

# Disclosure review and de-identification of quantitative data

Cristina Magder

UK Data Archive

UK Data Service

University of Essex

Managing and sharing research data: What is new with the GDPR?

4 May 2018

---

UK Data Service

---



University of Essex



# Protecting Confidentiality – the “5 Safes”

---

The ‘**Five Safes**’ framework enables **safe access to data** that meet the needs of **data protection** yet fulfils the demands for **open science and transparency**

**Safe data** - treat data to protect confidentiality

**Safe people** - educate researchers to use data safely

**Safe projects** - research projects for ‘public good’

**Safe settings** – Secure Lab system for sensitive data

**Safe outputs** – Secure Lab projects outputs screened



[5 Safes Animation](#)

---

UK Data Service



# Type of identifiers

---

**Direct identifiers:** directly identify participants such as:

- **names**
- **Addresses (physical, e-mail)**
- **telephone numbers**
- **account numbers**
- **vehicle identifiers**

**Indirect identifiers:** can provide enough information in combination to deduce identity of participants:

- **Sensitive information:** medical conditions; crime records; income etc. (consider access control)
- **Less sensitive?:** age; education; employment; household size; geographic area etc.

# Disclosure and Statistical Disclosure Control

---

**De-identification** – refers to a process of removing or masking *direct identifiers* in personal data

**Anonymisation** - refers to a process of ensuring that the risk of somebody being identified in the data is negligible. This invariably involves doing more than simply de-identifying the data, and often requires that data be further altered or masked. Anonymisation allows data to be shared ethically and legally while preserving confidentiality

1. Formal Anonymisation
2. Guaranteed Anonymisation
3. Statistical Anonymisation
4. Functional Anonymisation



# Statistical Disclosure Control and Disclosure Events

The notion of statistical anonymization is tied into a technical field called *statistical disclosure control (SDC)*

The basic principle of SDC is that it is impossible to reduce the probability of re-identification to zero, and so instead one needs to control or limit the risk of disclosure events.

**Disclosure** occurs when a person/organisation (also known as an **intruder**) **uses published data** in order to **find and reveal sensitive and/or unknown information** about a data subject



The **main goal of SDC is to minimize the potential risk of disclosure** to an appropriate level **while sharing as much data as possible**

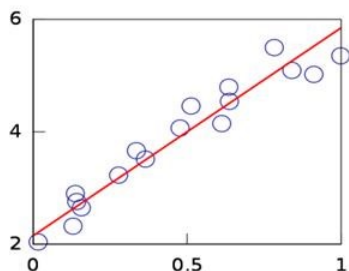
# Types of Disclosure



**Identity Disclosure** – the intruder successfully associates an individual/organisation with the released data – through direct identifiers, combinations of key identifying variables, linking with other available data (outliers and rare combinations increase the risk)



**Attribute Disclosure** – the intruder is able to determine unknown/sensitive features of an individual/organization based on the info in the released data – high certainty –



**Inferential Disclosure** - the intruder is able to determine more accurately sensitive features of an individual/organization with the use of the released data than it would be possible without - low certainty –

# Why is Disclosure Review important?

---



**Response Rates**



**Legal**



**Financial**



**Reputation**



**Future Research**

# SDC Freeware

---

- **R tool - sdcMicro (scripting + GUI)**
  - R package (and free dependable software R and RStudio)
  - [reference manual](#)
  - new Shiny GUI – in [detailed vignette](#)
  - [SDC methods and sdcMicro](#)
- **μ-Argus**
  - standalone software recommended by Eurostat for government statisticians
  - [software and manual](#)
- **ARX**
  - comprehensive open source software for anonymizing sensitive personal data
  - [software and documentation](#)





# Evaluation Tools

---

- Useful for **providing comparison** between SDC methods
- **Quick and easy** to explore what changes have biggest effect
- More **problematic** when trying to **define absolute risk** - the numbers might have no real meaning –



# Variables from the SDC Perspective

---

**Direct identifiers** – variables that identify directly an individual such as NINo, names, addresses, IP, etc. . Beware of string variables (open ended questions, “other” string variables)

**Key variables** – combinations of variables (indirect identifiers) that when taken together can identify a respondent. Geographical scope and sample size are of extreme importance as disclosure risk will be highly dependable on them

**Non-identifying variables** – neither direct nor indirect identifiers (however in the bigger picture each piece of data does contribute to the disclosure risk of a dataset)

# Exercise: Type of Variables

---

**Occupation**

**VAT Number**

**Ethnicity**

**Gender**

**Highest  
Qualification**

**Newspaper  
Readership**

**Local  
Authority**

**Car  
Ownership**

**Age**

# Exercise: Type of Variables

---

**Direct Identifiers:**

**VAT Number**

**Key Variables:**

**Occupation**

**Ethnicity**

**Highest  
Qualification**

**Gender**

**Local  
Authority**

**Age**

**Non-identifying  
variables:**

**Newspaper  
Readership**

**Car  
Ownership**



# De-identification

---

Currently no available automated software to detect direct identifiers.

However:

- Understanding and knowing the data (have **good metadata**)
- Paying attention to **string variables**
- Packages like SPSS provide the find (**find and replace**) function
- **Frequencies, frequencies, frequencies!**

# SDC Approaches

---

Types of SDC approaches:

**Perturbative methods** (*data-distortion controls*) - including an element of error

- adding noise - typically continuous variables
- micro-aggregation - continuous variables
- post-randomization (PRAM) - categorical variables
- cell or value suppression

**Non-perturbative methods** (*metadata-level controls*) - detail reduction

- recoding
- sampling
- choice of variables

**Hybrid method:** local suppression

# Exercise: Key Variables

| Occupation                                     |     |
|--|-----|
| Accountant                                     | 67  |
| Ambulance Officer                              | 56  |
| Baker  | 97  |
| Cook   | 109 |
| Electrical Engineer                            | 51  |
| Locksmith                                      | 32  |
| Middle School Teacher                          | 46  |
| Plumber  | 19  |
| Pressure Welder                                | 11  |
| Registered Nurse (Aged Care)                   | 22  |
| Registered Nurse (Child and Family Health)     | 14  |
| Registered Nurse (Community Health)            | 9   |
| Registered Nurse (Critical Care and Emergency) | 7   |
| Roof Plumber                                   | 2   |
| Secondary School Teacher                       | 36  |
| Software Engineer                              | 105 |
| Statistician                                   | 21  |
| Zookeeper                                      | 41  |

Which categories pose a disclosure risk?

# Exercise: Key Variables

| Occupation               |     |
|--------------------------|-----|
| Accountant               | 67  |
| Ambulance Officer        | 56  |
| Baker                    | 97  |
| Cook                     | 109 |
| Electrical Engineer      | 51  |
| Locksmith                | 32  |
| Middle School Teacher    | 46  |
| Plumber                  | 21  |
| Pressure Welder          | 11  |
| Registered Nurse         | 52  |
| Secondary School Teacher | 36  |
| Software Engineer        | 105 |
| Statistician             | 21  |
| Zookeeper                | 41  |



# Exercise: Key Variables

| Ethnicity   |     |
|---|-----|
| White - English/Welsh/Scottish/Northern Irish/British   | 339 |
| White - Irish   | 210 |
| White - Gypsy or Irish traveller                        | 1   |
| Any other White   | 3   |
| Mixed - White and Black Caribbean                       | 120 |
| Mixed - White and Black African                         | 170 |
| Any other Mixed   | 42  |
| Asian / Asian British - Indian                          | 87  |
| Asian / Asian British - Pakistani                       | 82  |
| Asian / Asian British - Bangladeshi                     | 63  |
| Asian / Asian British - Chinese                         | 61  |
| Any other Asian   | 22  |
| Black / African / Caribbean / Black British - African   | 89  |
| Black / African / Caribbean / Black British - Caribbean | 102 |
| Any other Black / African / Caribbean                   | 82  |
| Other ethnic group - Arab                               | 31  |
| Any other ethnic group                                  | 1   |

# Exercise: Key Variables

| Ethnicity   |     |
|---|-----|
| White - English/Welsh/Scottish/Northern Irish/British   | 339 |
| White - Irish   | 210 |
| White - Gypsy or Irish traveller                        | 0   |
| Any other White   | 0   |
| Mixed - White and Black Caribbean                       | 120 |
| Mixed - White and Black African                         | 170 |
| Any other Mixed   | 42  |
| Asian / Asian British - Indian                          | 87  |
| Asian / Asian British - Pakistani                       | 82  |
| Asian / Asian British - Bangladeshi                     | 63  |
| Asian / Asian British - Chinese                         | 61  |
| Any other Asian   | 22  |
| Black / African / Caribbean / Black British - African   | 89  |
| Black / African / Caribbean / Black British - Caribbean | 102 |
| Any other Black / African / Caribbean                   | 82  |
| Other ethnic group - Arab                               | 31  |
| Any other ethnic group                                  | 0   |

# Before applying any SDC methods

---

What is the **disclosure scenario**?

- **Who?**
- **What?**
- **Why?**

What are the **needs of secondary users**?

What are the key characteristics of the data:

- **Population vs Sample**
- **Geographical Coverage**
- **Level of non-response**

What is the **acceptable risk**?

Intended outcome: **Safe data with high utility**

# Key Variables – combinations and patterns

---

**Gender**

**Age**

**Ethnicity**

**Occupation**

**Highest  
Qualification**



# Uniqueness

---

**Uniqueness** - one of the fundamental concepts in disclosure review

**A record is unique on a set of key variables if no other record shares its combination of values for those variables.**

Two types of uniqueness on a set of key variables:

- **population uniqueness** — a unit is unique in the population (or within a population data file such as a census);
- **sample uniqueness** — a sample unit is unique within the sample file.

# K-anonymity

---

**K-anonymity** is a statistical method used to **evaluate whether or not a case is unique in the data**; is a hybrid disclosure risk assessment and disclosure control technique.

“A dataset is regarded as k-anonymised if – on all sets of key variables – each combination of possible values of those variables has at least k records that have that combination of values”

Depending on the level of access of a study the k is usually set to 3 or 5. Example: if k=3 all the patterns that are unique in the data (f=1) and the patterns that have a duplicate (f=2) will violate 2- and 3-anonymity assumptions and will be considered to pose a higher risk of disclosure.

patterns=records/cases with the same key identifiers

# K-anonymity

| Gender | Ethnicity   | Occup. | fk |
|--------|---|--------|----|
| Female | White - English/Welsh/Scottish/Northern Irish/British | Baker  | 4  |
| Female | Asian / Asian British - Indian                        | Baker  | 2  |
| Female | White - English/Welsh/Scottish/Northern Irish/British | Baker  | 4  |
| Female | Asian / Asian British - Indian                        | Baker  | 2  |
| Female | Asian / Asian British - Pakistani                     | Baker  | 2  |
| Female | White - English/Welsh/Scottish/Northern Irish/British | Baker  | 4  |
| Female | Asian / Asian British - Pakistani                     | Baker  | 2  |
| Female | White - English/Welsh/Scottish/Northern Irish/British | Baker  | 4  |
| Female | Black / African / Caribbean / Black British - African | Baker  | 3  |
| Female | Black / African / Caribbean / Black British - African | Baker  | 3  |
| Female | Black / African / Caribbean / Black British - African | Baker  | 3  |



ce



# K-anonymity

| Gender | Ethnicity   | Occup. | fk |
|--------|---|--------|----|
| Female | White - English/Welsh/Scottish/Northern Irish/British | Baker  | 4  |
| Female | Asian / Asian British                                 | Baker  | 4  |
| Female | White - English/Welsh/Scottish/Northern Irish/British | Baker  | 4  |
| Female | Asian / Asian British                                 | Baker  | 4  |
| Female | Asian / Asian British                                 | Baker  | 4  |
| Female | White - English/Welsh/Scottish/Northern Irish/British | Baker  | 4  |
| Female | Asian / Asian British                                 | Baker  | 4  |
| Female | White - English/Welsh/Scottish/Northern Irish/British | Baker  | 4  |
| Female | Black / African / Caribbean / Black British - African | Baker  | 3  |
| Female | Black / African / Caribbean / Black British - African | Baker  | 3  |
| Female | Black / African / Caribbean / Black British - African | Baker  | 3  |



# Exercise: Key Variables – k-anonymity

1 record (fk=1):



2 records (fk=2)



3 records (fk=3)



16 records (fk=16)



# Key Variables – combinations: sample and population



Basildon

Occupation and highest level of education

Sample 70% of a population of 170,000

fk=1

Male

84

Bangladeshi

Baker

MA, MSc

fk=5

Male

65 or  
over

Baker

MA, MSc



England

Ethnicity, occupation , highest level of education

Sample 10% of a population of 5.3 million

fk=1

Male

84

Bangladeshi

Baker

MA, MSc

Fk=23

Male

84

Bangladeshi

Baker

MA, MSc

UK Data Service



Where fk= sample frequencies and Fk=population frequencies

# L-diversity

**L-diversity** is an **extension of k-anonymity** - there has to be at least  $l$  different values for each sensitive variable within each pattern on the key variables. If no sensitive variables are present in the data  $l$ -diversity is not needed.

Examples of sensitive variables:

- stress level: low, normal, high
- HIV positive: yes, no

| Gender | Ethnicity   | Occup. | HIV | L-diversity | Count |
|--------|---|--------|-----|-------------|-------|
| Female | Asian / Asian British                                 | Baker  | no  | 2           | 4     |
| Female | Asian / Asian British                                 | Baker  | no  | 2           | 4     |
| Female | Asian / Asian British                                 | Baker  | yes | 2           | 4     |
| Female | Asian / Asian British                                 | Baker  | no  | 2           | 4     |
| Female | Black / African / Caribbean / Black British - African | Baker  | Yes | 1           | 2     |
| Female | Black / African / Caribbean / Black British - African | Baker  | Yes | 1           | 2     |



# Information Loss

---



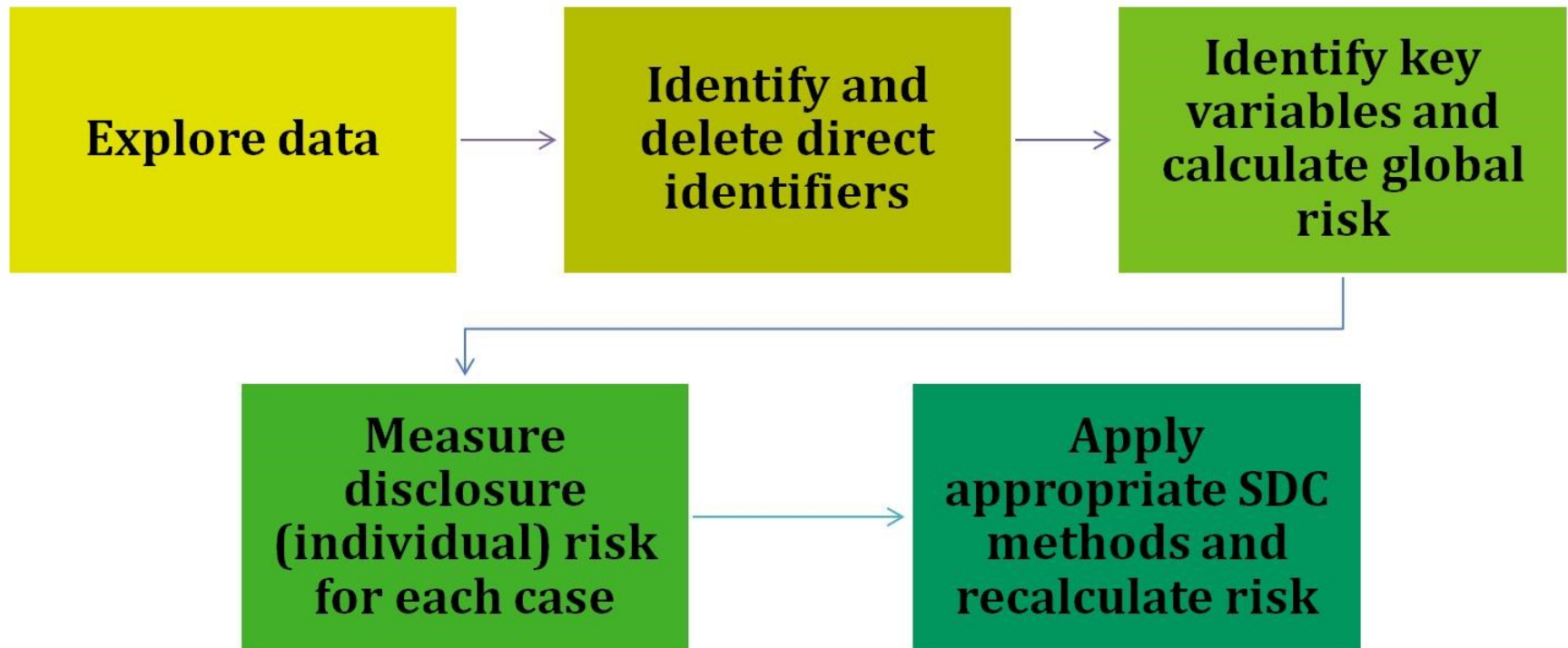
**Data quality** is important for all parties involved in this process from data producers to data providers and to users. So, while applying SDC methods the data utility and information loss must always be of **utmost importance**

The sdcMicro package/GUI **calculates the information loss each time a new SDC method is applied** to the data, which is one of the fantastic features that ease the process of SDC when using this package/GUI



# SDC Workflow

---



# Functional Anonymization

---

## Data does not exist in isolation

### A data environment

- **Data** – what (other) data exist in the data environment?
- **Data users** – who is accessing the data?
- **Governance processes** – are data access controls licensing arrangements and policies in place?
- **Infrastructure** – how is the data stored, exchanged and provided?



# Access conditions

---

## Open

- No real risk. Under open licence; almost no restrictions on reuse

## Safeguarded

- Zero to low real risk. Requires authentication and authorisation e.g. registered user and End User Agreement

## Controlled

- Real risk. Requires project approval, user vetting and training; access via a safe setting; output checking

# What is New with the GDPR?

---

Anonymisation is now part of the data-sharing toolkit, not only an ethical means to protect privacy

“A procedure that adds to the business case for data sharing” – Mark Eliot et al, 2018

Recital 26: ‘data rendered anonymous’



# Conclusion

---

SDC is a complex topic - **an active research field**

- Be aware of the **issues and considerations of statistical disclosure**
- Be able to **make principled judgements** about the disclosiveness of your data
- Be able to **find the balance** between minimizing disclosure risk and maintaining high data utility; what's the point in safe unusable data?

# Questions

---

