Seminar Series:

Facts and pitfalls of observational studies - How to plan and conduct HRO projects

Q&A from the session

"De-Identification"

June 25, 2025

- Differences between de-identification and pseudonymized/ anonymized datasets.
- What are the legal definitions and requirements of anonymisation, pseudonymisation and de-identification and what are the practical implications for research projects?
    - ➢ De-identification is the overarching term for pseudonymization/ coding and anonymization
    - ➢ Practical implications:
        - data protection in clinical trials and research projects according HRO chapter 2
        - data protection and possibility to apply/ make use of the General Consent (GC) by obtaining consent from the persons providing their data / samples to a further use research project according to HRO chapter 3

- What regulations beyond HRO do apply to HRO Projects? Federal / cantonal act on data protection?
    - Covered by the speaker Thomas Gruberski during the presentation part I:
        - ➢ Federal Data Protection Act (new version enacted on September 1st 2023):
            - Applies for federal authorities & private organisations (such as pharma companies)
        - ➢ Cantonal Data Protection Acts, e.g. IDG BS: applies for USB
        - ➢ Pro memoria: GDPR (DSGVO)
        - ➢ HRA plus its ordinances: "Lex specialis" = takes precedence over the Data Protection Acts
            - If no research (in terms of the HRA [see next slide]): Data Protection Acts still applicable.
            - Handling with anonymised data: Data Protection law no longer applicable.

- Does the federal data protection act also apply to public organisations?
    - Answered by the speaker Thomas Gruberski in the Zoom chat:
        - ➢ It applies to federal public organisations (such as ETH, EPFL...).

- Is there any guidance that would support me communicating the requirements to lay persons (e.g. students who have to draft consent forms or colleagues from the IT department)?
    - Answered by Thomas Gruberski after the session for the Q/A document:
        - ➢ I recommend the text concerning data protection in the ICF-templates of swissethics. There, the process of pseudonymisation is explained in a way that also lay persons should understand it. Example: Information about the use of health-related data and

samples for research purposes (https://swissethics.ch/assets/studieninformationen/vorlage_gk_e.pdf), page 1: «If your data and samples are used for a research project, they will be coded or anonymised. Coded means that all personal information such as your name or date of birth is replaced by a code. The key showing which code belongs to which person is kept safe by a professional who is not involved in the research project. People who do not have the code are not able to identify you. In case of anonymisation, the link between the biological material and/or associated data and the participant is definitely removed so that no specific participant can be re-identified.»

- Are the definitions of anonymization and pseudonymization internationally harmonized?
  - Answered by Thomas Gruberski after the session for the Q/A document:
    - ➢ As far as Europe is concerned, yes - for the most part. See, for example:
      - Clause 4 (5) EU General Data Protection Regulation (GDPR): «Pseudonymisation means 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.»
      - Old version of § 3 (6) of the German Federal Data Protection Act: «Anonymisation is the alteration of personal data in such a way that the individual information about personal or material circumstances can no longer be attributed to a specific or identifiable natural person or can only be attributed with an unreasonable amount of time, cost and labour.»
      - Clause 18, Module 4, footnote 2 of the Adoption by the European Commission of the Implementing Decisions (EU) 2021/914: «…requires rendering the data anonymous in such a way that the individual is no longer identifiable by anyone ... and that this process is irreversible.»

- When is it mandatory, to submit a risk-assessment form for de-identification together with the ethics application?
  - ➢ As mentioned by the speaker Thomas Gruberski during the part I of the session: it is always mandatory.

- What is the impact of the different types of de-identification for the planning of a further use research project?
  - ➢ It has an impact on the application/ usage of the General Consent as a type of consent obtained from the persons providing their data / samples to the further use research project

- The method used for anonymisation must be documented, including a description of the residual risk of re-identification. But why…?
  - Asked during the s by presentation part I by the speaker Thomas Gruberski to the audience.

- Commented/ reflected on that by participants:
  - Maybe it is useful if you publish the data and will be asked by the journal for the method of anonymisation.
  - It's not just about justifying something to the ethics committee or journals - it's about assessing and MANAGING risk!
  - Assessing the risk of re-identification helps better evaluate whether the re-identification effort is truly disproportionate or not, i.e. whether the data can be considered as fully anonymised.

- How do I assess proportionality with respect to re-identification? Is there any guidance? Do I need to develop a threat model?
  - Partially answered by the speaker Thomas Gruberski in the chat:
    - SPHN has developed a «De-Identification Risk Assessment Tool: https://sphn.ch/network/data-coordination-center/de-identification/

- Data should only be shared/published in anonymised way. How is this now possible, if to correctly anonymise data I need to delete the key, however, at the same time I need to store the Subject Identification Log (that includes the key) for 10/20 years?
  - Partially answered by the speaker Thomas Gruberski in the chat:
    - The storage requirements for research data and documents, etc. account for projects that fall into the scope of the Human Research Act (HRA), but anonymized data do not fall into the scope of the HRA and therefore do not have to be stored according to the requirements described in the law (Art. 2 HRA).
    - 20 years retention period do only apply to data / documents from clinical trials according to ClinO (Art. 45 ClinO)
    - 10 years retention period do only apply to project data from research projects according to HRO chapter 2 (Art. 23a HRO), not data / samples of further use research projects according to HRO chapter 3.
    - The term “subject identification log” refers to clinical trials only.

- Why should ethical approval not be required for health-related data which is anonymised only after data collection?
  - Answered by the speaker Thomas Gruberski during the Q/A part of the session:
    - Yeah, that's correct. And thank you for the question. It's, it's an important issue. I wasn't focusing on that, because it's rather a question concerning the fact when do you have to get an approval, that is in which cases and and which cases not. The common understanding is you have to get approval by the EC if you have a concrete question to be answered and you have already a protocol in which it is layed how to answer the question. But this is only half of the truth. Because if we are talking about the scope of the HRO, it also says that – I`m just going to read that for the purposes of the chapter about gaining extra material or gaining extra data for research purposes, then you have already a research project. If you just want to make further use for research purposes of the biological material or the health rates personal data. So what is meant by that is that even if I just specifically take data from people and I do not have the research question in mind already, I'm just going to build a bio bank

or a database, then - as long as it is not anonymized at that moment, because if I anonymize it right before collecting, then EC is out of the game.

- But if want to keep all the identifying factors and I'm just collecting the data, then it's a research project in the terms of the HRA and therefore I have to have an OK, an approval from the EC.
- So the difference is, am I anonymizing right from the beginning then the EC is out. Or am I thinking maybe about anonymizing, there is not a question in talking about the log file that this one has to be stored, but this is another consolation from my point of view, this means the fact or the consolation that I I'm thinking about anonymization at the end, not at the beginning, but we will come to that eventually.

- If the researcher is responsible of a registry containing identifying data, how is it possible for him/her to do research on the data even he/she has to keep the key for the registry?
- Interested on this topic for registry and clinical research project using registry-collected data. Identification seems to be necessary in the registry to avoid duplicates. Is it correct to de-identify just before analysis? Or are there ways to de-identify patients in the registry that still avoid duplicates?
    - Partially answered by the speaker Thomas Gruberski during the Q/A part of the session:
        - If a person is responsible for feeding and managing a registry, then he/she is allowed to work with identified / not pseudonymized/ not coded data. This person could be even the responsible person for extracting data for a research project team and code them appropriately and store the code list/ key
        - If one and the same person runs a registry and is also project leader for a further use study making use of data from this registry then he/she can do so and can even code/ pseudonymize the data, but from a regulatory perspective these data will NOT be considered as “pseudonymized/coded”, but rather “un-coded/ not pseudonymized”.
            - If these un-coded data are non-genetic data, then there is no problem by even applying the general consent (GC) obtaining consent (Art. 31 HRO)

- How can I ensure de-identification as I need the link to the informed consent and how can I ensure anonymisation during project running when I need the link to the participant. Is the article 26 meant after study end as during the study running this is not possible?
- Isn't the patient identification list (key list for the pseudonymization) usually kept by the Hospital trial team (e.g. in the "Investigator Site File"). As it is also needed for data entry, for example. However, the coding list should it be kept by a party not involved in the Research Project according Art. 26 Abs. 2? How exactly can this be done in practice?
    - It needs to be distinguished:
        - If we code participants in prospectively designed clinical trials or research projects according to HRO chapter 2, the Investigator with the study team or the project leader with the project team codes/ pseudonymizes the data and the code list by themselves - the key remains with the researchers being stored in the study folder/ Investigator Site File (ISF) → the Investigator/ study team/ project leader still knows which patient is behind the code!
        - The term pseudonymize/ coded in the context of further use research according to HRO chapter 3 is differently meant → here data/ samples are considered to be pseudonymized/ coded if the researcher/ project team does NOT know who is behind the code! → the data is extracted from a clinic information system and is

coded/ pseudonymized by an entity/ person/ organizational unit that is NOT involved in the research project (e.g. someone else from the clinic not working on the same project) and only transfers the data in a coded format, and eventually even stores the code list/ key

- We (Eye Clinic USB) work closely with a Research Institute in Ophthalmology. Often we are asked, why we cannot "just" provide them with data. Would it be possible if:
  - the Eye Clinic has an approved retrospective further use project for data collected in the Eye Clinic with permission to share the data,
  - has concluded a DTUA with that institute
  - provides the data to the institute in a pseudonymized/ coded (code list/ key will never leave the USB)
- Does the institue need an EC permission although the data would appear anonymized to them?
  - If it is an approved further use project of the Eye Clinic USB, then the Eye Clinic requires approval by the EC
  - If the Eye Clinic plans/ describes the intended sharing of the project data in coded/ pseudonymized form with the Ophthalmology Institute as a collaboration partner based on a DTUA, then the Ophthalmology Institute does not need an EC approval
  - If the researchers/ project team of the Eye Clinic USB extracted the data from the clinic information system by themselves and not an entity being not involved in this project, the data are not considered to be “coded/ pseudonymized” for the researchers at the Eye Clinic USB
  - But if they based on the DTUA are transferred to the Ophthalmology Institute in coded form and the code list/ key remains at the USB then the data are considered to be pseudonymized/ coded for the researchers at the Ophthalmology Institute

- How far must the key for coding be locked away? Would it be OK to have it on the same RedCap Instance two projects one the key and the other one the database?
  - Covered by the speaker Thomas Gruberski during the presentation part I
    - Coding must be effected using a method based on the current state of the art. The key must be stored separately from the biological material or personal data and in accordance with the principles of Article 5 paragraph 1, by a person or organisational unit to be designated in the application, not involved in the research project.

- Please also comment on anonymisation.
- In todays modern world, is the existence of an anonymized medical dataset not a wish, which never can be realised (from the data protection perspective of our patients), so better to forbid this term and talk about pseudonymized and de-identified datasets?
  - Covered by the speaker Thomas Gruberski during the presentation part I based on HRO Art. 25
    - For the anonymisation of biological material and health-related personal data, any association with a specific person must be rendered (1) impossible or (2) eliminated in such a way as to allow this association to be re-established only with disproportionate effort.
    - Anonymisation must be effected using a method based on the current state of the art.

- Impossibility of Re-Identification as 100% - not required by law (!) “Disproportionate effort” is sufficient.

- Merci d'indiquer ce qu'une Commission d'éthique doit vérifier en plus du terme adéquat (codé, pseudonymisé, anonymisé)
  - Covered by the speaker Thomas Gruberski during the presentation part I
    - The method used for pseudonymization / anonymisation must be documented, including a description of the residual risk of re-identification
    - The EC reviews if the coding process is correct and secure (HRO Art. 34)

- About data shift, wouldn’t a data as the real admission date be useful for the project? I mean that if the date of admission, for example, and you shift the data to code it or to anonymize it wouldn't that be for the researcher important information?
  - Answered by the speaker Matthias Joos during the Q/A part of the session:
    - Of course, yes, it is maybe very valuable information for some of the research projects when we receive data request and it's stated there that original date time fields are required, of course, we can leave them. So if we receive a data request, and of course the underlying ethics proposal also clearly states that original date values are needed, then of course we leave everything as it is.
    - So if the data are relevant for the project, then they are not modified nor deleted.

- How are the rules for coding/anonymization MRI/CT images?
- Remaining issues regarding anonymization of MRIs
- Pourriez-vous montrer un exemple concret de dé-identification sur une imagerie (radio, scan, etc.)
  - All three questions covered by the speaker Matthias Joos during the presentation part II:
    - Please see slides of the presentation part II (pdf of the presentation and / or video recording)
    - In radiology, images are stored in the so-called DICOM format. And there we actually need to perform 2 de-identification steps: so first the de-identification of the metadata (metadata is hidden data which is not visible on the image itself). There we perform more or less the same de-identification steps with replacing some original IDs with the hash key. And then the more, I'd say complicated second step is the removal of identifiable information on the image itself, because CT or MR devices, they store identifiable information directly on the image. And so historically, we tried to cut away this segment of the image, but then it turned out that the dimension were kind of cropped from the images and this made it unusable for further analysis. So we came up with a computational solution which actually blackens out the the corresponding segment or area of the image. And this is how we de-identify identifiable information on the image itself.
    - But even this sometimes is not enough because when we specifically talk about head CT or MR images…(so let's say we would deliver a series of a head CT or MRT images to a researcher, maybe consisting of 5000 single images or so)… And when a researcher uses standard tool to build from the single images 3D model again, then

maybe this reveals some characteristic anatomical features in the forehead. And that's why we had to came up also with computational method to de-identify had CT or MRT images. And together with a research group from University of Zurich, we implemented a tool which works with artificial intelligence or deep learning actually two years ago. This deep learning algorithm is trained on thousands of head images and it detects the segments of the forehead and sets kind of a boundary box around it. And in the next steps, it's kind of a brutal approach, but the whole segment of the forehead is caught away then. But still what remains is I think for researchers, the important areas of the image such as the whole brain structure, the vessel structure and so on.

- Are there approaches similar to the defacing method for other problematic data types (e.g. video taken during endoscopy where the face of the patient or the surroundings can be identified)?
- De-identification in the context of qualitative studies, with interviews and audio- and/or video recorded materials of participants
    - Answered by the speaker Matthias Joos during the Q/A part of the session:
        - Yes, I think it goes into the direction that, for example, our team needs to de-identify video material as well, then I think it would go into a similar direction which uses the the same technologies such as used for the defacing. So in other words «artificial intelligence». These are most likely algorithms which are trained really on video material, not only on single images.
        - But w haven't had such a de-identification request in our pipeline yet. I think our internal development team would be pushed if we receive such an enquiry for the de-identification of video material.

- Given recent advances in AI's ability to cross-analyse and correlate large health datasets, could such technologies be considered as tools to assess the reliability of de-identification methods in HRO projects? In parallel, while being mindful of potential re-identification risks, is there room within current or future Swiss regulation to support the use of AI for these purposes?
    - Partially answered by the speaker Matthias Joos during the presentation part II:
        - Nowadays I think everybody of you is using generative AI or large language models in their daily lives such as GPT and so on. And it's kind of clear that in the near future we need to replace this rule-based method (author`s note: tool to de-identify unstructured data, such as text) with a more, let's say, computationally sophisticated method. And I'm sure that there will be some developments that large language models which are specifically trained on that the identification of text will be available.