

Division of Computing Science and Mathematics
Faculty of Natural Sciences
University of Stirling

Streamlining Statistical Disclosure Control Methods for
Health Data Research Output

Graeme Diack

**Dissertation submitted in partial fulfilment for the degree of
Master of Science in Mathematics and Data Science**

September 2019

Abstract

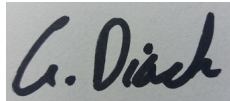
The National Health Service in Scotland collects a wealth of high quality health and administrative data that is ripe for use in research. A national network of *Safehavens* exist that facilitates access to the data for utilisation by research institutes, however a balance has to be struck between the release of informative and progressive research, and the rights of the data subjects to their privacy. *Statistical Disclosure Control* encapsulates the methods by which this balance is measured and handled. *Disclosure Checking* of research output is one of these methods, and is carried out by members of the *electronic Data Research and Innovation Service (eDRIS)*, a function within *National Services Scotland*, and part of NHS Scotland. This disclosure checking is a large time burden for eDRIS, and it was considered by the team leaders that there could be potential for streamlining the process via some method, either existing in the contemporary publications on the subject, or by invoking a creative solution based on these publications and the processes already in place within the team. Following a review of the existing material on Statistical Disclosure Control practices, it is clear that an incredible amount of work has gone into conceptual breakthroughs such as framing outputs within how-to guides[11], nurturing a safe and positive environment via the *Five Safes*[30] framework, and providing training[32] that informs all stakeholders of the risks and their responsibilities. However, the topic is burdened with complexities that confound the application of a standard format that could be captured by automation tools in the traditional sense. The achievements of this project are therefore a set of suggested *Administrative* changes that might be made to the existing processes within eDRIS, and an *R Package* that can be used to highlight one particular attack vector for disclosure from tabular data, that of *Differencing*, the combination of which may lead to some time savings for the team.

Attestation

I understand the nature of plagiarism, and I am aware of the University's policy on this.

I certify that this dissertation reports original work by me during my University project.

Signature:

A rectangular box containing a handwritten signature in black ink. The signature appears to be 'G. Diach' written in a cursive style.

Date: 20/09/2019

Acknowledgements

I would like to extend my thanks to many persons for their role in this project.

My partner for her guidance in the early months and patience in these closing months, without which I would not have been able to complete this degree.

Mrs. Jackie Caldwell, as my line manager during the placement but also a valued mentor, and Dr. David Bailey as my industrial supervisor, mentor and R Guru! The whole team at eDRIS for being so welcoming.

Staff at the University of Stirling, Dr. Andy Hoyle as course coordinator always ready to alleviate any worries I had about the course, Ms. Kate Howie for her excellent module on Statistics and for agreeing to be my academic supervisor for this project.

Datalab and MBN not only for providing my funding but also for their excellent workshops, networking events and social events.

Mr. Ahmed Mahmoud for allowing use of his PhD tabular output as a case study for the *R Package*.

Ms. Amy Tilbrook for helping me hit the ground running with lots of links, data and advice laying in wait on the network folders, Mr. Suhail Iqbal of the ADRC for his guidance, and Mr. Clifford Nangle of the Usher Institute for valuable help with my poster design.

Finally, all those who gave me their time to discuss their experiences with SDC: Dr. Richmond Davies of National Services Scotland NHS, Dr. Michael Fleming of the University of Glasgow, Dr. Matthew Iveson, Professor Chris Dibben, Dr. Lee Williamson, all of the University of Edinburgh, and Research Officer Shayla Leib at the Office of National Statistics.

A special mention for Dr. Nancy Burns, National Records Scotland, for the discussion and also the introduction to tracking differencing through a spreadsheet which was a key to the development of the *R Package* within this work.

This dissertation was produced in L^AT_EX, via Overleaf [26].

Contents

Abstract	i
Attestation	ii
Acknowledgements	iii
1 Introduction	1
1.1 Background and Context	1
1.2 Scope and Objectives	2
1.3 Achievements	2
1.4 Overview of Dissertation	2
2 Statistical Disclosure Control	3
2.1 A Long History	3
2.2 Input SDC	4
2.3 Output SDC	5
2.4 Statistical Disclosures	6
2.4.1 Data Linkage	6
2.4.2 Intruders	7
2.4.3 The Data Environment	7
2.4.4 Output Types	8
2.4.5 Attack Scenarios	8
2.4.6 Resolution	10
3 Existing Methodologies	11
3.1 Applying Controls	11
3.2 Framework Solutions	13
4 Proposed Solutions for eDRIS	15
4.1 Administrative Solutions	15
4.1.1 Three Potential Time Savers	15
4.1.2 Synthetic Data	16
4.2 Programmatic Solutions	20
4.2.1 Standardising Output	20
4.2.2 Machine Learning	20
4.2.3 Differencing Tool	20
5 R Package Development	22

5.1	High-Level Tool Design	22
5.2	Input Metadata Development	23
5.2.1	Output Data	23
5.2.2	Linked Data	24
5.3	Input Processing Development	25
5.3.1	Data Storage	25
5.3.2	Import Function	25
5.4	Output, Reports and Visualisation	26
6	A Case Study	28
6.1	Preparation	28
6.2	User: Analyst	30
6.3	User: Researcher	31
6.4	User: Research Coordinator	32
6.4.1	Disclosure Requests	32
6.4.2	Features Created Post-Testing	37
7	Conclusion	40
7.1	Summary	40
7.2	Evaluation	40
7.3	Future Work	41
	Appendix 1 – R Package Creation Notes	45
	Appendix 2 – User guide	46
	Appendix 3 – Installation guide	49

List of Figures

2.1	Google n-gram Viewer Search for ‘Disclosure’ followed by ‘Control’, ‘Analysis’ and ‘Limitation’	4
2.2	Risk-Utility Diagram	5
4.1	Density plots of numerical ‘Iris’ attributes from original and 3 synthetic datasets created with default options. Blue = Original, Red = Synthetic	19
4.2	Density plots of numerical attributes split by Species, original and 3 synthetic datasets created with default options. Blue = Original, Red = Synthetic	19
5.1	Differencing Tool User Journey	22
5.2	Simple Network of Tables (Nodes) and Variables (Edges)	26
6.1	Sample of Case Study Tables Metadata	29
6.2	Fictitious Dataset Based on Case Study	29
6.3	Dataset Metadata	30
6.4	<code>add_newdataset</code> function and effect	31
6.5	<code>add_newtable</code> function and effect	32
6.6	Console Output Report - First 7 tables	33
6.7	Network Representation - First 7 Tables	33
6.8	Network Specific to table 7	34
6.9	Variables Risk Report	34
6.10	Full 28 Table Console Risk Report	35
6.11	Full 28 Table Network	35
6.12	Table 5 Specific Network	36
6.13	Table 15 Specific Network	36
6.14	Include Param - <code>report_tablesRisk(include=c(‘Ethnic Group’,‘overall_cat’))</code>	38
6.15	Exclude Param - <code>report_tablesRisk(exclude=c(‘Frequency’,‘simd_quintile’))</code>	38
6.16	Exclude param - Default Colours	39
6.17	Exclude param - Custom Colours (‘red’,‘orange’,‘yellow’,‘green’,‘blue’)	39

1 Introduction

1.1 Background and Context

The National Health Service in Scotland has been collecting patient and administrative data since the 1960s. The data is of very high quality not only due to continual improvements to the collection process, but also largely because early on in the history of this data collection (circa 1968) a decision was made to store it in a machine readable format [15]. This foresight has led to Scotland's health service being one of the most data rich in the world [27], and an incredibly valuable resource to support health research studies. The Information Services Division (ISD, note: formerly known as Information and Statistics Division), part of NHS Scotland, retains records on over 5 million people, many of which cover an individuals entire life span from pre-birth antenatal care of their mother through to the recording of their death [28]. The collection and storage of this data brings with it the responsibility of ensuring it is used in an appropriate manner, in accordance not only with the most up to date data protection laws, but also with society's expectations of how their personal and sensitive information is treated. In Scotland, the infrastructure is in place that is designed to encourage the utilisation of the data whilst maintaining its confidentiality. This infrastructure is a national network of *Safehavens* that researchers can use to access the data via secure workstations. Built alongside these Safehaven environments is the procedural framework for gaining access to data under ethical and lawful best practice. This framework establishes the reasons for the research, what the exact data set or sets that are required and how the research can be shown to be of benefit to the public, before any data is released. The goal is to create a safe environment for researchers to access the data and carry out critically important research, while ensuring the confidentiality of that data, and minimising the risk of any data breach occurring. This environment consists of two primary teams, the Information Governance and Public Benefit and Privacy Panel (IG & PBPP), and the electronic Data Research and Innovation Service (eDRIS).

"Researchers desiring to work with Scottish Health data are supported by the eDRIS team. The eDRIS team offer guidance and assistance on submitting a research application. These applications are then considered for approval by PBPP, who assess the balance between the public benefit of the research and its privacy risk. If the PBPP panel approve the project then the eDRIS team assemble the required, anonymised, linked dataset. Once the researchers have completed the required IG training and signed the appropriate agreements, the data are made available to them through the National Safehaven." - David Bailey, Senior Analyst at eDRIS.

A key focus of the support structure is that of maintaining the trust of the public, the subjects of which the data concerns. If the public do not feel that their health data are being treated fairly and responsibly this could ultimately lead to radical reductions in what data is available for study. Due to the nature of the data collection, it is considered as 'unconsented', and therefore falls under certain laws regarding its use. In accordance with these laws, steps are taken to anonymise the data before any of it is released to researchers. However, anonymised data can be easily re-identified given the right circumstances. Any results that researchers would like to have released to the public must be reviewed for risks of 'Statistical Disclosures', that is disclosures of any kind that could lead to sensitive information about individuals being revealed [16]. The practice of checking for and eliminating these disclosure risks is termed *Statistical Disclosure Control (SDC)*, and it must be

applied to all outputs created from data held in the Safehaven. It is ultimately the responsibility of the *Data Controller*, those who initially collect the data, to ensure no confidentiality breaches occur via the data they curate. At the moment this function is carried out by members of the eDRIS team, and is termed internally as *Disclosure Checking*. The team do not apply any controls themselves, merely highlight the risks to the researchers and in some cases advise possible solutions. As this is a critical point at which decisions are made on whether to release data to the researcher or not, considerable time is spent examining the output to ensure it is safe.

1.2 Scope and Objectives

The aim of this project is to examine research in the field of Statistical Disclosure Control and investigate ways to reduce the time burden on the eDRIS team, with a focus on the aspect of *Disclosure Checking* the output that is to be taken out of the safe environment and used in publications.

1.3 Achievements

The main result of this project is a new R package that eDRIS will use to keep track of tabular data within a study, assisting the detection of potential disclosure via the method of *Differencing*. The tool has been written as an *R Package* because *R* is already an established tool within the team and it could be integrated easily with existing practices. Further achievements come in the form of recommendations of administrative changes that could reduce the burden of disclosure checks. The application of *Synthetic Data* could be introduced to the researchers, with a view to only releasing outputs based on this data instead of the original. The placing of more trust and responsibility on the researchers themselves by establishing *Dual Sign-off* practices. The introduction of a new pricing structure that is based on the quantity of outputs being produced during a study. A slight change in the methods used to transfer output to be checked between the Safehaven workstations and the eDRIS secure folders.

1.4 Overview of Dissertation

The following chapters will contain an overview of SDC from a historical perspective, establishing a context of how the field has grown from the simple controlling of small cell counts in tabular releases, to a fully fledged area of research. The concept of *Output Statistical Disclosure Control (OSDC)* will then be focused on, with an investigation of the methods currently being recommended as best practices. The final chapters will focus on the findings of the project, with detailed descriptions of the administrative recommendations for eDRIS and of the resulting *R Package* that has been created.

2 Statistical Disclosure Control

This term is generally used in a way that is actually the encapsulation of three distinct concepts, centred on the protection of sensitive data. The first is that of the disclosure itself. A *Statistical Disclosure* is a data confidentiality breach which required a level of statistical analysis to reveal. It encompasses not just individual re-identification, but also the ‘idea that confidential information is revealed’ [16]. The second is the practice of appraising data for the risk of it leading to the first, before that data is published. In some sense also evaluating the level of risk, as opposed to a binary measure of *risk vs no risk*. The third, and in fact the true meaning of the full term *SDC*, is the techniques that can be employed to reduce this risk whilst still maintaining a certain level of data utility. The overall meaning is that *SDC* is the practice of minimising the risk of disclosing new information about entities, such as individual people, groups of people, or organisations, via data that is expected to enter into the public domain. It is in fact the second of these meanings that is the focus of our problem and solution, and is primarily referred to as *Disclosure Checks* throughout this work, however it may be simply referred to as *SDC* if the context allows.

2.1 A Long History

National Statistical Institutes (*NSIs*) are tasked with the objective of gathering information and disseminating it in order to supply society with detailed and informative statistical outputs. They are also trusted to ensure the confidentiality of the underlying data is maintained in these outputs. The concept of disclosure control grew out of the balance of this goal against the management of data with increasing detail that was becoming available through improved gathering and storage methods. The creation of Statistical Institutes dates back to the mid 1800’s [18], however the earliest references to *SDC* practice, along with alternate terms ‘Disclosure Analysis’ and ‘Disclosure Limitation’, appear in earnest from around the mid to late 1960’s, with the earliest reference [29] (figure 2.1) coming from a 1954 publication by the US Statistical Bureau, within which it is clear that the concept is already well established [14]:

”In the field of industrial and business statistics, for example, employment data are shown only when three or more companies are included in a statistical total. For value figures, the rules are even more protective: Regardless of the number of companies involved, value data are withheld if one or two companies account for such a large proportion of the total that publication would be tantamount to disclosure.”

We can see that some of the core rules (described later 2.4) used in modern *SDC*, that of *Minimum Cell Counts* and *Domination*, are in use already. These early accounts primarily focus on the *SDC* techniques of *Cell Suppression (CS)* and *Categorical Aggregation* to deal with small cell counts in tabular outputs. Advances within the field steadily became more sophisticated throughout the 60’s and 70’s, and into the 80’s blossomed almost hand in hand with ‘Big Data’ as both concepts were freed by the falling price of compute and storage, and of course the world wide web [5][8]. Disclosure controls can be concerned with input data and output data, sometimes referred to as *Pre-Tabular* and *Post-Tabular*, respectively, with the two problems being quite distinct from each other.

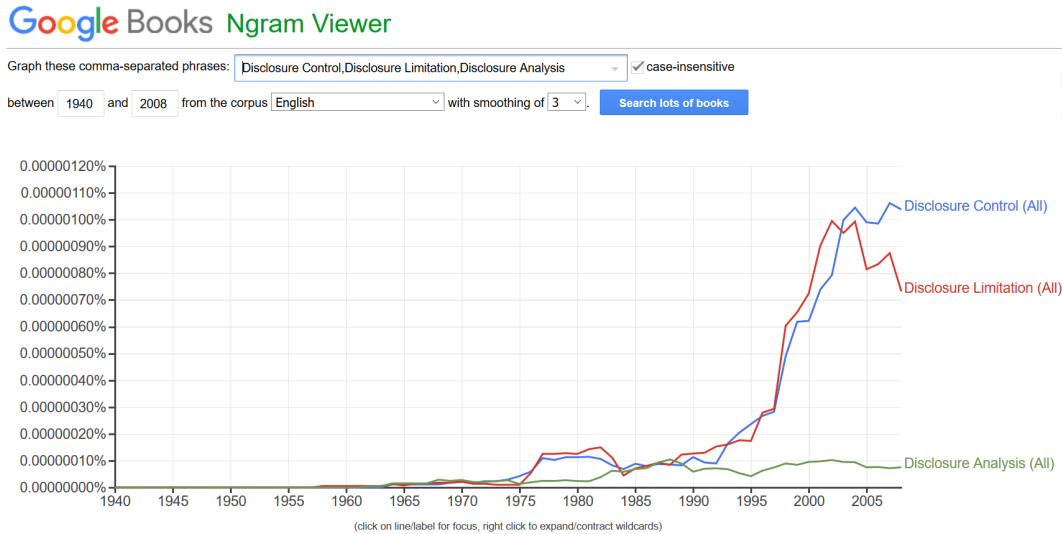


Figure 2.1: Google n-gram Viewer Search for ‘Disclosure’ followed by ‘Control’, ‘Analysis’ and ‘Limitation’

2.2 Input SDC

Applying controls to the input data is relatively well understood. A dataset that contains individual level observations, whether people or organisations, is generally referred to as *microData*. The data handled by ISD is *microData*, records of patients contact with NHS services throughout Scotland on an individual level. The most obvious control that can be applied to this format is *Anonymisation*. Simple anonymisation of data can be carried out by removing attributes that could be used to directly identify an individual in a dataset. Intuitively, attributes such as names, national ID numbers, and addresses form the bulk of these identifiers. A more sophisticated breakdown of attributes that could form direct identifiers are described by Professor Mark Elliott and colleagues [10], categorising them into a number of types such as *Unique*, covering national ID numbers among others, *Associative*, covering telephone numbers or car license plates, to *Social* which is where the traditional names and addresses find themselves. They allude to the fact that these traditional identifiers are actually not the most identifiable due to their lack of uniqueness, e.g. some names are very common. Anonymisation is considered a complex discipline in itself, with *SDC* being a tool of the trade [8]. *Pseudonymisation* is the act of replacing identifiers with a pseudonym, such as replacing a name with a serial number. Data held by ISD is pseudonymised, each patient is indexed by their Community Health Index (CHI). However it is well known that this kind of masking can be easily overcome by an *intruder*¹ when presented with datasets that contain large numbers of attributes, or wide data. *Unique Attribute Sets*, where a single observation has a combination of attributes that is unique in the data, is a problem in any size of dataset, but clearly increases as attribute numbers increase. There are a number of methods that could be described to reduce the disclosure risk in *microData* after anonymisation or pseudonymisation has been carried out. Greater detail will not be explored here as

¹In the field of SDC, any person or persons that are attempting to undo disclosure controls is referred to as an intruder.

our focus is on the disclosure control of output, however it will suffice to mention that methods for protecting *microData* fall into two broad categories, *perturbative* and *non-perturbative*. As can be inferred from these labels, the set of controls that fit into the latter does not involve modifying the data but rather masking it, whereas the former involves modifying the data by methods such as adding noise, swapping attributes and multiple imputation. The goal being to maintain the statistical properties of the data as much as possible whilst also creating uncertainty that an intruder could not overcome. The advantage of applying disclosure controls to input data comes with the fact that the data is static and in a regular format. The disadvantage is somewhat enshrined in the principle of *utility vs risk*. As *microData* is subjected to controls, the disclosure risk may well fall but also the utility of the data, how much use it will have for research for example, will also fall (figure 2.2)

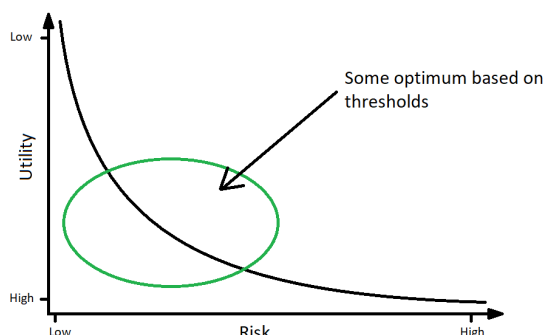


Figure 2.2: Risk-Utility Diagram

2.3 Output SDC

The health research community thrives on raw *microData* to enable studies to uncover trends or links that in turn inform healthcare policies. Therefore there is a demand for access to anonymised sensitive health data that has had no perturbative disclosure controls applied. This leads to the requirement of controls being applied at the output stage of these studies. Unfortunately, applying controls to the output of research does not come as easily as to the input data. The problem lies in the scale, variability and complexity of output produced by researchers. Not only are there many ‘types’ of statistical output, e.g. tabular, graphical and models, there is also the formatting to contend with, e.g. documents, spreadsheets, *CSV* files and software files from Stata, SPSS and the likes. Disclosure checks must be carried out on them all, and cross referenced against previous output produced in the same study. These problems are then compounded by the sheer volume of items that need checked. Once a study begins the researcher(s) will quickly start to produce output that they would like to have released from the safe environment. The output may be in one of two classes, that which is intended only for internal use, such as to illustrate progress/initial findings to stakeholders, and that which is intended for publication. The former is referred to as pre-publication or management output, and disclosure controls can in some instances be relaxed or even suspended given adequate disclaimers. The latter however must receive the full attention of the agent carrying out disclosure checks. There are categories [3.2] of output that can be quickly assessed, and there are those that require thorough appraisal, but due to the nature of progressive research there will also be outputs produced that take time for the agent to fully understand and even categorise before disclosive checks can be done. The variety of output that researchers can produce, though not infinite, is without doubt large.

”An analogy might be to imagine disclosure control as providing an enclosure for animals which keeps the animals safe and alive [...] SDC in a research environment is designing a zoo, not assessing a cage.” - Felix Ritchie, 2011 [22]

This is where the issue arises. Disclosure checks are an enormous time burden for eDRIS and teams in similar roles. Over the last 4 years (since April 2015) the ServiceNow² figures show that the eDRIS team reviewed on average over 400 items per month, and the time required per item can vary greatly from just minutes to days. The average time per item is around 0.5 days. This gives a feel for the requirement of some form of streamlining in the process. The team have been involved in almost 150 projects over this time period. Disclosure Checking, or *OSDC*, is considered the largest single burden that the teams in eDRIS have to contend with. The burden is multifaceted, in that not only does it take a large amount of time to appraise output for disclosure risks, but also the task requires a skilled worker to carry out and unfortunately is not considered particularly interesting or fulfilling. Currently the eDRIS team uses manual methods to apply checks, with an agent carrying out the work on a per project basis to allow some familiarity with the outputs to grow with the project. The lifetime of a project is generally measured in years, and can produce hundreds of output items, which itself can increase the time taken to assess disclosure risk as the agents must check each new output against those previously released.

”SDC can be a time consuming and onerous task for us, particularly when many researchers are working on a study over a number of years each requiring outputs. On average a study has 130 disclosure requests and the burden of differencing between each set of outputs is especially time consuming and can be prone to error. Anything that can be done to de-risk this will be extremely helpful.” - Jackie Caldwell, Information Commissioner at eDRIS.

2.4 Statistical Disclosures

As described, statistical disclosures occur when it is possible to ascertain previously unknown information about an individual from published data. In order to fully understand the scenarios in which these can occur, some concepts of the environment must first be described.

2.4.1 Data Linkage

Data linkage is the method by which multiple datasets for a set of data subjects are related to each other, with the goal of individuals’ records being matched across them. As data is collected by organisations the identifiers used between these organisations, and sometimes within them, will vary. Even within organisations such as the NHS, linking records isn’t an exact science since social identifiers as described above are commonly shared. Data linkage [15] relies on a probabilistic approach to capture how likely it is that two records with a certain combination of variables belong to the same individual. The data that eDRIS work with are indexed by a trusted third party and linked by EPCC in order to comply with information governance guidelines. National Records Scotland provide the third party indexing function, replacing the CHI number with a pseudonymised identifier in order to add a layer of anonymity. The indexed data is then passed

²ServiceNow is a third-party tool used by eDRIS to track requests on a per study basis

over to Edinburgh Parallel Computing Centre to be linked by a bespoke linking agent. The linked data are then checked and moved to an area in the National Safe Haven where they can be accessed by the approved researchers.

2.4.2 Intruders

Intruder is the term used to describe an entity who will deliberately attempt to undo any anonymisation that has been applied to data released to the public. This includes undoing any disclosure controls that may have been applied. An intruder is motivated in some way to disclose confidential information. This can be from simply a disgruntled individual, to an organisation, to even state level operations. An important factor is how well equipped an intruder is. An individual is unlikely to spend much time and resources undoing disclosure control, whereas a motivated organisation or state backed intruder may well spend months or years on the task. The understanding of how an intruder will attempt to uncover identities has led to researchers in the field of disclosure control to try to create a framework of ‘plausible intrusion scenarios’ [7].

2.4.3 The Data Environment

The term *Data Environment* covers the idea that there is already an incredibly large amount of data out in the public, potentially personal, which is accessible to anyone who cares to search for it. This environment is not static, slow changes in the public’s view of what data they are willing to share openly have a long term effect whilst dramatic changes can unfold quickly with innovations such as online social media. People have always been motivated to maintain a level of privacy surrounding themselves and their families, and this level was considered ‘intuitive’ enough that a general acceptance of what information was shareable and what was not existed. However, with the advent of social media in the early 2000s it suddenly became a lot less clear what people were willing to share about themselves freely, share for a small reward (e.g. social convenience, establishing personal brand), or what they wanted to keep private at all costs. This was mainly due to naivety of social media users during the early years of these new platforms. Very few truly understood how data was being collected and analysed, and it is only in the last few years that the power of this technique has become common knowledge and people are again thinking twice about what they share. It is however already a bit too late for most, what is on the internet stays on the internet after all. This has led to the concept of our *Data Environment* [9][16]. A motivated intruder will almost certainly gather information from this data environment and use it to cross reference any new sources of information. In theory this means that a data controller should expect the disclosure checking agents to do the same. In reality this is not a practice that could be sustainable. Thankfully data privacy laws exist that protect stakeholders who are acting responsibly and ethically. As long as data controllers have exercised ‘reasonable’ precautions with regard to their outputs they are protected by these laws, the assumption being that the intruder instead would be subject to prosecution should they attempt to circumvent such reasonable precautions [17]. The development of an understanding of how intruders will try to use the data environment and statistical releases is useful in developing strategies to help protect against statistical disclosure [7][16].

2.4.4 Output Types

As alluded to in previous sections, the variety of types of output created by researchers is large. To list and attempt to describe them all would be futile in this work, but conveying the idea of how they effect disclosure checks is important. With each output type comes its own potential vector for disclosing information, such as releasing descriptive statistics of a cohorts ‘Income’ or ‘Age’ may show exact figures for individuals in the cohort within the summary, or releasing scatterplots or boxplots can show individuals as points on the graphs, therefore increasing the chances of their identification in the Data Environment. Each output has its own weakness that must be identified and overcome by *SDC*.

2.4.5 Attack Scenarios

Two broad categories can cover the mechanisms leading to a disclosure, *Attribute Disclosure* and *Identity Disclosure*. Attribute Disclosure occurs when all members of a group have the same class within an attribute, or fall within a particular range for continuous data. This carries the risk of revealing sensitive information about all members of that group. Identity Disclosure is where the data reveals information which will relate to an individual. Identity disclosure can in fact come in two flavours; *Internal*, where there are two individuals who can identify each other and therefore gain knowledge on the other, and *external* where a single individual can be identified by all. It is simplest to represent these risks in the form of tabular data, though they can be found across most of the different types of output in one way or another. A *Dominance Attack* can also lead to both Internal and External Identity disclosure should a small number of subjects in the data make up a large proportion of an attribute, such as Income, when represented in a magnitude table.

Tabular Output

Disclosure control relating specifically to tabular output can be thought of as the best defined of all the output types, and in fact pre-dates the *microData* techniques as the early references to disclosure control described previously generally related to the release of statistical tables. Two main types of table are described in most literature relating to SDC, Magnitude tables and Frequency tables.

”Frequency tables display the count of respondents at the crossing of the categorical attributes, e.g. number of patients per disease and municipality. Magnitude tables display information on a numerical attribute at the crossing of the categorical attributes, e.g. Average age of patients per disease and municipality.” - Josep Domingo-Ferrer [5]

	Class 1	Class 2	Class 3	Class 4	Class 5	Totals
Group 1	6	4	4	1	2	17
Group 2	5	3	5	2	5	20
Group 3	6	7	0	0	0	13
Group 4	2	1	3	3	4	13
Group 5	10	0	0	0	0	11
Totals	29	15	12	6	11	73

Table 2.1: Unsafe Table 1

Tabular output can be disclosive if cells within the table contain counts that are too low or if sets of cells, such as those covering a whole category, are empty or full. Disclosive properties of basic tabular data are well defined, and simple to illustrate through example. The frequency table 2.1 demonstrates all three of the previously described categories/flavours of dis-

closure. The Groups may be groups of Individuals, such as Age Group or Ethnicity, and the Classes some attribute of the groups such as Income Bracket, Education Level, or some measure of Well being.

Cell Count Thresholds - Minimums

The cell count refers to the number of observations that appear in any single cell in tabular data. In some tables the cell count can be related directly to the number of individuals in that category, such as is the case in table 2.1. If the count is only 1 or 2 then this is considered disclosive in two different ways:

- 1 equates to the external disclosure, where the individual is potentially identifiable to everyone
- 2 equates to the internal disclosure, where two individuals are potentially identifiable to each other, but not to everyone.

A count of 3 is considered the minimum count that is statistically safe, however in practice most thresholds for cell counts are at least 5, often 10 and in cases where the data is especially sensitive this can be as high as 30.

Cell Count Thresholds - Maximums

There is also a maximum count that should be considered in the table cells. Where the individual cell counts approach or equal the margin total an Attribute Disclosure risk is created. This type of disclosure can reveal that an entire group resides within a small number of classes, therefore the class of that entire group is revealed either exactly, for example Group 5 in Table 2.1, or within a small range of classes, for example Group 3 in the table.

Differencing

Differencing attacks can occur when multiple outputs are created from the same source data, but the variable breakdowns between the outputs vary slightly or the cohort size changes by some factor. For example, table 2.2 shows a similar table as in 2.1, but the cohort has been reduced due to a binary factor, for example this table may now show only one Gender, or those who are Non-Diabetic. The resultant table when 2.2 is subtracted from 2.1 is shown in 2.3. Every cell in this table now presents a disclosure risk in one of the manners described previously, with an increase in overall risk due to the additional information of whatever binary attribute was used to divide the cohort.

CONTROLLED	Class 1	Class 2	Class 3	Class 4	Class 5	Totals
Group 1	4	4	4	1	2	15
Group 2	5	2	5	2	2	16
Group 3	6	7	0	0	0	13
Group 4	2	1	2	2	3	10
Group 5	7	0	0	0	0	7
Totals	24	14	11	5	7	61

Table 2.2: Unsafe Table 2

CONTROLLED	Class 1	Class 2	Class 3	Class 4	Class 5	Totals
Group 1	2	0	0	0	0	2
Group 2	0	1	0	0	3	4
Group 3	0	0	0	0	0	0
Group 4	0	0	1	1	1	3
Group 5	3	0	0	0	0	3
Totals	5	1	1	1	4	12

Table 2.3: Unsafe Table - Differenced

2.4.6 Resolution

These are obviously extreme examples which are easy to use as illustration of the risks. The real problem arises when the risks are much more subtle and obscure. The next chapter describes what techniques the field of *SDC* has evolved to detect and counter Disclosure Risks.

3 Existing Methodologies

Contemporary Research focusing on progressing *SDC* cover areas such as establishing standards and methods to formalise the problem, aligning terminology and definitions, and designing robust practices based on these developments, as well as progressing techniques to measure data utility vs risk, and techniques for applying controls.

3.1 Applying Controls

A number of methods for controlling disclosure risk have been devised to help deal with data at the output stage. Again these methods are either *Perturbative*, where the output is modified in some way that changes the underlying figures, or *Non-Perturbative* where elements are simply removed or masked.

Non-Perturbative

The simplest methods are those that do not change the data at all. *Cell Suppression* involves masking those cells that fall under a threshold set by the data controller, as described previously this can be from 3 to in excess of 30 depending on the sensitivity of the data. The tables in 3.1 and show two ways that this can be achieved, with a minimum count of 3. All values lower than 3 are suppressed and the marginal totals are adjusted. *Global Re-coding*, as illustrated in table 3.2, encourages the minimum cell count to be avoided altogether by aggregating groups or classes together until there are no cell values below the limit. This approach avoids disclosure risk and also ensures no data loss. This should be applied in a logical manner, where the groups/classes are ordinal rather than nominal, and consecutively clustered. These methods have the advantage of not changing the shape of the data and also being relatively easy to implement, even for much larger tables. The disadvantage is in the utility loss of the data by the reduction of the tables *resolution*.

ORIGINAL	Class 1	Class 2	Class 3	Class 4	Class 5	Totals
Group 1	6	4	4	1	2	17
Group 2	5	3	5	2	5	20
Group 3	6	7	0	0	0	13
Group 4	2	1	3	3	4	13
Group 5	10	0	0	0	0	11
Totals	29	15	12	6	11	73

CONTROLLED	Class 1	Class 2	Class 3	Class 4	Class 5	Totals
Group 1	6	4	4	-	-	14
Group 2	5	3	5	-	5	18
Group 3	6	7	-	-	-	13
Group 4	-	-	3	3	4	10
Group 5	10	-	-	-	-	10
Totals	27	14	12	3	9	65

CONTROLLED	Class 1	Class 2	Class 3	Class 4	Class 5	Totals
Group 1	6	4	4	<3	<3	14
Group 2	5	3	5	<3	5	18
Group 3	6	7	<3	<3	<3	13
Group 4	<3	<3	3	3	4	10
Group 5	10	<3	<3	<3	<3	10
Totals	27	14	12	3	9	65

Table 3.1: Unsafe Table 1 - Suppressed

ORIGINAL	Class 1	Class 2	Class 3	Class 4	Class 5	Totals
Group 1	6	4	4	1	2	17
Group 2	5	3	5	2	5	20
Group 3	6	7	0	0	0	13
Group 4	2	1	3	3	4	13
Group 5	10	0	0	0	0	11
Totals	29	15	12	6	11	73

CONTROLLED	Class 1	Class 2	Class 3	Class 4	Class 5	Totals
Group 1+2	11	7	9	3	7	37
Group 3+4+5	18	8	3	3	4	36
Totals	29	15	12	6	11	73

Table 3.2: Unsafe Table 1 - Re-code Groups

Perturbative

Perturbative methods include those that are of a deterministic nature and those that are stochastic. Deterministic methods include *Rounding*, all values are rounded to a defined base which is often selected to be 3, and *Controlled Rounding* which applies a more intelligent *Integer Linear Programming* algorithm to decide on which direction to round given a set of constraints set by the marginal totals of the original table.

ORIGINAL	Class 1	Class 2	Class 3	Class 4	Class 5	Totals
Group 1	6	4	4	1	2	17
Group 2	5	3	5	2	5	20
Group 3	6	7	0	0	0	13
Group 4	2	1	3	3	4	13
Group 5	10	0	0	0	0	11
Totals	29	15	12	6	11	73

CONTROLLED	Class 1	Class 2	Class 3	Class 4	Class 5	Totals
Group 1	6	3	3	0	3	18
Group 2	6	3	6	3	6	21
Group 3	6	6	0	0	0	12
Group 4	3	0	3	3	3	12
Group 5	9	0	0	0	0	12
Totals	30	15	12	6	12	72

CONTROLLED	Class 1	Class 2	Class 3	Class 4	Class 5	Totals
Group 1	6	6	3	0	3	18
Group 2	6	3	3	3	6	21
Group 3	6	6	0	0	0	12
Group 4	3	0	3	3	3	12
Group 5	9	0	0	0	0	9
Totals	30	15	9	6	12	72

Table 3.3: Unsafe Table 1 - Base 3 Rounding, Simple (middle) and Controlled (bottom)

Privacy. This technique could be described as at the cutting edge of *SDC* algorithmic methods, and some proponents describe it as the solution to a generalised method by which to apply disclosure controls to a wide range of data types [6][20].

Stochastic methods usually involve the addition of *Noise*. *Barnardisation* is a well established implementation of this, where the random addition of $+1$, 0 or -1 is applied to each cell, or *Controlled Tabular Adjustment (CTA)* which first applies this random noise only to the sensitive cells, then recovers the original marginal totals by applying noise to some non-sensitive cells. The tables in 3.3 portray the *Rounding* methods, and table 3.4 is the product of *Barnardisation*. Applying *CTA* to the example table is more advanced, requires the use of tools such as the *R Package sdcTable* and is not be demonstrated here. There are more advanced methods that require a high level of understanding not only of mathematics, but also concepts of *SDC* itself that will also not be demonstrated. However, it would be remiss not to at least mention *Differential*

ORIGINAL	Class 1	Class 2	Class 3	Class 4	Class 5	Totals
Group 1	6	4	4	1	2	17
Group 2	5	3	5	2	5	20
Group 3	6	7	0	0	0	13
Group 4	2	1	3	3	4	13
Group 5	10	0	0	0	0	11
Totals	29	15	12	6	11	73

CONTROLLED	Class 1	Class 2	Class 3	Class 4	Class 5	Totals
Group 1	6	3	5	1	3	18
Group 2	4	3	5	2	5	19
Group 3	5	8	0	1	1	15
Group 4	1	2	3	3	4	13
Group 5	9	0	0	1	1	11
Totals	25	16	13	8	14	76

Table 3.4: Unsafe Table 1 - Barnardisation

3.2 Framework Solutions

The current thinking on how to approach disclosure checking focuses on framework based solutions. So far there have been no breakthroughs with regards to a programmatic solution due to the unstructured nature of the problem, i.e. a lack of formatting standards and the variability of techniques required for output types.

Rules Versus Principles

In the context with which we are focused, health data research, the primary goal of both research institute and data controller is that of ensuring the data can be utilised as fully as possible for the benefit of society. It is therefore important that the relationship between the two entities remains stable and positive. The rules vs principles idea is borne out of this want. As described previously, there are sets of minimums and maximums that the data controller can require output to abide to, described as Rules. In some instances these rules could be applied to outputs with no leeway. However, it is understood that there will be circumstances where the rules should be tightened or relaxed depending on the nature of the output, such as taking into account the balance of public benefit and confidence. A principles based approach gives both the researcher and the OSDC checker flexibility around the rules, leading to a positive discussion on the release of the output rather than a yes/no option. This is defined as a Principles based approach, and though technically making the OSDC checking more cumbersome, has the added benefit of keeping the relationship between the two institutes positive.

Assessing Output for Disclosure Risks

Methods of assessing disclosure risk varies depending on the type of statistical output being reviewed. There are publications available detailing how this should be accomplished for a wide range of these outputs. An incredibly detailed guide was produced by the *European Statistical System Network For Excellence (ESSNet)* in 2010 [12] which led to the publication of the *Statistical Disclosure Control* book [13] written largely by the same group. These are both long and detailed works covering the full breadth of *SDC*, with the latter being considered mandatory reading for

those entering the field. In just the last month, the Safe Data Access Professionals working group [31] have released the first version of their handbook [11] which focuses on *Disclosure Checking* of outputs, and with its accessible layout is a good guide for this particular aspect of SDC.

Categories of Output

Although the variety of conceivable outputs can be described as continuous, they can by and large be categorised into a discrete set. Taking this discrete set and creating recommended best OSDC practices for each makes the work of an OSDC checker easier. Each category brings with it the rules that should be applied or considered for that type of output as well as how to apply them.

Safe and Unsafe Statistics

Work done in the Office of National Statistics, and carried on by Professor Felix Ritchie [23], has led to the concept of safe and unsafe statistics. This builds upon the categorisation approach as each output will fit into a category which is already declared safe or unsafe, leading to faster appraisal for the safe ones and more in-depth appraisal for the unsafe ones.

The Five Safes

The Five Safes [30] is a framework by which those involved in data control and access can follow to appraise how safe the overall environment of a study is. The five safes are:

- Safe Projects
- Safe People
- Safe Setting
- Safe Data
- Safe Outputs

Each one of these safes is a layer of strength that should be considered when thinking about establishing access to data for a study. The strengths of each are balanced between one another to create a safe environment. In the context of the Safehavens, the strengths of the 1st, 3rd and 5th are high, enabling the 2nd to be slightly relaxed and the 4th to be very relaxed.

OSDC Training

Training courses on the topic of OSDC bring together the above concepts and are designed to be delivered to all stakeholders in order to raise awareness of the situation. The Office of National Statistics (ONS) runs an *Approved Researcher Scheme* which includes a course and examination that cover *SDC* principles such as those described above. The goal of the scheme is to create ‘Safe Researchers’ who can access ‘...data that cannot be published openly, for statistical research purposes, as permitted by the Statistics and Registration Service Act 2007 (SRSA)’ [32]. A more technical course is currently being trialled by the ONS that focuses on applying controls to a spread of different output types, and is aimed at those who are at the front line of disclosure checking, such as eDRIS.

4 Proposed Solutions for eDRIS

4.1 Administrative Solutions

There is potential for implementing non-technical solutions that could harness time savings through decreasing the quantity of output that eDRIS have to deal with or simply make the output checking process a smoother. These solutions fall under the category of Administrative because in order to implement them changes would need to be made to the processes that the teams involved follow, including pricing structure, IT configurations and categorising the outputs themselves.

4.1.1 Three Potential Time Savers

Tokenise Output Checking

The Statistical Disclosure Control book [13] (section 6.6.4) describes a system that could encourage researchers to be more selective about the outputs that they request for release. This system would impose a charge based on the time burden that output checking creates. The idea is based on the principle that a large amount of the outputs that researchers put forward for release are not required directly by the study. They can often be intermediate outputs, not intended for publication, such as to present to certain stakeholders. This approach may appear distasteful, as if the data controllers were monetising on research. However, if the cost demonstrably covers the extra time burden and also reduces the initial price that must be paid to access the data, researchers may welcome such a system.

Reduce Steps in FTP Process

Part of the process of checking for disclosures involves obtaining the output that must be checked from the Safehaven workstation and placing it on to a secure network location that all agents have access to. At the moment this process involves multiple steps which have the potential to be condensed by means of automation. Currently, following the researcher contacting eDRIS to request OSDC checking on their output, the agent will:

1. Log on to Researchers Safehaven Workstation
2. Locate Output File(s) (Usually in a specific folder)
3. Open sFTP tool and Upload File(s)
4. Open sFTP tool on Local Workstation
5. Via FTP tool function, send File(s) to an External eMail Address
 - Auto-Forwarding then sends this on to eDRIS eMail Address
6. Save File(s) from this eMail on to Secure Network Location
7. Carry out OSDC Checks

This process will take around 3 or 4 minutes on average. In theory this could be more automated [13] (section 6.8.2):

1. Researcher places Output File(s) in to OSDC Folder on Safehaven WS
2. Agent Authorises Output files, entering file names and Study ID
3. sFTP server in Safehaven Environment detects existence of authorised files and picks them up
4. Automated query based in Production Environment detects existence of files on the sFTP server and moves to secure folder in Production Environment
5. Agent Carries out OSDC checks

Dual Sign-Off

During discussions with Professor Chris Dibben and Dr Lee Williamson, an idea was posed that more trust could be placed with the researchers themselves. Researchers are already required to attend the Safe Researcher training and pass an assessment before gaining access to sensitive data. Therefore, there is at present an investment in ensuring researchers know how to carry out basic disclosure checks on their own output before requesting it is released from the safehaven. This could be extended to a system that permits researchers to sign-off output as safe themselves. A method by which this could be allowed would be a Dual Sign-Off system. The principle researcher in the study would be required to verify output is safe, and another more senior representative of the same research establishment would also be required to confirm the same. This would have to be encapsulated in a robust audit system to detect failures and a set of deterrents to reduce inappropriate sign-off leading to release of disclosive output. Such deterrents could be in the form of added time barriers such as:

- A reduction of established trust. Output checking for the study would be returned to eDRIS team, increasing the time to release.
- Removal of access to the safehaven until the Safe Researcher Training and assessment was retaken

4.1.2 Synthetic Data

Synthetic data is produced by recreating new observations based on the statistical properties of the original, real, observations. Synthetic data has its primary advantage in being completely fabricated and therefore fully non-disclosive. It is technically possible that a synthetic observation is created which exactly matches an original observation. This is a risk that can be somewhat controlled at the synthesis stage by modifying how accurately the new data reflects the original data, adjusting parameters to prevent ‘overfitting’. Synthetic data generation began its life as an extension of existing Multiple Imputation techniques that were used to generate new data for missing values in datasets [21]. Multiple Imputation (MI) was developed to improve model accuracy where previously single imputation was used, such as filling in missing data with the mean, mode or random sample of the attribute. MI instead produces a vector of length n containing new values predicted using parametric statistical methods on the original attribute. This vector is then used to recreate n new datasets with the non-missing data and analysis is carried out on all datasets. This technique was extended to create larger vectors and cover all the attributes in a dataset, in effect creating a whole new imputed dataset, or synthetic data [24]. Since this application, research into

synthetic data production has exploded, driven in part by the Machine Learning community for creating larger training datasets and in part by the Data Confidentiality community for creating safe data. The techniques available have grown to include more parametric and non-parametric methods, some of which involve machine learning techniques such as Random Forests, and some even employing Genetic Algorithms [4], and have been shown to produce good quality synthetic data [19].

The limitation of using synthetic data in the environment being investigated is that it applies to microData, or input data, rather than the output data which is the primary burden of the eDRIS team. However, as a strategy to reduce the burden of disclosure assessment, this technology has potential. It is proposed that researchers who have access to the original data carry out their analysis as per usual, but any output that needs to be generated and released can be created using the synthetic data. This is rather like the opposite of the ‘Gold Standard Analysis’, which is a popular use of synthetic data in research [25]. As long as the results match the original closely, then the message that output is intended to convey within a paper or other publication could be represented just as effectively if based on synthetic data. Importantly, the study conclusions should always be reproducible by future researchers using the original data. This would not suit a situation where small numbers of observations were a key finding of the study, in which case the output would be high risk already and the normal principle based approach of disclosure control would need to be carried out.

synthPop Example

The *R Package*, ‘*synthPop*’, produced by Nowok et al [19], can be used to create synthetic data via a number of these techniques. *SynthPop* has many parameters that can be set, however they will not all be explored here. Rather, the aim of this demonstration is to use the default settings to create three synthetic sets and compare them to the original. The goal is to show that it is relatively easy to harness the power of *synthPop*, and in turn synthetic data, meaning that researchers could be shown that creating their own synthetic data is straight-forward.

Taking the sample *iris* dataset in *R Studio* for this demonstration. This dataset is simple, with four numerical attributes representing measurements of Iris flowers, and a single categorical attribute representing the Iris species. It is popular as an example dataset for demonstrating statistical tests and machine learning techniques as the numerical attributes combined are excellent predictors of the species. Comparisons of the original distributions against the synthetic will be made by way of kernel density plots and an appropriate analysis of variance test.

The *iris* data has the following attributes:

```
> names(iris)
[1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width" "Species"
```

Following the loading of the appropriate libraries, the *iris* data can be passed in to *synthPop*:

```
> synthetic_data <- syn(iris)
```

Once synthesis completes, the variable `synthetic_data` will contain a list of objects describing the parameters used to create the new data, including the data itself:

```
> names(synthetic_data)
[1] "call" "m" "syn" "method" "visit.sequence"
[6] "predictor.matrix" "smoothing" "event" "denom" "proper"
[11] "n" "k" "rules" "rvalues" "cont.na"
[16] "semicont" "drop.not.used" "drop.pred.only" "models" "seed"
[21] "var.lab" "val.lab" "obs.vars" "numtocat" "catgroups"
```

The synthetic data is contained in the `syn` list item.

Comparison

Figures 4.1 and 4.2 show the kernel density estimates for the data in two flavours, the overall view of each numerical attribute, and subdivided by Species category since the classic use case of this data has Species being the ‘target’ attribute. Due to bimodal distributions present in the overall data, a Kruskal-Wallis one-way ANOVA on ranks test was used to compare each synthetic distribution against the original.

- Kruskal-Wallis ANOVA Hypotheses:
 - H_0 : There is no significant difference between the distributions
 - H_1 : There is at least one difference between the distributions
- Significance level: 0.05

The plots include the p-value results of those tests.

From these results it can be seen that the synthetic versions of the Iris data created using default options in `synthPop` do maintain similar distributions and correlations to the original data. All p-values are significantly higher than the threshold of 0.05.

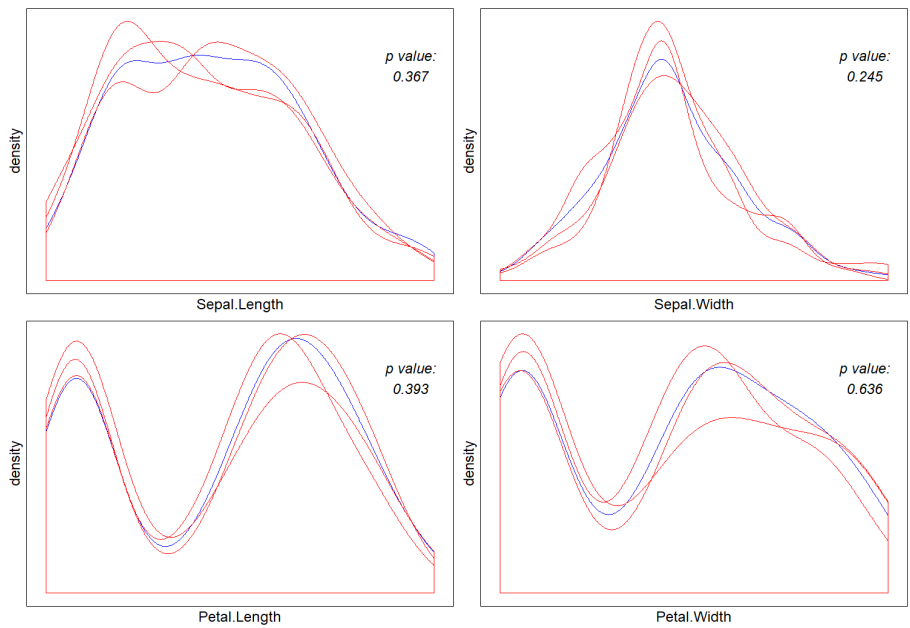


Figure 4.1: Density plots of numerical 'Iris' attributes from original and 3 synthetic datasets created with default options. Blue = Original, Red = Synthetic

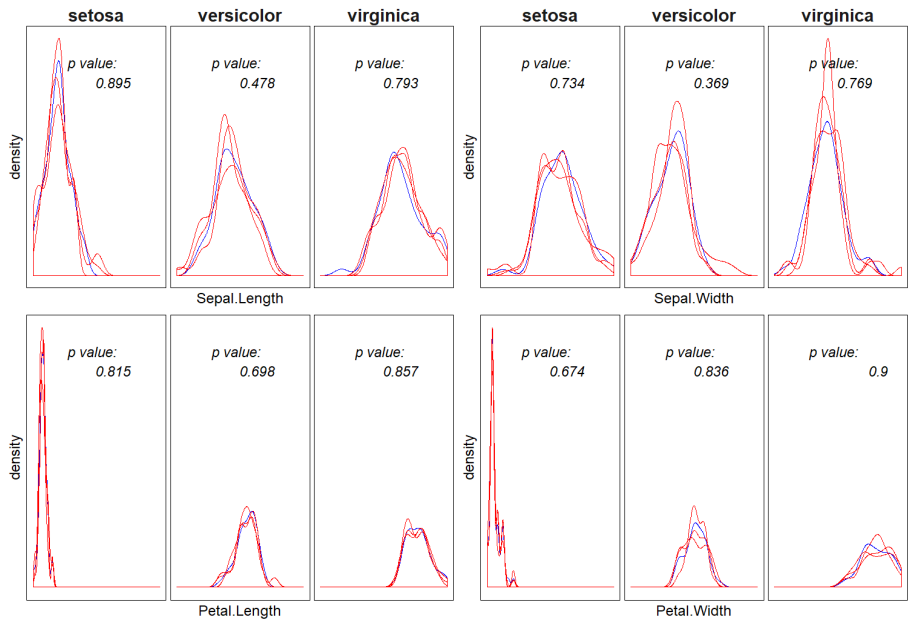


Figure 4.2: Density plots of numerical attributes split by Species, original and 3 synthetic datasets created with default options. Blue = Original, Red = Synthetic

4.2 Programmatic Solutions

4.2.1 Standardising Output

From a data analytics and automation point of view, ‘tidy’ data is a big step towards creating programmatic solutions for processing that data. The output that the SDC agents in eDRIS need to appraise are generally in a standard format throughout projects, but rarely across projects, and the standard format in use is often not simple enough for passing in to algorithms. The standardisation of output would be the first step in achieving tidy data. If standard formats for popular statistical outputs were accepted across research institutes this would make the development of automation more likely. Building upon the work done to classify output and SDC methods [11], there could be further work to agree upon how these outputs are presented to SDC agents in format and in attached metadata.

4.2.2 Machine Learning

Although standardisation would be a step in the right direction, an alternative could be found in the ‘Unstructured Data’ mining field. There may be potential in combining Natural Language Processing with Image Processing to, for example, classify the outputs into a type 3.2 that can then be appraised for disclosure risk by the appropriate method outlined in the Safe Data Groups guide [11], or even be automatically classed as safe [23].

4.2.3 Differencing Tool

A single idea stood out as achievable in the time frame of the project, and that was creating a tool that could highlight if tabular output is at risk of a differencing attack. Differencing was put forward by eDRIS staff as being a potential line of inquiry as a large amount of time is spent checking new outputs with previous outputs within a project 2.3. As described previously 2.4, the principle behind using differencing as an attack vector is obtaining multiple tables that have been produced on the same dataset but one or more variables have a slightly different breakdown. When one table is subtracted from another this could lead to small numbers being revealed and therefore increases the potential for a disclosure. During a conversation with Dr Nancy Burns of National Records Scotland, an approach to tracking these risks was described via use of a spreadsheet containing details on variables and their existing breakdowns. Using the spreadsheet, Dr. Burns is able to quickly compare a new breakdown with existing breakdowns of a variable, and visually assess which breakdowns would pose a differencing risk, and so know which tables would need extra attention when carrying out SDC checks. As a small illustration, Table 4.1 shows this in action. In the example, Tables 1 to 4 were initially created in sequence with a medium risk rating between the two breakdowns in use across them, table 5 introduced a high risk due to a third breakdown. It can be ascertained by visual inspection that this new breakdown, C, has a small overlap with breakdown B, and so table 5 should be checked against tables 2 and 3, but doesn’t need to be compared to tables 1 and 4 as there is no overlap between breakdowns A and C.

This method lends itself well to translation into an automated process that could keep track of large numbers of variables and breakdowns in use across a project. Currently the eDRIS analysts and RCs are accustomed to using *R* [35] tools to carry out daily tasks. The analysts also develop *R Packages* when the need arises for custom functions. Therefore the method chosen to create a tool to track and highlight differencing was via an *R Package*.

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																								
A	Table1,Table4																								
B	Table2,Table3																								
C	Table5																								
	1971-1980										1981-1990														
	1971-1974					1975-1978					1979-1982					1983-1986					1987-1990				
	1971-1975					1976-1980					1981-1985					1986-1990									

Table 4.1: Differencing Spreadsheet with risk areas highlighted: red-high, yellow-medium.

5 R Package Development

5.1 High-Level Tool Design

The proposed solution is an *R Package* containing functions used to keep track of tabular output on a per study basis. The key goal to produce a report on which groups of tables present a risk of differencing by comparing breakdowns. Some basic functions/features were considered for this goal:

1. An agreed set of inputs, metadata describing the Linked Data and the Research Outputs.
2. A method to store the metadata for use across sessions.
3. A function to add new metadata as it is created by the researcher.
4. A function to analyse and report on the existing outputs, highlighting differencing risks.

Also, a ‘User Journey’ was drafted to understand how the tool would be used by the eDRIS teams, Figure 5.1, and user profiles were considered when designing this workflow.

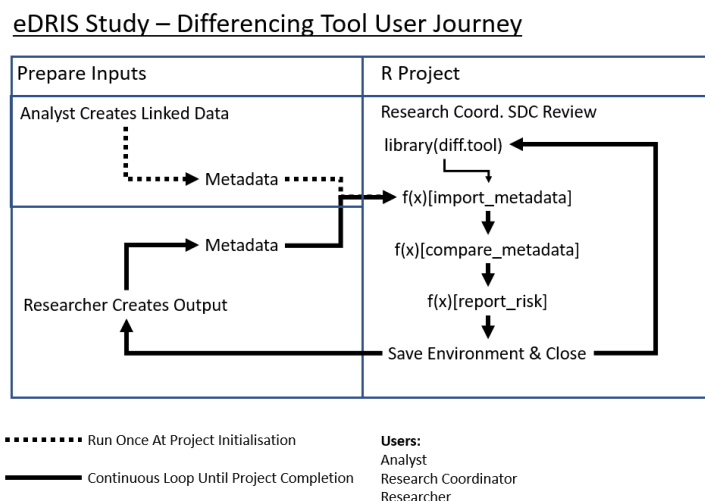


Figure 5.1: Differencing Tool User Journey

The users in this representation have varying levels of technical knowledge, and assumptions have been made with regard to their knowledge of *R*, their ‘profiles’:

- The Analysts were assumed have good to expert levels as they are required to use it daily as part of their function in the eDRIS team.
- Research Coordinators have at least an understanding of the console, loading libraries and running functions, but not necessarily any higher level of knowledge as it does not form a core part of the RC skill set.
- The Researchers were assumed to have no knowledge of *R* as they may use any number of statistical tools in their research, however it was assumed that they have knowledge of editing and saving excel spreadsheets in order to complete a template.

5.2 Input Metadata Development

5.2.1 Output Data

The implementation of item 1 of the features above, the inputs, required first an understanding of what information would be needed by the tool to fulfil item 4. This information would form the basis of metadata for both the datasets created by the analysts, and the tabular output created by the researchers. By working backwards from the key goal of the tool, it was decided that the minimum information covering each tabular output should be:

- Variable Name
- Table Name
- Variable Breakdown

The first 2 were self explanatory, they would simply be represented as short *string* objects in R, however the 3rd presented a complexity that had to be described more fully. The breakdown could take different forms depending on the type of data the variable represents. *Categorical* data would be represented most easily as *string* objects, and passed in to R as either a *list* of *strings* or a single *string*. *Numeric* data on the other hand could be a set of ‘bins’, e.g. ‘0-9,10-19,20-29...’, again as either a *list* of *strings* or a single *string*, or it could be formed of a *list* of *integers*, such as breaks in the *range* ‘10,20,30...’. As the latter of these options would require more complexity in the functions, by way of a single routine that must handle both *string* and *integer*, it was opted to represent both data type breakdowns as *strings* in the initial package. If the need to improve upon this arose later, say to add functionality that separated the data types, then it could be addressed at that point.

It was considered prudent to also identify information that might be useful for more in-depth analysis and include this in the table metadata at an early stage, so that it was available for future versions of the tool. Obvious features such as the size of the cohort, number of categories in the breakdown, the data type of each variable, and the originating datasets (should there be multiple datasets used in the study) were therefore included as items in the output metadata. A not so obvious feature, highlighted by the eDRIS team, was information relating to variables that do not appear in the original datasets but are instead derived from those original variables. An extended set of metadata was formulated to capture this additional information:

- Size of cohort in output (integer, n)
- Data Type (string, Categorical or Numerical)
- Data Width (integer, Number of Categories or Numerical Range)
- Original Dataset (string)
- Derived From (string, list of original variable names, can be NA)

Who would provide this metadata was a factor to consider. Asking the researchers to create it would mean the metadata would require formatting by the eDRIS team before entering into the

tool. Having the eDRIS team themselves do it would possibly cost extra time as the agents would not be as familiar as the researcher with the output. A good balance was by providing a template that the researcher could fill out and the tool could read in. The format for this template was chosen to be *XLSX* as it would allow the use of drop-down menus for the variable name field, should this be considered a useful feature at a later stage in development. The complexity overhead for choosing *XLSX* over *CSV* is minimal thanks to existing *R* packages for data import, such as those found in the *Tidyverse* [36] packages¹.

The initial template for researchers to enter their output metadata is therefore as in table 5.1.

table_name	variable_name	derived_from	original_dataset	data_type	data_width	breakdown	n
-	-	-	-	-	-	-	-

Table 5.1: Template for Output Metadata

5.2.2 Linked Data

The datasets that are made available to the researchers are in the form of pseudonym-ised micro-data. The metadata for these datasets needs to describe the features of this micro-data such as variables in the dataset, their data types, ranges or number of categories, and of course the dataset name. Rather than carry this out manually, a function is used by the Analyst to extract this information and place it into a *CSV* file, using a standard name convention for the file. The *CSV* file(s) is then imported in to the *R Project* environment at the project initialisation stage. Additional information was also identified by eDRIS staff as being useful to capture at this stage. The sensitivity of the variables and a risk level is encoded for each variable, to allow the reporting functions to convey this to the SDC agents. This was difficult to translate into the function described and so a manual approach was chosen instead. The function creates blank columns in the *CSV* file for the Analyst to then complete before releasing the file to import. The following information is therefore captured in the Linked Data Metadata, the first 4 by the function and the last three by the Analyst:

- Dataset Name (string)
- Variable Name (string)
- Data Type (string)
- Data Width (integer)
- Sensitive Variable (boolean yes/no or 1/0)
- Sensitive Geography (boolean yes/no or 1/0)
- Risk (factor, 1 to 5 for low to high)

¹The Tidyverse is not only a set of packages for R, it also promotes an etiquette for working with data in a ‘Tidy’ manner across all of those packages.

5.3 Input Processing Development

5.3.1 Data Storage

The metadata that is created during the study must somehow be passed in to the tool and stored across multiple *R* sessions. This is possible in *R studio* by saving the *environment* upon closing the session. To take advantage of this feature, the SDC agent must make use of the tool functions from within an *R Project*, rather than simply from an *R Console*, as this would ensure a standard folder structure is in place to retain the relevant files. Therefore it was decided to create an *R Project* at the initialisation of the study, along with the normal folder structure that is currently created at this time. Additional folders would be required to contain the project files and the metadata. The data is imported in to data frames, using the *Tidyverse* ‘*Tibble*’ object as it provides a more intuitive structure than the *Base R* ‘*data.frame*’ object.

5.3.2 Import Function

The metadata must be added in to the *R Project* environment via some *Functions* to be used by the RCs. The metadata comes in two flavours, Linked Data will be *CSV* format AND will only need to be imported ONCE at the outset of the project, whereas Output Data will be generated and imported continually throughout the study. Therefore two import functions handle the *CSV* and *XLSX* files that contain the project metadata.

Linked Data Import

The Linked Data import is simple and static, with a routine that handles single or multiple *CSV* files targeting a specific directory and specific name pattern. This function is designed to be run once, and passes all the *CSV* files available into a *tibble*.

Output Data Import

The Output Data import function adds new data to the existing *tibble* created by the *Initial Configuration Script*. This function will be used by the Research Coordinators to add new outputs as disclosure review requests are made by the researchers.

Data Correction/Removal

As an additional feature, a removal counterpart function exists for each of the two import flavours above in order to make corrections easier than rebuilding the *tibbles* from scratch. These functions target rows in the *tibbles* by dataset name or table name and removes those rows.

5.4 Output, Reports and Visualisation

The key goal of the tool is to create a report that Research Coordinators can use to appraise differencing risks across all the tables created during the study. The main route to this ‘risk detection’ is by comparison of tables and their properties such as variables in use, breakdowns, and cohort size. As described in the section 2.4, differencing can become an issue under a few distinct circumstances. The simplest is when two tables use the same variable but different breakdowns, and so this forms the basis of an initial reporting function. The report simply prints tables, grouped by variable, where unique breakdowns within the groups exceed 1. This format, though fulfilling the key goal, was considered to be too basic as an end point for the tool and so more advanced reporting, and the addition of visualisation, was developed.

Visualisation via Simple Networks

A popular method to visualise relationships between members of a group, or groups, is by way of a mathematical network, or graph. This encoding of relationships is used across many areas of research, from analysis of linguistics and word networks [2], to disease control and cattle movement [3], and has been found to reveal interesting aspects [1] of these relationships that are not immediately conceivable via other methods. The mathematical field of Network Theory covers many applications of network analysis designed to find features within a network, such as cliques, path lengths and density, but the core principle of the network is defined as follows:

A network can be formally defined as $G = (V, E)$, consisting of a set of vertices, V , and a set of *pairs of elements* from V , E . If these pairs are *Ordered* the network is *directed*, otherwise it is *undirected*.

For the purpose of conveying differencing risk across tables, the data available can be translated into a network in some different ways. The choice of which element to represent as the vertices and which to represent as the edges must be made, with the potential to allow different configurations depending on the preference of the user.

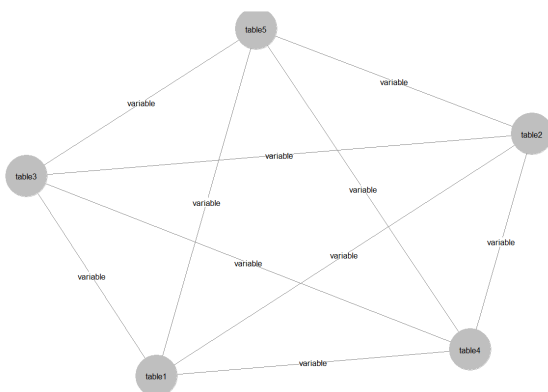


Figure 5.2: Simple Network of Tables (Nodes) and Variables (Edges)

A set of functions are used to carry out this translation from the output metadata *tibble* first into an *Adjacency Matrix* then into a *network* object using the *network* package [37]. These objects can then be drawn using *ggnet2*, via the *GGally* package [38], which is based upon the *Tidyverse GGplot2* package for visuals in *R*. Figure 5.2 shows an example of a simple, complete network plotted with these packages, which would be produced if we had five tables each using the same variable. The reporting function displays the current graph of all tables and their relationships by variable,

along with the groupings in the console described previously.

Additional Features Developed Post Testing

A function that creates a mini-report on risks in relation to a single table. The function simply strips the adjacency matrix of all edges other than those connected directly to the target table, specified by the user.

During the case study [6](#) it was observed that the graph would get cluttered with edges. Include and Exclude options were added for the graphing function, allowing the user to include in the graph a specific set of variables only, or conversely exclude a set of variables.

The addition of a colour matrix improved the visual by allowing the user to modify the colours of the edges and edge labels by passing a list of 5 colours representing risk levels 1 to 5.

6 A Case Study

In order to demonstrate the functionality and test the effectiveness of the Differencing Tool package, a particular ongoing study was put forward as a good candidate for a case study. The Child Smile [33] programme promoted by NHS Scotland has a wealth of ongoing research to shape its promotions, programmes and publications. One such study is being carried out as a PhD at the University of Glasgow [34] by Mr. Ahmed Mahmoud, entitled: *Investigation of the role of ethnicity and socioeconomic factors in relation to dental health among children in Scotland*. This study began in 2016 and has produced a number of tabular outputs so far. These outputs will be used in a simulation of how the tool would be used as part of a live, newly initiated study from the point where the linked datasets have been created and the study folder structure is in place.

6.1 Preparation

The following steps are purely in preparation for this example, and not those that would be carried out during a live case.

Tables Metadata

Before being able to fully test the tool, some basic preparation must be carried out to simulate the initiation of a new study. Unlike a *live* use case, this study already has a number of tables in existence that metadata must be produced for. The tables were received by eDRIS within *MS Word*, *MS Excel* and *CSV* files over the course of 14 months from September 2017. The files were examined in order of their creation and transcribed into metadata via the template devised, with names assigned to the table as *table1*, *table2*, *table3*, *etc...* for simplicity. As the table names are presented within the report, simple and ordinal names are recommended. A total of 28 tables were examined before it was felt there was enough material to start demonstrating the package functions. This was a time consuming process, but when done at the point of table creation by the researcher it would not be expected to present a noticeable time burden. Some examples of the table metadata files are shown in figure 6.1.

Linked Data

A ‘fictitious’ version of the real linked *microData* for this case study was also produced (rather than applying for access to the real one!) by taking all the variables discovered during the tables metadata collection, and by randomly sampling the breakdowns, creating a completely invented version. This file is purely used to demonstrate a function within the tool and therefore the only properties that matter are the variable names, types and their range/classes. Figure 6.2 shows the first few lines of this dataset.

table_name	variable_name	derived_from	original_dataset	data_type	data_width	breakdown	n
table1	simd_quintile	simd_quintile	1516-0368_NDIPP1	categorical		5 1,2,3,4,5	3000
table1	Frequency	Frequency	1516-0368_NDIPP1	numerical		1 Frequency	3000

A	B	C	D	E	F	G	H
table_name	variable_name	derived_from	original_dataset	data_type	data_width	breakdown	n
table2	simd_quintile	simd_quintile	1516-0368_NDIPP1	categorical		5 1,2,3,4,5	3000
table2	overall_cat	overall_cat	1516-0368_NDIPP1	categorical		6 A,B,C,F,N,X	3000

A	B	C	D	E	F	G	H
table_name	variable_name	derived_from	original_dataset	data_type	data_width	breakdown	n
table3	simd_quintile	simd_quintile	1516-0368_NDIPP1	categorical		5 1,2,3,4,5	3000
table3	overall_cat	overall_cat	1516-0368_NDIPP1	categorical		2 Obvious Decay, No	3000

A	B	C	D	E	F	G	H
table_name	variable_name	derived_from	original_dataset	data_type	data_width	breakdown	n
table4	simd_quintile	simd_quintile	1516-0368_NDIPP1	categorical		5 1,2,3,4,5	3000
table4	intervention_type	intervention_type	1516-0368_NDIPP1	categorical		4 CHSP 6-8 Week Ass	3000

Figure 6.1: Sample of Case Study Tables Metadata

ID	Ethnic Group	Inspection_Attendance	intervention_type	NDIP linkage success	NDIP year	NDIPEXam_YesNo	NDIPRecord_Yes	Occupational social	Occupational social class	overall_cat	Schor
1	Missing	Parental Refusal	Diet	high	200809	NDIP Examination	NDIP Record	1 - Higher manageri	3 - Routine and manual c	A	F
2	White Gypsy	inspected	FVA	high	200809	No NDIP Examinat	NDIP Record	1 - Higher manageri	5 - Semi-routine and rou	A	A
3	Other	Child Refusal	CHSP 6-8 Week As	high	201415	No NDIP Examinat	NDIP Record	2 - Intermediate occ	7 - Students - not stated	C	E
4	Asian Other	Remove from list	CHSP 6-8 Week As	low	201112	No NDIP Examinat	NDIP Record	3 - Routine and mar	2 - Intermediate occupat	F	D
5	White Gypsy	Remove from list	CHSP 6-8 Week As	low	200910	NDIP Examination	NDIP Record	1 - Higher manageri	2 - Intermediate occupat	A	C
6	Pakistani	inspected	FVA	high	201213	NDIP Examination	No NDIP Record	1 - Higher manageri	3 - Small employers and	C	B
7	Bangladeshi	Absent	FVA	low	200809	NDIP Examination	NDIP Record	0 - Children born ou	0 - Children born outside	C	E
8	Other	Not Attending	CHSP 6-8 Week As	low	200809	No NDIP Examinat	No NDIP Record	3 - Routine and mar	2 - Intermediate occupat	B	C
9	White Other	inspected	CHSP 6-8 Week As	high	200809	NDIP Examination	No NDIP Record	3 - Routine and mar	0 - Children born outside	F	C
10	Other Arab	Not Attending	Diet	high	201112	No NDIP Examinat	NDIP Record	3 - Routine and mar	2 - Intermediate occupat	X	E
11	White Irish	Child Refusal	OHI	high	200910	No NDIP Examinat	NDIP Record	0 - Children born ou	6 - Never worked and lor	B	G
12	Missing	inspected	FVA	high	201112	NDIP Examination	No NDIP Record	1 - Higher manageri	4 - Lower supervisory an	B	F
13	Asian Other	Not Attending	Diet	high	201112	NDIP Examination	NDIP Record	2 - Intermediate occ	3 - Small employers and	A	D
14	White Scottis	Not Attending	Diet	high	201314	No NDIP Examinat	NDIP Record	1 - Higher manageri	5 - Semi-routine and rou	X	I
15	Caribbean / B	Absent	CHSP 6-8 Week As	low	200910	NDIP Examination	No NDIP Record	2 - Intermediate occ	2 - Intermediate occupat	F	H
16	Pakistani	Not Attending	FVA	low	201415	No NDIP Examinat	NDIP Record	1 - Higher manageri	3 - Routine and manual	c	B
17	Missing	inspected	Diet	low	201314	No NDIP Examinat	No NDIP Record	0 - Children born ou	3 - Small employers and	F	I
18	Not Known	Not Attending	CHSP 6-8 Week As	low	201314	NDIP Examination	NDIP Record	1 - Higher manageri	1 - Higher managerial,	ad	F
19	White Irish	Child Refusal	OHI	low	200910	NDIP Examination	NDIP Record	1 - Higher manageri	0 - Children born outside	A	H
20	White Scottis	Parental Refusal	FVA	low	201213	NDIP Examination	NDIP Record	0 - Children born ou	1 - Higher managerial,	ad	A
21	Missing	Not Attending	OHI	low	201314	NDIP Examination	NDIP Record	1 - Higher manageri	3 - Routine and manual	c	F
22	Pakistani	Child Refusal	FVA	high	201011	NDIP Examination	No NDIP Record	3 - Routine and mar	6 - Never worked and lor	B	M
23	White	Parental Refusal	Diet	low	200909	NDIP Examination	No NDIP Record	0 - Children born ou	6 - Never worked and lor	B	M

Figure 6.2: Fictitious Dataset Based on Case Study

6.2 User: Analyst

Project Initialisation

An *R Project* is initialised via *R Studio* within the Safehaven work space, stored in the study folder structure under some appropriate name. Options for the project such as saving environment *.RData* file upon exit automatically are set to ensure that any steps taken by the RCs during Disclosure Checking with the tool, that modify the global variables, are saved. The Differencing Tool package must be loaded and an initial configuration script can be run:

```
> library(differencing.tool)
> init_project()
```

This will create two empty *tibbles* that will contain the dataset metadata and the tables metadata, `dataset_metadata` and `tables_metadata`.

Linked Data Metadata Creation

The linked datasets will be stored in a folder within the *R Project* folders. Once the analyst has placed the dataset(s) into this folder, the first function to be run is `create_dataset_meta()` which takes `filename` parameter as the name of the dataset file, and outputs a *CSV* file to the same directory containing the dataset metadata. This folder currently defaults to `./datasets/`, though can be modified via function parameter `filepath`.

```
> create_dataset_meta(filename = '1516-0368_NDIPP1.csv')
```

This must then be updated with the sensitivity and risk ratings via manual editing. Figure 6.3 shows this file, with sensitivity ratings completed.

dataset_name	variable_name	data_type	data_width	sensitive_var	sensitive_geog	risk
1516-0368_NDIPP1	Ethnic Group	character	17	0	0	3
1516-0368_NDIPP1	Inspection_Attendance	character	6	0	0	2
1516-0368_NDIPP1	intervention_type	character	4	1	0	5
1516-0368_NDIPP1	NDIP linkage success	character	2	1	0	5
1516-0368_NDIPP1	NDIP year	numeric	7	0	1	5
1516-0368_NDIPP1	NDIPExam_YesNo	character	2	0	0	3
1516-0368_NDIPP1	NDIPRecord_YesNo	character	2	0	0	1
1516-0368_NDIPP1	Occupational social class NS-SEC3	character	4	0	0	1
1516-0368_NDIPP1	Occupational social class NS-SEC5	character	9	0	1	4
1516-0368_NDIPP1	overall_cat	character	6	1	0	5
1516-0368_NDIPP1	School census dataset linkage success result	character	12	0	0	1
1516-0368_NDIPP1	Sex	numeric	2	0	0	2
1516-0368_NDIPP1	simd_decile	numeric	10	0	1	3
1516-0368_NDIPP1	simd_quintile	numeric	5	1	0	5
1516-0368_NDIPP1	Repeat NDIP inspection	character	2	0	0	2
1516-0368_NDIPP1	Occupational social class (P1)	character	11	1	0	5

Figure 6.3: Dataset Metadata

Linked Data Metadata Addition to Environment

The final action required for the Analyst is to add this information into the environment variable `dataset_metadata`, figure 6.4 shows the effect on the *tibble* `dataset_metadata`

```
> dataset_metadata
# A tibble: 0 x 7
# ... with 7 variables: dataset_name <chr>, variable_name <chr>, data_type <chr>, data_width <dbl>, sensitive_var <dbl>, sensitive_geog <dbl>, risk <dbl>
> add_newdataset('1516-0368_NDIPP1_meta.csv')
Parsed with column specification:
cols(
  dataset_name = col_character(),
  variable_name = col_character(),
  data_type = col_character(),
  data_width = col_double(),
  sensitive_var = col_double(),
  sensitive_geog = col_double(),
  risk = col_double()
)
> dataset_metadata
# A tibble: 16 x 7
  dataset_name variable_name data_type data_width sensitive_var sensitive_geog risk
  <chr>         <chr>         <chr>      <dbl>      <dbl>      <dbl> <dbl>
1 1516-0368_NDIPP1 Ethnic Group character 17 0 0 3
2 1516-0368_NDIPP1 Inspection_Attendance character 6 0 0 2
3 1516-0368_NDIPP1 intervention_type character 4 1 0 5
4 1516-0368_NDIPP1 NDIP linkage success character 2 1 0 5
5 1516-0368_NDIPP1 NDIP_year numeric 7 0 1 5
6 1516-0368_NDIPP1 NDIPExam_YesNo character 2 0 0 3
7 1516-0368_NDIPP1 NDIPRecord_YesNo character 2 0 0 1
8 1516-0368_NDIPP1 Occupational social class NS-SEC3 character 4 0 0 1
9 1516-0368_NDIPP1 Occupational social class NS-SEC5 character 9 0 1 4
10 1516-0368_NDIPP1 overall_cat character 6 1 0 5
11 1516-0368_NDIPP1 School census dataset linkage success result character 12 0 0 1
12 1516-0368_NDIPP1 Sex numeric 2 0 0 2
13 1516-0368_NDIPP1 simd_decile numeric 10 0 1 3
14 1516-0368_NDIPP1 simd_quintile numeric 5 1 0 5
15 1516-0368_NDIPP1 Repeat NDIP inspection character 2 0 0 2
16 1516-0368_NDIPP1 Occupational social class (P1) character 11 1 0 5
> |
```

Figure 6.4: `add_newdataset` function and effect

The two functions `create_dataset_meta` and `add_newdataset` must be run for all linked datasets that are available to the study. In this case, there is only one and so the Analyst activities are now complete, *R Studio* would be closed and the Analyst would log off the Safehaven workstation.

6.3 User: Researcher

The study will begin and the researcher will start creating output that they would like released from the Safehaven. Any tabular output will be accompanied by the metadata files as described in the preparation section 6.1. In this example there are already 28 tables metadata available so the remainder of the workflow will be demonstrated with the Researchers steps excluded for brevity. With each new table(s) addition, a new *Disclosure Request* initiated the activity.

6.4 User: Research Coordinator

6.4.1 Disclosure Requests

Add Tables Metadata

The Research Coordinator receives a Disclosure Request via email and ServiceNow, including details of the filenames containing the tables and the metadata. The RC logs on to the Safehaven workstation and verifies the files exist in the correct folder `tables_metadata`. RC launches *R Project* where the variables are already available, the Differencing Tool library be loaded via `library(differencing.tool)`. RC runs `add_newtable` function with the list of new metadata files from the request. In this example, tables 1 through to 7 are included in the first request and so are added to the differencing tool environment. Figure 6.5 shows the resulting modifications to *tibble* `tables_metadata`.

```
> tables_metadata
# A tibble: 0 x 9
# ... with 9 variables: table_name <chr>, original_dataset <chr>, variable_name <chr>, derived_from <chr>, derived_from_sum <dbl>, data_type <chr>,
#   data_width <dbl>, breakdown <chr>, n <dbl>
> add_newtable(c('table1.xlsx', 'table2.xlsx', 'table3.xlsx', 'table4.xlsx', 'table5.xlsx', 'table6.xlsx', 'table7.xlsx'))
> tables_metadata
# A tibble: 14 x 9
  table_name original_dataset variable_name   derived_from   derived_from_sum data_type   data_width breakdown
  <chr>      <chr>          <chr>      <chr>          <dbl> <chr>      <dbl> <chr>
1 table1    1516-0368_NDIPP1 simd_quintile simd_quintile   1 categorical 5 1,2,3,4,5
2 table1    1516-0368_NDIPP1 Frequency      Frequency       1 numerical   1 Frequency
3 table2    1516-0368_NDIPP1 simd_quintile simd_quintile   1 categorical 5 1,2,3,4,5
4 table2    1516-0368_NDIPP1 overall_cat   overall_cat     1 categorical 6 A,B,C,F,N,X
5 table3    1516-0368_NDIPP1 simd_quintile simd_quintile   1 categorical 5 1,2,3,4,5
6 table3    1516-0368_NDIPP1 overall_cat   overall_cat     1 categorical 2 Obvious Decay, No Obvious Decay
7 table4    1516-0368_NDIPP1 simd_quintile simd_quintile   1 categorical 5 1,2,3,4,5
8 table4    1516-0368_NDIPP1 intervention_type intervention_type 1 categorical 4 CHSP 6-8 Week Assessment, Diet, FVA, OHI
9 table5    1516-0368_NDIPP1 simd_quintile simd_quintile   1 categorical 5 1,2,3,4,5
10 table5   1516-0368_NDIPP1 NDIPRecord_YesNo NDIPRecord_YesNo 1 categorical 2 NDIP Record,No NDIP Record
11 table6   1516-0368_NDIPP1 simd_quintile simd_quintile   1 categorical 5 1,2,3,4,5
12 table6   1516-0368_NDIPP1 NDIPExam_YesNo NDIPExam_YesNo 1 categorical 2 NDIP Examination,No NDIP Examination
13 table7   1516-0368_NDIPP1 simd_quintile simd_quintile   1 categorical 5 1,2,3,4,5
14 table7   1516-0368_NDIPP1 overall_cat   overall_cat     1 categorical 4 A,B,C,No Exam
> |
```

Figure 6.5: `add_newtable` function and effect

Generate Overall Report

The RC will now want to view a report on the tables showing which need to be cross referenced for Differencing Risk. The `report_tablesrisk()` function generates a report in console and a network representations, shown in figures 6.6 and 6.7, respectively. The report highlights that tables 2, 3 and 7 are using the same variable and have at least one different breakdown between them. The network shows a full view of the tables as nodes and the variables they use as connections between them, but has no representation of the breakdowns.

Table Specific Report

A second report type can be generated that modifies the network to show only edges connected directly to a specific table. The `report_table()` function generates the same console output as


```

> report_tablesRisk()
# A tibble: 3 x 3
  Tables_At_Risk Risk_Variable Risk_Breakdown
  <chr>          <chr>          <chr>
1 table2        overall_cat    A,B,C,F,N,X
2 table3        overall_cat    Obvious Decay, No Obvious Decay
3 table7        overall_cat    A,B,C,No Exam
> |

```

Figure 6.6: Console Output Report - First 7 tables

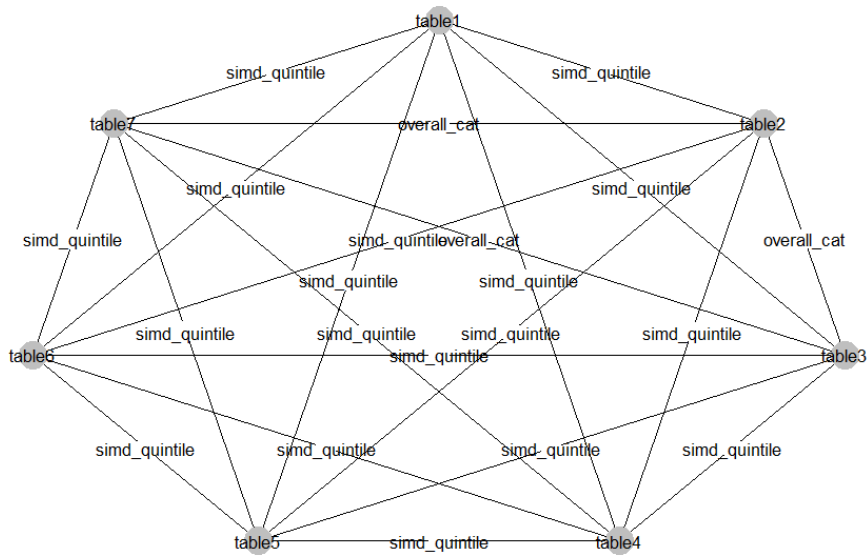


Figure 6.7: Network Representation - First 7 Tables

`report_tablesrisk()`, but the network representation is clearer with respect to the table passed to the report function, as shown in figure 6.8.

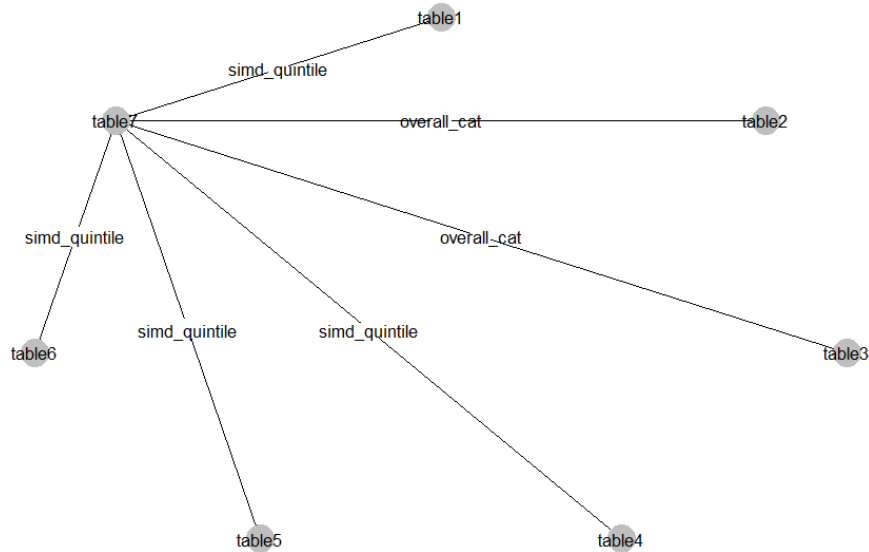


Figure 6.8: Network Specific to table 7

Variables Risk Report

A final report type is available via the `report_variablesrisk()` function that shows the risk levels of each variable in the datasets in use, based on the risk level found in the datasets metadata. This report has no network element, only an output to the console as shown in figure 6.9.

```

> report_variablesrisk()
[1] "Risk5"
# A tibble: 6 x 2
  dataset_name variable_name
  <chr>       <chr>
1 1516-0368_NDIPP1 intervention_type
2 1516-0368_NDIPP1 NDIP linkage success
3 1516-0368_NDIPP1 NDIP year
4 1516-0368_NDIPP1 overall_cat
5 1516-0368_NDIPP1 simd_quintile
6 1516-0368_NDIPP1 Occupational social class (P1)
[1] "Risk4"
# A tibble: 1 x 2
  dataset_name variable_name
  <chr>       <chr>
1 1516-0368_NDIPP1 Occupational social class NS-SEC5
[1] "Risk3"
# A tibble: 3 x 2
  dataset_name variable_name
  <chr>       <chr>
1 1516-0368_NDIPP1 Ethnic Group
2 1516-0368_NDIPP1 NDIPExam_YesNo
3 1516-0368_NDIPP1 simd_decile
[1] "Risk2"
# A tibble: 3 x 2
  dataset_name variable_name
  <chr>       <chr>
1 1516-0368_NDIPP1 Inspection_Attendance
2 1516-0368_NDIPP1 Sex
3 1516-0368_NDIPP1 Repeat NDIP inspection
[1] "Risk1"
# A tibble: 3 x 2
  dataset_name variable_name
  <chr>       <chr>
1 1516-0368_NDIPP1 NDIPRecord_YesNo
2 1516-0368_NDIPP1 Occupational social class NS-SEC3
3 1516-0368_NDIPP1 School census dataset linkage success result
  
```

Figure 6.9: Variables Risk Report

As the Study Progresses...

The outputs will continue to come via Disclosure Requests, and each time the above steps can be carried out to assess the differencing risk posed by new tables. With all 28 tables added to the environment, the view from the overall report is shown in figures 6.10 and 6.11, with some examples of table specific networks in figures 6.12 and 6.13. In this final example the benefit of the differencing tool becomes clearer. Let for example the most recent disclosure request cover tables 25 through to 28. The RC would previously browse through all the tables in the study, comparing the variables and confirming if a risk of differencing exists or not. With the tool, the RC can add the tables to `tables_metadata` and create the report which shows tables 25 and 26 pose no differencing risk within the study, whereas tables 27 and 28 must be checked against the groups reported in figure 6.10.

```
> report_tablesRisk()
# A tibble: 6 x 3
  Tables_At_Risk Risk_Variable Risk_Breakdown
  <chr>          <chr>          <chr>
1 table2        overall_cat    A,B,C,F,N,X
2 table3        overall_cat    Obvious Decay, No Obvious Decay
3 table7        overall_cat    A,B,C,No Exam
4 table8        overall_cat    A,B,C
5 table9        overall_cat    Obvious Decay, No Obvious Decay
6 table13       overall_cat    A,B,C,X
# A tibble: 2 x 3
  Tables_At_Risk Risk_Variable Risk_Breakdown
  <chr>          <chr>          <chr>
1 table4        intervention_type CHSP 6-8 Week Assessment, Diet, FVA, OHI
2 table10       intervention_type CHSP 6-8 Week Assessment, DHSW Contact, Diet, OHI, Dental Practice FVA, Nursery & School FVA, Toothbrushing Consent
# A tibble: 4 x 3
  Tables_At_Risk Risk_Variable Risk_Breakdown
  <chr>          <chr>          <chr>
1 table15       Ethnic Group    African,Asian Other,Bangladeshi,Caribbean / Black,Chinese,Indian,Mixed,Other,Other Arab,Pakistani,White Gypsy,White-
2 table24       Ethnic Group    African,Asian Other,Bangladeshi,Caribbean / Black,Chinese,Indian,Mixed,Other,Other Arab,Pakistani,White Gypsy,White Irish,White Other-
3 table27       Ethnic Group    African,Asian Other,Bangladeshi,Caribbean / Black,Chinese,Indian,Mixed,Other,Other Arab,Pakistani,White Gypsy,White Irish,White Other-
4 table28       Ethnic Group    African,Asian Other,Bangladeshi,Caribbean / Black,Chinese,Indian,Mixed,Other,Other Arab,Pakistani,White Gypsy,White Irish,White Other-
# A tibble: 2 x 3
  Tables_At_Risk Risk_Variable Risk_Breakdown
  <chr>          <chr>          <chr>
1 table19       Occupational    social class N~ 0 - Children born outside Scotland,1 - Higher managerial, administrative and professional occupations,2 - Intermediate-
2 table28       Occupational    social class N~ 0 - Children born outside Scotland,1 - Higher managerial, administrative and professional occupations,2 - Intermediate-
```

Figure 6.10: Full 28 Table Console Risk Report

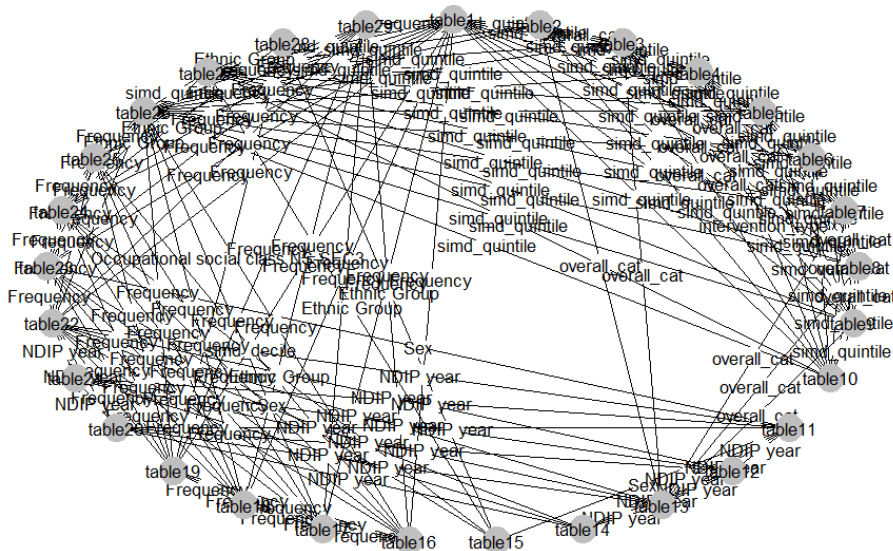


Figure 6.11: Full 28 Table Network

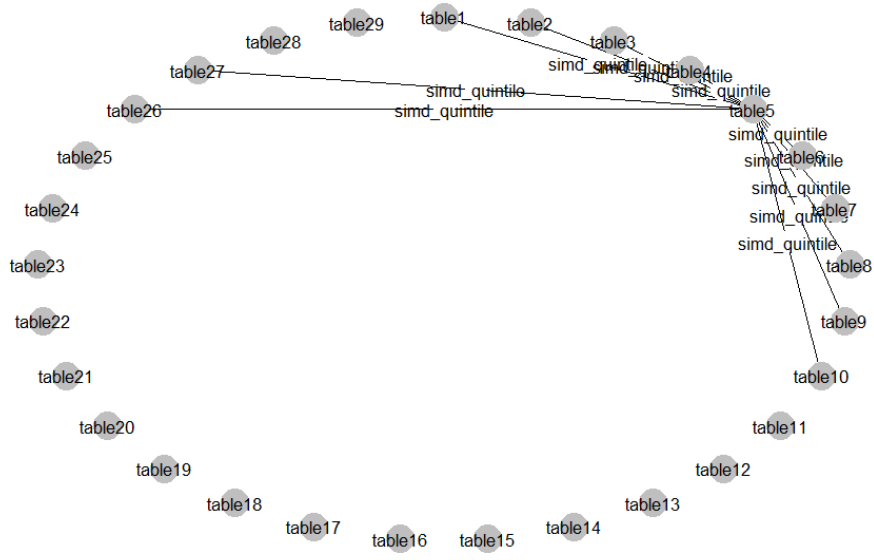


Figure 6.12: Table 5 Specific Network

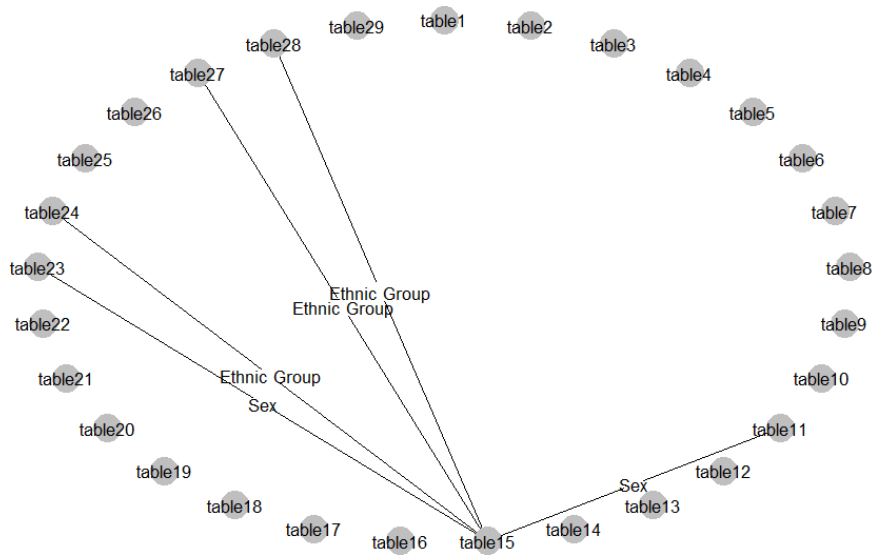


Figure 6.13: Table 15 Specific Network

6.4.2 Features Created Post-Testing

Some features were developed after the case study example was generated, in the final two days of the project.

Include and Exclude Options

As it was clear that the network visualisation was quickly getting cluttered, a method to modify this element was considered. The result was the creation of new parameters in the `report_tablesRisk` function to allow the user to strip out unwanted edges. The `include` parameter takes a list of strings representing those variables that the user would like included in the network, all other variables are ignored. The `exclude` parameter again takes a list of strings, however this list is removed from the network. Figures 6.14 and 6.15 show the effect of these modifications on the full network in 6.11.

The Colour of Magic

There was no element that made use of the variables *Risk* level in the visualisation, and it was considered that this may be a useful feature to give the network more life. The resultant development makes use of colours to represent the variables level of risk, a list of colours representing 1 to 5. A default colour scheme was created, where levels 1 to 3 would be *black* and levels 4 and 5 *red*. This default is modifiable via a parameter in `report_tablesRisk` function. Figures 6.16 and 6.17 shows this feature in conjunction with the `include` feature.

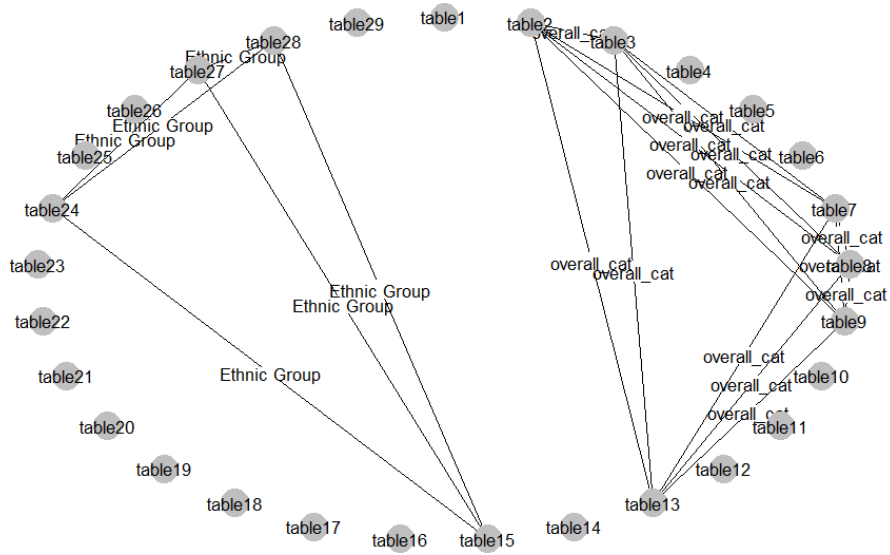


Figure 6.14: Include Param - report_tablesRisk(include=c('Ethnic Group','overall_cat'))

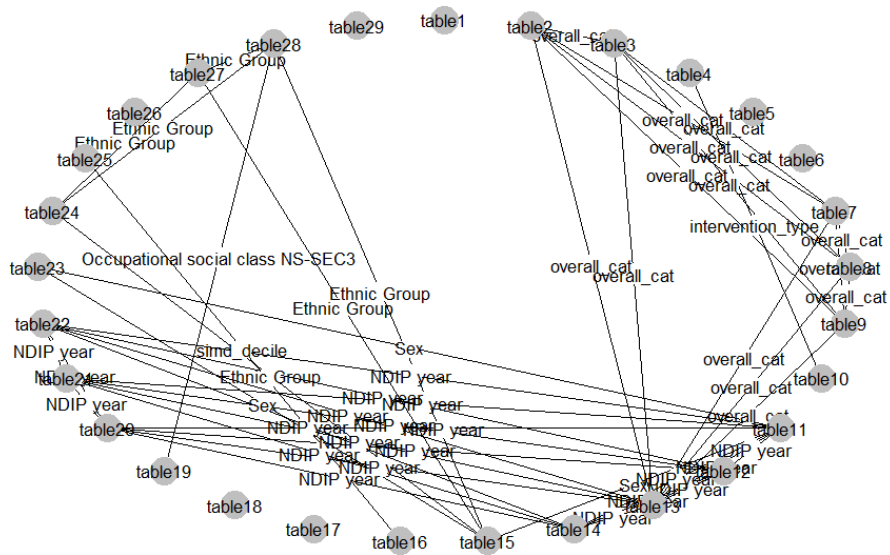


Figure 6.15: Exclude Param - report_tablesRisk(exclude=c('Frequency','simd_quintile'))

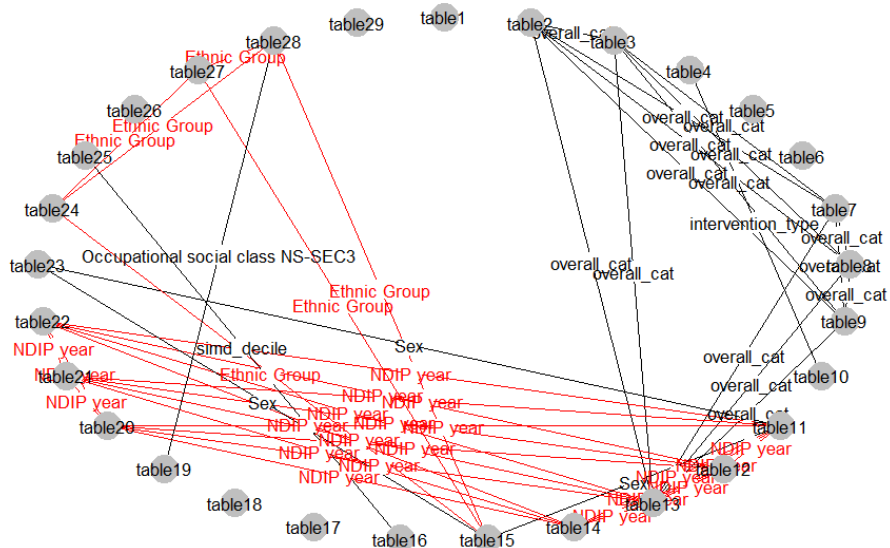


Figure 6.16: Exclude param - Default Colours

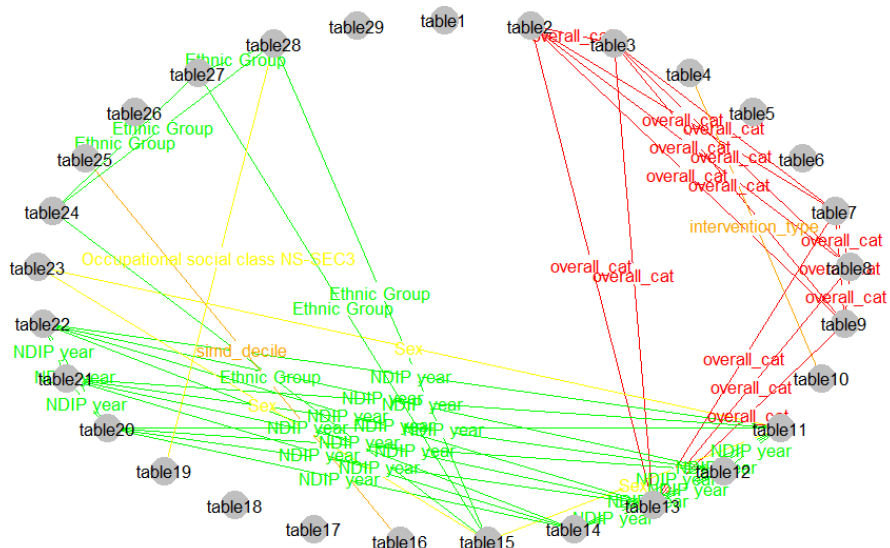


Figure 6.17: Exclude param - Custom Colours ('red','orange','yellow','green','blue')

7 Conclusion

7.1 Summary

The effective application of Statistical Disclosure Controls to the output produced by researchers is critical to ensuring the continuation of ethical and productive research in Health, and of course many other fields of research. It is highlighted in this work, and raised in most papers on the subject, that this is a difficult problem. The difficulty stems from the scale, variability and complexity of output produced, where no standards exist in terms of formatting and presentation, then taking into account the Data Environment it seems an almost insurmountable problem. Despite this, some incredible achievements have been made by groups around the world who are dedicated to bringing some order into this field. Allowing a categorisation approach to outputs, and identifying them as ‘Safe’ and ‘Unsafe’ so that appropriate time can be allocated to each achieves a good time saving for those involved in Disclosure Checking. The aim of this project was to examine the existing knowledge of SDC practices and find ways to reduce the time burden on the eDRIS team, with a focus on the aspect of *Disclosure Checking* the output that is to be taken out of the safe environment and used in publications. There are some recommendations that may make small savings in time for the team, and an *R Package* that will hopefully be useful to do the same, and potentially even a starting point for something bigger should the team have the resources to expand it.

7.2 Evaluation

The suggested solutions were received with interest by the team. Direct feedback on some of them was also received. The use of Synthetic Data in the way described may have potential, however it has a major weakness when dealing with data trends like those found in the some of the Scottish Morbidity Records. For example, SMR01 has relationships not only between the attributes, but also between the observations themselves. Currently synthetic data generation algorithms are not good at capturing this type of relationship in data. Potentially this will improve, as research into these algorithms is progressing quickly, driven somewhat by the Machine Learning field and the appetite for more training data. The Dual Sign-Off solution was supported by those I spoke to, but the difficulty lies in the fact that it implies an accepted risk that some disclosures may occur. With the sensitivity of the data in use, this is in fact not an acceptable risk for eDRIS. The R Package was received with enthusiasm:

‘The solution that Graeme has developed integrates into the tools and processes that we currently use and will very definitely reduce the burden of SDC.’ - Jackie Caldwell, Information Commissioner at eDRIS.

There are a number of improvements that were clear could be made given more time to develop the product. The tool does not handle tables that are based on the ‘frequency’ of a single variable, such as total counts of each ‘Ethnic Group’ within the study cohort. Currently these create a ‘false’ positive differencing risk as the ‘frequency’ variable is treated as the same variable across

the tables. This is in part solved by the *include* and *exclude* parameters that can be passed to the report function, but it would be an improvement if tables of this nature were handled separately. A solution that covered a larger portion of the disclosure checks that eDRIS carry out would have been more satisfying, however it is a starting point and introduces the power of visualisation techniques and how they may help keep track of output density within a study.

7.3 Future Work

The R package could be expanded in many ways. Currently the tool reports simply on tables that are using the same variable but with a different breakdown. The most important update, as described above, would be to separate ‘Frequency’ tables out of the core set and report on them based on the size of the cohort they present, removing the false positives. A further improvement would be in the extended use of the visualisations, to increase the information contained within the networks. Colours already represent variable risk, but they could also convey many aspects such as breakdown between variables, or tables with high degrees. The ‘node’ colour could be modified, as could the line thickness, all of which would improve the visualisation aspect of the tool.

Bibliography

- [1] Barabasi, A-L., Albert, R. *Emergence of Scaling in Random Networks*. Science vol 286. 1999
- [2] Cancho, R., Sole, R. *The Small World of Human Language*. Proceedings of the Royal Society B, Biological Sciences, vol 268. 2001
- [3] Chen, S., White, B. J., Sanderson, M. W., Amrine, D. E., Ilany, A., and Lanzas, C. *Highly dynamic animal contact network and implications on disease transmission*. Nature - Scientific Reports vol 4, Article 4472. 2014
- [4] Chen Y, Elliot M, Smith D. *The Application of Genetic Algorithms to Data Synthesis: A Comparison of Three Crossover Methods*, Privacy in Statistical databases. 2018
- [5] Domingo-Ferrer, J. *Data Anonymization: A Tutorial*, Unesco Chair in Data Privacy. 2014
- [6] Dwork, C. *Differential Privacy*. Encyclopedia of Cryptography and Security, 2011 Edition. 2006
- [7] Elliot, M. J., Dale, A. *Scenarios of Attack: The Data Intruder's Perspective on Statistical Disclosure Risk*, Netherlands Official Statistics, vol. 14. 1999
- [8] Elliot, M. J. and Domingo-Ferrer, J. *The Future of Statistical Disclosure Control*, The National Statistician's Quality Review. 2018
- [9] Elliot M. J., Lomax, S., Mackey E, Purdam K. *Data environment analysis and the key variable mapping system*, Privacy in Statistical Databases, vol. 6344. 2010
- [10] Elliot, M. J., O'Hara, K, Raab, C, O'Keefe, CM, Mackey, E, Dibben, C, Gowans, H, Purdam, K and McCullagh, K *Functional anonymisation: Personal data and the data environment*, Computer Law & Security Review, vol. 34. 2018
- [11] Griffiths, E., Greci, C., Kotrotsios, Y., Parker, S., Scott, J., Welpton, R., Wolters, A. and Woods C. *Statistical Disclosure Control Handbook version 1.0*. Guide produced by Secure Data Access Professionals Working Party. 2019
- [12] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Nordholt, E. S., Seri, G., de Wolf, PP. *ESSNet Handbook on Statistical Disclosure Control*, v1.2, Guide. 2010
- [13] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., de Wolf, PP. *Statistical Disclosure Control*, Book. 2012
- [14] Intensive Review Committee. *Appraisal of Census Programs*, Special Report. 1954
- [15] Kendrick, S. and Clarke, J. *The Scottish Record Linkage System*, Health Bulletin (Edinb). 1993
- [16] Mackey, E and Elliot, M. J. *Understanding the Data Environment*, XRDS: Crossroads, The ACM Magazine for Students - The Complexities of Privacy and Anonymity, vol. 20. 2013

- [17] Mourby, M., Mackey, E., Elliot, M. J., Gowans, H., Wallace, S. E., Bell, J., Smith, H., Aidinlis, S., Kaye, J. *Are pseudonymised data always personal data? Implications of the GDPR for administrative data research in the UK*, Computer Law & Security Review 34. 2018
- [18] Nixon, J. W. *History of the International Statistical Institute*, Book. 1960
- [19] Nowok, B., Raab, G. M., Dibben, C. *Bespoke Creation of Synthetic Data in R*, Journal of Statistical Software. 2017
- [20] Page, H., Cabot, C., Nissim, K. *Differential Privacy: An Introduction for Statistical Agencies*. A Contributing Article to the National Statistician's Quality Review into Privacy and Data Confidentiality Methods. 2018
- [21] Raghunathan, T. E., Reiter, J. P., Rubin, D. B. *Multiple Imputation for Statistical Disclosure Limitation*, Journal of Official Statistics. 2003
- [22] Ritchie, F. *Statistical Disclosure Detection and Control in a Research Environment*, Wales Institute of Social & Economic Research, Data & Methods. 2011
- [23] Ritchie, F. *Operationalising Safe Statistics: The Case of Linear Regression*, Economics Working Paper Series 1410. 2014
- [24] Rubin, D. B. *An Overview of Multiple Imputation*, Proceedings of the Survey Research Section, American Statistical Association 1988
- [25] Snoke, J., Raab, G. M., Nowok, B., Dibben, C. and Slavkovic, A. *General and Specific Utility Measures for Synthetic Data*, Journal of the Royal Statistical Society, Statistics in Society. 2018
- [26] **Website:** *Overleaf L^AT_EX Online*. Accessed September 2019 <https://www.overleaf.com>
- [27] **Website:** *Information Services Division Scotland*. Accessed August 2019 <https://www.isdscotland.org/About-ISD/Data-Collection/>
- [28] **Website:** *National Data Catalogue Scotland*. Accessed August 2019. <https://www.ndc.scot.nhs.uk/National-Datasets/>
- [29] **Website:** *Google Books N-Gram Viewer*. Access July 2019. <https://books.google.com/ngrams>
- [30] **Website:** *The Five Safes*. Accessed September 2019. <http://www.fivesafes.org>
- [31] **Website:** *Secure Data Access Professionals Working Party*. Accessed August 2019. <https://securedatagroup.org>
- [32] **Website:** *Office of National Statistics: Approved Researcher Scheme*. Access September 2019. <https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme>
- [33] **Website:** *Child Smile Program*. Accessed September 2019. <http://www.child-smile.org.uk/>

- [34] **Website:** *University of Glasgow Community Oral Health Research*. Accessed September 2019. <https://www.gla.ac.uk/schools/dental/research/communityoralhealth/>
- [35] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 2019 <https://www.R-project.org/>
- [36] Wickham, H. *Tidyverse: Easily Install and Load the 'Tidyverse'*. *R package version 1.2.1*. 2017 <https://CRAN.R-project.org/package=tidyverse>
- [37] Butts, C. *network: Classes for Relational Data. The Statnet Project* <http://www.statnet.org>. *R package version 1.13.0.1*. 2015 <https://CRAN.R-project.org/package=network>
- [38] Schloerke, B., Crowley, J., Cook, D., Briatte, F., Marbach, M., Thoen, E., Elberg, A., and Larmarange, J. *GGally: Extension to 'ggplot2'*. *R package version 1.4.0*. 2018 <https://CRAN.R-project.org/package=GGally>

Appendix 1 – R Package Creation Notes

Creating an *R Package* is not very different from creating a simple project in *R Studio*. The primary differences come in some extra configuration steps and the style in which external packages are used.

Steps: Create New Project, select Package, choose name and directory.

Initialise the environment using some tools that make package development easier, *devtools* and *usethis* libraries. To convert simple comment lines into formatted documentation:

```
devtools::document()
```

To utilise pipe operators, such as `\%$>$\%`, within the functions:

```
usethis::use_pipe()
```

To utilise any external libraries:

```
usethis::use_package('dplyr')
```

NOTE that this last command must be entered for each library that is going to be invoked by the package.

The environment is now ready. Good practice followed by the eDRIS *R Developers* is to group similar functions together in a single script file, therefore three script files contain the functions created for the differencing tool; Import Functions, Graphing Functions and Reporting Functions'

A principle to adhere to when writing the package is to ensure the users environment is not modified. The functions used in the differencing tool package employ functions from other R libraries, such as Tidyverse. For the differencing tool functions to work, the user must at least have the required libraries installed, but they should not need to have them loaded. In order to circumvent the loading of dependencies, R allows the libraries to instead be referenced directly within the function use double colon (:). For example, a function might use the *read_csv* function from *readr* library. To use this function without loading the whole *readr* library, the command *readr::read_csv* is used.

Appendix 2 – User guide

A number of functions exist in the Differencing.tool package. There are User Functions that will be invoked by the user directly, and Internal functions that are nested within the user functions and are not intended to be accessed directly. A typical study workflow using the tool will look like this:

1. Study Invoked, Linked Data and Folders created
2. R Project created within study folders
3. User runs `init_project()`
4. User runs `create_dataset_meta(filename,filepath)`
5. User edits datasets_metadata CSV files to include risk information
6. User runs `add_newdataset(filename,filepath)`
7. Researcher creates tabular metadata
8. User runs `add_newtable(filename,filepath)`
9. User runs `report_na()`
10. User requests any NA values are corrected by researcher, and if required repeats steps 8 and 9 again, remembering to `remove_table(tables)` any tables that are to be replaced
11. User runs `report_tablesRisk()`
12. User appraises report and carries out Disclosure Checks
13. steps 7 through to 12 repeat until study is complete

User Functions

- `init_project()`
 - Function: Creates the initial *tibbles* that will contain the metadata.
 - Param: No Input
 - Output: Two *tibbles*: `tables_metadata` and `dataset_metadata`

- `create_dataset_meta(filename,filepath)`
 - Function: Creates CSV file containing Dataset Metadata
 - Param: `filename,string` - the dataset filename
 - Param: `filepath,string` - dataset path (default: `./datasets/`)
 - Output: A single CSV file with the same name as *filename* but with 'meta' appended.

- `add_newdataset(filename,filepath)`
 - Function: Adds datasets metadata to `datasets_metadata` tibble
 - Param: `filenames,vector` - a vector containing the metadata filenames as strings
 - Param: `filepath,string` - (default: `./datasets/`)
 - Output: Modifies `datasets_metadata`

- `remove_dataset(datasets)`
 - Function: Removes a dataset from `datasets_metadata`
 - Param: `datasets,vector` - a vector containing dataset names as strings
 - Output: Modifies `datasets_metadata`

- `add_newtable(filename,filepath)`
 - Function: Adds table metadata to `tables_metadata` tibble

- Param: filenames,vector - a vector containing the metadata filenames as strings
- Param: filepath,string - (default: ./tables_metadata/)
- Output: Modifies tables_metadata

- `remove_table(tables)`
 - Function: Removes a table from tables_metadata
 - Param: tables,vector - a vector containing table names as strings
 - Output: Modifies tables_metadata

- `report_tablesRisk()`
 - Function: Produces a full report on tables vs variables and breakdowns risk
 - Param: include, vector - modifies network to display only those variables in include
 - Param: exclude, vector - modifies network to not display those variables in exclude
 - Param: colours, vector - modifies default colour scheme ('black','black','black','red','red')
 - Output: Prints to Console and creates network visualisation

- `report_table(table)`
 - Function: Produces same console report as `report_tablesRisk()` but has a reduced network based on table
 - Param: table, string - table name
 - Output: Prints to Console and creates network visualisation

- `report_na(metadata)`
 - Function: Reports if there are any NA values in the metadata
 - Param: metadata, tibble - target metadata to report on
 - Output: Prints report to console

- `report_variablesrisk()`
 - Function: Produces a report categorising the variables in use by their risk level
 - Param: None
 - Output: Prints report to console

- `report_full(outfile)`
 - Function: Produces a full study report file based on an Rmarkdown template
 - Param: outfile, string - name of the output file
 - Output: HTML file

Internal Functions

- `reset_matrix(metadata,node_column,filler)`
 - Function: Creates a square matrix for building network object
 - Param: metadata,tibble - the target metadata
 - Param: node_column,integer - which column in metadata should be used as the row/-column labels
 - Param: filler,string or integer - value which will fill the matrix.
 - Output: Returns matrix object

- `adjmatrix_complete(metadata,edge_column,node_column,include,exclude,colours)`
 - Function: Create Adjacency Matrix
 - Param: metadata,tibble - the target metadata

- Param: `edge_column,integer` - which column in metadata should be used as the network edge information
 - Param: `node_column,integer` - which column in metadata should be used as the row/-column labels
 - Param: `include,vector` - list of values to include from the `edge_column`
 - Param: `exclude,vector` - list of values to exclude from the `edge_column`
 - Param: `colours,vector` - list of 5 colours as strings representing Risk level of variable
 - Output: Returns a list of two matrices, an adjacency matrix for creating a network and an equivalent matrix for labelling the edges.
- `adjmatrix_singletable_impact(M,table)`
 - Function: Create Adjacency Matrix with edges only connecting to table
 - Param: `table,string` - the table to target
 - Output: Adjacency Matrix
- `labelled_graph(M_edges,M_labels)`
 - Function: Draws a labelled network
 - Param: `M_edges,matrix` - the adjacency matrix for the network
 - Param: `M_labels,matrix` - the equivalent matrix with edge labels
 - Output: Draws Network

Appendix 3 – Installation guide

The *R Package* is contained within a `tar.gz` source file and is installed with the following command:

```
install.packages("<pathToFile/example.tar.gz"> ,  
                repos=NULL, type="source")}
```

Which can be issued from within *R Studio* or any *R Console*.