Streamlining Statistical Disclosure Control Methods for Health Data Research Output

Graeme Diack
MSc in Mathematics and Data Science

UNIVERSITY of STIRLING

Health Data – Balancing Confidentiality with the Benefits of Research

The use of health data in research leads to insights that benefit society. Protecting the identity of the subjects involved is critical in maintaining public trust. The field of Statistical Disclosure Control provides a framework for identifying disclosure risks.

SDC In Action

Simple anonymisation will remove identifiers from a dataset. However, this can be easily overcome by a motivated intruder. The table at the bottom is a toy 'pseudonymised' individual level health dataset. The table on the right represents publicly available data, information found on social media for example.

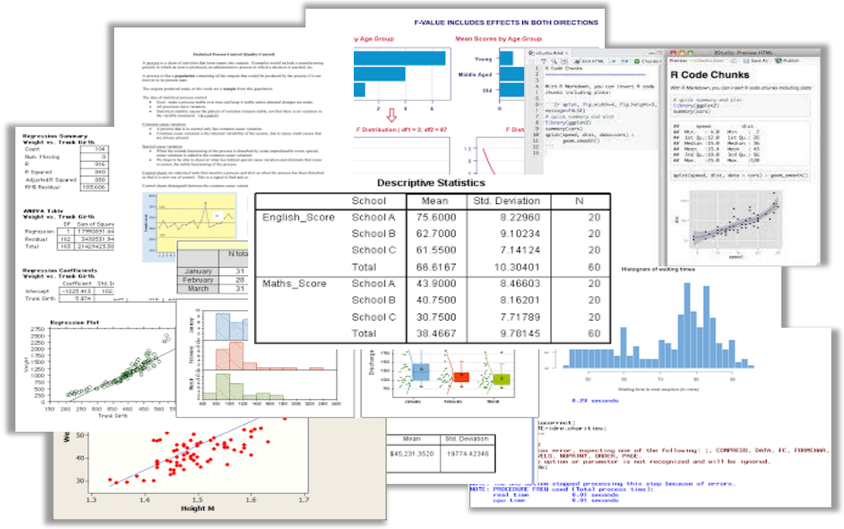
With these two datasets it is possible for an intruder to re-identify subjects 14 and 22, expose a group of three (10, 24, 28) all with 'Ailment 2', plus subjects 04 and 20 can identify each other.

Data Environment			
Name	Sex	Occupation	Smoker
Jane Doe	F	Engineer	n
Jyn Erso	F	Engineer	n
Rebecca Buck	F	Engineer	n
Lydia Swanson	F	Scientist	n
Lois Lane	F	Scientist	n
Alicia Ferguson	F	Scientist	n
Jessica Jones	F	Super Hero	y
Arthur Dent	M	Engineer	n
Bilbo Baggins	M	Engineer	n
Han Solo	M	Engineer	n
Joe Bloggs	M	Scientist	n
Ford Prefect	M	Scientist	n
Clark Kent	M	Scientist	y

Attributes					Sensitive Health Data		
ID	Age	Sex	Occupation	Smoker	Ailment 1	Ailment 2	Ailment 3
06	38	F	Engineer	n	n	n	n
16	19	F	Engineer	n	y	y	y
26	46	F	Engineer	n	n	n	y
10	40	F	Scientist	n	y	y	y
24	37	F	Scientist	n	n	y	y
28	29	F	Scientist	n	y	y	n
14	45	F	Super Hero	y	y	n	n
08	26	M	Engineer	n	y	n	n
12	20	M	Engineer	n	y	n	n
18	43	M	Engineer	n	n	n	n
04	57	M	Scientist	n	n	y	y
20	50	M	Scientist	n	n	n	n
22	49	M	Scientist	n	y	y	n

The Problem

The problem lies in the time it takes to apply disclosure controls to the variety of output produced by researchers.



SDC Agents, such as those based in the eDRIS team, spend a considerable amount of time appraising the outputs of projects in order to ensure they are non-disclosive. This task covers the lifetime of the project, where outputs can number into the hundreds. Their workflow software shows that each output can take on average half a day to appraise.

Carried out in collaboration with the electronic Data Research and Innovation Service (eDRIS)

Streamlining Opportunity: Detecting Differencing Attack Potential

A disclosure risk can be created if a number of tables are released based on the same data but with slightly different variable breakdowns. This is referred to as "Differencing", where one table subtracted from another reveals small counts.

Variable Breakdown Tracker																				
Unit Breakdown		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Breakdown	Tables in Use	Year of Birth																			
1	Table1	1971-1980										1981-1990									

One table created with Year of Birth broken down by decade

Variable Breakdown Tracker																				
Unit Breakdown		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Breakdown	Tables in Use	Year of Birth																			
1	Table1, Table4	1971-1980										1981-1990									
2	Table2, Table3	1971-1974				1975-1978				1979-1982				1983-1986				1987-1990			

A new set of tables follow, some with a new breakdown of four year intervals, introducing a risk of differencing across the four tables

Variable Breakdown Tracker																				
Unit Breakdown		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Breakdown	Tables in Use	Year of Birth																			
1	Table1, Table4	1971-1980										1981-1990									
2	Table2, Table3	1971-1974				1975-1978				1979-1982				1983-1986				1987-1990			
3	Table5	1971-1975					1976-1980					1981-1985					1986-1990				

A further fifth table is created with the variable broken down by five year intervals, adding a higher risk of differencing across tables 2, 3 and 5

Keeping track of variable breakdowns, as demonstrated in the table series above, is an aid to detecting this scenario. There are three breakdowns of the variable Year of Birth, creating potential differencing attacks on the tables that use them.

Solution

A tool that keeps track of this data and alerts the user (Researcher or SDC Agent) to the risk of creating a new breakdown could achieve a reasonable time saving.

For further information contact: Graeme Diack
gad00033@students.stir.ac.uk