no longer modify the original dataset. Instead, a differentially private mechanism returns a perturbed response, decreasing the risk of an adversary to infer anything from the answers. Its definition is independent of the prior knowledge of adversaries. The idea behind differential privacy is that no individual represented by a record in the database will significantly impact the response to the query. So, the presence, or absence of any individual in the dataset, will not change the probability of the query's output.

Definition 3.1 *Differential Privacy.* A mechanism M is ϵ-differentially private (DP) if for any datasets D_1 and D_2 that differ in one element, and for any set S of all possible outputs of M,

$$Pr[M(D_1) \in S] \leq \exp(\epsilon) \times Pr[M(D_2) \in S] \quad (1)$$

A DP Mechanism is responsible for adding random controlled noise to the query response to guarantee ϵ-Differential Privacy. The budget ϵ limits the impact of adding or removing any individual in a dataset. A small budget indicates that any individual will introduce a minimum change on the mechanism distribution, giving higher protection.

The Laplace mechanism [Dwork et al. 2006] has been widely used in differential privacy approaches, particularly in queries about numerical data that return aggregate count values, for example, queries that return the sum, count, or an attribute histogram of a dataset. Given a numeric function $f : D \rightarrow \Re^k$, where k is the dimension of the output, for ϵ-DP to be valid for all possible outputs of the function f, the concept of global sensitivity is needed. The global sensitivity is the maximum variation of f between two neighboring datasets D_1 and D_2, which differ by at most one element.

Definition 3.2 *Global sensitivity.* Let D_1 and D_2 neighboring datasets. The global sensitivity of a f function, denoted by Δf, is given as:

$$\Delta f = \max_{D_1, D_2 \in D} \parallel f(D_1) - f(D_2) \parallel_1 \tag{2}$$

In essence, the Laplace mechanism [Dwork et al. 2006] calculates the value of f with a dataset as input and adds to that value a noise that follows the Laplace distribution. The amount of noise required is calculated according to the sensitivity of the query f applied to the dataset D. The noise scale will be calibrated based on the global sensitivity of f (divided by ϵ, which is the privacy parameter).

Definition 3.3 *The Laplace Distribution.* The Laplace Distribution (centered at 0) with scale b is the distribution with probability density function:

$$Laplace(x|0, b) = \frac{1}{2b}\exp(-\frac{|x|}{b}) \tag{3}$$

Definition 3.4 *Laplace Mechanism.* Given any function $f : D \rightarrow \Re^k$, the Laplace mechanism is defined as:

$$M_f(D) = f(D) + (Y_1, ..., Yk), \tag{4}$$

where Y_i are independent and identically distributed random variables drawn from Laplace$(0, \frac{\Delta f}{\epsilon})$.

Theorem 3.5. *The Laplace mechanism satisfies ϵ-differential privacy.*

Different from numerical attributes, categorical ones do not allow an approximation of their value. Therefore, applying the Laplace mechanism to answer queries over categorical attributes is not a good strategy. For this purpose, the exponential mechanism is more suitable and satisfies ϵ-differential privacy. [McSherry and Talwar 2007] proposed the exponential mechanism to situations where we wish for the "best" answer. Unlike the Laplace mechanism, which adds random noise to the response, the exponential seeks to randomly choose a response from the set of possible answers for the query. Therefore, the exponential mechanism is ideal for answering queries where there is a score for the answer utility while preserving differential privacy.

The probability that a response is defined by an utility function, or score function, which will map all possible datasets D and all possible outputs O to a utility score, $u : (D \times O) \rightarrow \Re$, for a budget ϵ. Δu is the sensitivity function for the query in terms of the score function u:

$$\Delta u = \max_{o \in O} \max_{D_1, D_2 : ||D_1 - D_2||_1 \leq 1} |u(D_1, o) - u(D_2, o)|,$$

for all D_1, D_2 differing by at most one element.

Definition 3.6 *Exponential Mechanism.* For any function $u : (D \times O) \rightarrow \Re$, and a privacy budget ϵ, the exponential mechanism $M_u^\epsilon(D)$ produces o as output with probability proportional to $\exp(\frac{\epsilon u(D,o)}{2\Delta u})$, where Δu is the sensitivity of the utility function. That is:

$$Pr[M_u^\epsilon(D) = o] = \frac{\exp(\frac{\epsilon u(D,o)}{2\Delta u})}{\sum_{o' \in O}\exp(\frac{\epsilon u(D,o')}{2\Delta u})} \tag{5}$$

3.2 Usability

The Laplace mechanism adds random noise to the query response, ensuring that an opponent will not extract knowledge that allows him to identify individuals within the dataset. For example, consider

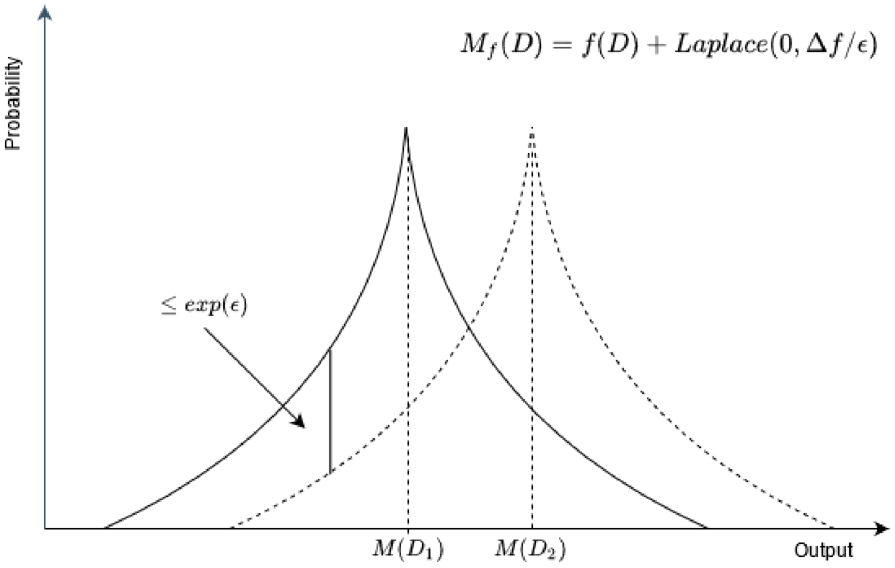


Fig. 1. Laplace Mechanism (adapted from [Brito and Machado 2017]).

the following query to the dataset: "The total number of individuals over 50 years of age". Suppose the correct answer for this query is 20 people. The Laplace mechanism will calculate a noise r to add to the query response. Thus, the response returned would be $20 \pm r$.

Figure 1 illustrates the difference between the Laplace distribution over two neighboring data sets when the Laplace mechanism is applied. This difference is limited by $\exp(\epsilon)$ [Dwork et al. 2014].

The goal of the Exponential mechanism is to select a random response with the probability based on its score. For instance, consider a query about the predominant race of the individuals who died from COVID in 2020. When applying the exponential mechanism, for each possible response, i.e., Asian, White, Mixed Race, Afro-Brazilian[1], the utility function will calculate a score. Assuming that the score function is the number of individuals with that race and the predominant group is White people. The mechanism will give the highest probability for the response referring to the White race.

Figure 2 illustrates the exponential mechanism, which receives the privacy budget, the score function, the sensitivity of the score function, and the dataset as input. Then, performing the query anonymization calculates the score of all outputs used to infer its probability by the mechanism. The analyst will receive a differentially private response.

4. ANONYMIZATION OF HEALTH DATA VIA DIFFERENTIAL PRIVACY

The Integra SUS [SUS 2020] platform contains data collected from patients from different institutions in the health network of the State of Ceará, Brazil. It provides a series of visualizations as a product of the analysis of the data collected, e.g., lethality rate, number of infected, progression in the number of deaths per day, among others. Also, it makes available the dataset used for that analysis. This data has gone through a process of anonymization with the suppression of identifying attributes [Fung et al. 2010b], like CPF – a kind of Brazilian SSN – and personal name. Moreover, information such as the location of the patients' residence, affection of comorbidities, and some dates are still available and are subject to attacks that may lead to the exposure of sensitive information of individuals [Narayanan and

[1] In the dataset, we have the races "Amarelo", "Branco", "Pardo" and "Preto", we translate them for the races Asian, White, Mixed Race, Afro-Brazilian, respectively.

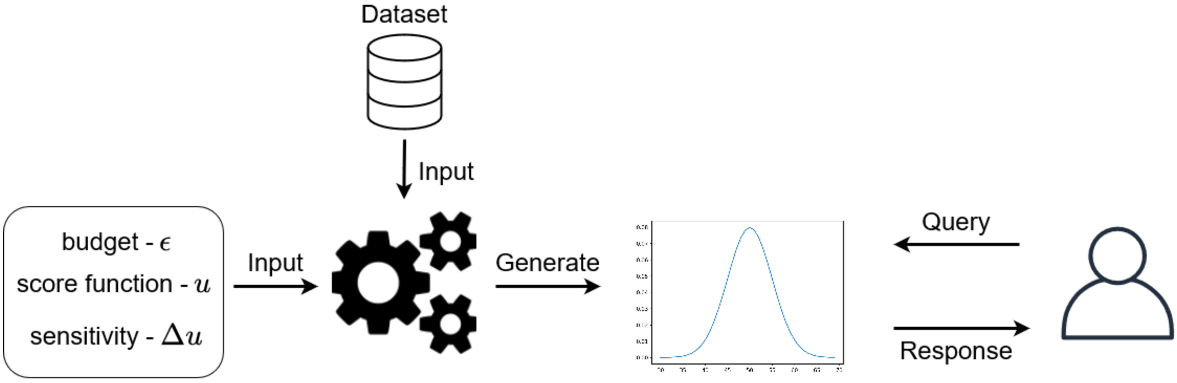


Fig. 2. Exponential Mechanism.

Shmatikov 2006]. Suppressing such attributes would prevent a necessary analysis of the population, such as the distribution of cases by neighborhood, the incidence of comorbidity by gender or ethnicity, the total number of infected people in Fortaleza, among others. Thus, the addition of a differential privacy noise seems to be the most suitable strategy to guarantee a helpful analysis without loss of privacy.

The COVID-19 dataset from the State of Ceará, Brazil, collected on January 25, 2021, contains 1,687,344 records and 62 attributes. It presents information about patients and the results of their exams. First, we performed a data cleaning phase: removing individuals who do not reside in the State of Ceará; renaming the patients' neighborhoods to have a pattern in the names of the neighborhoods and removing those who did not give valid information. In addition, we remove medical records that did not present information about the patient's gender and race. Thus, we obtained 855,803 records to carry out the experiments.

We aim to ensure that a potential attacker could not identify individuals from the database when correlating query responses with external information. We will answer the following questions with our experiments:

—What should be private in the responses to the queries?

—Which mechanism should one apply for numerical and categorical queries?

—What is the impact on the utility of the noisy response to the queries performed?

—What is the best privacy budget to reach a good balance between privacy and utility?

We want everyone, including their information, to be kept private from potential attackers. Since the responses to the queries are the entire dataset in an aggregated way, the individual's record remains private. Adjusting the budget ϵ used by the differential privacy mechanism and performing a series of queries presented in Table I, we evaluate the balance between privacy and utility on the responses. A lower budget indicates a higher level of privacy and consequently a lesser utility of the answer. Conversely, a larger budget indicates a lower guarantee of privacy, and therefore a better answer utility. To apply differential privacy and analyze its impact on the usefulness of the published information, we applied the two most used mechanisms in the literature for numerical and categorical queries, the Laplace and the Exponential Mechanism, respectively, following the type of query. We use budget values, $\epsilon \in \{0.01, 0.05, 0.1, ln(2)\}$ for the Laplace mechanism, while for the exponential mechanism, we applied $\epsilon \in \{0.01, 0.05, 0.1, 0.25, 0.5\}$.

We measure the utility using the mean square error for numerical queries and the accuracy between

Table I. Queries carried out, their IDs and mechanisms applied respectively to each one of them.

ID	Query	Mechanism
Q1L	Total number of patients infected by COVID-19 alive by neighborhood on Fortaleza-CE	Laplace
Q2L	Total number of confirmed cases between males and females by race in the State of Ceará	Laplace
Q1E	Which are the three comorbidities that have the highest number of COVID-19 cases	Exponential

the correct response and the noise response for categorical queries. For mean square error, we have:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_i})^2,$$

where n is the number of neighborhoods in Fortaleza-CE, x_i is the number of occurrences in the neighborhood, and $\overline{x_i}$ is the number of occurrences added by noise for the same neighborhood. To measure accuracy, we have:

$$Acc = \frac{1}{k}\sum_{i=1}^{k} 1(\overline{c_i} = c_i),$$

where k is the number of comorbidities in the query response. We rank all comorbidities based on the number of people who have the comorbidity and tested positive for COVID-19. From that, c_i is the position of the comorbidity ranking of the correct response to the query, and $\overline{c_i}$ is the position of the comorbidity ranking of the differentially private query response.

4.1 Numeric queries with Laplace Mechanism

To demonstrate the mechanism's effectiveness, we add Laplace's noise on the responses of query Q1L: "Total number of patients infected by COVID-19 alive by neighborhood in Fortaleza-CE" applied to the COVID-19 dataset in the State of Ceará. The query Q1L requires the combination of three attributes: the neighborhood of each patient, the exam result, and whether the person died. It needs an anonymization process, or otherwise, it may leak private information. It allows an analysis of Fortaleza's neighborhoods, which helps in the elaboration of disease prevention policies.

The figures below showing the experimental results contain histograms with the total number of COVID-19 positive cases reported by each neighborhood. In blue, we have the correct answer, and in yellow, the noise answer. To simplify the visualization, we show the results in three neighborhood groups representing positive cases. Since it is a counting query, the difference by the presence or absence of a record in the database is at most 1. The global sensitivity of the query is $\Delta f = 1$.

Figure 3 presents the results on the twenty neighborhoods with the fewest occurrences, with values between 1 and 23 cases. With such a low number of cases, these are the neighborhoods most likely to expose their residents. We can see from Figure 3(a) that the noise needed to guarantee privacy is high when the ϵ is very low. In order to preserve patient privacy, the privacy mechanism adds a high noise to confuse the opponent. In these cases, adjusting the ϵ is essential. Figure 3(c) shows a performance when we applied $\epsilon = 0.1$, where we obtained a good balance between privacy and utility.

Figure 4 presents the results on the twenty neighborhoods with a median number of occurrences, with values between approximately 24 and 55 cases. We obtained a decent utility level when ϵ is close to or greater than 0.5. In these cases, the amount of noise necessary to guarantee the privacy of patients is low. In Figure 4(a), where $\epsilon = 0.01$, it requires a large amount of noise for neighborhoods with the lowest number of occurrences in the series, as, as shown in the Figure 3, we want a high level of privacy when we have values low budget for small amounts of occurrences. However, when we compare those results to the neighborhoods in Figure 3(a), the utility gain is already significant.

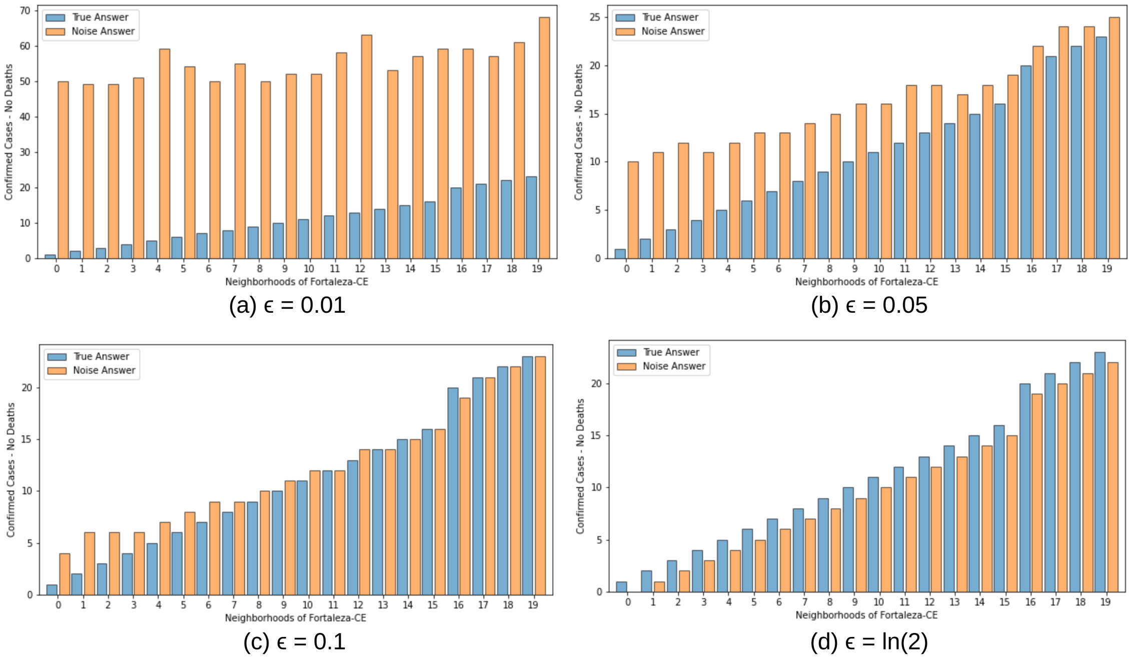
In Figure 5, we have the twenty neighborhoods with the highest number of occurrences, with values between 512 and 1345. In this Figure 5, the noise necessary to guarantee the privacy of patients is low, this is due to the number of occurrences. In this case, the presence or absence of any patient in these groups will not have a great impact on privacy. Even for a small budget of 0.01, the correct response to the query is very close to the noisy response.

Figure 6 shows the mean square error for all neighborhoods in the State of Ceará. As we can see, the error is significantly reduced when $\epsilon = 0.05$. Analyzing Figure 6, it is evident that the budget choice is essential to obtain a good utility level. The mechanism's restrictions decrease with the increase of ϵ. Therefore, less noise is needed to ensure privacy. In all three scenarios, with $\epsilon = 0.05$, the mechanism presented a low mean squared error, which is sufficient to reach a good balance between privacy and utility. Thus, our experiments indicate a budget of 0.05 for the mechanism to answer privately the previously described query on the COVID-19 dataset published by the State of Ceará.

To complement our experiments, we carry out more complex queries, combining several semi-identifying attributes. Thus, with the same Covid-19 database, we performed the analysis of the following numerical query Q2L: "Total number of confirmed cases between males and females by race in the State of Ceará." Once more, query Q2L requires the combination of three attributes: gender, race, and the result exam of each patient. The response analysis is essential for the health authorities to understand if the number of cases may be related to race and gender. However, to ensure no privacy leak, applying an anonymization process over the response is necessary.

Figures 7 and 8 present histograms with the number of positive cases of COVID-19 in each of the races (Asian, White, Mixed Race, Afro-Brazilian) for each of the patients' genders (Male and Female). We have the correct answer for each race in blue, and in yellow, the noisy answer. As in the previous query, the global sensitivity Δf is also 1, due to the impact of the absence or presence of a record, in the database for the query execution.

As in the query Q1L, when we have the neighborhoods with the highest number of occurrences, we can observe in general that the Laplace noise necessary to protect the response is small. However, even for these queries, the noise increases when we decrease the mechanism's budget, making the



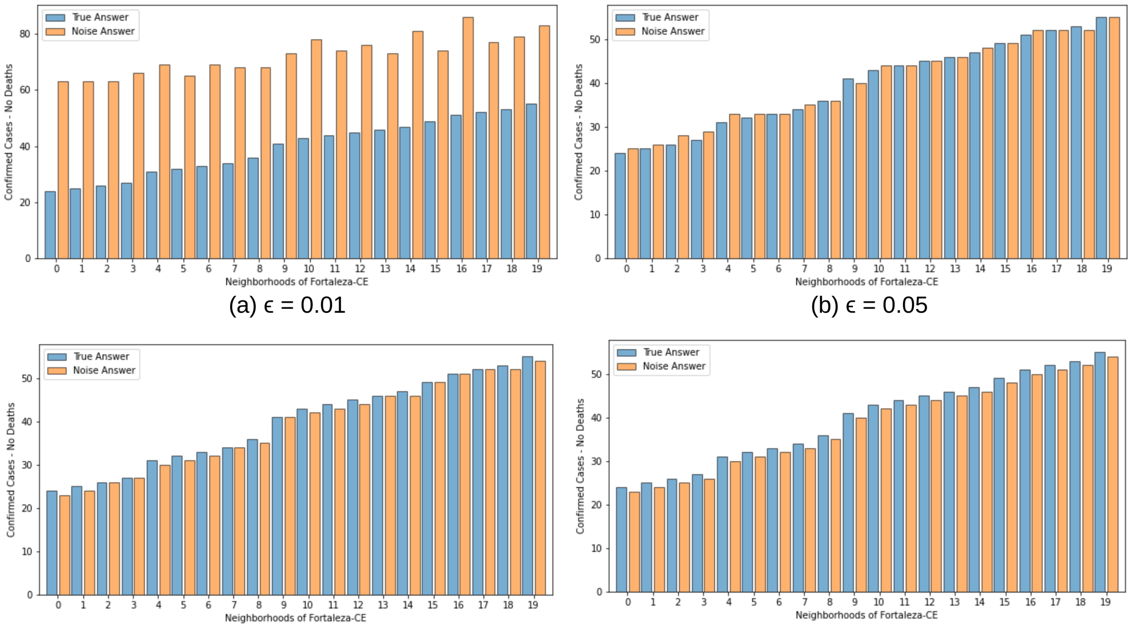
(a) ϵ = 0.01

(b) ϵ = 0.05

(c) ϵ = 0.1

(d) ϵ = ln(2)

Fig. 3. Number of positive cases in the 20 neighborhoods with the fewest occurrences in Fortaleza - CE.



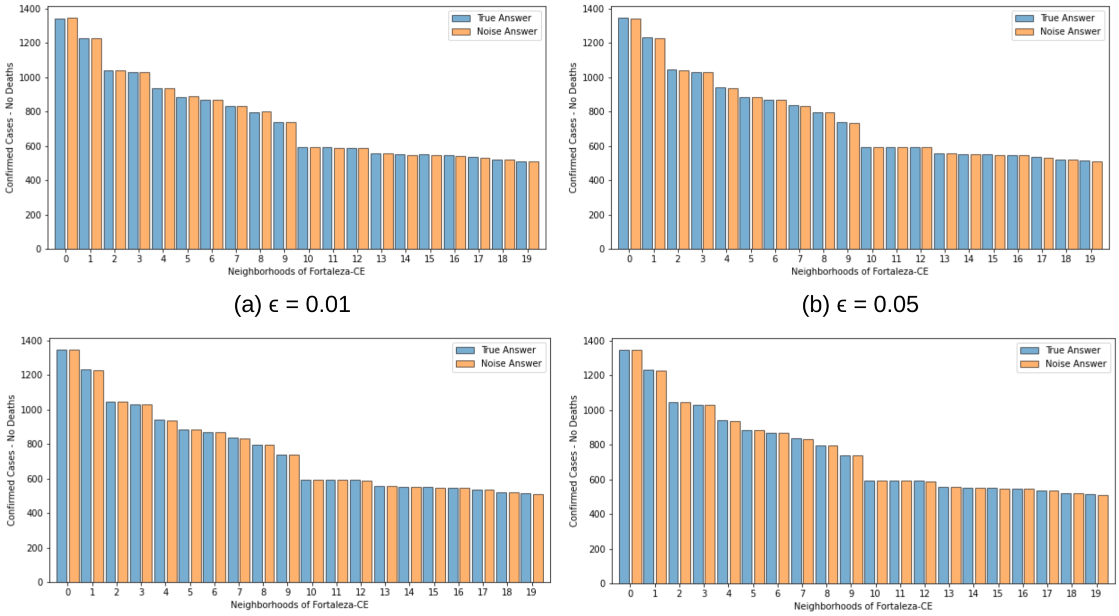
(a) ϵ = 0.01

(b) ϵ = 0.05

(c) ϵ = 0.1

(d) ϵ = ln(2)

Fig. 4. Number of positive cases in the 20 neighborhoods with median occurrences in Fortaleza - CE.



(a) ϵ = 0.01

(b) ϵ = 0.05

(c) ϵ = 0.1

(d) ϵ = ln(2)

Fig. 5. Number of positive cases in the 20 neighborhoods with the highest occurrences in Fortaleza - CE.

error more noticeable. We can see the results in Figures 7 and 8 for each gender, male and female, respectively. To find the best budget selection that would guarantee a balanced trade-off between privacy and utility when the data is published, we vary the value of the budget ϵ to perform the

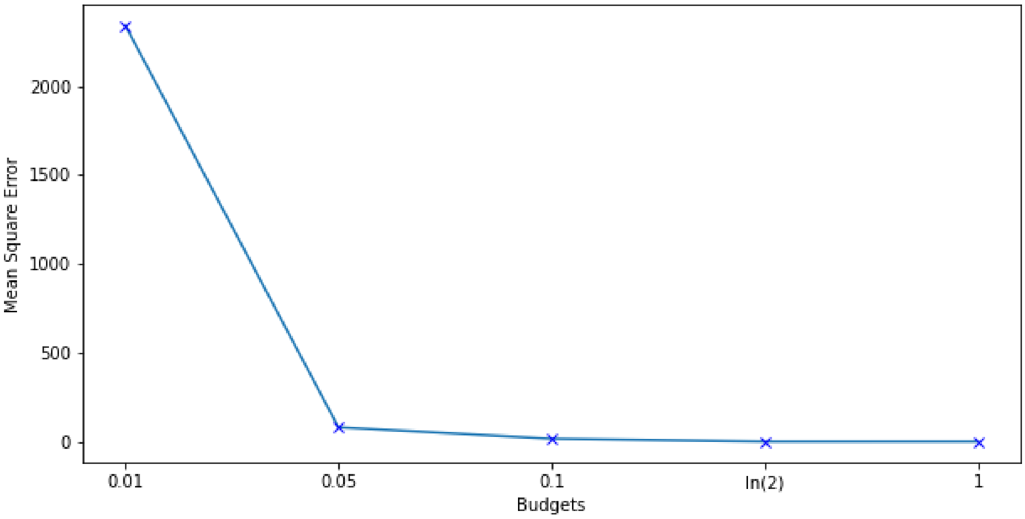
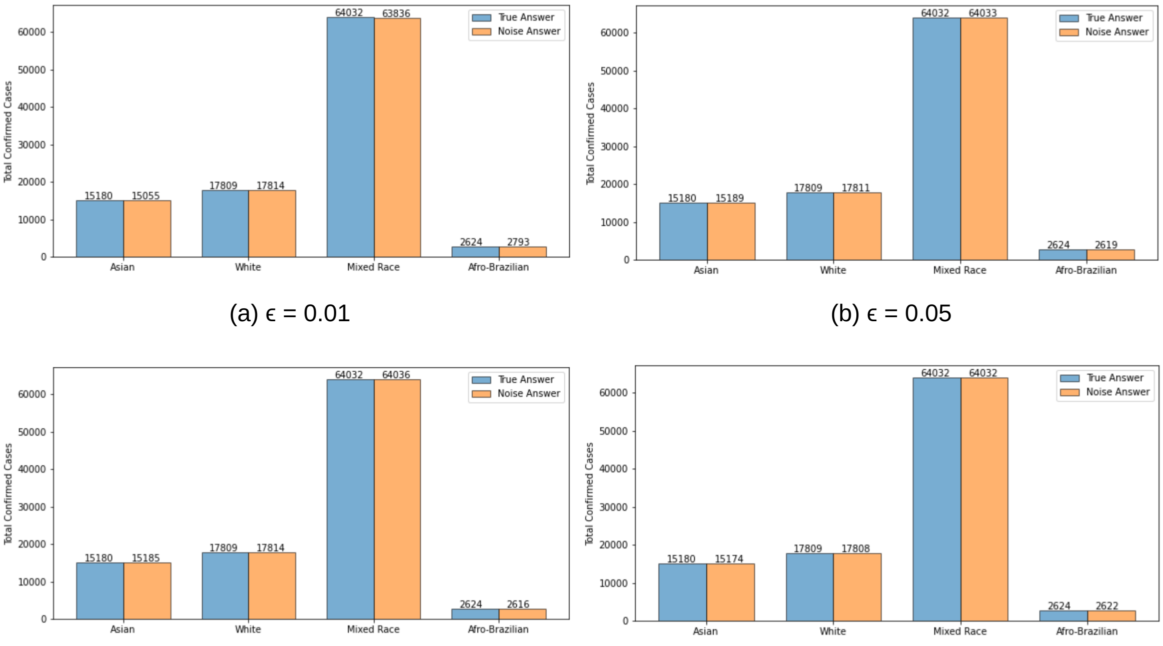


Fig. 6. Mean squared error for query responses Q1L.



(a) ϵ = 0.01

(b) ϵ = 0.05

(c) ϵ = 0.1

(d) ϵ = ln(2)

Fig. 7. Number of positive male gender cases for each race.

Laplace mechanism to answer the query, as we can see in Figures 7 and 8.

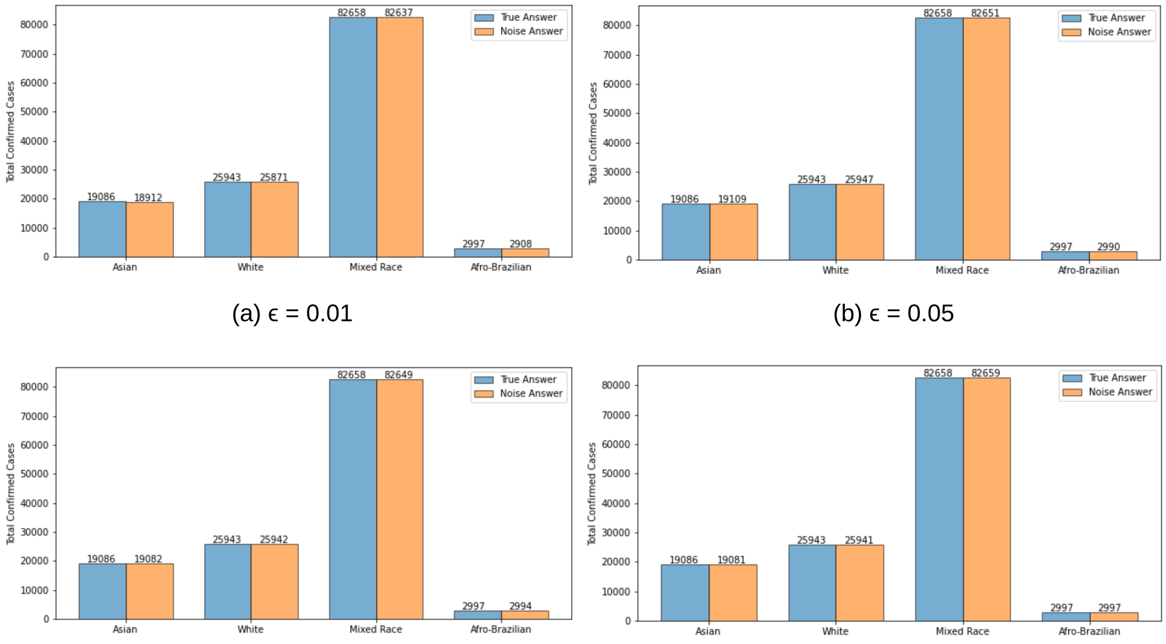
Even though the change is not noticeable from the observation of Figure 7 that indicates the number of confirmed cases when analyzing the error in Figure 9, it is possible to observe that the error decreases with the budget increase, improving the utility of the data. We can see a significant decay with the budget between 0.01 and 0.05. This behavior is very similar to that observed in the error analysis of the query Q1L in Figure 6.

The result mentioned above can also be analyzed when comparing the histograms of confirmed case numbers for the privacy budgets 0.01 and 0.05, in Figures 7(a), 8(a) and 7(b) and 8(b), for each gender and budget, respectively. In the first case, it is possible to analyze the difference in the Laplacian noise between men of the Afro-Brazilian race, in which we have a variation of about 300 cases more than the real value. This difference is more noticeable because it is one of the lowest values in the histogram. Therefore, there is a need to add more noise to guarantee the privacy of individuals in that category.

Analyzing the results when the budgets used are 0.1 and $ln(2)$, the noisy answer is very close to the correct answer. In Figure 9, we see the error related to these privacy budgets. When the budget increases, the error obtained answering the numerical queries decreases, and consequently, the data utility increases. When that happens, the data is more susceptible to privacy attacks.

Taking a closer look at the results of the experiments conducted over the queries presented in this section, we can conclude that the noise added by the Laplace mechanism depends on the total number of occurrences when answering a query. When this number is low, the noise necessary to ensure privacy is higher with a few individuals. This behavior is expected since the mechanism adjusts the noise necessary to maintain differentiated privacy restrictions, protecting individuals more subject to re-identification in a possible malicious attack.

With a significant number of values involved in a query answer, such as the neighborhoods shown in Figure 5 and the number of positive cases of the Mixed Race race regardless of gender in Figures 7 and 8, the needed noise is more negligible, despite the budget of privacy considered. Furthermore, there is less possibility of re-identification of these patients. That is due to the high values of occurrences, as it is easier to guarantee the privacy of individuals belonging to large groups in these situations. In



(a) ϵ = 0.01

(b) ϵ = 0.05

(c) ϵ = 0.1

(d) ϵ = ln(2)

Fig. 8. Number of positive female gender cases for each race.

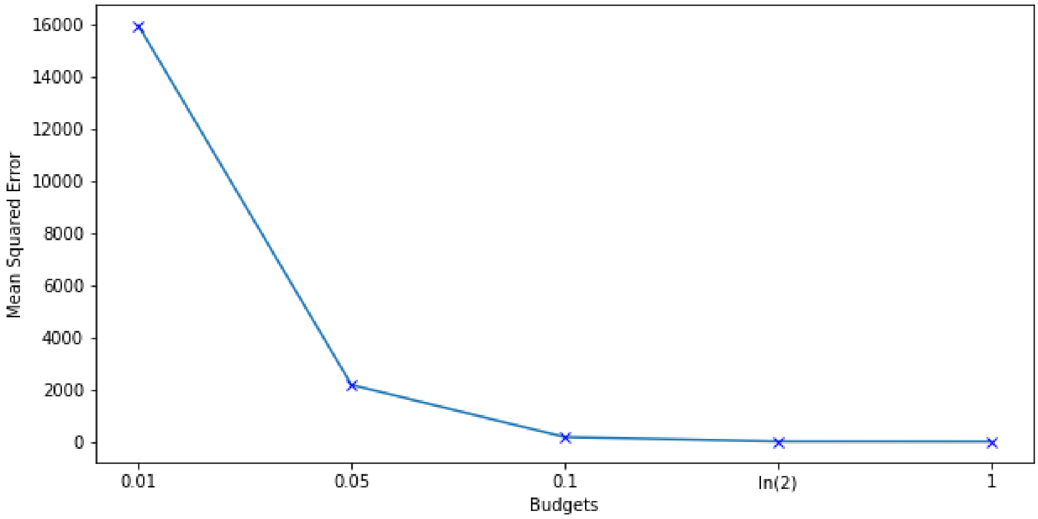


Fig. 9. Mean squared error for query responses Q2L.

small groups, more significant noise must be added to ensure privacy and make it harder to re-identify the patients.

4.2 Categorical queries with Exponential Mechanism

For categorical queries, we applied the exponential mechanism to answer query Q1E: "Which are the three comorbidities that have the highest number of COVID-19 cases". Since patients' risk of more severe symptoms is more significant for COVID-19 patients with comorbidities, it is necessary to identify which ones are more common among COVID-19 patients. It is important to remember that the same patient may have more than one comorbidity, which increases the risk of complications. The individuals' comorbidities are sensitive information, but they are essential to verify factors influencing COVID-19. Disclosing comorbidity information could lead to improper use of patient data without its authorization.

Query Q1E can return three of the comorbidities present in the database provided by IntegraSUS: Puerpera, Cardiovascular, Hematology, Down's Syndrome, Hepatitis, Asthma, Diabetes, Neurology, Pneumopathy, Immunodeficiency, Renal, Obesity, Hiv, and Neoplasms. The correct answer for the query is Cardiovascular, Diabetes, Renal. Any comorbidity may be present when applying the exponential mechanism to introduce privacy to the answer since we have a probability linked to each possible response.

As shown in Definition 3.6, the exponential mechanism requires a score function. We conducted several tests to find an appropriate score function that could correctly map the possible query outputs to a value that captures the answer quality. We ended up defining the score function that comprises 20% of the count of comorbidities. With low scores, the probability for each possible output is approximately the same. In contrast, the cost in choosing the outputs increases with the scores when applying the exponential mechanism.

Another relevant factor in calculating the probabilities of possible responses in the exponential mechanism is the sensitivity of the utility function, according to Definition 3.6. Analyzing the proposed categorical query and all possible answers, the greatest impact on the score function happens by adding

Table II. Average results from 100 runs of the exponential mechanism.

True response
Cardiovascular, Diabetes, Renal

Budget - ϵ	Query response
0.01	Diabetes, Hepatitis, Neoplasms
0.05	Cardiovascular diseases, Pneumopathy, Diabetes
0.1	Cardiovascular diseases, HIV, Puerperal
0.25	Cardiovascular diseases, Diabetes, Neoplasms
0.5	Cardiovascular diseases, Diabetes, Neurology

or removing the individual who has the greatest number of comorbidities. When executing the score function for each possible response of all neighborhood datasets, we obtained the sensitivity Δu is 9. This value is used to obtain the query output from the execution of the mechanism for budgets $\epsilon \in \{0.01, 0.05, 0.1, 0.25, 0.5\}$. We then compare the results obtained for each privacy budget, based on the output given by mechanism and the accuracy achieved.

The results in Table II refer to 100 executions of the exponential mechanism for each of the budgets. Note that as the budget increases, the response is more similar to the real one. For each query performed, the comorbidities returned can change according to the probabilities of the possible outputs calculated by the exponential mechanism. Thus, no answer is considered wrong. It only has a different probability of occurrence.

We apply accuracy to compare the results obtained for different privacy budgets and infer helpful information about these results. Thus, we sort the counts of all comorbidities. Then, we rank them in descending order. In the first position, we have the comorbidity with more confirmed cases of COVID-19, and in the last, the comorbidity with fewer numbers. In the next step, we map each pair, noise and correct response, to an array that would indicate the comorbidity rank. Then, we calculate the accuracy obtained in applying each privacy budget by comparing the answer given by the exponential mechanism and the true answer from query Q1E. Figure 10 shows the results. It confirms our previous analysis. As the privacy budget used on the exponential mechanism increases, the answer is closer to the correct response, thus, achieving a greater utility. For a lower number of comorbidities, more noise would be added to the query output, increasing the error, decreasing the accuracy, and therefore, generating less utility.

From the results, it is possible to infer that, as well as for numerical queries, the use of the budget

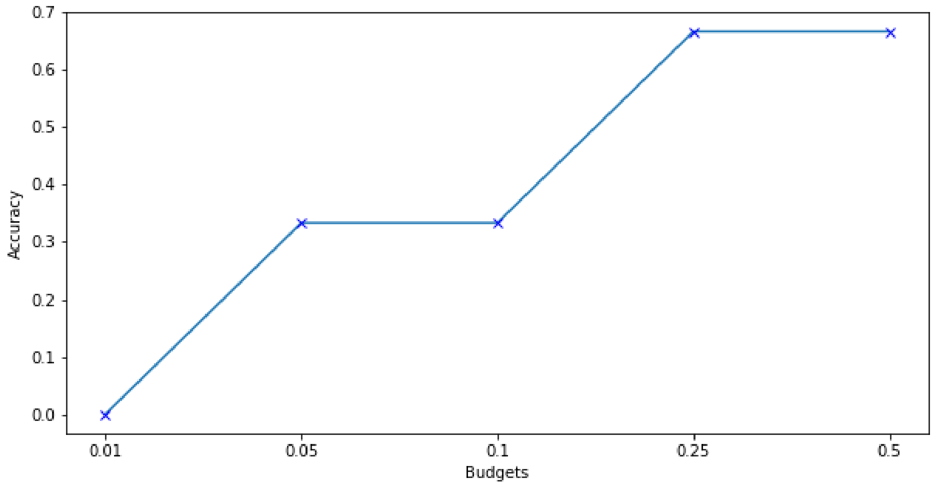


Fig. 10. Accuracy for query responses Q1E.

around $[0.05, 0.1]$ achieves a high level of privacy and a good level of utility. For those budgets, the exponential mechanism ensures ϵ-differential privacy with responses to provide a good utility level.

5. CONCLUSION

The patient data analysis is essential to fight the pandemic of COVID-19, both to help in the treatment and to contain the disease progress. Privacy is a right of every citizen and, therefore, must be preserved. It is possible to guarantee a private analysis with quality using differential privacy.

The Laplace mechanism presented good performance in answering the numerical queries about the COVID-19 dataset. Even with a small budget, it met the privacy requirements without a significant utility loss. Our experiments demonstrated that budgets between $[0.05, 0.1]$ performed the best for the queries we analyzed. We applied the exponential mechanism to preserve the individuals' privacy in the dataset answering categorical queries. Each possible output for the queries has a probability of occurring. The mechanism meets the required privacy level with the combination of score function, sensitivity, and privacy budget. Using ϵ between $[0.05, 0.1]$, we can guarantee individual's privacy and a relevant data utility in future analyses over the published data.

Finally, we can conclude that both mechanisms can achieve an excellent balance between privacy and utility on responses to numerical and categorical queries, especially with a budget between $[0.05, 0.1]$, decreasing the risk of malicious attacks and enhancing the quality of the analysis.

In the future, we intend to add to our analysis other types of aggregated queries using other categorical attributes, for example, age range, type of exam, to make more complex queries. Studying other noise addition mechanisms, such as Randomized Response [Warner 1965], Local Dampening [Farias et al. 2020], and comparing them with each other, can be helpful to assess the appropriate balance between privacy and utility in health data.

REFERENCES

Aktay, A., Bavadekar, S., Cossoul, G., Davis, J., Desfontaines, D., Fabrikant, A., Gabrilovich, E., Gadepalli, K., Gipson, B., Guevara, M., Kamath, C., Kansal, M., Lange, A., Mandayam, C., Oplinger, A., Pluntke, C., Roessler, T., Schlosberg, A., Shekel, T., Vispute, S., Vu, M., Wellenius, G., Williams, B., and Wilson, R. J. Google COVID-19 community mobility reports: Anonymization process description (version 1.0), 2020.

Bento Filho, M. E., Neto, E. R. D., and Machado, J. Publicação diferencialmente privada de dados de pacientes de covid-19. In *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*. SBC, SBC Open Lib, Fortaleza, pp. 247–252, 2020.

Brito, F. and Machado, J. Preservação de privacidade de dados: Fundamentos, técnicas e aplicações. In *36o JAI – Jornadas de Atualização em Informática*, F. Delicado, P. Pires, and I. Silveira (Eds.). SBC, 2, pp. 91–130, 2017.

Chaudhuri, A. and Mukerjee, R. *Randomized response: Theory and techniques*. Marcel Dekker, New York, 1988.

Dwork, C. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, M. Agrawal, D. Du, Z. Duan, and A. Li (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–19, 2008.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, S. Halevi and T. Rabin (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 265–284, 2006.

Dwork, C., Naor, M., Pitassi, T., and Rothblum, G. N. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*. Association for Computing Machinery, pp. 715–724, 2010.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9 (3-4): 211–407, 2014.

El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J.-P., Walker, M., Chowdhury, S., Vaillancourt, R., et al. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association* 16 (5): 670–682, 2009.

Erlingsson, Ú., Korolova, A., and Pihur, V. RAPPOR: randomized aggregatable privacy-preserving ordinal response. *CoRR* vol. abs/1407.6981, pp. 1054–1067, 2014.

Fahey, R. A. and Hino, A. Covid-19, digital privacy, and the social limits on data-focused public health responses. *Int. J. Inf. Manag.* vol. 55, pp. 102181, 2020.

Farias, V. A., Brito, F. T., Flynn, C., Machado, J. C., Majumdar, S., and Srivastava, D. Local dampening: Differential privacy for non-numeric queries via local sensitivity. *VLDB* 14 (4): 521,533, 2020.

Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.* 42 (4): 14:1–14:53, 2010a.

Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys* 42 (4): 1–53, June, 2010b.

Haas, S., Wohlgemuth, S., Echizen, I., Sonehara, N., and Müller, G. Aspects of privacy for electronic health records. *International journal of medical informatics* 80 (2): e26–e31, 2011.

Kifer, D. and Machanavajjhala, A. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. ACM, pp. 193–204, 2011.

Kuhn, C., Beck, M., and Strufe, T. Covid notions: Towards formal definitions - and documented understanding - of privacy goals and claimed protection in proximity-tracing services. *Online Soc. Networks Media* vol. 22, pp. 100125, 2021.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, pp. 656–672, 2019.

Lee, B., Dupervil, B., Deputy, N. P., Duck, W., Soroka, S., Botticchio, L., Silk, B., Price, J., Sweeney, P., Fuld, J., Weber, T., and Pollock, D. Protecting privacy and transforming COVID-19 case surveillance datasets for public use, 2021.

Lenert, L. and McSwain, B. Y. Balancing health privacy, health information exchange, and research in the context of the COVID-19 pandemic. *J. Am. Medical Informatics Assoc.* 27 (6): 963–966, 2020.

Li, N., Li, T., and Venkatasubramanian, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, IEEE, pp. 106–115, 2007.

Machado, J. C., Neto, E. R. D., and Filho, M. E. B. Técnicas de privacidade de dados de localização. In *XXXIV SBBD, Fortaleza, CE, Brazil, October 7-10, 2019*. Tópicos em Gerenciamento de Dados e Informações, 2019.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1 (1): 3, 2007.

McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, pp. 94–103, 2007.

McSherry, F. D. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*. SIGMOD '09. Association for Computing Machinery, New York, NY, USA, pp. 19–30, 2009.

Narayanan, A. and Shmatikov, V. How to break anonymity of the netflix prize dataset, 2006.

Nergiz, M. E., Atzori, M., and Clifton, C. Hiding the presence of individuals from shared databases. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. Association for Computing Machinery, pp. 665–676, 2007.

Reiter, J. P. Differential privacy and federal data releases. *Annual review of statistics and its application* vol. 6, pp. 85–101, 2019.

Samarati, P. Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.* 13 (6): 1010–1027, 2001.

SUS. Boletim epidemiológico novo coronavírus (covid-19), 2020. Acessado: 19-06-20.

Sweeney, L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (05): 557–570, 2002.

Vadrevu, P. K., Adusumalli, S. K., and Mangalapalli, V. K. Personal privacy preserving data publication of covid-19 pandemic data using edge computing. *Journal of Critical Reviews* 7 (1): 8103–8111, 2020.

Wang, Y., Wu, X., and Hu, D. Using randomized response for differential privacy preserving data collection. In *EDBT/ICDT Workshops*. Vol. 1558. Workshop Proceedings of the EDBT/ICDT 2016 Joint Conference, pp. 0090–6778, 2016.

Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* 60 (309): 63–69, Mar, 1965.

Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G., and Winslett, M. Differentially private histogram publication. *The VLDB Journal* 22 (6): 797–822, 2013.

Yu, K., Tan, L., Shang, X., Huang, J., Srivastava, G., and Chatterjee, P. Efficient and privacy-preserving medical research support platform against COVID-19: A blockchain-based approach. *IEEE Consumer Electron. Mag.* 10 (2): 111–120, 2021.