Using Maximum Entropy to Extend a Consent Privacy Impact Quantification

Arno Appenzeller

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
arno.appenzeller@kit.edu

Abstract

Due to the progress of digitization in the medical sector digital consent becomes more and more common. While digital consent itself has a huge number of benefits for the researcher it can impose a lot of questions for the individual giving it. One of those questions is what impact the consent to sharing data with a research project has on the individual's privacy. The Consent Privacy Impact Quantification (CPIQ) provides a quantification to help the user making a consent decision based on the potential data sharing risk and his individual acceptance preferences for a research project. While this quantification provides a good first estimation it has some limitations especially in the method the re-identification risk is calculated for a member of a dataset. This paper presents a method using the Maximum Entropy principle. This principle provides a way to measure the maximum unbiased distribution using limited background knowledge, which is provided by epidemiological data. This distribution can then be used to see how much higher the re-identification risk based on a sensitive attribute is compared to the uniform distribution. In addition, the first promising results of the method will be shown based on an experimental setting.

1 Introduction

Through the ongoing digitization medical research has potential access to an enormous amount of data. The recently soft-launched German "Elektronische Patientenakte" (ePA) also offers functionality of a research platform. While the main purpose is to provide a safe and secure storage for health data that is created during medical treatment, such data could potentially be useful for medical research. Having access to the digital treatment records of millions of people could lead to huge benefits, such as Big Data analysis of medical data. Analyzing large scale data is one of the most promising techniques to improve future treatment and make huge progress in medical care. Besides the obvious benefits there remain open questions regarding the privacy of the processed data. The European General Data Protection Regulation (GDPR) considers medical data to be highly sensitive which is prohibited to process by default. But there are several exceptions where one is the explicit consent of the affected person. Currently, this is the usual way to use data of a patient in medical research. Through the digitization the paper-based consent is more and more replaced and first digital consent systems are coming close to productive use [2, 3, 4]. On the one hand digital consent makes giving consent easier but it does not necessarily make the actual decision easier. In fact, many parties or research projects to grant access can make the decision even harder. Additionally, every consent made for medical data should be an informed consent. While the definition of what an informed consent is remains a research topic on its own, there are systems needed that support patients when making consent decisions. Such a system is the Consent Privacy Impact Quantification (CPIQ) [1]. CPIQ provides a way to measure different properties of a research project and considers the potentially shared data to support the affected patient with their a consent decision. CPIQ considers many different properties but lacks an actual quantification of the shared data of an individual in regard to the database where it is shared to. Such a look can make a huge difference in terms of privacy because unique and striking data can make re-identification a lot easier. Unfortunately, such things are often only noticed after the data is added. This could be too late to protect the privacy and it is too late to avoid a risky sharing decision. In this paper we will use the Maximum Entropy principle to provide a conservative estimation of the actual privacy risk of the data release. Therefore, we use epidemiological observations that are used as constraint for the Maximum Entropy methods. It is shown that using such methods can provide an accurate estimation how likely the re-identification of an individual is in a dataset. The remainder of this paper is structured as follows: Section 2 looks on related work of this topic. Section 3 introduces theoretical preliminaries that are needed to understand CPIQ and the method itself. Section 4 then describes our extension of CPIQ. Section 5 looks at our experiments. With this discussion the paper will be concluded and an outlook on future work will be provided.

2 Related Work

There is a multitude of papers that contribute to the topic of quantification of medical data in terms of privacy in various ways. Veeningen et al. describe a formal model for pseudonymization [14]. They use an exemplary digital health infrastructure where different data is shared across various sites. Each party is allowed to have different parts of data of an individual. The paper presents a so-called coalition graph which shows which party can combine which data to gain all data of a patient. This graph can be used to compare different data protection concepts. In contrast to this report Veeningen's approach focuses on pseudonymization architecture. While the idea for the formal model is very interesting this report considers the patient's view on its data and what impact on individual privacy data sharing has. The authors of "Quantifying the costs and benefits of privacy-preserving health data publishing" introduce a cost model for personal health records [9]. The approach tries to measure the cost of privacy and utility by comparing the cost of anonymization with the costs of a potential data breach. With the provided formulas a detailed comparison is possible but this approach is not suitable to measure individual data. Wan et al. look at a game theoretic approach to measure re-identification risks [15]. The authors try to weigh the factor between the monetary value of health data and the potential fine for a violation of privacy rules. They use different properties like generalization strategies for the data and their costs to create their model. An attacker is described that attempts re-identification when the benefit outweighs the costs. The paper concludes that it is possible to find something like zero risk if there is no incentive to attempt an attack. A game theoretical approach is not compatible with the main idea of CPIQ which is to measure an individual risk and provide decision support when sharing personal health data. Additionally, our attacker tries to re-identify regardless of the costs to express that individual risk of re-identification. Another work by Wang et al. presents methods to measure the privacy level of a dataset [16]. Their method evaluates the privacy impact of data with quantitative and qualitative factors. The factors will be assessed using hierarchical decision making with the help of expert knowledge to rate data sensitivity. The result is then combined into a so-called privacy score. In contrast to our work the need for expert knowledge can be a high obstacle in terms of real-world execution. A paper that is closer to our approach is "Privacy-MaxEnt" by Du et al. [5]. The authors consider a scenario where quasi-identifiers are bucketized with sensitive attributes. An example is gender and age as quasi-identifiers and diseases as sensitive data. The main principle would be that the probability that a sensitive attribute belongs to the individual is distributed equally. It is shown that this is not the case for certain sensitive attributes (e.g., gender specific diseases). This background knowledge is then modeled by using Maximum Entropy to show the probability given the sensitives attributes. While the paper discusses a sophisticated approach it lacks a real use case. It also makes no decision on what background knowledge should be used to model the constraints. In our work we use the core idea of the paper and extend our CPIQ technology with concrete examples. None of the here presented approaches describe a complete quantification that can support the decision of an individual to share its personal health data.

3 Preliminaries

In this section the preliminaries needed for MaxEnt CPIQ are described. We first introduce some common privacy preserving techniques which are used in CPIQ and explain the motivation to mitigate re-identification attacks. MaxEnt CPIQ uses Maximum Entropy to provide a more accurate privacy impact quantification. The concept of Maximum Entropy will be also explained in this section. Finally, the formal consent model needed for CPIQ is introduced and CPIQ itself is explained.

3.1 Privacy Preserving Technologies

The motivation behind the consent privacy impact quantification of CPIQ is to provide the affected person with an estimation of what risk comes with sharing its medical data. Even if the data is anonymized or pseudonymized before it is given to a third party there is still the risk of re-identification with background knowledge. Obvious examples are a large data set where only one person has a very rare disease. Such cases can be easily identified. However, there are several more sophisticated re-identification approaches that were shown in several studies and go beyond purely academic examples [12, 11]. Considering personal health data as highly sensitive data it should be clear that measures are needed to mitigate this risk. Therefore, different privacy preserving technologies exist. Besides technologies like homomorphic encryption, which is used in more and more cases, or statistical guarantees like differential privacy (DP) more traditional approaches rely on suppressing or generalizing quasi-identifiers. One of them is k-anonymity which requires that in a dataset there needs to be at least $k-1$ other individuals with the same quasi-identifiers [13]. This helps to reduce re-identification based on background knowledge about quasi-identifiers which could be de-facto public knowledge. To reach this goal suppression, where quasi-identifiers are removed, or generalization, where quasi-identifiers are grouped into more general categories, is used to form equivalence classes. One weakness of k-anonymity is that it does not consider the sensitive attribute itself. This is where l-diversity comes into place [10]. l-diversity requires that at least $l-1$ distinct sensitive attributes exist in each equivalence class. This mitigates the risk for cases where re-identification would be trivial. For example, k-anonymity would allow equivalence classes where everyone has the same sensitive attribute. This would be a privacy leak itself.

3.2 Maximum Entropy

The principle of maximum entropy follows the idea to define a maximum unbiased distribution given some constraints, which would be the distribution with the largest entropy. This information theory concept itself was introduced in 1957 by Jaynes [7, 8]. What is called constraints above can also be described as testable information. This information gives a mathematical statement of the probability distribution. A testable information can be that the sum of two event probabilities p_1 and p_2 is smaller than 0.5. Depending on the definition of Maximum Entropy there always is the universal constraint that the sum of all event probabilities is 1. Given those constraints equations can be formed under that the distribution fulfills the constraints and maximizes entropy. To solve the equations the so-called Lagrange multipliers can be used. Those mathematical details can be found in the original publication and are not looked at in this paper.

3.3 Formal Consent Model

The foundation for CPIQ is the formal consent model which was introduced in the original publication. The formal model defines the properties required to describe consent for secondary usage (e.g., research). The model is based on the technical consent model of the German ePA and is combined with properties out of the research consent template of the German "Medizin Informatik Initiative" which is widely accepted by data regulation authorities. Table 3.1 gives an overview of the properties. It consists of the subject S who can be a patient or a legal guardian. The researcher is referenced as the authorized party AP. Every declaration of consent is required to have a timespan $TS = (TS_{Start}, TS_{End})$ during it is valid. The consent then is defined through policies which also contain which documents or categories R are shared. A full policy consists out of P = [(AP, TS, R, A)] where A is the action allowed on the resources. For secondary usage this is limited to a read action. The next properties are focused on the concrete research project. The purpose PU of a project is considered as well as potential personal or social benefits PBE and SBE which are listed in BE. Furthermore, the degree of anonymization DA_D and

Identifier	Explanation
S = Patient | Legal Guardian	Subject
AP = [Researcher]	Authorized Party
TS_{Start} = Date	Starting Date
TS_{End} = Date	End Date
TS = (TS_{Start}, TS_{End})	Timespan
R = [Document | Category]	Resource
A = Read (r)	Action
P = [(AP, TS, R, A)]	Policies
PU = [Purpose] | Broad Consent	Research purpose
PBE = [Personal Benefit]*	Personal Benefit
SBE = [Social Benefit]*	Social Benefit
BE = [PBE | SBE]*	Benefit
DA_D = (k-Anonymity, l-Diversity)	Degree of anonymization for D
PS = Low | Medium | High	Processing security
D = (PS, DA_D)	Data processing
DA_{PUB} = (k-Anonymity, l-Diversity)	Degree of anonymization for P
I = ($false$ | $true$)	Information
PUB = (($false$ | $true$), DA_{PUB})	Publication
T = (I, PUB)	Transparency
RI = (PU, BE, D, T)	Research information

Table 3.1: Identifiers for the formal consent model

the processing security PS are part of the data processing $D = (PS, DA_D)$ of a project. Another factor is transparency T which consists out of information value I and the publication value PUB which states if there is a publication and if yes what kind of anonymization DA_{PUB} is used. This is then listed in the research information $RI = (PU, BE, D, T)$.

3.4 Consent Privacy Impact Quantification (CPIQ)

Based on the formal consent model mentioned in Section 3.3 CPIQ provides a consent privacy impact quantification that consists out of two main parts: acceptance and risk of a consent decision. The idea is that CPIQ calculates a score (in most times from 0-100) which indicates the higher it is the more the acceptance factors (AF) outweigh the potential risks. For the acceptance factors we refer to the original publication. They consist of the user-weighted formal model properties of purpose, personal and social benefits, information, publication, and trust.

The risk will be explained in more detail because this is where our model extension comes into place. At first, we assume an attacker that has access to all publicly available data of a patient. The attacker's goal is to gain knowledge about a potential sensitive attribute of an individual. This goal can be reached through a classical re-identification attempt. To quantify this risk two main attack points are identified. The first is the attack on the stored private data of the research project. For this the risk probability of a data breach DLP also needs to be considered. The second way is to try re-identification on the data that are part of the published data of a project. This depends on the publication factor probalitity PF. In both cases the re-identification itself will be measured through the sensitive attribute exposure probability $SAEP$. The assumption for CPIQ is that every project uses a certain degree of l-diversity as privacy preserving technology. So $SAEP = Min(1, \frac{|R|}{l})$ with $|R|$ as number of resources an individual has in the dataset. We also assume that a patient can have more than one sensitive attribute in the dataset, so this "row" in a dataset is mapped to the same set of quasi-identifiers and will weaken the l-diversity property. This leads to $\frac{|R|}{l}$ as probability. If there are more properties than ensured through l-diversity re-identification is obvious and therefore the probability is one. This combined with the two attack vectors results in the re-identification probability due to data leakage $RPD = P(\text{Damage}_{\text{Data leakage}}) = DLP * SAEP_{DP}$ and re-identification probability due to publication $RPP = P(\text{Damage}_{\text{Publication}}) = PF * SAEP_{PUB}$. CPIQ also considers the location of processing as a risk factor and requires that all processing locations have a regulation with similar standards as GDPR which is defined by the GNF factor. The Total Re-Identification Risk Probability $TRRP = P(\text{Damage}_{\text{Total}}) = 1 - ((1 - RPD) * (1 - RPP) * (1 - GNF))$ is the result of this. Combined with the acceptance factors the CPIQ score will be calculated with the following equation:

$$CPIQ = AF * (\frac{L}{2} * (1 - \frac{1}{s})) + (1 - TRRP) * (\frac{L}{2} * (1 + \frac{1}{s}))$$

with $s \geq 1$ as weighting factor between risk and acceptance and L as maximum value.

4 Using Maximum Entropy with CPIQ

After we introduced the preliminaries, this section will point out why the extension with Maximum Entropy should be done and how it can be implemented. In addition, we show some experiments with the new approach.

4.1 Model extension

Section 3.4 described the factors used to calculate the risk for CPIQ. One assumption made is that l-diversity is used as privacy preserving technology. This is used for $SAEP$ which is one of the main factors. While this assumption can be made it is not clear that this always can be implemented in practice. While anonymization and pseudonymization technologies are a common practice in medical research, it does not seem too realistic that every research project uses a suitable degree of l-diversity or that l-diversity can be applied in a good way. This could lead to the case that CPIQ does not give a very accurate consent evaluation. The goal for the extension was to consider more the uniqueness of

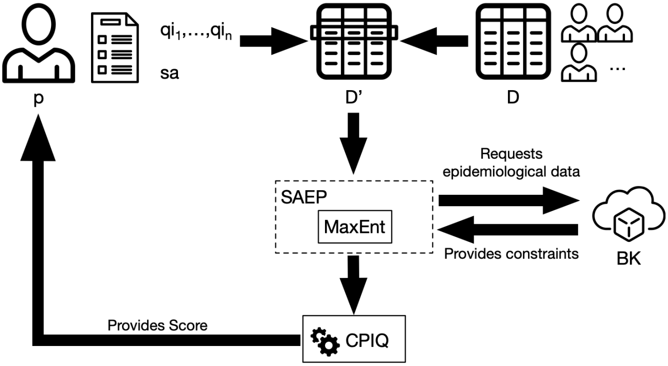


Figure 4.1: MaxEnt CPIQ Workflow

a sensitive attribute of an individual. While l-diversity provides a method for this on a database level it does not consider the background knowledge of a potential attacker. It also requires the assumption that every sensitive attribute is uniformly distributed. Especially with medical data this is not the case. There are certain diseases that are more common than others. In addition, different factors like age or gender can heavily affect the frequency of a disease. An obvious but good example for this is breast cancer, which occurs in female and male persons. However, breast cancer is very rare for males so that a database with different cancer types from individuals with both genders could lead to an easier linking of the sensitive attributes to the individuals. We found that Maximum Entropy suits best to include such information. This also deals with the facts that no background knowledge can be complete. For this the Maximum Entropy principle provides the best non-biased estimation given the currently available information.

To replace $SAEP$ with l-diversity an individual $p = (\{qi_1, ..., qi_n\}, sa)$ is introduced. The individual wants to give its sensitive attribute sa with its quasi-identifiers to a dataset D, which already includes other patients with sensitive data. We also introduce a background knowledge source BK which uses publicly available medical information like disease incidence per gender, age, region, and more epidemiological data to provide background knowledge constraints bkC_i. Figure 4.1 shows the workflow for Maximum Entropy CPIQ. The individual p provides its data to be combined before sharing with the dataset D of the research project. It remains to be noted that it is an open question where this combination and processing happens. One option would be to do it locally at the patient's device or at a trusted third party. This combined data set D' will be then used to calculate the $SAEP$. Therefore, epidemiological data for the given sensitive attributes and quasi-identifiers will be requested from BK. In our case this data is the incidence for the given diseases per gender and per region. This incidence is then used as constraints for a Maximum Entropy distribution. To calculate the risk the difference between the uniformly distributed re-identification risk of all individuals is compared with the constrained Maximum Entropy distribution re-identification risk. This factor is then divided by a custom threshold. This risk threshold defines the factor of how much higher the risk can be tolerated compared to uniformed distribution. The minimum of the received value or 1 will be returned as $SAEP$ value and the rest of the CPIQ process can continue as described.

Definition 4.1.1 (MaxEnt CPIQ)**.** Let UD be the uniform distribution of D. uR is the share of any individual in the distribution ($uR = 1/n$) where n is the number of individuals in the dataset. CD is the constrained distribution of D. This is calculated by using the given constraints for D with the Maximum Entropy principle. cR_i is the constrained distribution of a given individual p_i. The personal risk factor is then $pRF_i = cR_i/uR$. Let $\perp$ be the risk threshold. The weighted risk ratio is then $rr_i = min(1, pRF_i/\perp)$. rr_i is a value between 0 and 100% so that 1 (100%) is the maximum.

4.2 Experiments

To show the feasibility of the extension some experiments were done. Some exemplary scenarios with small sample datasets were created to show the approach in a comprehensible way. For this the technique was implemented in Python[1]. Maximum Entropy was implemented by using the Python Package

[1] https://www.python.org

ICD10 C*	C00-C14	C15	C16	C18-C21	C22	C50	...
Male Overall	17.2	9.0	14.8	51.5	9.4	0.7	...
Female Overall	6.9	2.2	7.5	35.1	3.6	109.2	...

Table 4.1: Exemplary excerpt of incidence data for different cancer types as ICD-10 Codes per gender

maxentropy[2]. The scenario is a database with of several individuals that have different types of cancer. For the background knowledge data on cancer we used the population wide cancer incidence provided by the German Center for Disease Control the Robert Koch Institut [6]. Table 4.1 shows an excerpt of the aggregated data. The actual data set also includes age specific and region-specific incidences. For this scenario we only differentiated by a higher risk age for breast cancer (ICD-10 code C50; higher risk with age older than 60) and the lower risk age. Since in our dataset every person has a disease and the incidence is a population wide metric, we also calculated a total share depending on the incidence and the complete data set. The labels in the dataset have the format: (qi_1, sa, qi_2) where qi_1 is the gender, qi_2 is the age and sa is the ICD-10 code for the type of cancer. As risk threshold $\bot = 3$ is used.

4.2.1 Scenario 1: Adding a lower risk person

Figure 4.2 shows the constrained distribution of the scenario dataset D before the additional subject is inserted. The male person with breast cancer (C50) has a very low risk to be relinked to its sensitive attribute which is obvious because it is rare for males. In addition, the individual with prostate cancer (C61) has a very high risk since this is one of the most common cancer types for males. Furthermore, this disease does not exist for females because of biological reasons. Next a new subject wants to share its data. The data is from a female with breast cancer in the lower risk age range $p_1 = ("Female", "50", 40)$. Figure 4.3(a)

[2] https://github.com/PythonCharmers/maxentropy

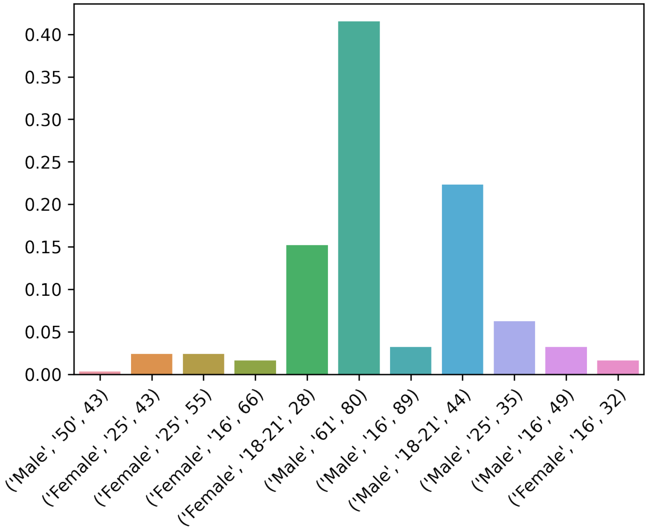
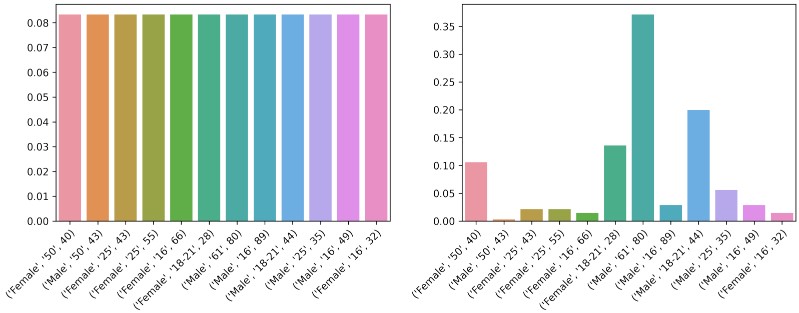


Figure 4.2: Scenario Dataset D before insertion with constrained distribution



(a) Uniform distribution

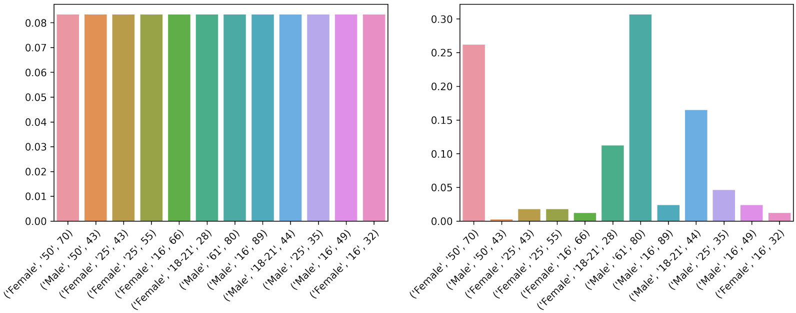
(b) Constrained distributed

Figure 4.3: Scenario 1 Dataset D' after insertion of p_1

shows the dataset with the inserted data and an assumed uniform distribution. This is needed to calculate the difference between the constrained distribution that can be seen in Figure 4.3(b). The constrained distribution shows that the re-identification is higher than with the uniform distribution but only slightly. The weighted risk ratio for $SAEP$ is 0.32, which can be considered as lower risk.

4.2.2 Scenario 2: Adding a higher risk person

The starting situation is the same as in Scenario 1. This time a higher risk for breast cancer person $p_2 = ("Female", "50", 70)$ is added. Figure 4.4(a) shows the uniform distribution and Figure 4.4(b) the constrained distribution. The risk is much higher than for the low-risk person. In fact, this has now the second highest risk which can also be seen in the $SAEP$ risk ratio which is the maximum with 1.



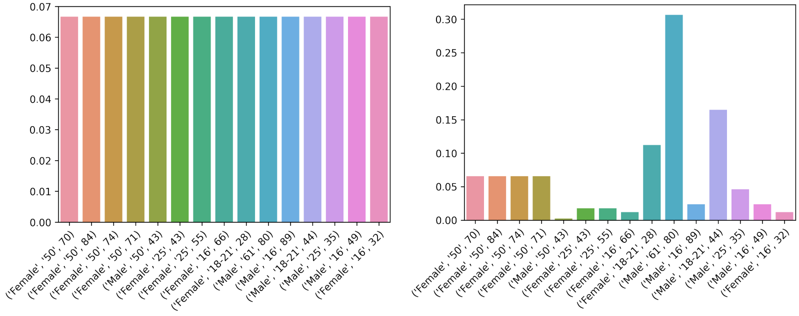
(a) Uniform distribution

(b) Constrained distributed

Figure 4.4: Scenario 2 Dataset D' after insertion of p_2

4.2.3 Scenario 3: Adding three higher risk persons

In contrast to Scenario 2 it needs to be looked at what happens when risk is more evenly distributed. Therefore, three higher risk persons are added. Figure 4.5(a) and 4.5(b) show that the data sharing for the individual higher risk person has now a smaller risk than before. The individual risk ratio for one of the three persons is $0,33$ which is a bit higher than in Scenario 1 but much smaller than in the second experiment.



(a) Uniform distribution

(b) Constrained distributed

Figure 4.5: Scenario 3 Dataset D' after insertion of three higher risk persons

5 Discussion

The experiments in Section 4.2 showed a proof of our concept. However, the experiment was done with a small sample size and is no complete proof for the principle. Nevertheless, Maximum Entropy provides a good way to include background knowledge. Publicly available data like the cancer registry data can be included easily as constraints for a re-identification metric. While more traditional metrics like k-anonymity or l-diversity provide a concrete way for the data owner to improve the privacy impact of a dataset those values remain vague for the affected person. In addition, not every dataset can implement any value for k or l. Furthermore, the generalization or suppression to receive several specific sized equivalence classes is no trivial task. From this point of view the Maximum Entropy model has fewer requirements depending on the data structure. Instead, it requires a specific form of data input for the model. Any database should be able to be mapped to the format of quasi-identifiers and the sensitive attribute which should be measured for uniqueness in the dataset. Public databases to form constraints out of epidemiological data should be also widely available. On the other hand, there are open questions where to process the evaluation. Expecting the individual to let the data process by a third party could easily require the same level of trust as it would to give the consent and share the data with the researcher. Another idea would be to provide the current dataset D to the potential participant to do the CPIQ evaluation. This could be unacceptable for research institutes for privacy reasons or even for intellectual property reasons. A trusted third party by the potential participant and the researchers would solve this but it would require high standards to gain this trust. While there were no user studies, we think that our risk calculation is more natural by using a metric that considers how much higher the risk depending on the quasi-identifier and epidemiological background knowledge for a sensitive attribute is compared to if every sensitive attribute is distributed equally in the population. As our experiments show there can be a hen egg problem with smaller datasets or rare diseases. While adding one individual that has a high risk for breast cancer would lead to bad CPIQ recommendation from the risk side it would be better if there were three persons with the same sensitive attribute. This imposes the questions where the additional persons should come from and if it is ok to assume that there are some privacy risk friendly persons that share their data regardless of the CPIQ score. The same applies to small datasets. Another thing that should be noted is that our experiments are limited. There is no complete comparison against the l-diversity version of $SAEP$ or an evaluation with real world data.

6 Conclusion and Outlook

This technical report presents an extension to the risk model of the consent privacy impact quantification CPIQ. The original form of CPIQ uses l-diversity to measure the individual privacy risk for a patient that wants to share his data. This imposes many issues and may be an impracticable requirement. Therefore a method that does not impose any requirements on the data structure or certain anonymization methods was needed. The Maximum Entropy principle is a promising method for this. It can be used to measure a maximum unbiased distribution based on limited background knowledge. As source for background knowledge epidemiological data which is publicly available for the potential disease as sensitive attribute is suggested. With this the Maximum Entropy principle can be used to measure the difference between a uniform distribution and the constrained distribution. This difference can then be used as weighted risk ratio which replaces l-diversity in the CPIQ method. An experimental evaluation is presented, and the results are discussed. While this paper does not provide a complete analysis of this extension the first results look very promising, and the Maximum Entropy extension seems to be a feasible method with less requirements than the original method.

As described before this paper does not provide a full analysis of our suggested extensions. For future work a complete evaluation against l-diversity is needed. It is important to measure the difference between l-diverse tables and their results in $SAEP$ and when the Maximum Entropy principle is used on this data. A limitation would be that not every assumption that was made for the original form of CPIQ can also be made for the extension. This also needs to be analyzed in depth. Furthermore, a real-world dataset evaluation would be very interesting. Our experiments only used a very limited and small sample data set. It would be interesting to obtain a real-world data set from for example a hospital or a cancer registry and measure the constrained distribution in this dataset. This could also be used to analyze the acceptance of such a method. Finally, the question for an optimal distribution and size of a dataset should be looked at. Our experiments showed that the size of a dataset and the distribution of it has a large influence on the risk estimation. While this is a question itself the here introduced method can be used to recommend optimal datasets that have a high acceptance for the potential data donors.

References

[1] Arno Appenzeller et al. “CPIQ - A Privacy Impact Quantification for Digital Medical Consent”. In: *The 14th PErvasive Technologies Related to Assistive Environments Conference*. PETRA 2021. New York NY, USA: Association for Computing Machinery, 2021, pp. 534–543.

[2] Arno Appenzeller et al. “Enabling Data Sovereignty for Patients through Digital Consent Enforcement”. In: *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA ’20. Corfu, Greece: Association for Computing Machinery, 2020. ISBN: 9781450377737. DOI: `10.1145/3389189.3393745`. URL: `https://doi.org/10.1145/3389189.3393745`.

[3] M. Bialke et al. “MOSAIC - A Modular Approach to Data Management in Epidemiological Studies”. In: *Methods of Information in Medicine* 54.04 (2015), pp. 364–371.

[4] M Bialke et al. “A workflow-driven approach to integrate generic software modules in a Trusted Third Party.” In: *J Transl Med* 13 (2015), p. 176.

[5] Wenliang Du, Zhouxuan Teng, and Zutao Zhu. “Privacy-MaxEnt: Integrating Background Knowledge in Privacy Quantification”. In: SIGMOD ’08. Vancouver, Canada, 2008.

[6] Robert Koch-Institut (Hrsg) und die Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. *Krebs in Deutschland fr 2015/2016. 12. Ausgabe*. In German. 2019.

[7] E. T. Jaynes. “Information Theory and Statistical Mechanics”. In: *Physical Review* 106.4 (May 1957), pp. 620–630. DOI: `10.1103/PhysRev.106.620`.

[8] E. T. Jaynes. “Information Theory and Statistical Mechanics. II”. In: *Physical Review* 108.2 (Oct. 1957), pp. 171–190. DOI: `10.1103/PhysRev.108.171`.

[9] RH Khokhar et al. “Quantifying the costs and benefits of privacy-preserving health data publishing.” In: *J Biomed Inform* 50 (2014), pp. 107–121.

[10] A. Machanavajjhala et al. “L-diversity: privacy beyond k-anonymity”. In: *22nd International Conference on Data Engineering (ICDE’06)*. 2006, pp. 24–24. DOI: 10.1109/ICDE.2006.1.

[11] Yves-Alexandre de Montjoye et al. “Unique in the Crowd: The privacy bounds of human mobility”. In: *Scientific Reports* 3 (2013).

[12] Arvind Narayanan and Vitaly Shmatikov. “How To Break Anonymity of the Netflix Prize Dataset”. In: *CoRR* abs/cs/0610105 (2006). arXiv: cs/0610105. URL: http://arxiv.org/abs/cs/0610105.

[13] Latanya Sweeney. “k-Anonymity: A Model for Protecting Privacy”. In: *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10.5 (Oct. 2002), pp. 557–570. ISSN: 0218-4885. DOI: 10.1142/S0218488502001648. URL: https://doi.org/10.1142/S0218488502001648.

[14] Meilof Veeningen, Benne de Weger, and Nicola Zannone. “Formal Modelling of (De)Pseudonymisation: A Case Study in Health Care Privacy”. In: *Security and Trust Management: Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 145–160.

[15] Z Wan et al. “A game theoretic framework for analyzing re-identification risk.” In: *PLoS One* 10.3 (2015), e0120592.

[16] Dan Wang, Bing Guo, and Yan Shen. “Method for measuring the privacy level of pre-published dataset”. In: *IET Inf. Secur.* 12.5 (2018), pp. 425–430.