PharmaSUG China 2021 - Paper AD-008

Risk of Re-identification Assessment

Lanlan Peng, Sanofi

ABSTRACT

In recent years, data de-identification and anonymization of individual privacy data are always popular topics around the pharmaceutical companies so far. The European Medicines Agency (EMA) announced that protect patient privacy and balance the utility of de-identified data which are the keys for clinical data sharing.

As data de-identification approaches are almost complete, the new considers have been brought out. Could the anonymized data be shared while keeping it secure in non-public environment? How to evaluate risk of re-identification for clinical data? This paper will discuss two different concepts to calculate the risk of re-identification.

INTRODUCTION

Sharing clinical trial data has potential benefits to stimulate new ideas for clinical researchers and to maximize data utility to improve the safety and effectiveness of therapies for patients. At meantime, marking sure that shared clinical data should be de-identified, and it doesn't have risk of re-identification.

The paper discusses scenarios in which clinical trial data has already been de-identified, the risk of re-identification of anonymized data should be evaluated before sharing. The scope of this paper is to present the two popular methods in clinical data protection which are *k-anonymity* and *l-diversity*.

RISK OF RE-IDENTIFICATION ATTEMPT

The re-identification risk classified into two categories, one is external attack and the other is internal attack. The external attack depends on the data control such as deliberate attack, inadvertent re-identification, data breach and public data release. This paper focuses more on the internal attack which comes from de-identified data itself. Therefore, *k-anonymity* and *l-diversity* approaches will be used in the re-identification risk calculation.

QUASI-IDENTIFIER SELECTION

Before evaluating risk of re-identification with *k-anonymity* and *l-diversity*, the first step is that find quasi-identifiers which might link other information to identify an individual in anonymized clinical data, even if they have already generalized. Usually, age, sex, geographic information, race, ethnicity should be considered in the re-identification risk assessment.

RISK OF RE-IDENTIFICATION CALCULATION

K-ANONYMITY

Definition

Each de-identified clinical data satisfies *k-anonymity* if every patient in this data cannot be distinguished from at least *k-1* other patients by the combination of value of quasi-identifiers. By *k-anonymity* method, the probability of being identified becomes less than *1/k*.

K-anonymity calculation

The following is an anonymized clinical data with one sensitive variable and two quasi-identifiers.

- Quasi-identifier: Age, ZIP code
- Sensitive variable: Disease, this variable isn't sufficient to identify a patient, but they contain health information

	ZIP Code	Age	Disease
Equivalence class A	476*	2*	Heart Disease
	476*	2*	Heart Disease
	476*	2*	Heart Disease
	476*	2*	Heart Disease
Equivalence class B	4790*	>40	Flu
	4790*	>40	Heart Disease
	4790*	>40	Cancer
Equivalence class C	476*	3*	Heart Disease
	476*	3*	Cancer
	476*	3*	Cancer

Figure 1. Example for *k-anonymity*

In this example, the de-identified clinical data are categorized by two quasi-identifiers into three equivalence class.

- Equivalence class A: 4 records
- Equivalence class B: 3 records
- Equivalence class C: 3 records

The *k* value is the smallest number of records in equivalence classes. It presents that this de-identified clinical data has *3-anonymity* with respect to the combination of quasi-identifiers Age and ZIP code. The probability of being identified becomes less than or equal to one third.

Possible attacks of k-anonymity

Sometimes a *k*-anonymized data has two possible attacks which can disclose personally identifiable information. We use the anonymized clinical data from the above example to present the details.

Homogeneity attack

The attackers can discover the values of sensitive variables when it has little diversity, it's possible to re-identify an individual.

	ZIP Code	Age	Disease
Equivalence class A	476*	2*	Heart Disease
	476*	2*	Heart Disease
	476*	2*	Heart Disease
	476*	2*	Heart Disease
	4790*	>40	Flu
	4790*	>40	Heart Disease
	4790*	>40	Cancer
	476*	3*	Heart Disease
	476*	3*	Cancer
	476*	3*	Cancer

Figure 2. Example for homogeneity attack

The equivalence class A only consists of heart disease patients. If an attacker knows patient A participates in this clinical trial and his ZIP code is 4768 and he is 29 years old, he must have heart disease.

Background knowledge attack

The attackers often have background knowledge, it means that *k*-anonymized data has a risk of privacy information exposure while using background knowledge.

	ZIP Code	Age	Disease
	476*	2*	Heart Disease
	476*	2*	Heart Disease
	476*	2*	Heart Disease
	476*	2*	Heart Disease
	4790*	>40	Flu
	4790*	>40	Heart Disease
	4790*	>40	Cancer
Equivalence class C	476*	3*	Heart Disease
	476*	3*	Cancer
	476*	3*	Cancer

Figure 3. Example for background knowledge attack

If an attacker knows patient B is a 31-year-old Japanese who lives in ZIP code 4760, but he isn't sure whether patient B has cancer or has heart disease. However, there is background knowledge about Japanese has an extremely low incidence of heart disease. So, the attacker concludes with near certainty that patient B has cancer.

With these two simple examples, *k-anonymity* cannot guarantee privacy against using those two attacks. Therefore, we propose another powerful approach called *l-diversity* that can defense such attacks and better remind *k*-anonymized data maybe exist risk of re-identification.

L-DIVERSITY

Definition

Each equivalence class satisfies *l-diversity* if there are at least *l* "well represented" value for the combination of sensitive variables.

These are three possible ways to define "well represented":

- Distinct *l-diversity*
- Entropy *l-diversity*
- Recursive *(c-l)-diversity*

In this paper, it presents the details of distinct *l-diversity* which is the simplest way to calculate *l* value.

Distinct l-diversity calculation

The following example describes an anonymized clinical data with two quasi-identifiers and one sensitive variable.

- Quasi-identifier: Sex and Age
- Sensitive variable: Disease

	Sex	Age	Disease
Equivalence class A	F	2*	Heart Disease
	F	2*	Heart Disease
	F	2*	Heart Disease
	F	2*	Flu
Equivalence class B	M	>40	Flu
	M	>40	Heart Disease
	M	>40	Cancer
Equivalence class C	F	3*	Heart Disease
	F	3*	Cancer
	F	3*	Cancer

Figure 4. Example 1 for distinct *l-diversity*

This anonymized clinical data contains three equivalence class.

- Equivalence class A: 4 records, 2 different diseases
- Equivalence class B: 3 records, 3 different diseases
- Equivalence class C: 3 records, 2 different diseases

The *l* value is equal to the smallest number of different diseases in equivalence class. It means that this de-identified clinical data has *3-anonymity* and *2-diversity* with respect to the combination of value of Sex and Age. An attacker cannot link those two quasi-identifiers with sensitive variable Disease with probability greater than or equal to one half.

Through this example, it can be concluded this 3-anonymized clinical data has risk of re-identification with high probability. In order to minimize the risk of re-identification, Age need to be replaced with a broader category. The value 2* and 3* are replaced by <40, three equivalence classes are regrouped to two equivalence classes.

- Equivalence class A: 7 records, 3 different diseases
- Equivalence class B: 3 records, 3 different diseases

	Sex	Age	Disease
Equivalence class A	F	<40	Heart Disease
	F	<40	Heart Disease
	F	<40	Heart Disease
	F	<40	Flu
	F	<40	Heart Disease
	F	<40	Cancer
	F	<40	Cancer
Equivalence class B	M	>40	Flu
	M	>40	Heart Disease
	M	>40	Cancer

Figure 5. Example 2 for distinct *l-diversity*

The probability of being identified becomes less than or equal to one third.

CONCLUSION

The *l-diversity* approach can compensate for the weaknesses of *k-anonymity* to minimize the risk of re-identification of *k*-anonymized clinical data. Typically, when the sensitive variables have little diversity value, *l-diversity* model will be a defense. The risk of re-identification assessment not only evaluate risk but also it will help statistical programmer to find the best way to anonymize clinical data, specially, for the aggregated variables.

REFERENCES

V. Ciriani., S. De Capitani di Vimercati., S. Foresti., P. Samarati. 2007. “k-Anonymity”.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam. 2007. “l-Diversity: Privacy Beyond k-Anonymity”.

PHUSE Data Transparency Working Group. “Data Anonymisation and Risk Assessment Automation”. 09 June 2020. Available at Data+Anonymisation+and+Risk+Assessment+Automation.pdf (phuse.s3.eu-central-1.amazonaws.com)

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Lanlan Peng
Enterprise: Sanofi
Phone: (+86)15178819923
E-mail: lanlan.peng92@gmail.com