De-identification Guidance

Prepared by the Portage Network, COVID-19 Working Group on behalf of the Canadian Association of Research Libraries (CARL)

Kristi Thompson (Western University)
Erin Clary (Portage)
Lucia Costanzo (University of Guelph)
Beth Knazook (Portage)
Nick Rochlin (University of British Columbia)
Felicity Tayler (University of Ottawa)
Jane Fry (Carleton University)
Chantal Ripp (University of Ottawa)
Kathy Szigeti (University of Waterloo)
Qian Zhang (University of Waterloo)
Roger Reka (University of Windsor)
Minglu Wang (York University)
Rebecca Dickson (Coppul)
Mark Leggott (RDC-DRC)
Melanie Parlette-Stewart (Portage)

SEPTEMBER 2020

Portage Network
Canadian Association of Research Libraries
portage@carl-abrc.ca

www.carl-abrc.ca

portage
SERVICES PARTAGÉS POUR LES DONNÉES DE RECHERCHE
SHARED STEWARDSHIP OF RESEARCH DATA

CARL ABRC
CANADIAN ASSOCIATION OF RESEARCH LIBRARIES
ASSOCIATION DES BIBLIOTHÈQUES DE RECHERCHE DU CANADA

Table of Contents

De-identification Guidance 2

Identify and Remove Direct Identifiers 3

How do I remove this information? *3*

Identify and Evaluate Indirect or Quasi-Identifiers based on Perceived Risk and Utility 4

How do I figure out what combination of quasi-identifiers are a problem? *4*

How do I assess the sensitivity of non-identifying variables in dataset? *7*

Considerations for Qualitative Data De-identification 8

Brief Considerations for Social Media, Medical Images, and Genomics Data 10

Data collected from social media or social networking platforms (e.g., Twitter, Facebook). *10*

Medical Images *11*

Genomics data, and other biomedical samples *12*

Appendix 1: Code for Checking *K*-Anonymity 14

Appendix 2: Free de-identification software packages 16

Appendix 3: Fee-based services for de-identification 18

Resources 19

References 19

De-identification Guidance

The guidance below is intended to help you minimize disclosure risk when sharing data collected from human participants. If you use any of the following techniques to anonymize your data, please include this information in your documentation and README file.[1] For transparency, it should be clear how the dataset was modified to protect study participants.

Before proceeding, please note that not all human participant data needs to be de-identified, or stripped, of direct and indirect identifiers. Please review your consent form and prepare your data to share only what participants have agreed to share. If you are unsure whether you need to de-identify your data, please see the Portage help guide Can I share my Data? and consult with your institution's Research Ethics Board.[2] For help selecting a repository for your data, please see Portage's Recommended Repositories for COVID-19 Research Data guide or consult with librarians at your institution to see if further support is available.[3]

[1] Learn more about creating appropriate documentation for depositing your datasets in the Portage COVID-19 Working Group's "Documentation and Supporting Materials Required for Deposit," September 25, 2020, https://doi.org/10.5281/zenodo.4042034.
[2] Portage COVID-19 Working Group, "Can I Share My Data?" September 25, 2020, https://doi.org/10.5281/zenodo.4041661.
[3] Portage COVID-19 Working Group, "Recommended Repositories," September 25, 2020. https://doi.org/10.5281/zenodo.4042037.

Identify and Remove Direct Identifiers

Direct identifiers are those which place study participants at immediate risk of being re-identified. Unless explicit consent was received from study participants, they must be removed from any published version of your dataset. The following list is based on various sources, including guidance from major international funding agencies, the US Health Insurance Portability and Accountability Act (HIPAA) and the *British Medical Journal*. See Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers and List of 18 items considered under HIPAA to be identifiers.[4]

Direct identifiers are:

- Names or initials, as well as names of relatives or household members
- Addresses, and small area geographic identifiers such as postal codes / zip codes
- Telephone numbers
- Electronic identifiers such as web addresses, email addresses, social media handles, or IP addresses of individual computers
- Unique identifying numbers such as hospital IDs, Social Insurance Numbers, clinical trial record numbers, account numbers, certificate or license numbers
- Exact dates relating to individually-linked events such as birth or marriage, date of hospital admission or discharge, or date of a medical procedure
- Multimedia data: unaltered photographs, audio, or videos of individuals
- Biometric identifiers including finger or voice prints, and iris or retinal images
- Human genomic data, unless risk was explained and consent to share data or consent for secondary use of data was received from study participants
- Age information for individuals over 89 years old

How do I remove this information?

Removing direct identifiers from your data is relatively straightforward. You may either record this personal information in a separate document, spreadsheet, or database and link this to the other data points via a series of codes that can be removed before publishing, or choose to delete the identifying data points entirely at the end of the project. Refer to your consent forms to determine how to proceed. If you are unsure whether data can simply be unlinked or if it must be destroyed, consult your local Research Ethics Board.

[4] Iain Hrynaszkiewicz, Melissa L. Norton, Andrew J. Vickers, and Douglas G. Altman, "Preparing Raw Clinical Data for Publication: Guidance for Journal Editors, Authors, and Peer Reviewers." *BMJ* 340 (January 29, 2010): c181, https://www.bmj.com/content/340/bmj.c181; and Steve Alder, "What is Considered PHI Under HIPAA Rules?" *HIPAA Journal* (December 28, 2017), https://www.hipaajournal.com/considered-phi-hipaa/.

Identify and Evaluate Indirect or Quasi-Identifiers based on Perceived Risk and Utility

Indirect or quasi-identifiers are characteristics (such as demographic information) relating to individuals that could be linked with other data sources to violate the confidentiality of individuals. Quasi-identifiers may not be identifying on their own but can be disclosive in combination. For instance, identifying a participant's home community size within an overall limited geographic study area may allow someone to infer that participant's location more precisely. A variable should be considered a quasi-identifier if someone could plausibly match that variable to information from another source. See the International Household Survey Network Anonymization Principles and the Information and Privacy Commissioner Ontario De-identification Guidelines for Structured Data.[5]

A list of potential quasi-identifiers:

- Geographic identifiers (census geography, town name, urban/rural indicator) of home, place of birth, place of treatment, place of schooling, or other geography linked to individuals
- Sex / gender identity, orientation
- Ethnic background, race, visible minority, or Indigenous status
- Immigration status
- Membership in organizations
- Use of specific social networks or services
- Socioeconomic data, such as occupation or place of work, income, or education
- Household and family composition, marital status, number of children / pregnancies
- Criminal records and other information that may link to public records
- Generalized dates linked to individuals, e.g. age, graduation year, immigration year
- Some full-sentence responses
  - Note: These must be checked individually. For instance, the comment "The library should be open longer" is not identifying; however, a comment like "As chair of a research group that uses the library,…" is potentially identifying.
- *Some* medical information (e.g. permanent disabilities or rare medical conditions) may be identifying; temporary illness or injury is less likely to be so. The test is whether this is information that can be found elsewhere and therefore could be used to re-identify the person.

How do I figure out what combination of quasi-identifiers are a problem?

[5] "Anonymization Principles," International Household Survey Network, accessed August 4, 2020, https://ihsn.org/node/137; Information and Privacy Commissioner of Ontario, *Deidentification Guidelines for Structured Data*, Information and Privacy Commissioner of Ontario, June 8, 2016. https://www.ipc.on.ca/privacy-organizations/de-identification-centre/.

1. Observe the possible combinations

A good first step may be to look at the demographic variables in the dataset and consider describing an individual to a friend using only the values of those variables. Is there any likelihood that the person would be recognizable? For example, "I'm thinking of a person living in Toronto who is female, married, has a University degree, is between the ages of 40 and 55 and has an income of between 60 and 75 thousand dollars." Even if there is only one such person in the dataset, this is likely not enough information to create risk UNLESS contextual information about the dataset narrows things down further. For instance, if your data is limited to a specific, narrow group of individuals, such as the referees for the Ontario Hockey Association, the list of quasi-identifiers given above may be enough to uniquely identify an individual. Quasi-identifiers need to be evaluated in the context of what is known or what may be reasonably inferred about the survey population.

2. Assess these combinations mathematically

K-anonymity is a mathematical approach to demonstrating that a dataset has been anonymized, where *k* is an integer selected by the researcher that represents a group of records with the same information across all quasi-identifiers.[6] Within your dataset, a set of '*k'* records (e.g., a set of 3 or 5 records) is called an equivalence class. To achieve *k*-anonymity, it should not be possible to distinguish one record from the other records in its equivalence class. For example, if you choose a *k* value of 5, each record in your dataset must have the exact set of quasi-identifiers that are present in at least 4 other records in order to achieve *k*-anonymity.

K-anonymity only works to precisely estimate risk if a dataset is a complete sample of some population. *K*-anonymity considerably overestimates risk in the case of a dataset that is a subsample of a population. When determining the appropriate *k* value to use, consider:

- A lower *k* value of 3 may be sufficient in datasets that contain small samples from a large population.
- A higher (or more conservative) *k* value should be used if a dataset is a complete sample of a population.

Keep in mind that a dataset that is a complete sample of a known population may have additional risk factors. Imagine that all the respondents in a particular equivalence class answered a question the same way - you would know how each person in the survey belonging to that equivalence class answered the question. Respondents to surveys are generally told that their responses will be kept confidential, not merely that no one will know which line of data contains their specific answers. A *k*-anonymous dataset that is a complete sample may not fulfill that promise.

The code in Appendix 1 can be used with your preferred statistical software package to create equivalence classes based on the quasi-identifiers in the dataset and to list them by size. If any

[6] Khaled El Emam and Fida Kamal Dankar, "Protecting Privacy Using k-Anonymity," *Journal of the American Medical Informatics Association* 15, no. 5 (September 2008): 627–637, https://doi.org/10.1197/jamia.M2716.

equivalence class has fewer members than the value of *k* you selected, use the data reduction techniques below to further reduce dataset risk.

For more on *k*-anonymity, see International Household Survey Network (IHSN)'s Measuring the Disclosure Risk and the UK Anonymisation Network's Anonymisation Decision-Making Framework section 2.2.2, Guaranteed anonymisation.[7]

3. Use data reduction techniques to address dataset risk

Univariate frequencies and bivariate crosstabs can be used to identify small[8] categories of quasi-identifiers. Data reduction techniques can be used to mitigate risk once you have identified these small groups. There are three simple types of data reduction you may wish to consider:

1. The simplest is to completely **drop risky variables** from the dataset. This is an option for variables with relatively high risk that are not considered to be of high research value. (For example, in some datasets geography may be considered relatively less important than ethnicity or language.)
2. The second is **global re-coding**, or aggregating the observed values into a defined set of classes, such as transforming a variable with years of age into a variable of ten-year age categories, or top-coding a high income category to "$100,000 and above".
3. A third option for unusual cases is to use **local suppression**. For example, a very young married respondent might have their marital status set to 'missing' as an alternative to globally re-coding the otherwise non-risky age variable into a larger group.

After each exercise in data reduction, repeat the test for *k-anonymity* described above and check equivalence classes until all groups are larger than your selected value for K.

For more information, including information about more complex types of data reduction, see UK Anonymisation Network's Anonymization Decision-Making Framework section 2.5, Anonymisation solutions.[9]

[7] "Measuring the Disclosure Risk," International Household Survey Network, accessed August 4, 2020, https://ihsn.org/anonymization-risk-measure; and Mark Elliot, Elaine Mackey, Kieron O'Hara, and Caroline Tudor, *The Anonymisation Decision-Making Framework*. UK Anonymisation Network (UKAN), University of Manchester, 2016, https://ukanon.net/ukan-resources/ukan-decision-making-framework/.

[8] 'Small' is relative; as a first pass, groups smaller than 5% of the dataset or containing fewer than 20 cases could be considered.

[9] Elliot, Mackey, O'Hara, and Tudor, *The Anonymisation Decision-Making Framework*.

How do I assess the sensitivity of non-identifying variables in dataset?
Non-identifying information includes survey responses and measurements that are not likely to be recognizable as coming from specific individuals. Examples include opinions, rankings, scales, or temporary measures such as resting heart rate after meditation or the number of times an individual ate breakfast in a week.

It is possible for non-identifying information to be highly sensitive as well. Information that could be used to stigmatize or discriminate against an individual, such as a criminal record, sexual practices, illicit drug use, mental health and psychological well being, and other sensitive medical information all increase the risk of the dataset and should be considered when deciding whether to release the data at all. You may wish to remove or modify these variables to create a less sensitive version of the data.

Considerations for Qualitative Data De-identification

Qualitative data describes qualities or characteristics that can be observed, but not necessarily measured. This type of data is collected through interviews, surveys, or observations, and may be in the form of transcripts, notes, video and audio recordings, images, and text documents. As with quantitative data, direct identifiers may appear in the form of names, date and place of birth, other locations, and even photos. These direct identifiers can be used along with indirect or quasi-identifiers, such as medical, education, financial, and employment information, to trace or determine an individual's identity.

The process for removing identifying information in a video recording, audio interview, or oral transcript is very different from that used to de-identify a medical record. For one, it is harder to do programmatically. Extremely detailed field notes or audiovisual information often requires someone to read or watch the content thoroughly.

General Advice

- Avoid asking for identifying information in the first place.
    - It is easier to edit the information at the point of capture than it is to remove information after it has been recorded.
    - If you require identifying information at the research stage, try to capture it within the first few minutes of an interview or recording, so that it is easy to edit it out quickly. Alternatively, transcribe the information in a separate document that can be removed from a person's file.
- Make de-identification a part of the process of informed consent.
    - Ensure that study participants are aware of your planned use of the data, and the fact that their information may be anonymized to protect them. Make it clear in your consent forms how extensively they will be de-identified (i.e., what elements will be replaced or removed). While direct identifiers may be eliminated (name, address, birthday, etc.), there may be other subtle clues to their identities that remain within the recording or transcript.
    - Agree in advance with participants which type of identifying information can be revealed in an interview. (For example, the participant may not wish to mention an employer's name). This is easier than removing information after the fact.
    - Keep in mind that not all data needs to be de-identified or anonymized. In some circumstances, you may be recording deeply personal accounts and should be mindful of a participant's right to have their story told in their own words. Some participants may have a personal interest in staying identified.

De-identification Guidance

- Use pseudonyms and change identifying details to protect anonymity.
    - If changing the person's name, location of residence, or occupation can be done without compromising the dataset, this can help to protect their anonymity. Be advised that this could influence the utility of a dataset as it may alter a future researcher's perception of the interviewee's socio-economic status or behaviour.
- If necessary, remove blocks of sensitive text or edit out portions within audio-visual data.
    - Some portion of the research may need to be redacted. Be wary of using search and replace techniques as it is easy to replace the wrong piece of information.
    - Voices in audio recordings may need to be masked by altering pitches.
    - Faces in visual data may need to be pixelated.
- Restrict access.
    - This is not preferred, but some datasets will not remain useful if all identifiers are removed. It may be possible to allow researchers seeking secondary access to request that queries be performed by the original research team, who can then share results if they are non-disclosive or can be appropriately de-identified.

For more information, see the UK Data Service's Guide to Anonymisation of Qualitative Data.[10]

[10] "Anonymisation: Qualitative Data," UK Data Service, last modified June 30, 2020, https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/qualitative.aspx.

Brief Considerations for Social Media, Medical Images, and Genomics Data

Data collected from social media or social networking platforms (e.g., Twitter, Facebook).
Although information on social networking sites may be free to access or view, it does not automatically follow that it is free to redistribute. Many platforms have terms of use that you will need to abide by, and the people who use the platform may have an expectation of privacy which must be respected. Some platforms require users to register before content is visible, and others may have terms that prohibit data collection, data scraping, or republishing content elsewhere.

Here are a series of questions to consider before you deposit social media data:

- Could the topic you are studying be considered sensitive?
- Could your data lead to stigmatization of, or discrimination against, the content author?
- Is the study population vulnerable?
- What expectation of privacy might the individual users of this platform have?
- Is it possible or reasonable to obtain informed consent?
- Can or should the data be anonymized?
- Do the platform's terms of use allow you to redistribute content?

For example, Twitter allows the content author to maintain control over their tweets. As part of Twitter's policies, only numeric Tweet IDs and User IDs should be redistributed.[11] If you have weighed the questions above and decide to deposit your dataset, the Tweets must first be 'dehydrated' (distilled down to just the Tweet ID) using a tool such as DocNow's twarc.[12] Any secondary use of the data would then require an end-user to "rehydrate" the Tweet IDs using the Twitter REST API or an external tool such as DocNow's Hydrator.[13] Content will not be returned for tweets that have since been deleted.

The following resources provide more in-depth guidance:

- Zeffiro and Brodeur, Social Media Research Data Ethics and Management (slides from a workshop presented at McMaster University).[14]
- Ryerson University Research Ethics Board's Guidelines for Research Involving Social Media.[15]

[11] "Developer Terms: More About Restricted Uses of the Twitter APIs," Twitter, accessed August 4, 2020, https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases.

[12] Documenting the Now, "DocNow/twarc." GitHub, accessed August 4, 2020, https://github.com/docnow/twarc.

[13] Documenting the Now, "DocNow/hydrator." GitHub, accessed August 4, 2020, https://github.com/DocNow/hydrator.

[14] Andrea Zeffiro and Jay Brodeur, "Social Media Research Data Ethics and Management." Workshop presented April 5, 2018, Sherman Centre for Digital Scholarship, McMaster University, http://hdl.handle.net/11375/25327.

[15] Ryerson University Research Ethics Board, "Guidelines for Research Involving Social Media," Ryerson University, November, 2017, https://www.ryerson.ca/content/dam/research/documents/ethics/guidelines-for-research-involving-social-media.pdf.

- Mannheimer and Hull, Sharing Selves: Developing an Ethical Framework for Curating Social Media Data.[16]

North Carolina State University's Social Media Archives Toolkit, which contains guidance on the legal and ethical implications of sharing social media data, and an annotated bibliography with further resources.[17]

Medical Images

Before you archive medical images, remove any direct identifiers you do not have explicit consent to share, such as name, patient ID, and exact dates from the image header or embedded metadata, and black out any pixels in the image that contain identifying information. Neuroimages must also be defaced using a tool such as PyDeface.[18]

The following resources provide more guidance for de-identifying DICOM files:

- The Cancer Imaging Archive (TCIA) De-identification Overview.[19]
    - See specifically "Table 1 - DICOM Tags Modified or Removed at the source site" for a list of DICOM tags deemed to be unsafe.
- The Radiological Society of North America (RSNA) International Covid-19 Open Radiology Database (RICORD) De-identification Protocol.[20]
- The DICOM standard itself provides important guidance for de-identifying header information. Specifically, DICOM Part 15: Security and System Management Profiles, Appendix E: Attribute Confidentiality Profiles may be useful.[21]

[16] Sara Mannheimer and Elizabeth Hull, "Sharing Selves: Developing an Ethical Framework for Curating Social Media Data," *International Journal of Digital Curation* 12, no. 2 (April 18, 2018), https://doi.org/10.2218/ijdc.v12i2.518.

[17] "Social Media Archives Toolkit," North Carolina State University Libraries, accessed August 4, 2020, https://www.lib.ncsu.edu/social-media-archives-toolkit.

[18] Some repositories may be able to assist you or recommend tools for defacing. For example, the International Neuroimaging Data-Sharing Initiative (INDI) can help researchers who plan to share their data on the INDI platform. For further information, see the INDI *Data Contribution Guide*, accessed August 31, 2020, http://fcon_1000.projects.nitrc.org/indi/indi_data_contribution_guide.pdf. See also, Omer Faruk Gulban, Dylan Nielson, Russ Poldrack, John Lee, Chris Gorgolewski, Vanessasaurus, and Satrajit Ghosh, "Poldracklab/pydeface: V2.0.0." October 31, 2019. http://doi.org/10.5281/zenodo.3524401.

[19] Kirby, Justin. "Submission and De-identification Overview." The Cancer Imaging Archive (TCIA), University of Arkansas for Medical Sciences, April 27, 2020, https://wiki.cancerimagingarchive.net/display/Public/Submission+and+De-identification+Overview.

[20] "RSNA International Covid-19 Open Radiology Database (RICORD) De-identification Protocol," Radiological Society of North America, International COVID-19 Open Radiology Database, accessed August 10, 2020, https://www.rsna.org/-/media/Files/RSNA/Covid-19/RICORD/RSNA-Covid-19-Deidentification-Protocol.pdf.

[21] Medical Imaging & Technology Alliance, DICOM Standards Committee, "DICOM Part 15: Security and System Management Profiles." *DICOM Standard* (Arlington, VA: National Electrical Manufacturers Association), accessed August 4, 2020,
https://www.dicomstandard.org/current/.

- These profiles attempt to balance the need to protect privacy with the need to retain information so the data remain useful.
- If it is necessary to retain identifiers, your REB application will have ideally referenced the profile you intend to use, and your consent form should clearly state what information will be shared.

De-identification of DICOM files may be done programmatically, using a software to strip identifiers from the header.

- TCIA recommends the Clinical Trial Processor (CTP) software developed by RSNA.[22]
- RSNA's Covid-19 Open Radiology Database (RICORD) recommends another RSNA software called Anonymizer, and has published instructions on how to install and use it. [23]Anonymizer implements RICORD's de-identification protocol.[24]
- There are many other non-commercial options available, such as the DicomCleaner™ tool.[25]
- As with all de-identification software, results may be variable, and you should confirm that identifying information was removed before you share your images. Note that:
    - Vendors or end-users may not have always used DICOM elements in a way that conforms to the standard.
    - Private elements or private tags may have been used to store personal information, and the use of these tags may not be well-defined in the vendor documentation.

Genomics data, and other biomedical samples

Because each person's DNA sequence is unique, human biological materials can never be truly anonymous. Before you archive or biobank these data, please review your consent form. Ideally the consent process will have:

- provided participants with information about how their data will be used, analyzed, stored and shared,
- identified what information will be stored alongside the data,
- communicated what level of privacy or confidentiality a participant may expect, and who may have access to the data,
- indicated whether the data/samples will be stored in Canada or outside of Canada,
- acknowledged whether there is a possibility that the data will be used for commercial purposes,

[22] "Clinical trial processor (CTP)," Radiological Society of North America, Medical Imaging Resource Community (MIRC), accessed August 4, 2020, https://www.rsna.org/research/imaging-research-tools.

[23] "RSNA COVID-19 DICOM Data Anonymizer," Radiological Society of North America, International COVID-19 Open Radiology Database, accessed August 10, 2020, https://www.rsna.org/-/media/Files/RSNA/Covid-19/RICORD/RSNA-Anonymizer-Program-Instructions.pdf.

[24] "RSNA International Covid-19 Open Radiology Database (RICORD) De-identification Protocol," Radiological Society of North America, International COVID-19 Open Radiology Database.

[25] Clunie, David A., "DicomCleaner™," PixelMed Publishing™, accessed July 16, 2020, http://www.dclunie.com/pixelmed/software/webstart/DicomCleanerUsage.html.

- clearly explained the risks of disclosure.

Further information is available in TCPS 2 (2018), Chapter 12: Human Biological Materials Including Materials Related to Human Reproduction (sections A and D specifically), and Chapter 13: Human Genetic Research.[26] See also Thorogood (2018) Canada: will privacy rules continue to favour open science?[27]

The NIH Privacy in Genomics webpage provides a concise overview of some of the benefits and risks of sharing genetic information.[28] For an example of how genetic information was used to identify study participants, see Identifying Personal Genomes by Surname Inference, or a summary of the study in the 2013 *Nature* editorial on Genetic privacy.[29]

For further information on ethics and consent in genomics, see the Global Alliance for Genomics and Health Regulatory & Ethics Toolkit resources, such as Data Privacy and Security Policy and Consent Policy.[30]

[26] Government of Canada (Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada, and the Social Sciences and Humanities Research Council), *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans,* December 2018, https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2018.html.

[27] Adrian Thorogood, "Canada: Will Privacy Rules Continue to Favour Open Science?" Human Genetics 137: 595–602 (July 16, 2018), https://doi.org/10.1007/s00439-018-1922-z.

[28] "Privacy in Genomics," National Human Genome Research Institute, February 24, 2020, https://www.genome.gov/about-genomics/policy-issues/Privacy.

[29] Gymrek, Melissa, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich. "Identifying Personal Genomes by Surname Inference." *Science* 339, no. 6117 (Jan 18, 2013): 321-324. https://doi.org/10.1126/science.1229566; and "Genetic privacy" [Editorial], *Nature* 493 (January 24, 2013): 451, https://doi.org/10.1038/493451a.

[30] Global Alliance for Genomics & Health. *Genomic Toolkit: Regulatory & Ethics Toolkit.* Toronto, ON: Global Alliance for Genomics and Health, accessed July 20, 2020, https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/.

Appendix 1: Code for Checking *K*-Anonymity

-- Stata --

```
* Stata code for checking k-anonymity

* Kristi Thompson, May 2020


* create the equivalence groups

egen equivalence_group= group(var1 var2 var3 var4 var5)

* create a variable to count cases in each equivalence group

sort equivalence_group

by equivalence_group: gen equivalence_size =_N

* list the ID numbers of equivalence groups containing 3 or fewer cases

tab equivalence_group if equivalence_size < 3, sort

* list the values of the quasi-identifiers for each small equivalence
class.

list var1 var2 var3 var4 var5 if equivalence_group == 1
```

--- R --

```
# R code for checking k-anonymity

# Carolyn Sullivan and Kristi Thompson, May 2020


# install plyr, a useful data manipulation package.

install.packages("plyr")

# Load the library.

library('plyr')
```

```
datafile <- " location of the data file - csv format -  "

# Read the csv  file.

df <- read.csv (datafile)


# Figure out what equivalence classes there are, and how many cases in
each equivalence class.

dfunique <- ddply(df, .(var1, var2, var3, var4, var5), nrow)

dfunique <- dfunique[order(dfunique$V1),]

View(dfunique)
```

The UK Anonymisation Network's Anonymisation Decision-Making Framework, appendix B has code for doing this in SPSS.[31]

[31] Elliot, Mackey, O'Hara, and Tudor, *The Anonymisation Decision-Making Framework*.

Appendix 2: Free de-identification software packages

Many of these tools take a hierarchical approach to de-identifying data, which means that you will need to pre-define possible generalizations for the quasi-identifiers in the dataset, and the program will search for possible solutions and recommend a set of the generalizations to use to best meet anonymization goals. For datasets with a large number of quasi-identifiers, or cases where several datasets with similar quasi-identifiers need to be de-identified, this might be a useful approach. For smaller datasets, it may be more straightforward to work in a statistical package. The software packages included here all have some usability issues, and fairly steep learning curves. Amnesia and the graphical user interface to sdcMicro may be the most user-friendly.

Recommended tools:

- Amnesia
  - This software has both online and desktop versions, however, uploading sensitive data to a third-party web site is not generally recommended. If possible, install the software locally (Windows or Linux only).
  - Amnesia supports *k*-anonymity and k^m-anonymity (a slightly more flexible approach to anonymity when the number of quasi-identifiers in a dataset is very high, as it allows for combinations up to *m* quasi-identifiers to appear at least *k* times in the published data).
  - A few limitations: there is not currently a way to specify missing values; documentation could be more thorough, for instance, defining hierarchies is not straightforward.
  - This software may work best for clinical data, or data which are not survey data.
- sdcMicro
  - An R package for statistical disclosure control (microdata anonymization). This software can read many data types (e.g., csv, sav, dta, sas7bdat, xlsx) and can be used in Windows, Linux or Mac operating systems. Implements muArgus code.
  - A graphical user interface is available, and there is a vignette with guidance called 'Using the interactive GUI - sdcApp' linked from the sdcMicro landing page in CRAN repository.[32]
  - Please be aware that large datasets take time to load, and computation time for large or complex datasets may be lengthy.[33]
  - The *Statistical Disclosure Control for Microdata* practice guide section on SDC with sdcMicro in R may be helpful if you need further guidance installing and using the *sdcMicro* package, or see Benschop's sdcMicro GUI manual Documentation.[34]

[32] "Using the interactive GUI – sdcApp, The Comprehensive R Archive Network (CRAN), accessed August 31, 2020, https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdcMicro.html.

[33] "Computation time," SDC with sdcMicro in R: Setting Up Your Data and more, SDC Practice Guide, 2019, https://sdcpractice.readthedocs.io/en/latest/sdcMicro.html#computation-time.

[34] "Statistical Disclosure Control (SDC): An Introduction," SDC Practice Guide, 2019, https://sdcpractice.readthedocs.io/en/latest/SDC_intro.html; and Thijs Benschop and Matthew Welch. *Statistical Disclosure Control for Microdata: A Practice Guide for sdcMicro,* International Household Survey Network, accessed August 10, 2020, https://sdcpractice.readthedocs.io/en/latest/index.html.

Other tools that may be useful:

- ARX
    - Open source anonymization tool for use in Windows, Linux, and Mac. Provides support for SQL databases, xlsx and csv files, and has a graphical user interface.
    - Supports various privacy models including k-anonymity, and variants ℓ-diversity, t-closeness, β-Likeness, and more.[35]
    - Allows end-users to categorize, top and bottom code, generalize, and transform data in more complex ways.
    - Large datasets take time to load, and computation time for large or complex datasets may be lengthy.
- mu-Argus
    - Software to apply Statistical Disclosure Control techniques. The program takes a hierarchical approach to de-identifying data.
    - JAR file should be executable in Windows or Mac OS.
    - A tester found that getting data loaded and correctly defined was a bit of a challenge and advised that the program could use better documentation on setting up hierarchies.
- The University of Texas at Dallas Anonymization Toolbox
    - The toolbox currently supports 6 different anonymization methods and 3 privacy definitions, including k-anonymity, ℓ-diversity, and t-closeness.
    - Algorithms can either be applied directly to a dataset or can be used as library functions inside other applications.
    - This is a set of Java routines. Data curators who prefer to do their statistical programming in Java might find it useful.

[35] "Privacy Models," ARX – Data Anonymization Tool, accessed August 31, 2020, https://arx.deidentifier.org/overview/privacy-criteria/.

Appendix 3: Fee-based services for de-identification

A few fee-based services that researchers may opt to use for de-identification are included below:

- d-wise (American & European offices)
    - Offering free anonymization services to anyone working on a COVID-19 vaccine.[36]
    - Offering free anonymization services to researchers who deposit individual participant-level data from COVID-19 clinical trials in Vivli.[37]
- Inter-university Consortium for Political and Social Research (ICPSR) (Archive headquartered at University of Michigan)
    - If you wish ICPSR to conduct disclosure analysis of your data, you will need to purchase the Professional Curation package. Cost is based on the number of variables and complexity of the data. Contact ICPSR Acquisitions at deposit@icpsr.umich.edu for additional information (information obtained from Open ICPSR FAQ under Pricing and Sensitive Data sections).[38]
- Privacy Analytics (Ottawa-based company)
    - Privacy Analytics can review datasets as part of their Data Privacy Validation Services.[39]
    - Methodology based on the HIPAA Expert Determination De-identification Standard.
    - To find out more about their services, please fill in the form at the bottom of their "Certification" webpage.[40]

[36] "d-wise Offers Free Transparency Services Accelerating COVID-19 Vaccine Research," Cision PRWeb, March 10, 2020, https://www.prweb.com/releases/d_wise_offers_free_transparency_services_accelerating_covis_19_vaccine_research/prweb16970368.htm.

[37] "d-wise offers anonymization services available on Vivli COVID-19 portal," Center for Global Clinical Research Data, April 13, 2020, https://www.prweb.com/releases/d_wise_offers_free_transparency_services_accelerating_covis_19_vaccine_research/prweb16970368.htm.

[38] "FAQs," OpenICPSR, accessed August 31, 2020, https://www.openicpsr.org/openicpsr/faqs.

[39] "Clinical Trial Transparency Services," Privacy Analytics, accessed on August 31, 2020, https://privacy-analytics.com/clinical-trial-transparency/ctt-services/.

[40] "Double-check your data and leverage it with confidence," Privacy Analytics, accessed on August 31, 2020, https://privacy-analytics.com/health-data-privacy/health-data-services/expert-data-opinion-services/.

Resources

1. Amnesia https://amnesia.openaire.eu/
2. ARX https://arx.deidentifier.org/overview/
3. d-wise https://www.d-wise.com/de-identification-services
4. mu-Argus https://github.com/sdcTools/muargus
5. Inter-university Consortium for Political and Social Research (ICPSR) https://www.openicpsr.org/openicpsr/
6. Privacy Analytics https://privacy-analytics.com/services/certification/
7. sdcMicro https://cran.r-project.org/web/packages/sdcMicro/index.html
8. The University of Texas at Dallas Anonymization Toolbox http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php

References

1. Alder, Steve. “What is Considered PHI Under HIPAA Rules?” HIPAA Journal, December 28, 2017. https://www.hipaajournal.com/considered-phi-hipaa/.
2. Benschop, Thijs, and Matthew Welch. “Statistical Disclosure Control for Microdata: A Practice Guide for sdcMicro.” International Household Survey Network. Accessed August 10, 2020. https://sdcpractice.readthedocs.io/en/latest/index.html.
3. Clunie, David A. “DicomCleaner™.” PixelMed Publishing™. Accessed July 16, 2020. http://www.dclunie.com/pixelmed/software/webstart/DicomCleanerUsage.html.
4. Documenting the Now. “DocNow/hydrator.” GitHub. Accessed August 4, 2020. https://github.com/DocNow/hydrator.
5. Documenting the Now. “DocNow/twarc.” GitHub. Accessed August 4, 2020. https://github.com/docnow/twarc.
6. El Emam, Khaled, and Fida Kamal Dankar. “Protecting Privacy Using k-Anonymity.” *Journal of the American Medical Informatics Association* 15, no. 5 (September 2008): 627–637. https://doi.org/10.1197/jamia.M2716.
7. Elliot, Mark, Elaine Mackey, Kieron O’Hara, and Caroline Tudor. *The Anonymisation Decision-Making Framework.* UK Anonymisation Network (UKAN). University of Manchester. 2016. https://ukanon.net/ukan-resources/ukan-decision-making-framework/.
8. “Genetic privacy.” [Editorial]. *Nature* 493 (January 24, 2013): 451. https://doi.org/10.1038/493451a.
9. Global Alliance for Genomics & Health. *Genomic Toolkit: Regulatory & Ethics Toolkit.* Toronto, ON: Global Alliance for Genomics and Health. Accessed July 20, 2020. https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/.
10. Government of Canada (Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada, and the Social Sciences and Humanities Research Council). *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans.* December 2018. https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2018.html.

11. Gulban, Omer Faruk, Dylan Nielson, Russ Poldrack, John Lee, Chris Gorgolewski, Vanessasaurus, and Satrajit Ghosh. “Poldracklab/pydeface: V2.0.0.” Zenodo. October 31, 2019. http://doi.org/10.5281/zenodo.3524401.
12. Gymrek, Melissa, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich. “Identifying Personal Genomes by Surname Inference.” *Science* 339, no. 6117 (Jan 18, 2013): 321-324. https://doi.org/10.1126/science.1229566.
13. Hrynaszkiewicz, Iain, Melissa L. Norton, Andrew J. Vickers, and Douglas G. Altman. “Preparing Raw Clinical Data for Publication: Guidance for Journal Editors, Authors, and Peer Reviewers.” BMJ 340 (January 29, 2010): c181. https://www.bmj.com/content/340/bmj.c181.
14. Information and Privacy Commissioner Ontario. *Deidentification Guidelines for Structured Data*. Information and Privacy Commissioner of Ontario. June 8, 2016. https://www.ipc.on.ca/privacy-organizations/de-identification-centre/.
15. International Household Survey Network. “Measuring the Disclosure Risk.” Accessed August 4, 2020. https://ihsn.org/anonymization-risk-measure.
16. International Neuroimaging Data-Sharing Initiative (INDI). *Data Contribution Guide*. Accessed August 4, 2020. http://fcon_1000.projects.nitrc.org/indi/indi_data_contribution_guide.pdf.
17. Kirby, Justin. “Submission and De-identification Overview.” The Cancer Imaging Archive (TCIA), University of Arkansas for Medical Sciences. April 27, 2020. https://wiki.cancerimagingarchive.net/display/Public/Submission+and+De-identification+Overview.
18. Mannheimer, Sara, and Elizabeth Hull. “Sharing Selves: Developing an Ethical Framework for Curating Social Media Data*.” International Journal of Digital Curation* 12, no. 2 (April 18, 2018). https://doi.org/10.2218/ijdc.v12i2.518.
19. Medical Imaging & Technology Alliance, DICOM Standards Committee. “DICOM Part 15: Security and System Management Profiles.” In DICOM Standard. Arlington, VA: National Electrical Manufacturers Association. Accessed August 4, 2020. https://www.dicomstandard.org/current/.
20. Moore, Stephen M., David R. Maffitt, Kirk E. Smith, Justin S. Kirby, Kenneth W. Clark, John B. Freymann, Bruce A. Vendt, Lawrence R. Tarbox, and Fred W. Prior. “De-identification of Medical Images with Retention of Scientific Research Value.” *RadioGraphics* 35, no. 3 (May 13, 2015). https://doi.org/10.1148/rg.2015140244.
21. National Human Genome Research Institute. “Privacy in Genomics.” February 24, 2020. Accessed August 10, 2020. https://www.genome.gov/about-genomics/policy-issues/Privacy.
22. North Carolina State University Libraries. “Social Media Archives Toolkit.” Accessed August 4, 2020. https://www.lib.ncsu.edu/social-media-archives-toolkit.
23. Portage COVID-19 Working Group, “Can I Share My Data?” September 25, 2020. https://doi.org/10.5281/zenodo.4041661.
24. Portage COVID-19 Working Group. “Documentation and Supporting Materials Required for Deposit.” September 25, 2020. https://doi.org/10.5281/zenodo.4042034.
25. Portage COVID-19 Working Group. “Recommended Repositories.” September 25, 2020. https://doi.org/10.5281/zenodo.4042037.
26. International Household Survey Network. “Anonymization Principles.” Accessed August 4, 2020. https://ihsn.org/node/137.
27. Radiological Society of North America, International COVID-19 Open Radiology Database. “RSNA International Covid-19 Open Radiology Database (RICORD) De-identification Protocol.” Accessed August 10, 2020. https://www.rsna.org/-/media/Files/RSNA/Covid-19/RICORD/RSNA-Covid-19-Deidentification-Protocol.pdf.

28. Radiological Society of North America, International COVID-19 Open Radiology Database. “RSNA COVID-19 DICOM Data Anonymizer.” Accessed August 10, 2020. https://www.rsna.org/-/media/Files/RSNA/Covid-19/RICORD/RSNA-Anonymizer-Program-Instructions.pdf.
29. Radiological Society of North America, Medical Imaging Resource Community (MIRC). “Clinical trial processor (CTP).” Accessed August 4, 2020. https://www.rsna.org/research/imaging-research-tools.
30. Ryerson University Research Ethics Board. “Guidelines for Research Involving Social Media.” Ryerson University. November, 2017. https://www.ryerson.ca/content/dam/research/documents/ethics/guidelines-for-research-involving-social-media.pdf.
31. Thorogood, Adrian. “Canada: Will Privacy Rules Continue to Favour Open Science?” *Human Genetics* 137: 595–602 (July 16, 2018). https://doi.org/10.1007/s00439-018-1922-z.
32. Twitter. “Developer Terms: More About Restricted Uses of the Twitter APIs.” Accessed August 4, 2020. https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases.
33. UK Data Service. “Anonymisation: Qualitative Data.” Last modified June 30, 2020. https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/qualitative.aspx.
34. Zeffiro, Andrea, and Jay Brodeur. “Social Media Research Data Ethics and Management.” Workshop presented April 5, 2018. Sherman Centre for Digital Scholarship. McMaster University. http://hdl.handle.net/11375/25327.