Anonymizing Vaccine Records

Pranav Srinivasan

JHU APL

Background

- Medical patient data is required for many research applications, but said patient data often contains large amounts of confidential personally identifiable information
- Current anonymization methods aim to solve this through a process of suppression and generalization of identifiers and quasi-identifiers
- This project aims to generate a fictional yet semi-realistic COVID-19 vaccine record data set and anonymize said dataset using the popular anonymization method *k*-anonymity

k-anonymity

- ***k*-anonymity:** A method of anonymization where information for one entry in a dataset is identical to at least *k*-1 other entries in that field, where *k* is a variable

Below is an example of a 4-anonymized dataset

Anonymous DB

id	Zipcode	age	nationality	disease
1	13053	28	Russia	Cardiac disease
2	13068	29	US	Cardiac disease
3	13068	21	Japan	Infectious dis.
4	13053	23	US	Infectious dis.
5	14853	50	India	Cancer
6	14853	55	Russia	Cardiac disease
7	14850	47	US	Infectious dis.
8	14850	49	US	Infectious dis.
9	13053	31	US	Cancer
10	13053	37	India	Cancer
11	13068	36	Japan	Cancer
12	13068	35	US	Cancer

4-anonymity DB

id	Zipcode	age	nationality	disease
1	130**	<30	*	Cardiac disease
2	130**	<30	*	Cardiac disease
3	130**	<30	*	Infectious dis.
4	130**	<30	*	Infectious dis.
5	1485*	≥40	*	Cancer
6	1485*	≥40	*	Cardiac disease
7	1485*	≥40	*	Infectious dis.
8	1485*	≥40	*	Infectious dis.
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

l-diversity and *t*-closeness

- ***l*-diversity:** Property of anonymized data that occurs if each *q**-block in a dataset contains at least *l* distinct values for a sensitive attribute

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

- ***t*-closeness:** Property of anonymized data that occurs if the distance between the distribution of a sensitive attribute in a class and the distribution of that attribute across the entire data set is no higher than a threshold *t*

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

COVID Vaccine Record Generation

- Generated COVID-19 vaccine record data using Python
- Data generated includes 34 different fields of information, all included in the Vaccine Administration Management System (VAMS) used for storing vaccine record information
- Information in dataset was created to resemble a real patient record

Anonymization

- Anonymization performed through Python
- Modified Nuclearstar's *k*-anonymity script to suit and work effectively with my generated dataset
- Implements *k*-anonymity, *l*-diversity, and *t*-closeness
- Followed HIPAA Privacy Rule's Safe Harbor Method to suppress 18 different identifiers and quasi-identifiers to protect privacy
- *k*=5 for *k*-anonymity for optimal balance between information loss and privacy protection

Conclusion

- Anonymization of publicly released medical data is incredibly important in order to protect patient privacy and anonymity
- *k*-anonymity can be used to effectively anonymize data for a variety of use cases
- Although different use cases of data may require different levels of anonymization and different types and amounts of information, *k*-anonymity is generally an effective way of properly anonymizing data to protect patient privacy

AOS/QKT

W Bruce Cottle