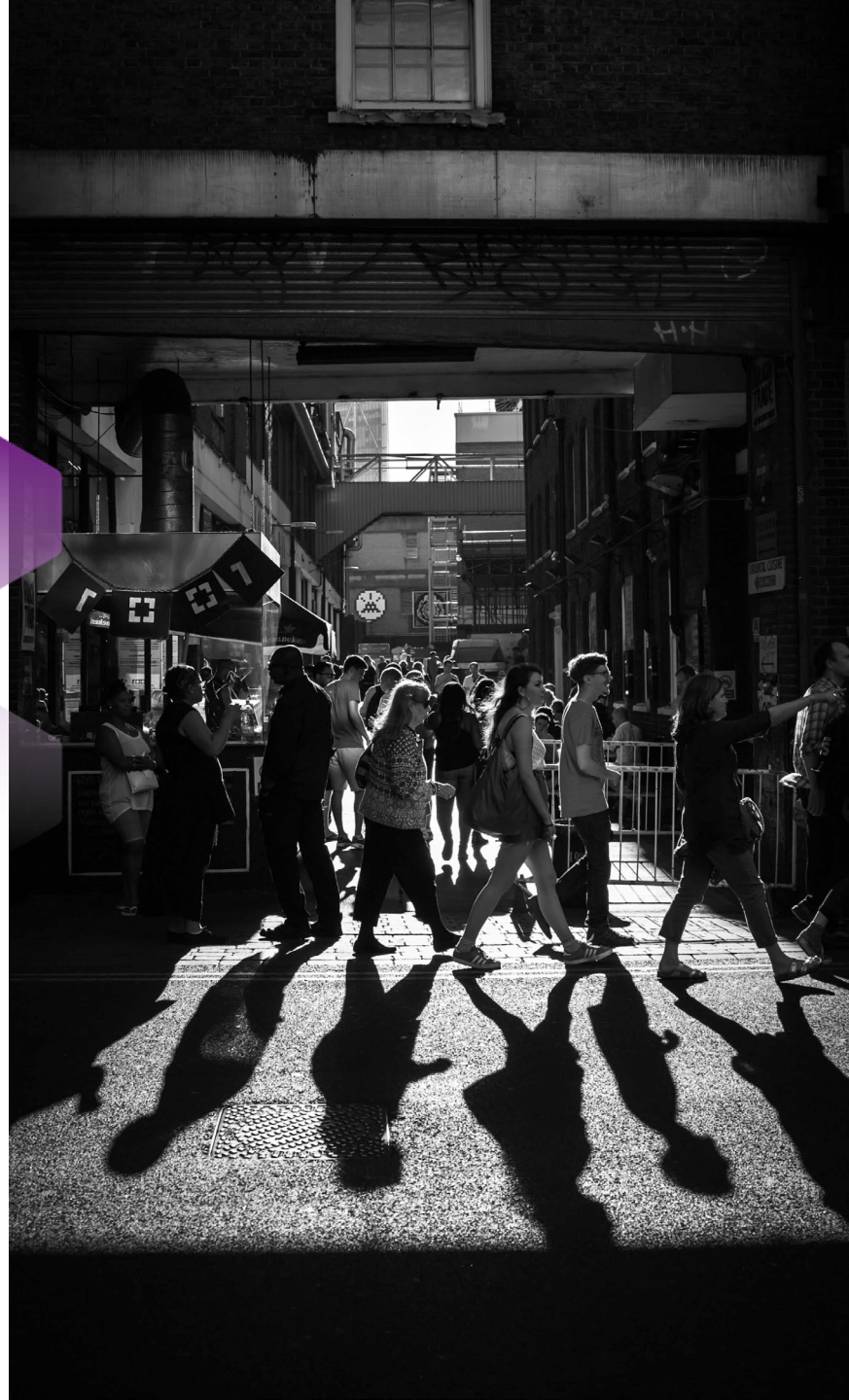
UK Data Service

How to anonymise qualitative and quantitative data

Anca Vlad

Maureen Haaker



Economic and Social Research Council

Overview

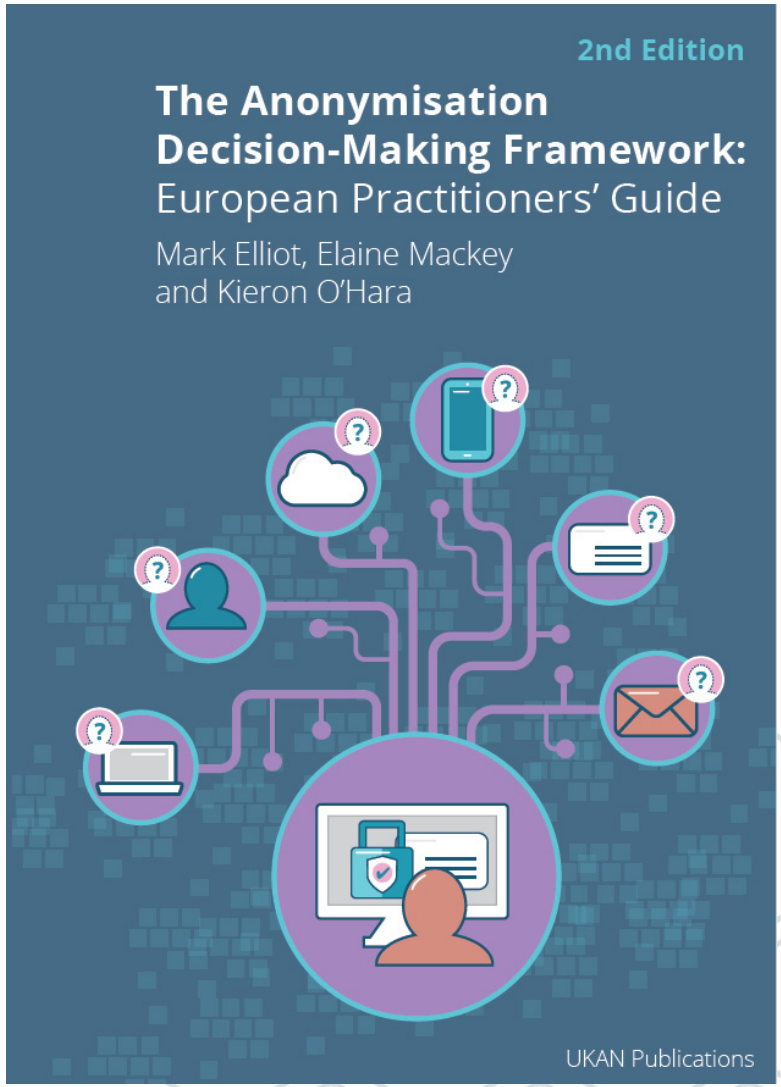
- Introduction: Why anonymise?
- A very short introduction to anonymisation theory.
- Anonymisation or Pseudonymisation?
- Access restrictions.
- Steps to anonymization.
- A couple considerations for qualitative data.
- Anonymisation planning.
- Exercise: Mentimeter.
- Further resources and questions.

Anonymisation Theory

Data situation audit

Risk analysis and control

Impact management



What is disclosure and why do we need anonymisation?

- Disclosure = identification.
- Disclosure happens when someone is able to identify a data subject from data or information they have access to from one source or multiple sources.
- Different types of disclosure: identity, attribute, inferential.
- Anonymisation is a process that attempts to prevent disclosure or identification of data subjects from a specific dataset.
- Anonymisation and pseudonymisation is part of Statistical Disclosure Control (SDC): the aim of SDC is to minimise/mitigate the risk of identification to an acceptable level that still allows researchers to maximise data use (use the data to it's full potential or as close to it as possible).

- When disclosure risk ⬇ information loss ⬆

Anonymisation / Pseudonymisation

Anonymised data

'...information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable." (Recital 26, GDPR).

- Cannot re-identify data subjects (even the data owner).

Pseudonymised data

“…the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.” (Article 4, GDPR).

- identifiable data has been removed or redacted so that cannot be traced back to the real values. Re-identification of data can only be achieved with knowledge of the de-identification key or by combination.

Cont...Anonymisation / Pseudonymisation

ICO: ‘re-identification’: describes the process of turning anonymised data back into personal data through the use of data matching or similar techniques.

The DPA does not prohibit the disclosure of personal data, but any disclosure has to be fair, lawful and in compliance with data protection principles.

To consider:

- the age of the information (less sensitive over time, but consider ethical)
- level of detail
- context: private life or about more public matters, such as their working life, or life satisfaction?
- Rule of thumb: try to assess the effect – if any - that the disclosure would have on any individual concerned
- Data environment – functional anonymisation

Classifying information (variables)

A. Identifying variables

1. Direct identifiers - information that –directly- identifies data subjects;
   - examples: social insurance number, names, address, national insurance number, IP address etc.

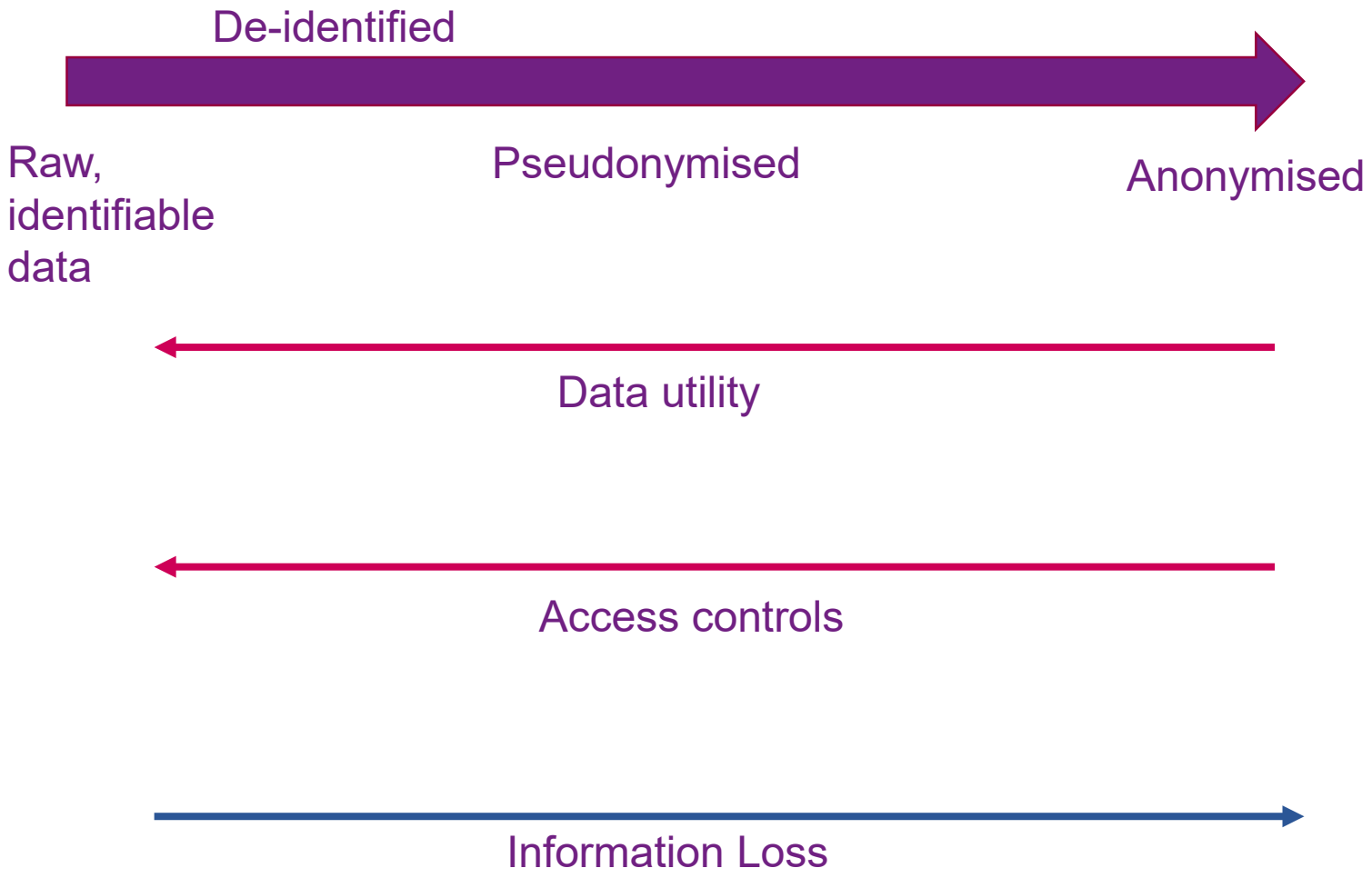
2. Indirect (key) identifiers - information that in combination, may uniquely identify data subjects;
   - can potentially be linked to other sources of data. (such as the electoral register)
   - examples: gender, age, region, occupation, income.

B. Sensitive variables - information that is often subject to legal and ethical concerns;
- examples: criminal history, sexual preferences and behaviour, political affiliations, medical records, income
- can lead to secondary (attribute) disclosure even if identity disclosure is prevented.

One variable can be both identifying and sensitive. Example: income.

You are not so anonymous!

Cont…Anonymisation / Pseudonymisation



Legal obligations, or when you need to break confidentiality

How will the data be used?

I'm asking for permission to use anonymised quotations and narrative themes, along with any photographs and video you provide in the interviews or diaries for research purposes. All diaries and interviews from all participants will be analysed together for common themes about what everyday life is like when pregnant. As I work through my analysis, I will transcribe any audio recordings or handwritten diary entries. As I transcribe, I'll anonymise any identifying details, such as your name and address. All digital files will be saved on a password-protected computer at University of Essex and all paper documents will be stored in a locked drawer at my office at the University of Essex, to which only I have access.

Throughout the project, I will be the only one with access to un-anonymised data, and my supervisors will have access to anonymised data. Since this project has gone through ethical approval from the Health Research Authority, NHS Trust staff may also be audit this project to ensure I am protecting your information appropriately, and may ask to see relevant sections of data.

I may need to break this confidentiality if you disclose illegal or criminal activity to me or I become aware of an issue that puts you or a child's safety at risk. In this instance, I will aim to first discuss the issue with you, but I may be legally obliged to share this information with the appropriate authorities.

Accessing data

Access Options

OPEN

Available for download/online access under open licence without any registration

•SAFEGUARDED

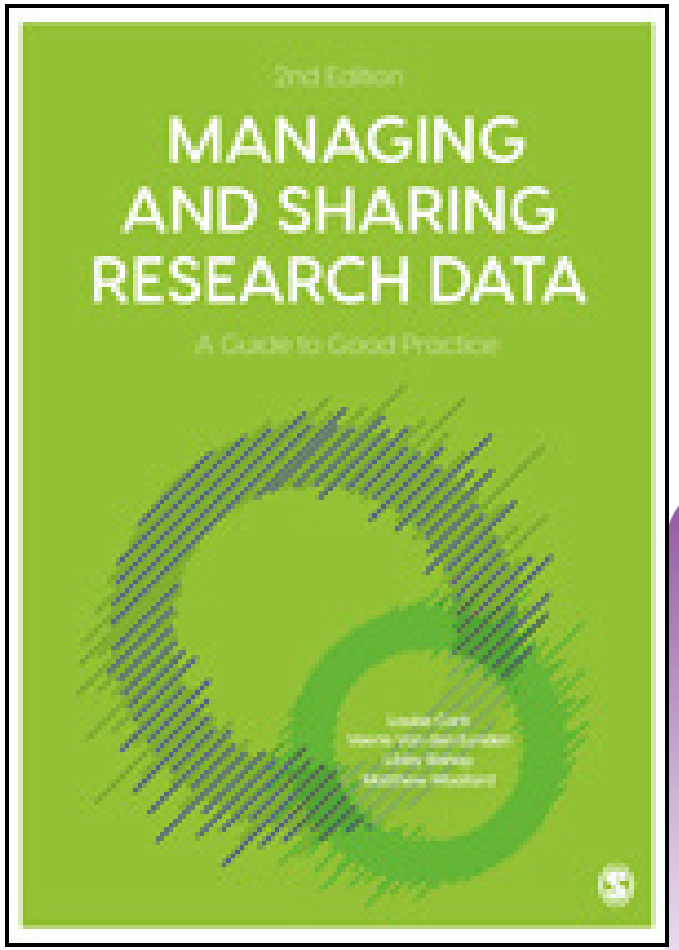
•Available for download / online access to logged-in users who have registered and agreed to an End User Licence; special agreements (e.g. depositor permission or approved researcher); embargo for fixed time period

•CONTROLLED

•Available for remote or safe room access to authorised and authenticated users whose research proposal has been and who have received training

UKDS data management guidance

- Best practice guidance
- *Managing and Sharing Research Data: A Guide to Good Practice* (Sage Publications Ltd)
- Training
- Twitter: @UKDSRDM



Anonymising qualitative data: some tips

- Plan or apply editing at time of transcription (*except: longitudinal studies to ensure linkages)*
- Consistency within research team and throughout project
- Identify replacements, e.g. with [brackets]
- Keep anonymisation log of all replacements, aggregations or removals made – keep separate from anonymised data files
- Avoid blanking out; use pseudonyms or replacements
- Avoid over-anonymising - removing/aggregating information in text can distort data, make them unusable, unreliable or misleading

Controlling access a better option than over-anonymizing!

In practice: example anonymisation

Ex 1. Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003 (study 5407 in UK Data Archive collection) by M. Mort, Lancaster University, Institute for Health Research.

Date of Interview: 21/02/02

Interview with Lucas Roberts, DEFRA field officer
Date of birth: 2 May 1965
Gender: Male
Occupation: Frontline worker
Location: Plumpton, North Cumbria

Lucas was living at home with his parents, "but I'm hoping to move out soon" so we met at his parents' small neat house. We sat in a very comfortable sitting room with an open fire and Lucas made me coffee and offered shortbread. Although at first Lucas seemed a little nervous, quick to speech and very watchful he seemed to relax as we spoke and to forget abut the tape.

I will just start by asking you to tell me a little bit about yourself and your background.

Well it is an agricultural background. I grew up on the farm where my brother is now. After I left school I did work on the farm but went to college and did exams, did land use recreation, sort of countryside/ environmental management course. So I obviously left agriculture, did the course and came back [to the farm] at weekends.

Comment [v1]: Replace: Ken
Comment [v2]: delete
Comment [v3]: delete
Comment [v4]: Replace: Ken
Comment [v5]: Replace: Ken
Comment [v6]: Replace: Ken

In practice: wording in documentation

- We expect to use your contributed information in various outputs, including a report and content for a website. Extracts of interviews and some photographs may both be used. We will get your permission before using a quote from you or a photograph of you.
- After the project has ended, we intend to archive the interviews at …. Then the interview data can be disseminated for reuse by other researchers, for research and learning purposes.

How will the data be used?

I'm asking for permission to use anonymised quotations and narrative themes, along with any photographs and video you provide in the interviews or diaries for research purposes. All diaries and interviews from all participants will be analysed together for common themes about what everyday life is like when pregnant. As I work through my analysis, I will transcribe any audio recordings or handwritten diary entries. As I transcribe, I'll anonymise any identifying details, such as your name and address. All digital files will be saved on a password-protected computer at University of Essex and all paper documents will be stored in a locked drawer at my office at the University of Essex, to which only I have access.

Throughout the project, I will be the only one with access to un-anonymised data, and my supervisors will have access to anonymised data. Since this project has gone through ethical approval from the Health Research Authority, NHS Trust staff may also be audit this project to ensure I am protecting your information appropriately, and may ask to see relevant sections of data.

In practice: data with access conditions

- Health and Social Consequences of the Foot and Mouth Disease Epidemic in North Cumbria, 2001-2003 (study 5407 in UK Data Archive collection) by M. Mort, Lancaster University, Institute for Health Research.
- Interviews (audio and transcript) and written diaries with 54 people
- 40 interview and diary transcripts are archived and available for re-use by registered users
- 3 interviews and 5 diaries were embargoed until 2015
- audio files archived and only available by permission from researchers

discover.ukdataservice.ac.uk/catalogue/?sn=5407

doc.ukdataservice.ac.uk/doc/5407/mrdoc/pdf/q5407userguide.pdf

In practice: Pioneers of Social Research (SN 6226)

Conducted by pioneering oral historian, Paul Thompson and his colleagues, this collection contains 43 life story interviews with pioneering social researchers, covering family and social background and key influences with detailed accounts of major projects.

1. Sara Arber

Biography

Sara Arber is a sociologist who has played an influential role in developing standards of survey practice and analysis through her publications (*Doing Secondary Analysis,* 1988) and through her teaching and research practice based at Surrey University.

In practice:
Managing Suffering at the End of Life (SN 850749)

Some dying people experience symptoms in the last hours or days of life that do not respond well to conventional therapies. In such circumstances, sedation may be given to induce a coma until death occurs. This practice is known as 'continuous deep sedation until death' or as 'palliative' or 'terminal' sedation. These data describe the care of dying people with refractory symptoms and include sensitive issues such as balancing symptom control with avoidance of hastening death.

Anonymisation: Step 1

Similar for both quantitative and qualitative data, the first step is always to identify and remove or redact identifying information (direct identifiers) in line with what participants agree to.

- Easier for quantitative data – removal/recode of variables.
- Can vary for qualitative data - replace with pseudonyms or not redact out.

Anonymisation: Step 2

Step 2: Identify all indirect identifiers

- Age/Date of Birth.
- Gender.
- Occupation.
- Income.
- Geography (area/county/city/village etc.)
- Ethnic Background/Ethnicity.
- Religion.

- Note here how important good quality metadata can be for this process (variable labels, value labels).

Anonymisation: Step 3

- Check frequencies to identify potentially disclosive information (small counts).
- Check outliers (if any).
- Check (any) string variables (other open text) to identify if they contain any personal, potentially disclosive or sensitive information (“I worked for X company for 30 years” or “my brother has a rare type of disease” or “I was a victim of domestic abuse and I used charity x for support”).

Anonymisation techniques

- Aggregate or reduce the precision (village -> town -> city).
- Recode categorical variables (indirect identifiers) into fewer categories.
- Suppressing specific values of indirect identifiers for some units.
- Generalise meaning of text variables - replace potentially disclosive free-text responses with more general text.
- Restrict the upper or lower ranges of a continues variable to hide outliers.
    - E.g. age – recode into 70+.
    - How to decide? -> Check frequencies for indirect identifiers.
- Anonymise geo-referenced data - replacing point coordinates with non-disclose variables.

Anonymisation techniques: Example 1

Age	Gender	Profession	Ethnicity
27	Male	Builder	Black (Caribbean)
118	Female	MD ®	White (Irish)
89	Female	Teacher®	Black (African)
56	Female	Customer Service	Asian (Pakistani)
48	Male	Builder	White (Scottish)
31	-99	Electrician	Black
57	Female	Customer Service	White

Cont..Anonymisation techniques: Example 1

Age	Gender	Profession	Ethnicity
27	Male	Builder	Black
80+	Female	MD ®	White
80+	Female	Teacher®	Black
56	Female	Customer Service	Asian
48	Male	Builder	White
31	-99	Electrician	Black
57	Female	Customer Service	White

Top-coded to hide (potential outliers)

Recode into fewer categories (less precision)

Cont..Anonymisation techniques: Example 2

Raw/Source

- Anna Thomson (she/her), 45, went to her chemotherapy treatment on 5 April 2020, at Bakersfield Hospital

De-identified

- Charlie, woman, 45, went to her chemotherapy treatment on 5 April 2020, at Bakersfield Hospital

Pseudonymised

- Charlie, woman, 45, went to her chemotherapy treatment in April 2020, at a hospital in Oxfordshire

Anonymised

- Charlie, 40-50, went to her chemotherapy treatment in 2020, at a hospital in England

Useful software

- sdcMicro – R package (free) – has a user friendly interface so minimal coding skills needed.
- QAMyData - UK Data Service developed a free (GitHub) easy-to-use open source tool, that provides a health check for numeric data. The tool uses automated methods to detect and report on some of the most common problems in survey or numeric data, such as missingness, duplication, outliers and direct identifiers.
- ARX - a comprehensive open source software for anonymizing sensitive personal data. It supports a wide variety of (1) privacy and risk models, (2) methods for transforming data and (3) methods for analysing the usefulness of output data.
- µ-Argus – developed by Statistics Netherlands; User Manual

In practice: Anonymisation plans

- Project background.
- File management.
- Mandatory anonymisation:
  - Direct Identifiers (names, contact details).
  - Places.
  - Ages and dates.
- Possible anonymisation:
  - Medical information about others not taking part in study.
  - Sensitive information (unfavourable opinions of others, details of legal cases, etc).

3-prong approach to protecting participants: Consent, anonymisation, and access

- Ask for consent to share –researchers must be informed about risks ***and*** benefits of data sharing.
- Anonymise – only if damage to data is minimal (not images)
- Regulate access:
  - End User Agreement (UK Data Archive).
  - Embargo.
  - For selected sensitive or disclosive data – registered users; permission from data depositor.

These strategies enable most data to be shared.

Tools and templates

- Model consent form and survey consent statement.
- Transcription template.
- Transcription instructions.
- Transcription confidentiality agreement.
- Data list template.

Further resources

- Anonymising Research Data - ESRC National Centre for Research Methods, Working Paper 7/06.
- Guide to Social Science Preparation and Archiving from the Inter-University Consortium for Political and Social Research.
- Anonymisation and Social Research, Ruth Geraghty.
- Timescapes anonymisation guidelines, University of Leeds.
- Anonymisation: managing data protection risk - ICO code of practice.
- The Anonymisation Decision-Making Framework - Mark Elliot, Elaine Mackey Kieron O'Hara and Caroline Tudor.
- Jisc guidance on anonymous data.
- Advice from med.data.edu on anonymization.

Get connected

UK Data Service

Jisc mail group

@UKDataService Twitter

UK Data Service YouTube channel

Powerpoint slides will be available on our website in due course and you can catch up on the recording on our Youtube channel. Check out our Twitter for more updates.

Upcoming events

Recurring Workshops:

- Data management basics
- Ethical and legal issues in data sharing.
- Introduction to copyright: Copyright and publishing.
- Getting started with secondary analysis.
- Consent issues in data sharing.
- Computational social science drop-in.
- Data documentation.
- Depositing your data with ReShare.
- Safe researcher training.

For more information and registration, see our Events Page

UK Data Service

Thank you.

Maureen Haaker
mahaak@essex.ac.uk

Anca Vlad
advlad@essex.ac.uk