National
Coordination Point
Research Data
Management

Dealing with pseudonymization and key files in small-scale research

A few basic steps

VERSION: DECEMBER 2019

LCRDM

The National Coordination Point Research Data Management (LCRDM) is a national network of experts on research data management (RDM) in the Netherlands. The LCRDM connects policy and daily practice. Within the LCRDM experts work together to put RDM topics on the agenda that ask for mutual national cooperation.

more information: www.lcrdm.nl

COLOPHON

Dealing with pseudonymization and key files in small-scale research
A few basic steps

PUBLICATION DATE | December 2019
DOI | 10.5281/zenodo.3571046

LCRDM Task Group Pseudonymization | **Simone van Kleef** (St. Antonius Hospital), **Jan Lucas van der Ploeg** (University Medical Center Groningen - UMCG), **Martiene Moester** (Leiden University Medical Center - LUMC), **Henk van den Hoogen** (University Maastricht/liaison LCRDM advisory group), **Erik Jansen** (University Maastricht/liaison DataverseNL), **Tineke van der Meer** (University of Applied Sciences Utrecht), **Francisco Romero Pastrana** (University Groningen), **Jolien Scholten** (Free University/VU), **Leander van der Spek** (University for Humanistic Studies), **Ingeborg Verheul** (LCRDM)

Consultancy Group | **Derk Arts** (Castor), **Marlon Domingus** (Erasmus University), **Laura Huis in 't Veld** (DANS), **Nicole Koster** (University of Twente), **Karin van der Pal** (Leiden University Medical Center – LUMC), **Alfons Schroten** (University Maastricht/MEMIC).

HAND OUT | Boudewijn van den Berg (LCRDM)
DESIGN | Nina Noordzij, Collage, Grou
TRANSLATION | Gosse van der Leij

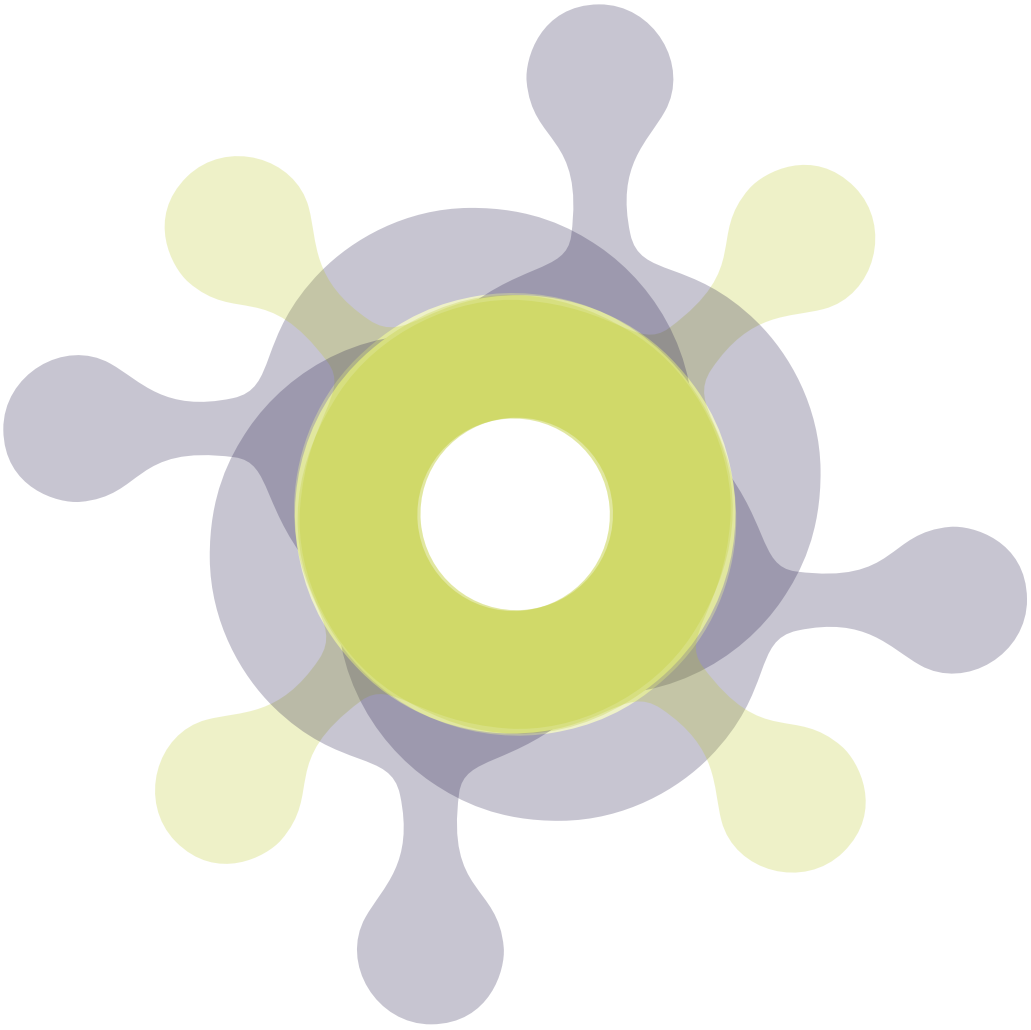
LCRDM

LCRDM IS SUPPORTED BY

Dealing with pseudonymization and key files in small-scale research

A few basic steps

Contents



Dealing with pseudonymization and key files in small-scale research

A few basic steps

1] Introduction

Everyone who carries out scientific research with (sensitive) personal data, faces the same questions: How can we guarantee to participants in research that their personal data will be stored safely? How do we ensure that directly identifiable personal data, necessary for communication and organization of research, are separated from research data?

It is common for datasets with (sensitive) personal data for research, to be stored in a data management system in pseudonymized form. For major research projects, part of the budget is often reserved for pseudonymization by Trusted Third Parties (TTP's), but in the case of small-scale research this is not possible. There is less time and money available.

What guidelines should be followed in such circumstances? Especially managing key(files) to link directly identifiable personal data and research data requires special care. How to store it? Who has access? Where to store it? And how can you make sure that access to the key file does not depend on the knowledge of one person and will also be available in the future in one form or another? Is there already a software application that has a satisfactory solution for these issues?

In the period between February and June 2019, a LCRDM task group investigated if there are practical ways of pseudonymizing data for small-scale research which can also be used - relatively simply - by other institutions. In the absence of such, suggestions could be made about how best to take the first steps to apply pseudonymization in small-scale research.

2] What is pseudonymization?

In a large number of research disciplines, pseudonymization has been common practice for some time. During research it may be necessary to identify research participants, for example, in order to verify source data or to monitor persons over a longer period of time. In such cases research data cannot be anonymized. Researchers then opt for pseudonymization; not only to protect the privacy of research participants, but also with regard to scientific integrity (for research, researcher and research institute). The recent publicity surrounding the implementation of the AVG (the Dutch equivalent of the GDPR) has generated extra interest for this subject.[1]

Pseudonymization is here understood to mean: replacing the directly identifiable variables in a dataset with a pseudonym. In some circles it is also called *coding*.[2] This way of working does not mean that a whole dataset has been pseudonymized. If the dataset contains free text fields, they may contain potentially directly traceable data. In addition, a combination of other (not directly identifiable) variables that are important for the research in question, may lead to the identification of a person.

The purpose of pseudonymization in thís form is therefore not to obtain an anonymous dataset (or as anonymous as possible[3]), but to protect the privacy of research participants from the onset, during the collection of data. Moreover, within the context of scientific integrity, the directly identifiable variables must be withheld from the researcher.

1 There was an overwhelming demand for clarity concerning this subject as proven by the speed with which people signed up for the task group. Within 24 hours, a ten-member task group had been formed together with a consultation group comprising a number of interested parties.

2 Appendix 1 contains a list of relevant definitions.

3 As opposed to pseudonymized data, anonymous data cannot be traced back to a person in any manner. A different task group is concerned with the issues of working with anonymous data.

3] What is small-scale research?

The task group focused on guidelines for pseudonymization during small-scale and quantitative research. We define small-scale research as: research with a limited number of participants and/or restricted financial means.

The limitation to quantitative data stems from the fact that pseudonymization of qualitative data (e.g. video and audio files and transcripts thereof) necessitate other measures than simply replacing directly identifiable variables in a data file with a code. Videos can show recognizable images of people or during an interview that is subsequently quoted in the accompanying transcript.

4] Approach

First, based on our own experience and available literature, the task group has made an inventory of how pseudonymization is applied in small-scale research. A survey has been drawn up to gain a better and more complete picture of current pseudonymization practices in research institutes. It was distributed via different channels, especially among data management support staff. The survey asked how data was pseudonymized and who was responsible; if and what kind of software was used; where key files were stored; who had access and what happened to the key files once the project had been completed; if the institute had a policy for pseudonymization and which problems it faced. Of 32 respondents, including (several) researchers and research support staff, 26 used (a form of) pseudonymization.

The most important conclusions that can be drawn from this survey are:

1. Most institutes do not use specific pseudonymization software for pseudonymizing data. Some institutes do have certain tools but these cannot be directly deployed outside their own research, or institute. Currently these tools are therefore not useful on a national level.
2. Most institutes do not have policy concerning pseudonymization or a subfield thereof, for example, dealing with key files. The variety of answers also show that opinions about whether or not something is permitted, differ widely.

Appendix 2 contains a more comprehensive description of the survey results.

To complement the survey, use cases in the daily practice at institutes of the task group members were examined. Relevant laws and regulations and other related documentation were also studied (see appendix 3). This led to the formulation of a list of basic steps for the pseudonymization of data.

5] Basic steps

This report is aimed at research support staff, researchers and/or research institutes that have little knowledge of pseudonymization and lack sufficient tooling/infrastructure.

Before going through the following basic steps, it is advisable to first check if there already is an existing policy for pseudonymization at the institute where you work. Contact a specialist within your own organization if you have questions about implementation of (one of) the measures described below. Institutional policy always takes precedence over the general basic steps listed here.

The task group identifies the following basic steps that researchers and research support staff can follow when pseudonymizing datasets for small-scale research.

1. In the data management plan, describe why and how you're going to pseudonymize data, how access to the separately stored key file and the dataset is regulated and what happens to the key file and the data when the project is completed.

2. Identify the following categories in your data:
   - Data necessary for identification, to organize research or to communicate with research participants
     - »» Store these in a key file
   - Data required for analysis
     - »» Preferably stored in a data management system[4]
   - Data not needed (e.g. in case of a supplied dataset)
     - »» This data should be deleted.

3. Pseudonymize the data as quickly as possible, i.e. immediately when collecting data. If you are sent a dataset with identifiable data by another party, pseudonymize the data immediately after receiving it.

4. Use different pseudonyms for different datasets. This prevents that data from participants who feature in multiple datasets can be linked via the pseudonym.

5. Store the key file separately from the research data.

6. Access to the key file should preferably be managed by someone who is not involved in the research project.

7. Make sure that the key file and the data are adequately backed up and secured.

4 A research data management system (DMS) is a programme which allows you to store and manage research data. A high-quality DMS records all activity in the research data base (audit trail) and ensures adequate security.

8. Take technical and organizational measures to prevent unauthorized people from linking the key file to the research data. After the data collection, persons in the role of the researcher should be denied access to the key file.

9. Limit access to the key file, but ensure that within the organization there is always someone who can has access.

6] In conclusion

This report does not intend to describe the full spectrum of pseudonymization and its proper application. It has a limited scope (pseudonymization for small-scale research). As it turned out, there is considerable variation in what institutes consider adequate measures for pseudonymization. Moreover, there is insufficient clarity about what pseudonymization implies exactly.

On the basis of an inventory, the task group initially set itself the goal of formulating demands and wishes for safe management of key files. The focus ultimately shifted to the formulation of basic steps. The reason for this alteration is that the inventory did not result in clear solutions or best practices that can be directly applied by other institutes. Tools referred to in the survey answers were often not specific pseudonymization tools and mostly they were developed for large research institutes (university medical centres, universities). The inventory therefore failed to yield the kind of results that small-scale researchers could profit from straight away. The most important conclusion to be derived from the survey is that expertise and experience of pseudonymization and good management of key files is insufficient among many researchers and research support staff. With these basic steps we've tried to offer some prerequisites for getting started with pseudonymization and managing key files.

7] Recommendations

In addition to the basic steps for researchers and research support staff, the task group makes the following recommendations:

1. Research institutions need clear and manifest policy for pseudonymization and in particular for the management of key files during and after research. In addition, there should be infrastructure in place where research data and identifiable data can be stored separately, preferably in two independent, adequately secured environments.

2. It would be beneficial if lcrdm organized a network day with privacy experts, policy makers and research support staff to draw up clear-cut definitions for privacy-related concepts, particularly for the terms *pseudonymization* and *anonymization*. There are no clear definitions of privacy-related concepts that are broadly accepted in the research world. Depending on the context of the data/research and the background of participants in the discussion, this can lead to very divergent definitions and irreconcilable viewpoints. For the one pseudonymization is the same as coding but for the other encoded data is not necessarily pseudonymized data. Some say it depends on the techniques used (encryption, secured environments) while others maintain that pseudonymized data is anonymous and yet others argue that anonymity depends on the user of the data. The different definitions not only lead to confusion; it's also not always clear if the correct precautionary measures have been taken before embarking on research. It would be good to agree on broadly accepted definitions and starting points with research support assistants, policy makers and privacy experts.

3. A following task group could make a comprehensive inventory of tools for pseudonymization and the storage of data keys. On the basis of this, requirements for a generally available tool could be formulated. This calls for a different approach and a task group of which the members possess specific expertise, namely sufficient knowledge of privacy and pseudonymization, both technically and functionally.[5]

5 Based on current experiences, we think that directly approaching dpo's or ciso's (corporate information security officers) is likely to yield more than a survey among support staff. They often have knowledge of policy matters and regulations concerning privacy and can point the way to people in the organization who concern themselves with pseudonymization.

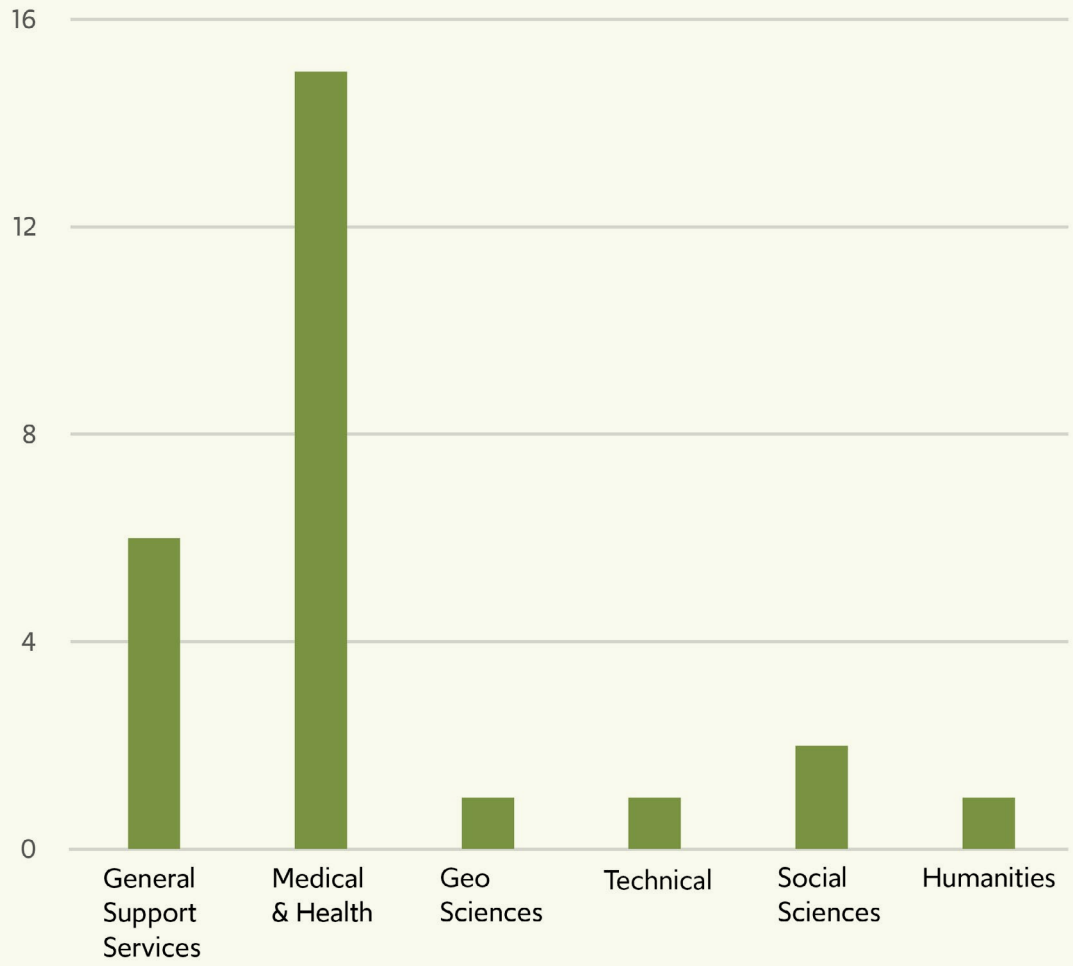
Appendix 1 Definitions

Personal data	Any information about an identifiable natural person ("the person concerned"); identifiable is considered to refer to a natural person who can be identified directly or indirectly, particularly by way of an identifier like a name, an identity number, location details, an online identifier or one or more elements that are characteristic of the physical, physiological, genetic, psychic, economic, cultural or social identity of the natural person in question (GDPR, article 4)
Pseudonymization according to the GDPR	The processing of personal data in such a way that personal details can no longer be linked to the specific person concerned without the use of supplementary data, on condition that these supplementary data are stored separately and technical and organizational measures are taken to ensure that the personal data are not linked to an identifiable natural person (GDPR, article 4). In pseudonymization, identifying data are separated from non-identifying data and replaced with artificial identifiers (GDPR Guidebook, pg. 27)
Pseudonymization according to the definition of the task group	Replacing directly identifiable data with a pseudonym. In medical research this is known as coding. This definition corresponds with that given in the GDPR Guidebook (see above), but strictly speaking not with the GDPR definition. Replacing directly identifiable data does not guarantee that the specific person concerned cannot be traced. Other data in the dataset, whether or not in combination with each other, can allow this to happen.
Code list or key file	File with the combination of code/pseudonym and corresponding directly identifiable data.

Appendix 2] Survey results

Most respondents have a background in medical science and research support. The reason for pseudonymization and the type of data used for research is diverse as shown by the information in the following graphs.

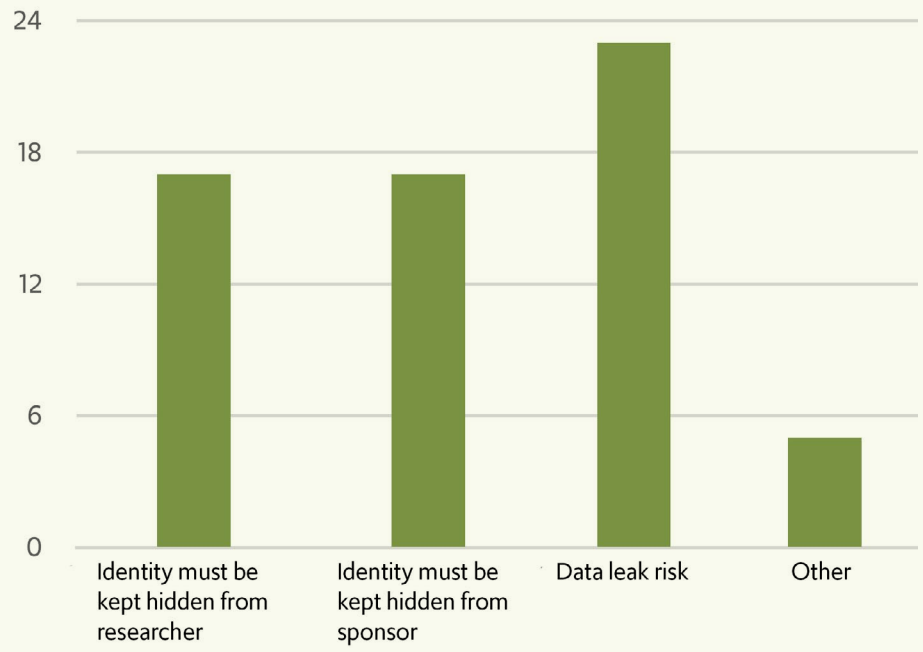
1] In which research discipline are you active?



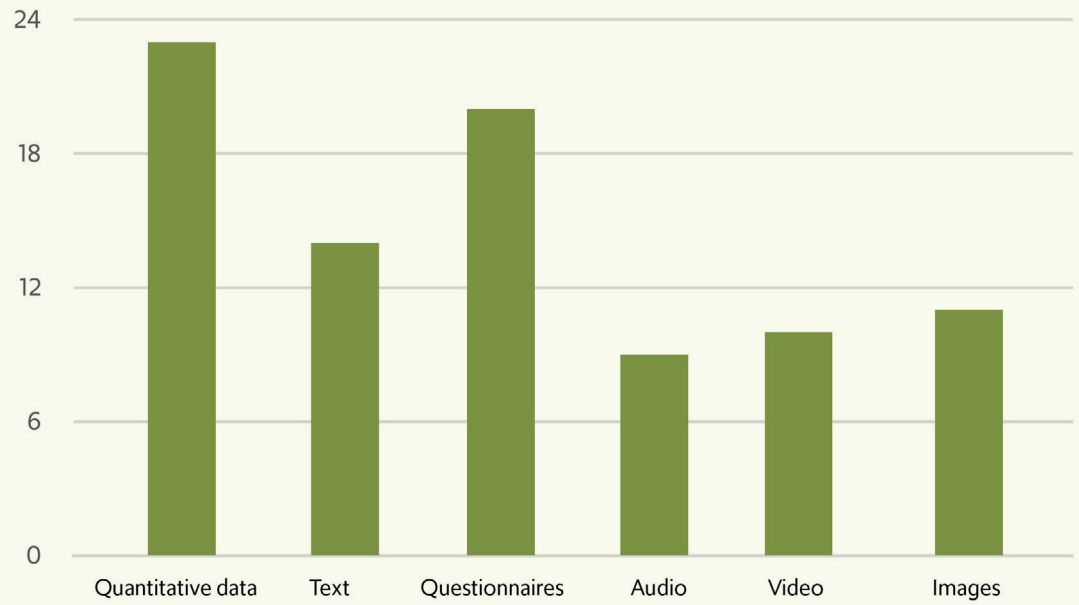
2] Which type of research uses coded personal data? (e.g.: qualitative/quantitative, WMO/nWMO, how many human subjects)

Most respondents answered that pseudonymization is used in all types of research, with qualitative and/or quantitative data, WMO or nWMO and involves twenty to ten thousand trial subjects.

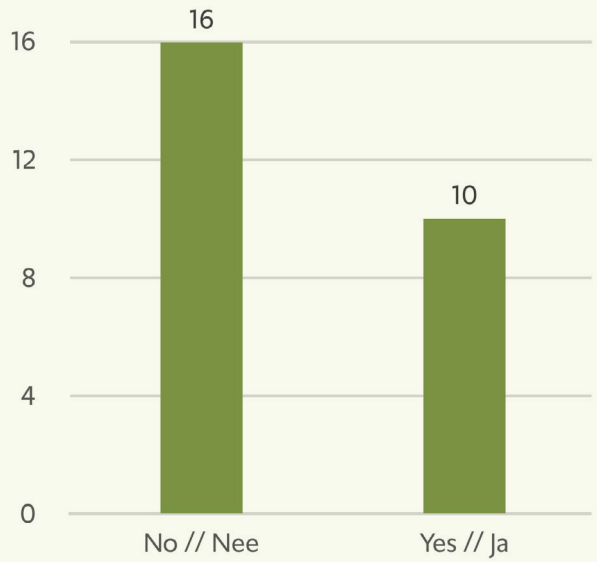
3] Why are personal data encoded in your organization?



4] What type of data are encoded?



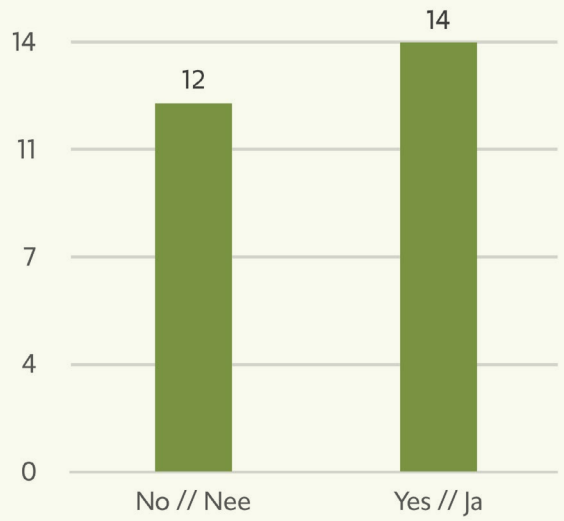
5] Has your organization formulated policy about how to encode personal data?



6] Where does your organization store research data?

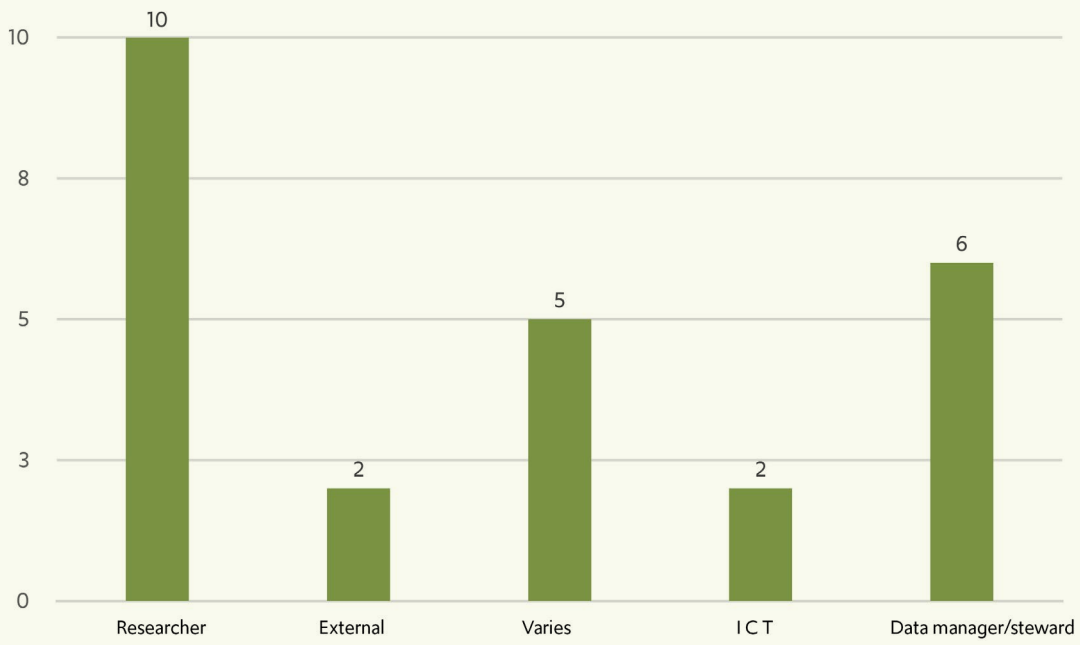
The most frequently cited answer is the network drive: 75% of respondents indicated they used this. Other locations often mentioned are: SURFdrive, eCRF/DMS and repositories. A number of respondents state explicitly that they suspect researchers also store private data on personal drives or in a dropbox.

7] Is software or other technical means used to encode data?



By far the most cited answer is that an institute specific research platform provides this functionality. Other options often mentioned are: SAS, encryption software like VeraCrypt and a ttp service.

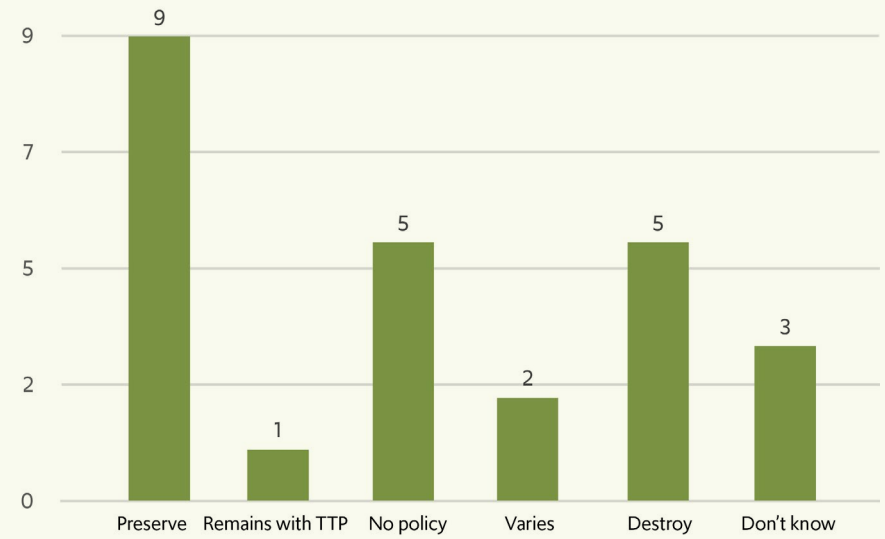
8] What is the role of the person responsible for encoding the dataset?



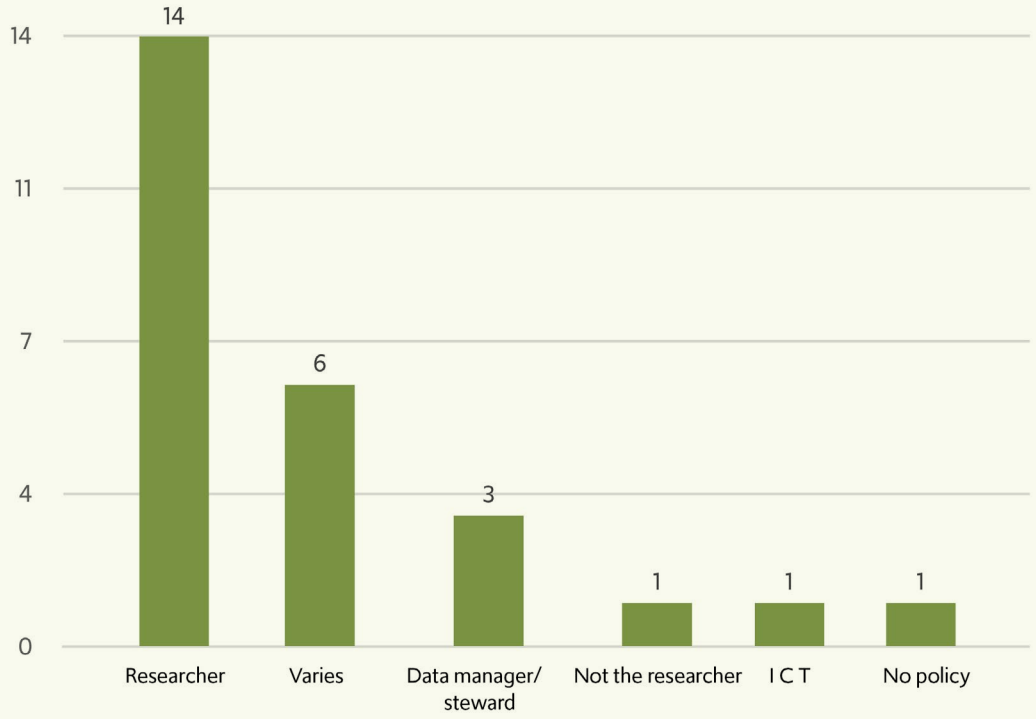
9] How and where is the key file stored?

Answers to this question are very diverse. Only five cases explicitly state that the key file is stored separately from the data. Six respondents name the same location as where they said data was stored. Nearly everyone claims that the file or directory is secured and can only be accessed by authorized personnel. However, there is no consensus about who is authorized. For some respondents only data managers are authorized, but in most cases the entire research team has access to the key file.

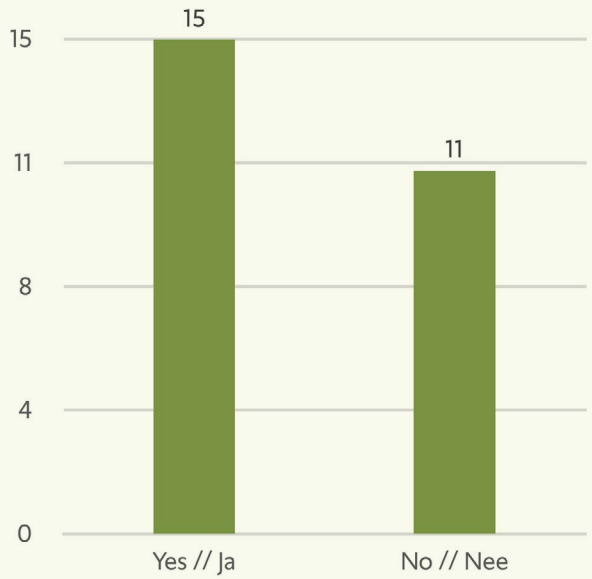
10] What happens to the key file after the research is completed?



11] Who has access to the key file?



12] Do you encounter problems during this process?



When asked to specify which problems they encounter with pseudonymization, respondents reply:

– Ambiguous guidelines: What is a good way to code personal data? How and where should you store key files?
– What to do with key files after the research has been completed?
– Who is responsible for the key file?
– Key files become lost
– Key files are insufficiently secured
– No possibilities for coding audio and video
– Multiple versions of key files
– People re-identify participants (or attempt to)
– Absence of proper monitoring; therefore it is not clear how researchers deal with data and key files.

Appendix 3] List of references

– European General Data Protection Regulation (GDPR)
– Gedragscode gezondheidsonderzoek (Code of conduct for healthcare research)
– Infographic What is personal data?
– ISO 25237:2017 - Health informatics - Pseudonymization
– Wet medisch-wetenschappelijk onderzoek met mensen (WMO) (Social Support Act)
– Wet op de geneeskundige behandelingsovereenkomst (WGBO) (Medical Treatment Agreement Act)
– Whitepaper on pseudonymization by the Data Protection Focus Group
– Pseudonimization guide for research data - Concept (F. Romero Pastrana, RUG)