Singapore Healthcare Management 2019

Data De-Identification Support System

Sudha Harikrishnan, SingHealth, HSRC
Lam Shao Wei , SingHealth, HSRC

On Behalf of the SingHealth Data Deidentification Workgroup

SingHealth
Defining Tomorrow's Medicine

Background

With the implementation of Human Biomedical Research Act (HBRA) (2017), the researchers who conduct HBR were to de-identify the data if the patient consent for the study is not obtained or the informed consent doesn't meet the relevant requirement specification. Thus, the SingHealth HSRC(Health Science Research Center) was tasked to develop a decision support system to de-identify the direct identifiers and the residual risks associated with the data in compliance with HBRA and existing policies and processes in a healthcare industry

Aim

To provide the seamless, one-stop platform for the end user, in order to complete the data de-identification process with the maximum automation.

Methodology

A synthetic dataset that is representative of medical data was used for this project. Electronic medical data mostly contains personal identifiers, quasi identifiers and sensitive data. The confidentiality and sensitivity of the data is highly correlated with the data type. Based on the data type, de-identification techniques like cryptographic hashing, masking, generalization, suppression, K-anonymity, etc. were applied to render the data non-identifiable before it is realized to researchers and collaborators.

Project Design

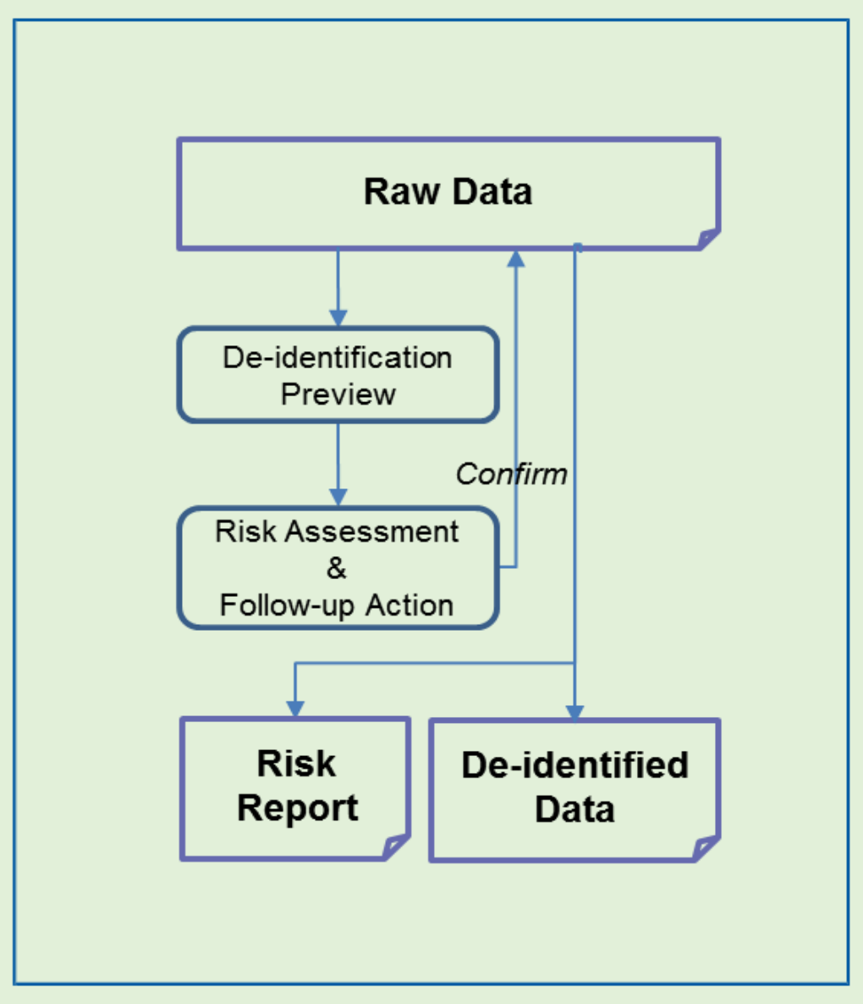
Privacy Property	k-anonymity l-diversity
Cryptography	Encryption Masking
Data Process	Automated De-identification Quasi-identifier Risk Assessment

Data Process

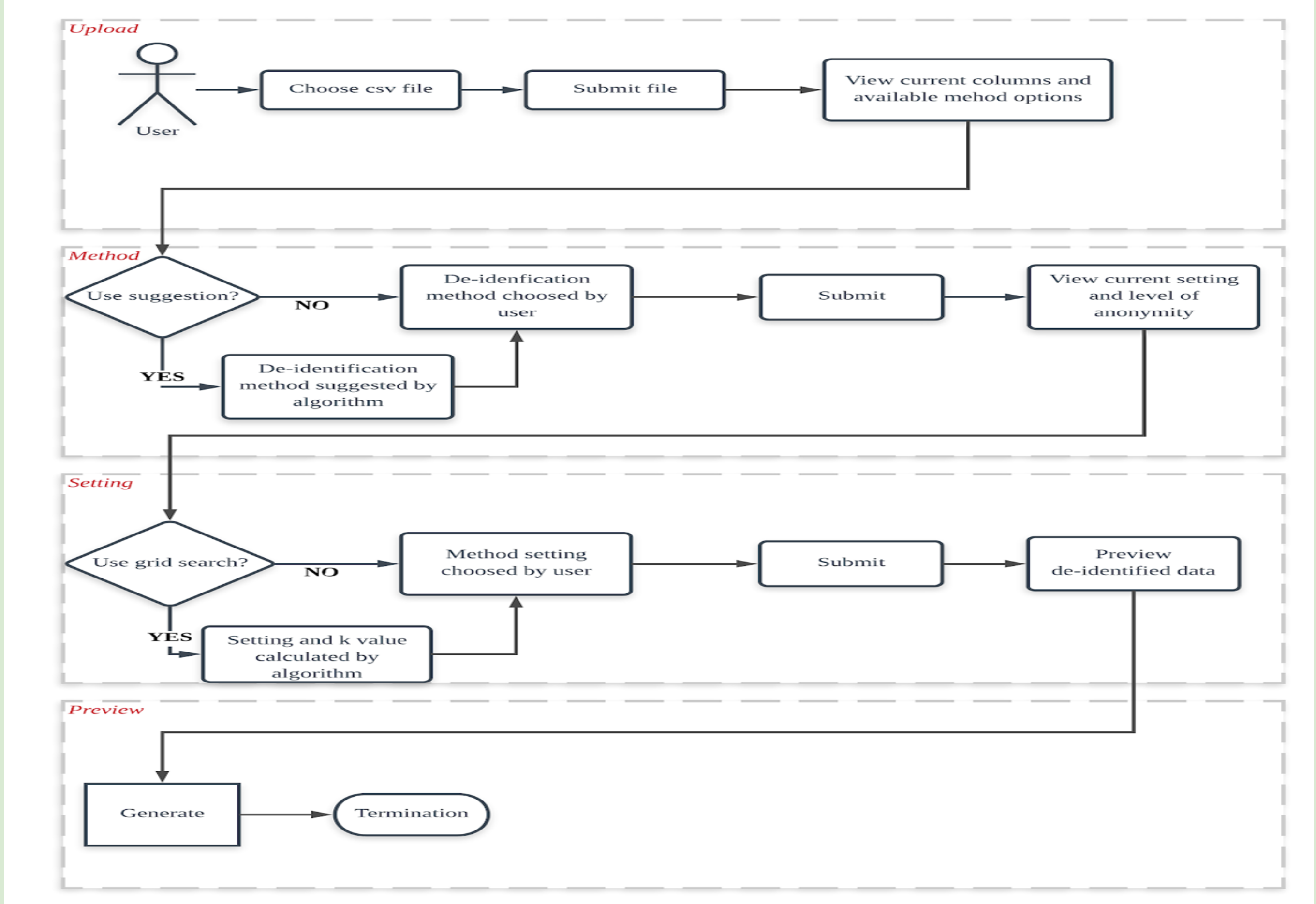
Key Components

1. **Automated De-identification**
   - Suggest De-identification method based on field type
   - Generate de-identification result file in the process ending
2. **Quasi-identifier Risk Assessment**
   - Masking & Grouping on Quasi-identifier
   - Analysis on Privacy Risk
3. **Decision Supporting Report**
   - Generate standardized privacy risk report
   - Record for future re-identification and digital watermark

Process Flow

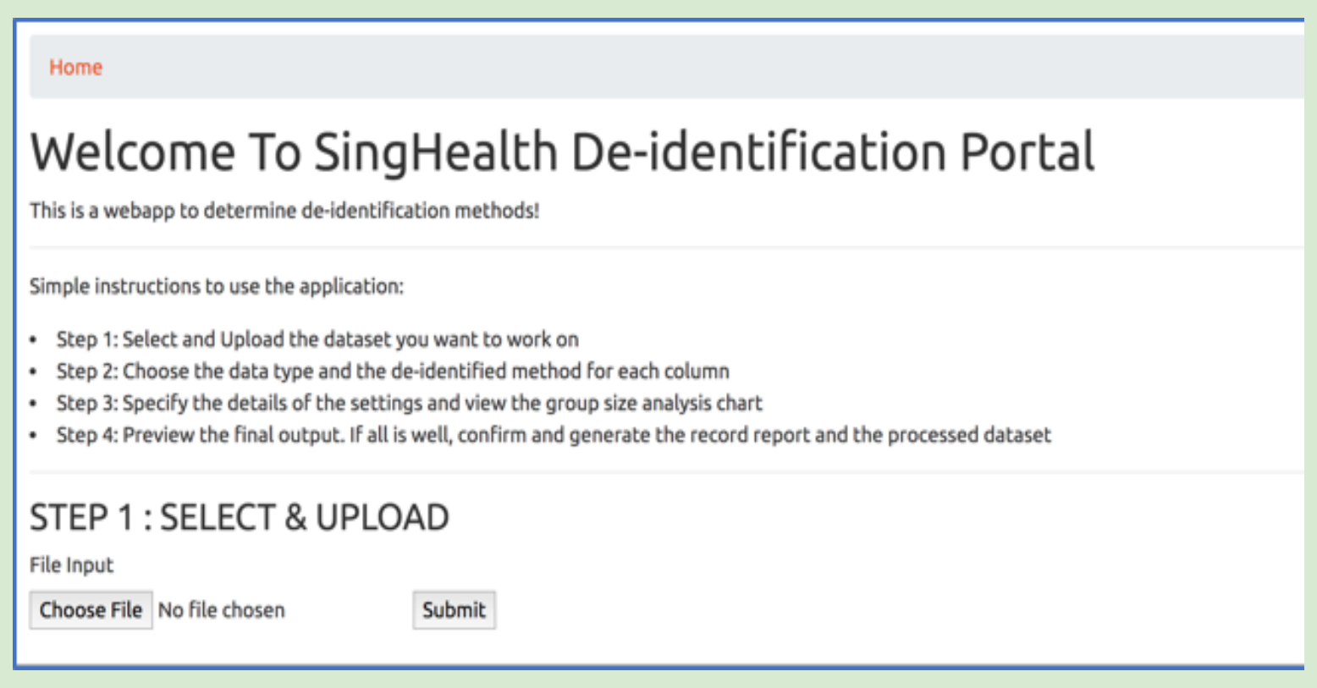


System Illustration


Upload
User
Choose csv file
Submit file
View current columns and available mehod options
Method
Use suggestion?
NO
YES
De-idenfication method choosed by user
De-identification method suggested by algorithm
Submit
View current setting and level of anonymity
Setting
Use grid search?
NO
YES
Method setting choosed by user
Setting and k value calculated by algorithm
Submit
Preview de-identified data
Preview
Generate
Termination

It is clearly planned by guiding user from step by step actions. During total 4 phases of processes, user will receive real-time feedback from each step's updates, and help for following steps.

System Preview



Welcome To SingHealth De-identification Portal

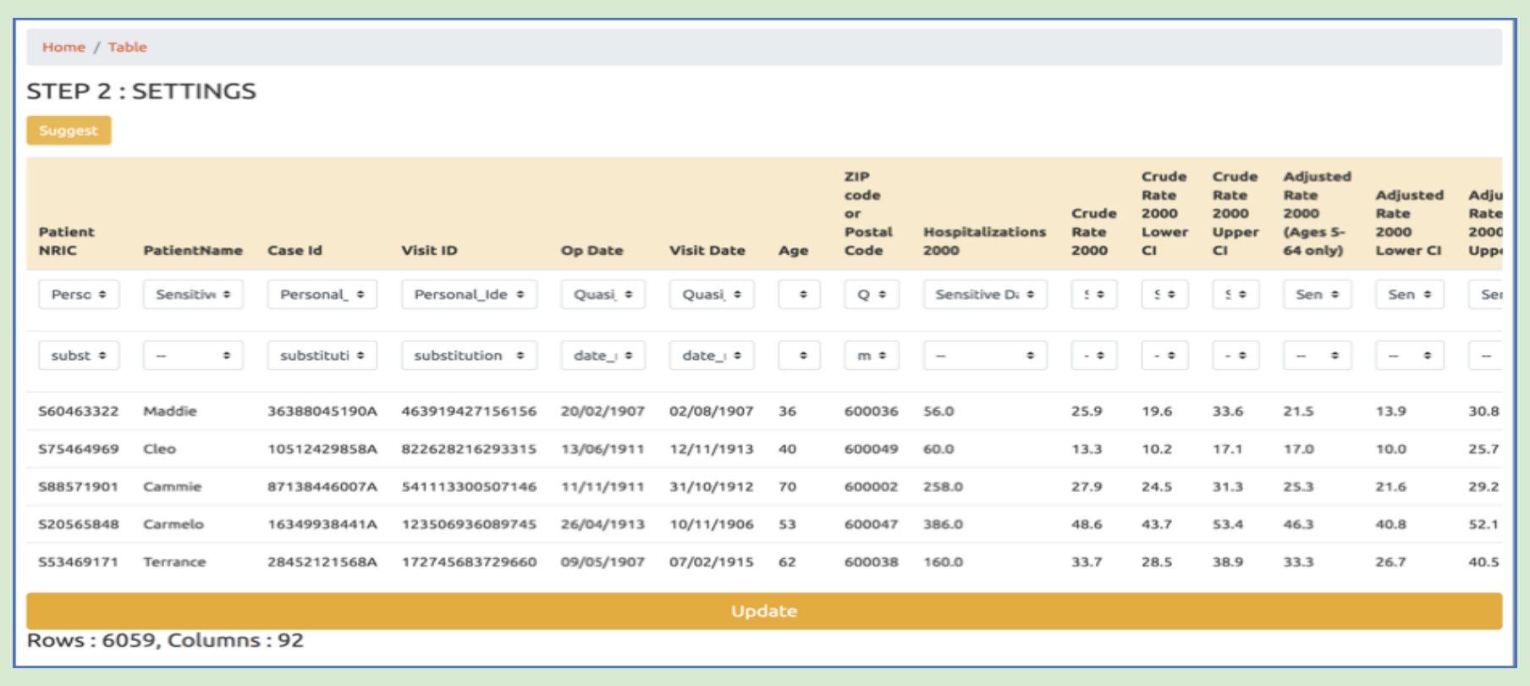
This is a webapp to determine de-identification methods!

Simple instructions to use the application:

- Step 1: Select and Upload the dataset you want to work on
- Step 2: Choose the data type and the de-identified method for each column
- Step 3: Specify the details of the settings and view the group size analysis chart
- Step 4: Preview the final output. If all is well, confirm and generate the record report and the processed dataset

STEP 1 : SELECT & UPLOAD

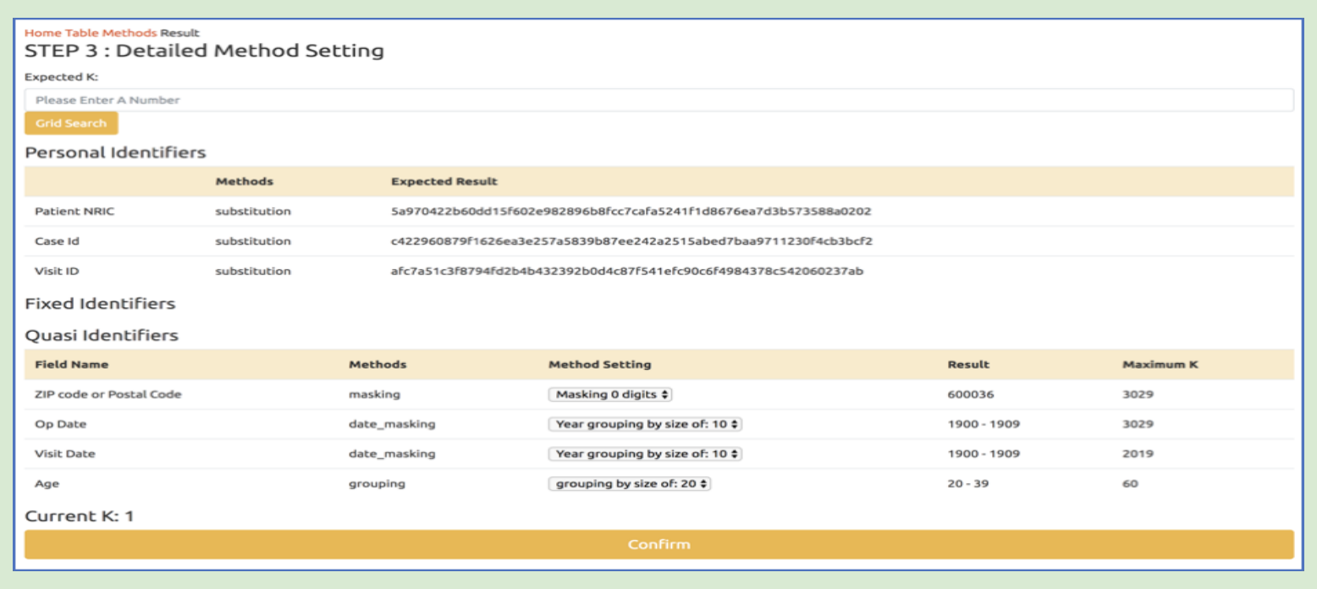
File Input



STEP 2 : SETTINGS

Patient NRIC	PatientName	Case Id	Visit ID	Op Date	Visit Date	Age	ZIP code or Postal Code	Hospitalizations 2000	Crude Rate 2000	Crude Rate 2000 Lower CI	Crude Rate 2000 Upper CI	Adjusted Rate 2000 (Ages 5-64 only)	Adjusted Rate 2000 Lower CI	Adju Rate 200(Upp
Persc	Sensitiv	Personal_	Personal_Ide	Quasi_	Quasi_		Q	Sensitive D:	!	!	!	Sen	Sen	Ser
subst	–	substituti	substitution	date_	date_		m	–	-	-	-	–	–	–
S60463322	Maddie	36388045190A	463919427156156	20/02/1907	02/08/1907	36	600036	56.0	25.9	19.6	33.6	21.5	13.9	30.8
S75464969	Cleo	10512429858A	822628216293315	13/06/1911	12/11/1913	40	600049	60.0	13.3	10.2	17.1	17.0	10.0	25.7
S88571901	Cammie	87138446007A	541113300507146	11/11/1911	31/10/1912	70	600002	258.0	27.9	24.5	31.3	25.3	21.6	29.2
S20565848	Carmelo	16349938441A	123506936089745	26/04/1913	10/11/1906	53	600047	386.0	48.6	43.7	53.4	46.3	40.8	52.1
S53469171	Terrance	28452121568A	172745683729660	09/05/1907	07/02/1915	62	600038	160.0	33.7	28.5	38.9	33.3	26.7	40.5

Rows : 6059, Columns : 92



STEP 3 : Detailed Method Setting

Expected K:

Please Enter A Number

Personal Identifiers

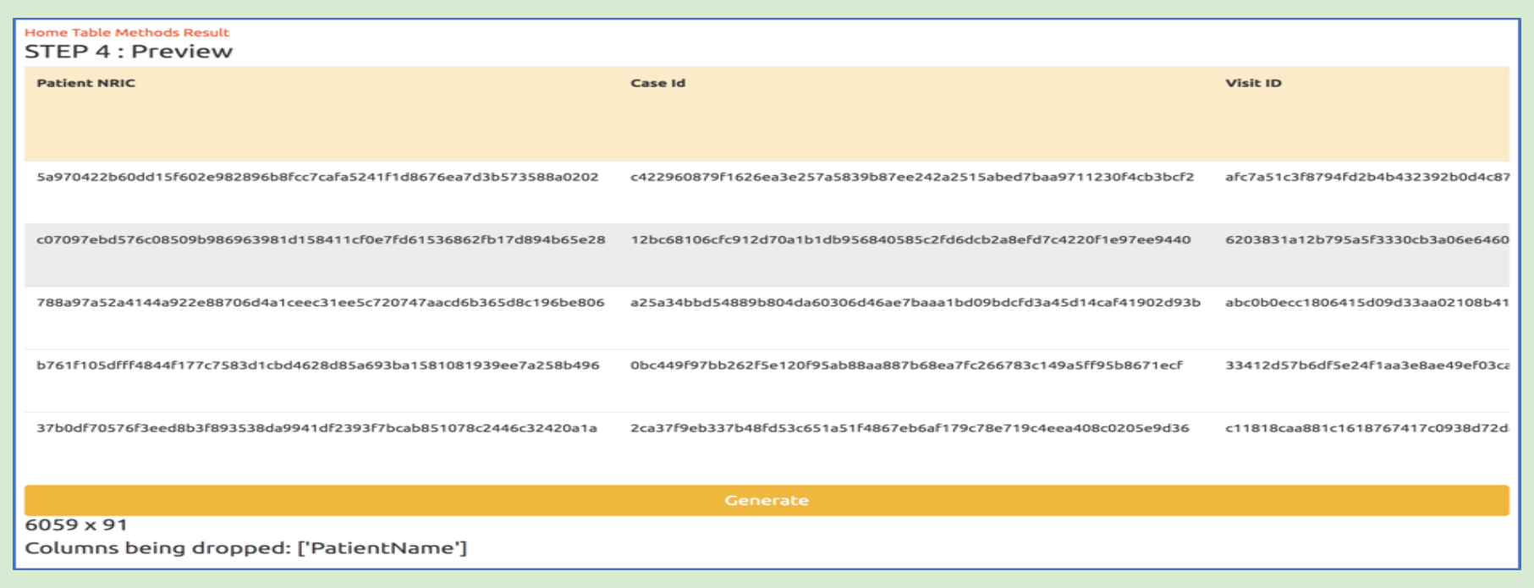
	Methods	Expected Result
Patient NRIC	substitution	5a970422b60dd15f602e982896b8fcc7cafa5241f1d8676ea7d3b573588a0202
Case Id	substitution	c422960879f1626ea3e257a5839b87ee242a2515abed7baa9711230f4cb3bcf2
Visit ID	substitution	afc7a51c3f8794fd2b4b432392b0d4c87f541efc90c6f4984378c542060237ab

Fixed Identifiers

Quasi Identifiers

Field Name	Methods	Method Setting	Result	Maximum K
ZIP code or Postal Code	masking	Masking 0 digits	600036	3029
Op Date	date_masking	Year grouping by size of: 10	1900 - 1909	3029
Visit Date	date_masking	Year grouping by size of: 10	1900 - 1909	2019
Age	grouping	grouping by size of: 20	20 - 39	60

Current K: 1



STEP 4 : Preview

Patient NRIC	Case Id	Visit ID
5a970422b60dd15f602e982896b8fcc7cafa5241f1d8676ea7d3b573588a0202	c422960879f1626ea3e257a5839b87ee242a2515abed7baa9711230f4cb3bcf2	afc7a51c3f8794fd2b4b432392b0d4c87
c07097ebd576c08509b986963981d158411cf0e7fd61536862fb17d894b65e28	12bc68106cfc912d70a1b1db956840585c2fd6dcb2a8efd7c4220f1e97ee9440	6203831a12b795a5f3330cb3a06e6460
788a97a52a4144a922e88706d4a1ceec31ee5c720747aacd6b365d8c196be806	a25a34bbd54889b804da60306d46ae7baaa1bd09bdcfd3a45d14caf41902d93b	abc0b0ecc1806415d09d33aa02108b41
b761f105dfff4844f177c7583d1cbd4628d85a693ba1581081939ee7a258b496	0bc449f97bb262f5e120f95ab88aa887b68ea7fc266783c149a5ff95b8671ecf	33412d57b6df5e24f1aa3e8ae49ef03ca
37b0df70576f3eed8b3f893538da9941df2393f7bcab851078c2446c32420a1a	2ca37f9eb337b48fd53c651a51f4867eb6af179c78e719c4eea408c0205e9d36	c11818caa881c1618767417c0938d72d

6059 x 91
Columns being dropped: ['PatientName']

Conclusion

In this project, we conceptualize, decide and implement a data de-identification system for Singhealth. The system aims to provide a seamless solution to current data requesting process. With the goal in mind, we develop a Python-based solution for data de-identification practice, which could achieve:

1) Field type and method auto suggestions
2) Risk Assessment using K Anonymity
3) Grid Search for partial optimal solution
4) Report Generating

We build our UI under web-based flask framework, to serve a better flexibility for our client. The solution solves most of the current pain point in the data requesting process, and could help the healthcare to both protect users privacy and enhance data-driven research capacity.

Acknowledgement: This project was supported by the NUS- School of Computing. Special thanks to He YingXu, Sheng Yu, Xiao ZuoLing, Yu ZongDong.