FORS

explore.understand.share.

Anonymisation of research data

Open Science Summer School 2022

Marieke Heers

Outline

1. Anonymisation: Concepts and definitions
2. Anonymising quantitative data
3. Anonymising qualitative data

Anonymization in the current research environment

- Digitalisation of data: more and more data are produced
- New research fields, including new types of data (Big Data)
- Computational power allows for analysis of increasingly rich datasets
- Facilitated access to data by the community
- New analytical/data extraction tools

Requirements

From funders:

- Data management plans (DMPs)
- Data sharing (in FAIR repositories)

From journals:

- Deposit of data used in publications
- Sufficient documentation

Anonymisation in data management

- Anonymisation is a key practice for protecting respondents and allowing data sharing.
- Anonymisation needs to be understood in light of the different legal and ethical requirements, but also in combination with other data management practices.

What are the main challenges you face regarding anonymisation?

Concepts and definitions

Anonymisation – A definition

- The notion of anonymisation refers to the process by which the elements allowing the identification of a person are definitively deleted from a dataset, a document, an interview transcript, etc.
- As a result, an individual cannot be identified *without significant effort*.
- Represents a principal solution for complying with data protection requirements.
- This is irreversible!

Anonymisation – A difficult promise

- Individuals are more unique than we might think!
- Crossing three simple variables, namely date of birth, postal code and gender, 63% of the US population can be identified (Golle, 2006).
- The collection of big data (via apps etc.) makes identification very easy due to the massive nature of unique data collected.
- The ability to cross-reference research data with other datasets, information from social networks, blogs, websites, etc. greatly facilitates (re)identification.
- Particularly relevant when working on small populations.

Anonymisation – Requirements

- The data itself and all options of recreating the original data are eliminated completely.
- The person can no longer be identified and the process is irreversible.
- Fully anonymized data is no longer considered personal data
- The effort to identify the data subject is too big
  - in terms of know-how
  - in terms of cost
- It is very difficult to have fully anonymised data.

Anonymisation vs. pseudonymisation

- Refers to the removal or replacement of identifiers with pseudonyms, which are kept separately and protected by technical and organisational measures.
- The data remain pseudonymous as long as the original identifying information exists.
- Is a means to enhance the security of the data you process by making the subject identifiable instead of directly identified.
- Pseudonymized data remain personal data.

https://www.fsd.uta.fi/aineistonhallinta/en/anonymization-and-identifiers.html

Anonymising data: Factors to be considered

1) The nature and type of personal data to anonymise
2) The future users of the data and conditions of use
3) Balancing utility and data protection
4) Risk management
5) What was promised to respondents

1) The nature and type of personal data to anonymise

- Sensitivity
- Sampling
- Duration
- Data from other sources

Sensitivity

How you anonymise will depend on the sensitivity of your data, that is, the extent to which your individual respondents might be harmed by their identification and the details that are revealed.

Sampling

- Sampling method
- Target population (size, exceptional or unique information)
- Response rate

Duration

- The older the data, the harder it is to identify respondents
- Longitudinal data present greater risks
- Longitudinal data also require consistent approaches

Linking to other data sources

- Information and research data about the same target population available elsewhere
- Publicly available information (e.g., public registers, social media)
- Local knowledge (e.g., what residential locations look like and what goes on in the area)
- Personal information about other people (what do I know, e.g., about my neighbours)
- Mixed methods require specific measures

2) Future users of the data and conditions of use

- Public release or restricted access
- Likely expertise of users
- Access conditions (e.g., with prior approval only, with user contract)

3) Balancing utility and data protection

- Increased protection implies decreased utility
- Expected analyses
- Variable level assessment
- What can be sacrificed?

4) Risk management

a. Motivations for an attack
b. Consequences of a disclosure
c. Disclosure without malicious intent (e.g., spontaneous identification)
d. How other data/knowledge might be linked to the data in question

There is no risk-free scenario. The goal should be to identify the acceptable level of risk for your project or your institution. This will help you define the needed levels of utility in relation to data protection issues.

5) Promises to respondents

- What you promise to respondents in a consent form is ethically and legally binding.

Setting up an anonymisation strategy

The strategy should be developed early in the project and include at least:

- an evaluation of disclosure risk, and
- a description of the anonymisation measures and their rationale.

This plan will serve as documentation, and should be updated after anonymization has been completed.

Setting up an anonymisation strategy

Relevant questions:

- What types of direct or indirect identifiers do my materials contain? Are there rare/unique information in the data?
- What combinations of variables can allow identification of an individual?
- Can information from other sources be linked to the data making identification possible? (social networks, blogs, etc.).
- What characteristics of the data do I want to retain (if possible) and which ones can be “sacrificed” in the anonymisation process?

Your experience

Thinking about your research projects

- What is the disclosure risk?
- What anonymisation measures have you applied? Why?

General principles and considerations

- Different anonymisation techniques are appropriate with different types of data.
- Different anonymisation techniques modify the dataset in different ways
- Risk should be reduced to an acceptable level.
- Preference to lighter techniques.
- Choosing the appropriate technique requires expertise with the subject matter.
- Each technique has advantages and limitations.

Quantitative data anonymisation

Direct and indirect identifiers

- Direct identifiers alone are sufficient to identify people (e.g., name, AVS number)
- Strong indirect identifiers allow fairly easy identification (e.g., home address, telephone number)
- Weak indirect identifiers allow identification through *combinations* of variables

Indirect identifiers: Socio-demographic variables

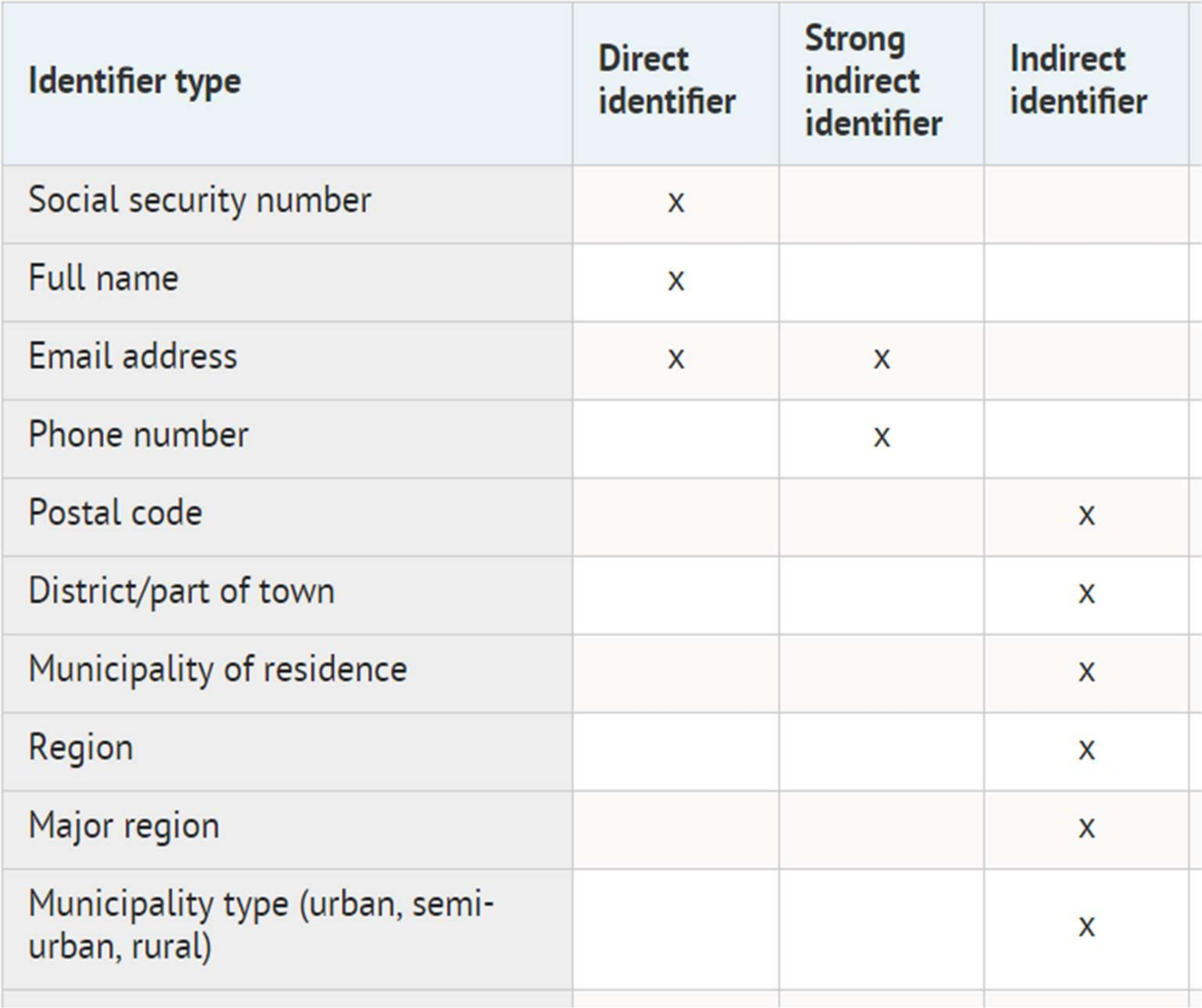
- Gender
- Age (DOB, MOB, YOB)
- Location (municipality, canton, main region, linguistic region)

- Civil status
- Nationality
- …

Basic approach

- Removal of direct and strong indirect identifiers
- Assessment of weak indirect identifiers and appropriate techniques
- Starting with a categorisation of variables

Categorisation of variables



Identifier type	Direct identifier	Strong indirect identifier	Indirect identifier
Social security number	x		
Full name	x		
Email address	x	x	
Phone number		x	
Postal code			x
District/part of town			x
Municipality of residence			x
Region			x
Major region			x
Municipality type (urban, semi-urban, rural)			x

Specific anonymisation techniques

- Variable suppression
- Record suppression
- Character masking
- Pseudonymisation
- Generalisation
- Data perturbation
- Swapping

Variable suppression

- Removal of an entire variable
- Extreme loss of information, so should be last resort
- First technique to apply
- Often used with sensitive open-ended questions
- Why collected information in the first place?

Record suppression

- Removal of an entire record that cannot easily be anonymized (e.g., an exceptional and easily identifiable individual)
- First assess whether other techniques might handle the problem (e.g. generalisation)
- In some cases, you can just suppress or alter a value for a variable within a record (e.g., an outlier)

What does this imply for your sample?

Character masking

- Change of the characters of a data value, using a constant symbol (e.g. “*” or “x”)
- Partial hiding within a string
- Replace a fixed or variable number of characters

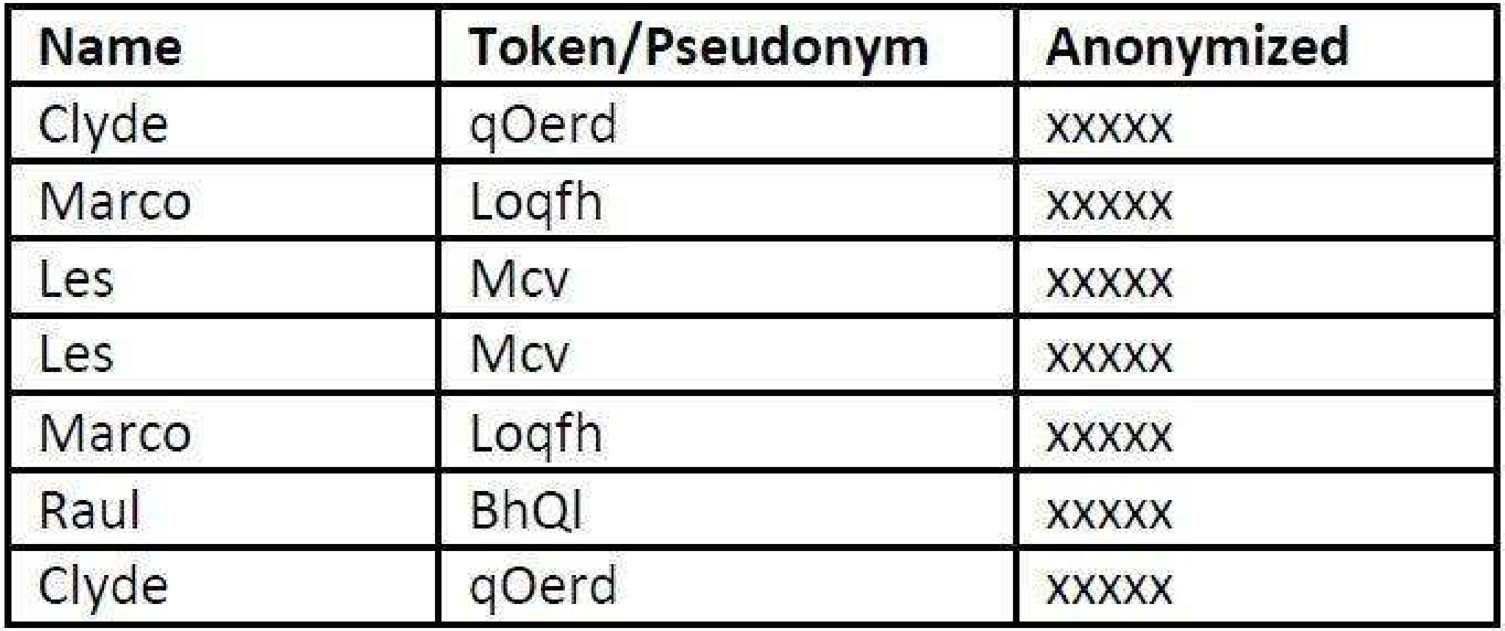
Example:

079 259 67 00 -> xxx xxx 67 00
078 452 83 14 -> xxx xxx 83 14

Pseudonymisation

- Replace identifying information with made-up values
- For cases where values must be uniquely distinguished
- Made-up values must be arbitrary and unique
- Can be reversible or irreversible
- Can be generated by software
- Often used to link individuals across datasets

Pseudonymization - example



Name	Token/Pseudonym	Anonymized
Clyde	qOerd	xxxxx
Marco	Loqfh	xxxxx
Les	Mcv	xxxxx
Les	Mcv	xxxxx
Marco	Loqfh	xxxxx
Raul	BhQl	xxxxx
Clyde	qOerd	xxxxx

Generalisation

- Reduction of precision of a variable
- Create discrete categories from a continuous variable
- Combine values into broader categories, e.g. age, professions, income, …

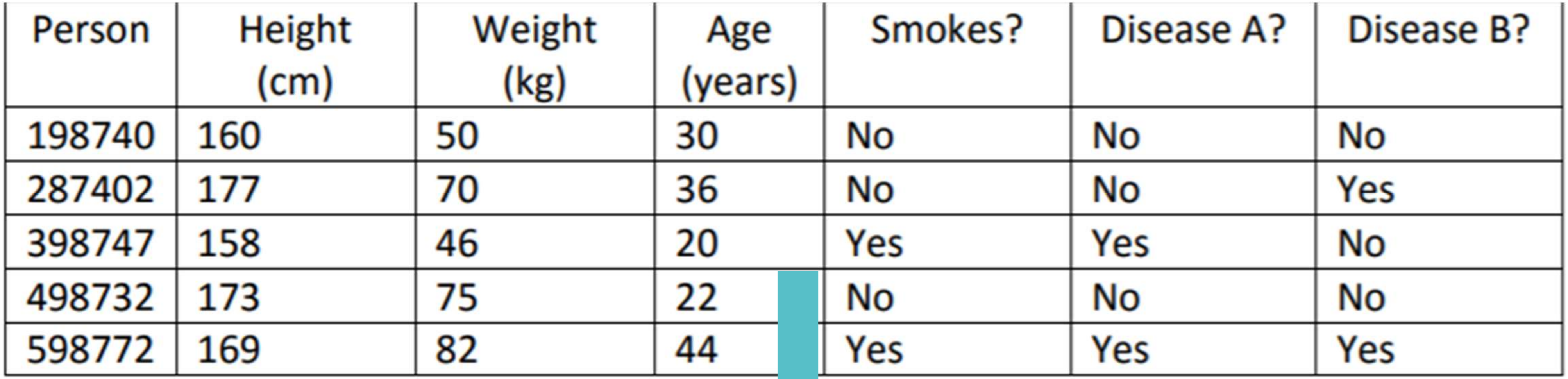
Generalisation – example

Commune	District
Aclens	Morges
Agiez	Jura - Nord vaudois
Arnex-sur-Orbe	Jura - Nord vaudois
Arzier-Le Muids	Nyon
Assens	Gros-de-Vaud
Ballaigues	Jura - Nord vaudois
Belmont-sur-Lausanne	Lavaux-Oron
Belmont-sur-Yverdon	Jura - Nord vaudois
Cheseaux-Noréaz	Jura - Nord vaudois
Jongny	Riviera - Pays-d'Enhaut
Jorat-Mézières	Lavaux-Oron
Moiry	Morges
Penthalaz	Gros-de-Vaud

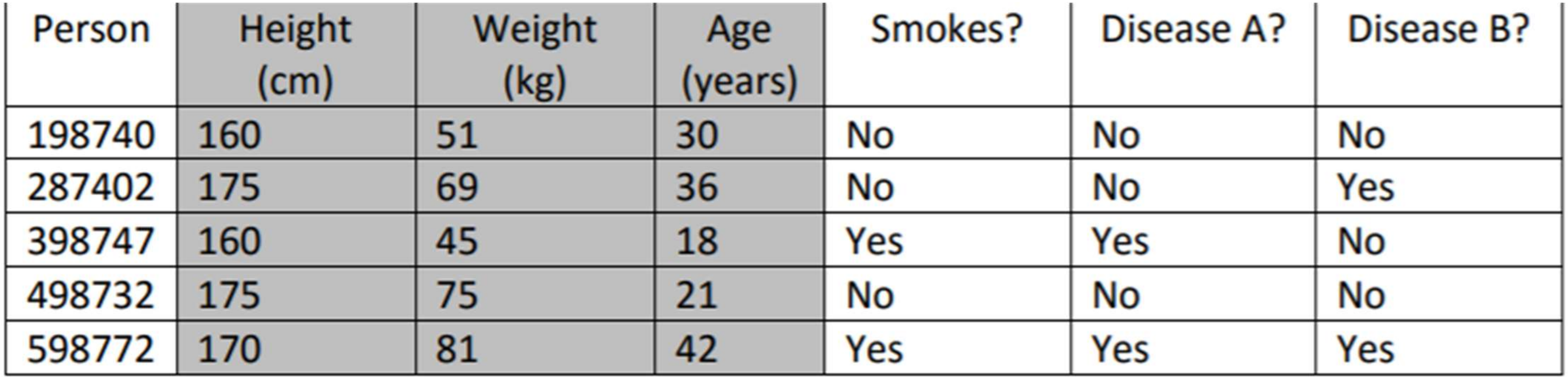
Data perturbation

- Modification of values to be slightly different
- Where small changes of value do not significantly affect analysis and accuracy
- Examples include base-x rounding and adding random noise

Example – base-x rounding



Person	Height (cm)	Weight (kg)	Age (years)	Smokes?	Disease A?	Disease B?
198740	160	50	30	No	No	No
287402	177	70	36	No	No	Yes
398747	158	46	20	Yes	Yes	No
498732	173	75	22	No	No	No
598772	169	82	44	Yes	Yes	Yes

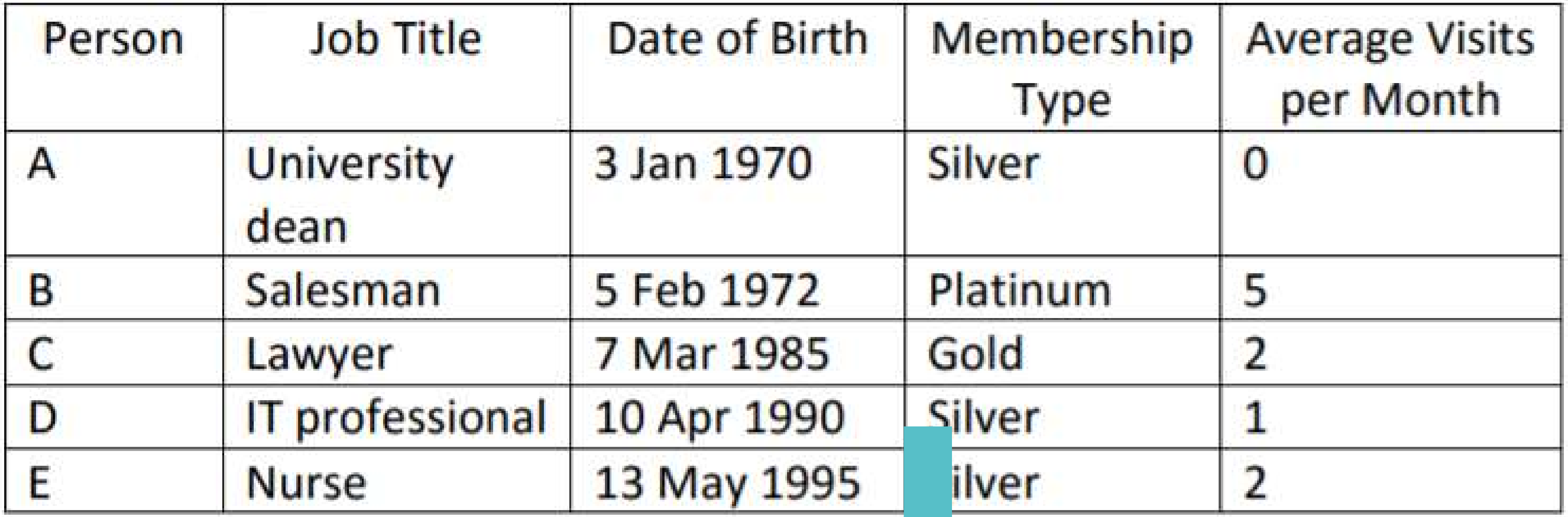


Person	Height (cm)	Weight (kg)	Age (years)	Smokes?	Disease A?	Disease B?
198740	160	51	30	No	No	No
287402	175	69	36	No	No	Yes
398747	160	45	18	Yes	Yes	No
498732	175	75	21	No	No	No
598772	170	81	42	Yes	Yes	Yes

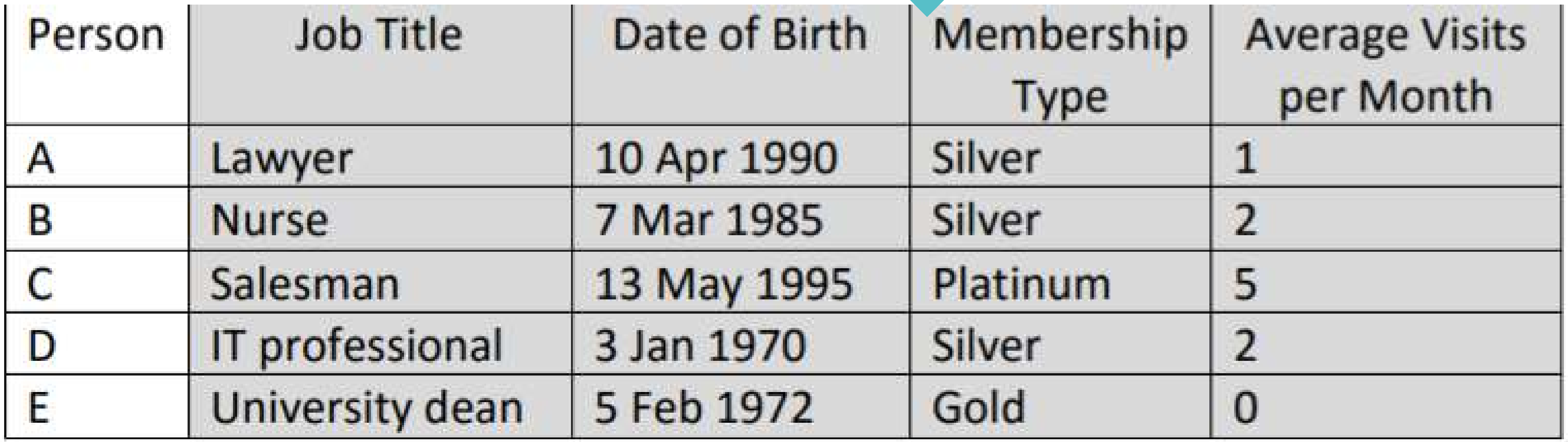
Swapping

- Rearrange data across records such that the individual variable values are still represented in the dataset
- Only to be used when analysis is on aggregate level, i.e., where there is no need to examine relationships between variables

Swapping



Person	Job Title	Date of Birth	Membership Type	Average Visits per Month
A	University dean	3 Jan 1970	Silver	0
B	Salesman	5 Feb 1972	Platinum	5
C	Lawyer	7 Mar 1985	Gold	2
D	IT professional	10 Apr 1990	Silver	1
E	Nurse	13 May 1995	Silver	2



Person	Job Title	Date of Birth	Membership Type	Average Visits per Month
A	Lawyer	10 Apr 1990	Silver	1
B	Nurse	7 Mar 1985	Silver	2
C	Salesman	13 May 1995	Platinum	5
D	IT professional	3 Jan 1970	Silver	2
E	University dean	5 Feb 1972	Gold	0

Qualitative data anonymisation

What is qualitative data anonymisation?

Qualitative data anonymisation is about rendering research participants anonymous by removing identifying information from the research data.

It tends to be more complex than anonymisation of quantitative data.

Data can be anonymized for at least two purposes:

- Publication
- Secondary analyses

Anonymisation techniques

- Replacing personal names with aliases
- Categorising proper nouns
- Changing or removing sensitive information
- Categorising background information
- Changing values of identifiers

Replacing names with aliases

- Changing proper nouns into aliases is the most common anonymisation technique.
- It is always a better option to use aliases rather than simply delete the names or replace them by a letter [x].
- It is important to be consistent in the selection and use of aliases throughout a research project.
- The same aliases should be used in both the data and the published excerpts.

Categorising proper nouns

Names of people who have no essential importance in understanding the data content can be removed from the data without creating aliases. These names can be replaced with broader categories such as:

[woman], [man],
[sister], [father],
[colleague, female], [neighbour male]

The same may apply to other proper nouns, such as institutions, names of places, etc.

[Lower secondary school], [restaurant]
[hometown], [residential area]

Categorising proper nouns (2)

- Large towns can usually remain [e.g., London].
- If it has been decided that the place of residence will not be revealed, remember to check the background information as well (e.g., the name of a specific restaurant could help reveal the place of residence).

Changing or removing sensitive information

- Identifying sensitive information should be removed, categorised or classified.
  - For example, if relevant to the subject matter, a rare disease could be recoded to [severe long-term illness] and thereafter referred to as [illness].
- Removing sensitive data is justified if the respondent mentioned it incidentally, if the information is not relevant to the subject matter, or if it constitutes a disclosure risk.

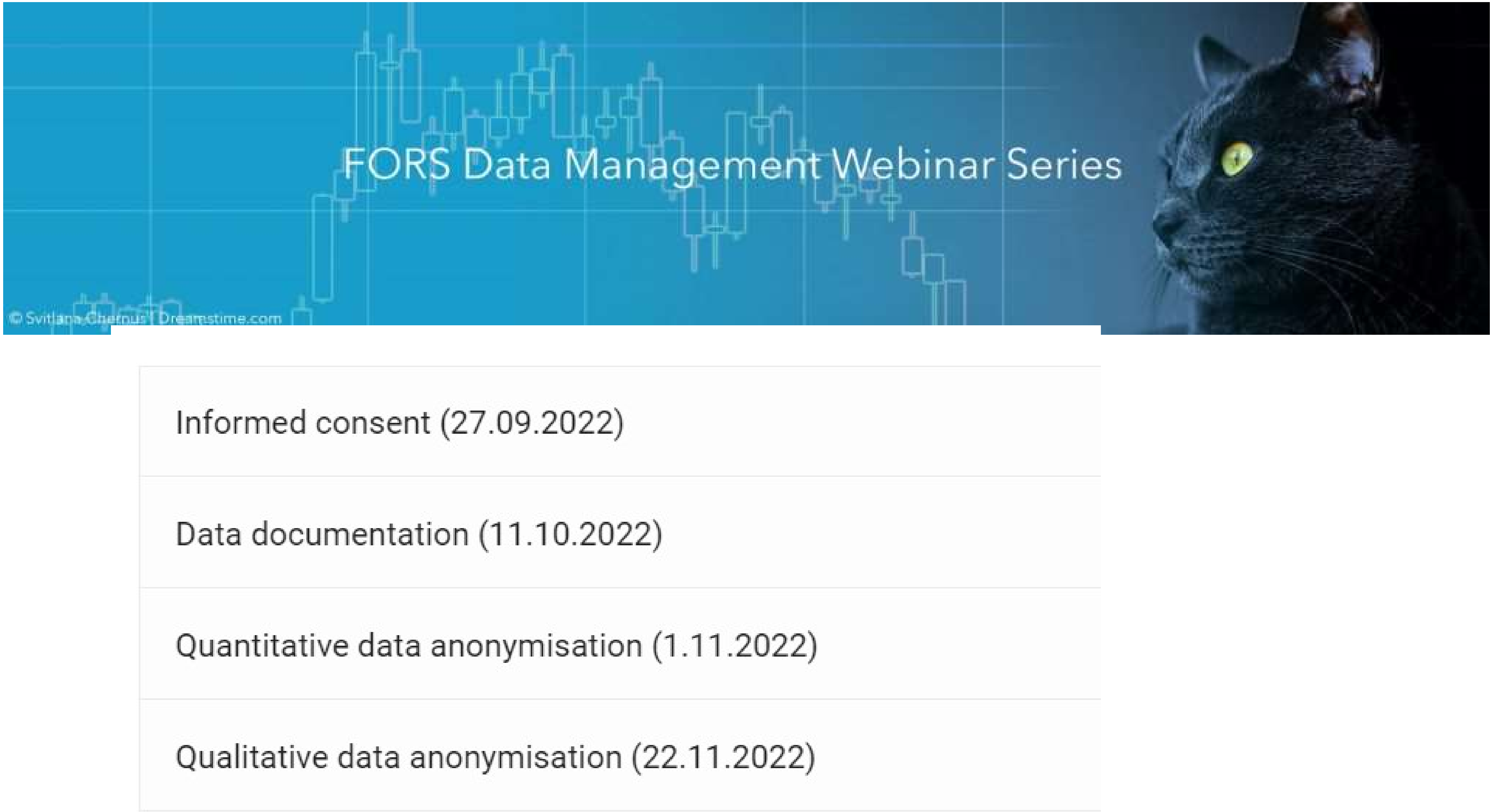
Best practices for anonymisation of qualitative data

- Do not collect disclosive data unless necessary
- Never disclose personal data – unless consent for disclosure
- Plan or apply editing at time of transcription
- Avoid blanking out; use pseudonyms or replacements
- Identify replacements, e.g. with [brackets]
- Mark the text that has been anonymised: XML tag <seg> or symbols @@ (at the start) and ## (at the end)
- Mark sensitive text that might need to be anonymised at a later date: $$ (at the start) and ## (at the end)
- Avoid over-anonymising – maintain maximum meaningful information

Some additional points to consider

- Anonymisation of research data should be considered together with consent agreements and access restrictions
- Regulating/restricting user access may offer better solution than anonymising
- Data that need anonymisation should be avoided in data collection
- Direct identifiers should be removed, masked or changed
- A maximum of information should be maintained
- Unedited versions of data should be retained for preservation
- Anonymisation should be planned from at the beginning of the research, not at the end

Questions?



FORS Data Management Webinar Series

© Svitlana Chernus | Dreamstime.com

Informed consent (27.09.2022)

Data documentation (11.10.2022)

Quantitative data anonymisation (1.11.2022)

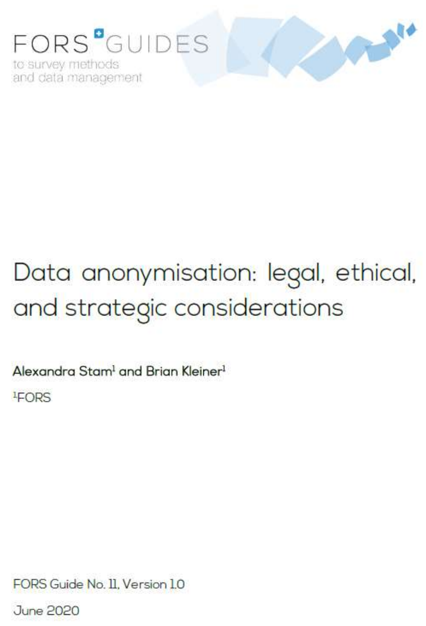
Qualitative data anonymisation (22.11.2022)

Registration here: https://forscenter.ch/data-management-webinar-series/

FORS

(Further) resources

- https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/5.-Protect/Anonymisation
- https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-anonymization_v1-(250118).pdf



FORS GUIDES
to survey methods
and data management

Data anonymisation: legal, ethical, and strategic considerations

Alexandra Stam[1] and Brian Kleiner[1]

[1]FORS

FORS Guide No. 11, Version 1.0

June 2020