De-identification of health-related data

Recommended phased approach

Guidance for de-identification of health-related data in compliance with Swiss law requirements

Developed by the Swiss Data De-identification Project Task Force in the realm of the Swiss Personalized Health Network (SPHN), namely by Julia Maurer (Swiss Institute of Bioinformatics, Personalized Health Informatics, PHI), Marc Vandelaer (wega Informatik AG), Jean-Louis Raisaro (CHUV), Katie Kalt (USZ), Antje Thien (USZ), Fabian Prasser (BHI at Charite, Germany), Bradley Malin (Vanderbilt University, USA) and in collaboration with additional Swiss university hospital representatives.

Version 1.0 (06-May-2022)

A project of SAMW ASSM
Schweizerische Akademie der Medizinischen Wissenschaften
Académie Suisse des Sciences Médicales
Accademia Svizzera delle Scienze Mediche
Swiss Academy of Medical Sciences

SIB Swiss Institute of Bioinformatics

SIB | Swiss Institute of Bioinformatics
PHI | Personalized Health Informatics Group
www.sphn.ch | dcc@sib.swiss

Table of contents

1. Scope and Purpose

Health-related data governance practices in Swiss hospitals allow data sharing within a research project provided that certain conditions and criteria are fulfilled. Health-related research projects are projects in which biological material is sampled or health-related personal data is collected from a person in order to a) answer a scientific question or b) make further use for research purposes of the biological material or the health-related personal data[1].

For most of the Swiss research projects this includes the availability and approval of a set of documents (i.e., Project plan; Patients' informed consent; De-identification strategy (pseudonymization or anonymization) of project personal data; Ethical committee approval/statement; Legal agreement among project partners in case of cross-institutional data transfer) and measures related to information security and patients' privacy.

The de-identification of health-related data, which leads to pseudonymized or anonymized data, establishes together with other conditions an essential approach to protect patient privacy and is a mandatory prerequisite for data sharing among a broader research community. Even though there exist international guidelines concerning the de-identification of data[2,3] there is no guidance for the de-identification of health-related data specific to the conditions of the Swiss law and data protection regulations. Defining consolidated de-identification rules, however, appears to be crucial for the conduct of multi-center projects. Often research projects documenting the de-identification process are referring to the Safe Harbor methodology described in the Privacy Rule of the United States (U.S.) Health Insurance Portability and Accountability Act of 1996 (HIPAA)[4]. The HIPAA Safe Harbor methodology establishes, for the U.S., a rule-based approach to de-identify individuals' protected health information, i.e. it defines information that qualifies as (potentially) identifying and suggests suppression or rules for pseudonymization of these so-called identifiers. Protected health information is information that 1) relates to: i) the individual's past, present, or future physical or mental health or condition, ii) the provision of health care to the individual, iii) the past, present, or future payment for the provision of health care to the individual, and that 2) identifies the individual or for which there is a reasonable basis to believe that it can be used to identify the individual. Protected health information includes many common identifiers (e.g., name, address, birth date, Social Security Number) that can be associated with the health information listed above. Data sharing organizations in the U.S. need to attest that they do not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information. However, the 18 identifiers defined by the Safe Harbor method are criteria that are considered directly or reasonably indirectly identifying under U.S. law and cannot be executed as such in every country. They should thus be adapted to cope with the legal and data protection frameworks of the respective countries. For example, in Finland, where the EU General Data Protection Regulation (GDPR)[5] applies, the Finnish Social Science Archive (FSD) established a guideline to assess the choice of de-identification technique and the robustness of the outcome[6]. For the de-identification of health-related data in Switzerland, HIPAA cannot be executed as such and the characteristics of the Safe Harbor method need to be adopted as identifiers under Swiss Law (see also section 2).

[1] Human Research Ordinance SR 810.301, Art 6, https://fedlex.data.admin.ch/eli/cc/2013/642
[2] https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf
[3] Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington (DC): National Academies Press (US); 2015 Apr 20. Appendix B, Concepts and Methods for De-identifying Clinical Trial Data. Available from: https://www.ncbi.nlm.nih.gov/books/NBK285994/
[4] https://aspe.hhs.gov/reports/health-insurance-portability-accountability-act-1996
[5] https://gdpr.eu/
[6] https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/

Under such conditions, the Personalized Health Informatics (PHI) Group has launched the Swiss data de-identification project[7] and steered the development of a guidance document for de-identification of health-related data elaborated within the Swiss Data De-identification Project Task Force[8] in the realm of the Swiss Personalized Health Network (SPHN) initiative. The aim of the recommendations assembled in this guidance document is to further enhance secure sharing of health-related data among the Swiss research community by harmonizing the de-identification approach and the documentation within the Swiss biomedical research community. Even if steered by the SPHN, these recommendations have been developed in a generic manner, such that their principle approach is applicable to all Swiss health-related research projects that involve the further use of data. The recommendations will be aligned with swissethics[9] in order to establish a national harmonized approach for reducing the risk of re-identification by de-identifying data.

As a consequence, this guidance document should be considered as a helper dedicated to providing researchers and data providers a structured approach to the evaluation of the risk of re-identification stemming from processing and sharing health-related data during their research project and the applicable measures for reducing such risk. Data providers (mainly hospitals) benefit from this harmonized and structured approach to further improve the evaluation and privacy of data sharing in the scope of multi-center research projects.

International experiences and available publications addressing de-identification and data protection aspects for the further use of data in research have been taken as a source of inspiration while primarily focusing on compliance with Swiss legal and data protection framework (section 2 below)[10,11,12,13,14,15].

The Swiss Data De-identification Project Task Force has elaborated a methodology that is based on a risk assessment approach for health-related data de-identification. This type of approach aims at evaluating in a project-specific case-by-case manner the risk of re-identification and applying data de-identification rules and potentially other safeguards (contractual and/or technical) in order to reduce such a risk under an acceptable threshold. Risk assessment can be performed by relying on formal quantitative mathematical models or on heuristics that are based on practical methods and determination by experts[16]. The literature provides several examples of both approaches. As the former approach typically relies on assumptions that do not hold under all circumstances, preventing their application in many real-world scenarios, the Task Force decided to focus on the second approach. This strategy does not prevent the complementary use of formal and quantitative

[7] https://sphn.ch/network/data-coordination-center/de-identification/

[8] The Swiss Data De-identification Project Task Force consists of Julia Maurer (Swiss Institute of Bioinformatics Personalized Health Informatics (PHI), Marc Vandelaer (wega Informatik AG), Jean-Louis Raisaro (CHUV) and Katie Kalt (USZ), Antje Thien (USZ), Fabian Prasser (BHI at Charite, Germany), Bradley Malin (Vanderbilt University, USA) and in collaboration with additional Swiss university hospital representatives.

[9] https://swissethics.ch/

[10] https://www.ncbi.nlm.nih.gov/books/NBK285994/

[11] Foufi, V., Gaudet-Blavignac, C., Chevrier, R., & Lovis, C. (2017). De-Identification of Medical Narrative Data. *Studies in Health Technology and Informatics*. https://doi.org/10.3233/978-1-61499-824-2-23.

[12] Angiuli, O., Blitzstein, J. O. E., & Waldo, J. I. M. (2015). How to De-identify Your Data. *Privacy and Rights*, *13*, 1–20. https://queue.acm.org/detail.cfm?ref=rss&id=2838930.

[13] Mainz, J. G. (2014). Leitfaden zum Datenschutz in medizinischen Forschungsprojekten (Issue May 2020).

[14] Malin, B. (2012). Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

[15] Institute of Medicine. (2015). The Clinical Trial Life Cycle and When To Share Data. In Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk.

[16] El Emam K. Guide to the de-identification of personal health information. CRC Press; 2013 May 6.

methods for specific data sharing use cases when strong and formal guarantees are required (e.g., for publishing data on the internet[17]). We emphasize that this document provides recommendations and guidance for health-related data de-identification, and it may not be read as a technical specification. The implementation of de-identification rules and other data protection mechanisms remain in the sole responsibility of each data provider.

[17] Jakob CE, Kohlmayer F, Meurers T, Vehreschild JJ, Prasser F. Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. Scientific data. 2020 Dec 10;7(1):1-0.

2. Compliance with Swiss legal framework

Prior to the elaboration of the de-identification recommendations described in this document, the PHI Group requested an independent legal opinion to outline the Swiss legal framework applicable for the de-identification of health-related data. This legal opinion was provided by the Homburger AG and supports the establishment of guidelines in order to ensure that such de-identification is performed in accordance with Swiss law requirements. However, it must not be considered as an evaluation of the methodology described hereunder.

The legal opinion provided by Homburger AG[18] is available in a full memorandum[19] dated January 5, 2021. The following sections are derived from this memorandum and refer to the Swiss law requirements and its interpretation.

2.1. Specific terminology

In accordance with the Homburger AG memorandum, the following terms will be used to distinguish whether certain data is, or is not, linked to an identified or identifiable person:

- **Personal data** is data that relates to an identified or identifiable person; thus, the person with access to such personal data will be able to directly or indirectly identify the person concerned. It involves information concerning the health or disease of an identified or identifiable person, including genetic data[20].
- **De-identified data** is data for which identifying attributes to an identified or identifiable person has been either suppressed, replaced or modified so that the person with access to de-identified data (but not to the original identifying data) is, in principle, not able to identify the person concerned. De-identified data encompasses both anonymized and pseudonymized data. Please note that, although the umbrella term “de-identified data” is commonly used in international considerations, the Swiss law does not explicitly mention the term but solely refers to “anonymisiert” (= anonymized) and “verschlüsselt” (= pseudonymized).
- **Anonymized data** is data for which the de-identification is, in principle, irreversible, because no key or code exists to re-link the data to an identified or identifiable person[21]. The Human Research Ordinance (HRO) states in particular that name, address, date of birth and explicitly identifiable information has to be masked or deleted[22]. However, note that according to experts, sustainable anonymization requires more than only substituting the identifiers with pseudonyms and/or deleting the key and needs a careful case-by-case evaluation.
- **Pseudonymized data** is data for which the de-identification is, in principle, reversible because there is a key or code to re-link the data to an identified or identifiable person. Data is correctly pseudonymized (coded), if, from the perspective of a person who lacks access to the key, data is characterized as anonymized[23]. In the Human Research Act (HRA), the term "coded data" is used for pseudonymized data. Given that the German term for "coded data" (i.e., "verschlüsselte Daten") used in the HRA is misleading as it may be misinterpreted to refer to "encrypted data", the term "pseudonymized

[18] https://www.homburger.ch/en
[19] https://sphn.ch/wp-content/uploads/2021/04/Homburger-memorandum_Swiss-Legal-Framework-for-De-identification-of-Health-Related-Data_20210105.pdf
[20] Human Research Act, Art. 3, https://www.fedlex.admin.ch/eli/cc/2013/617/en
[21] Handkommentar DSG-ROSENTHAL, Art. 3 n. 35.
[22] Human Research Ordinance SR 810.301, Art 25, https://fedlex.data.admin.ch/eli/cc/2013/642
[23] Human Research Ordinance SR 810.301, Art 26, https://fedlex.data.admin.ch/eli/cc/2013/642

data" will be uniformly used instead of "coded data" in this document. Such term is also more frequently used in literature regarding the Data Protection Act (DPA)[24].

Additional terminology used in this document are listed in section 4.

2.2. De-identification methodology proposal

In the conclusion of its memorandum Homburger AG highlights that:

A. Swiss law does not provide specific methods or processes that are to be applied in order to de-identify personal data, including health-related data. It only defines what anonymized and pseudonymized data is (i.e. de-identified data), and it does so on an abstract level: In order to assess whether data is de-identified, it needs to be considered whether there is a reasonable risk that a person with access to the data could re-identify the data, considering all relevant circumstances.
B. The sole application of the "Safe Harbor" method (= 'rule-based approach'), which is provided by the HIPAA, does not per se result in anonymized or pseudonymized data as is understood under Swiss law. However, the development and use of a list of identifiers to be removed/modified from a data set can be helpful to provide guidance as to which data in particular, but not exclusively, must be removed or modified for de-identification. A reasonably flexible list of identifiers may therefore serve as a starting point to de-identify data.
C. In addition to such a rule-based approach, however, a risk assessment is needed in order to ensure that the de-identification meeting Swiss law. Such risk assessment will have to take into account the specific context of the individual case, because whether or not a given data set can be considered as de-identified depends on a case-by-case assessment (= 'risk-based approach').

As a consequence of those statements, the present de-identification proposal follows a combined approach relying on both risk-based and rule-based methodologies to ensure that the residual risk that a person with access to the data could re-identify the data is acceptable, considering all relevant circumstances.

[24] SHK HFG-RUDIN, Vor Art. 32–35 n. 9 ff.

3. De-Identification phased approach

3.1. Overview

In agreement with conclusions of Homburger AG memorandum and considering (international) publications on the requirements of de-identification, the Swiss Data De-identification Project Task Force developed recommendations for a phased de-identification approach.

The aim of the de-identification workflow composed of three phases is to combine both risk-based and rule-based approaches as schematized in Figure 1.

The 1st phase is dedicated to assessing and mitigating patient re-identification risks. The risks are inherent to both the research project's control measures (e.g., data storage location, contracts and policies, cohort profile, IT infrastructure and security) and the data set itself (data types and specific variables). As such, this phase seeks to define and, subsequently, reduce the research project's risk profile by introducing appropriate control measures in the project's context and specifying accurate de-identification rules on dataset variables.

The 2nd phase consists of the implementation of de-identification rules defined during the 1st phase (e.g., replacement of variable value by a pseudo identifier, suppression of a variable value. It is in the responsibility of the data provider (i.e. individual hospital) to specify the implementation of these rules in detail as they depend on the provider's internal IT requirements and constraints (i.e., data privacy, information security, etc.). Nevertheless, to provide guidance, examples of de-identification rules that could be applied are listed in Appendix 7 below.

Since a research project's lifecycle frequently requires adaptations of data exchanges between the provider and the recipient (e.g., new variables required) or even of the project context (e.g., new processor involved), a 3rd phase completes the de-identification workflow. This phase is dedicated to a periodic review of the project and of any modification which may require the overall de-identification workflow to be run again (phases 1 and 2). Modifications to be considered should be those inherent to the research project, but also external ones related, for example, to technological or organizational evolutions impairing the initially assessed re-identification risks (see also 3.4).

The following sections describe the three phases visualized in Figure 1 in more details in terms of expected input and output as well as in terms of recommended methodology.

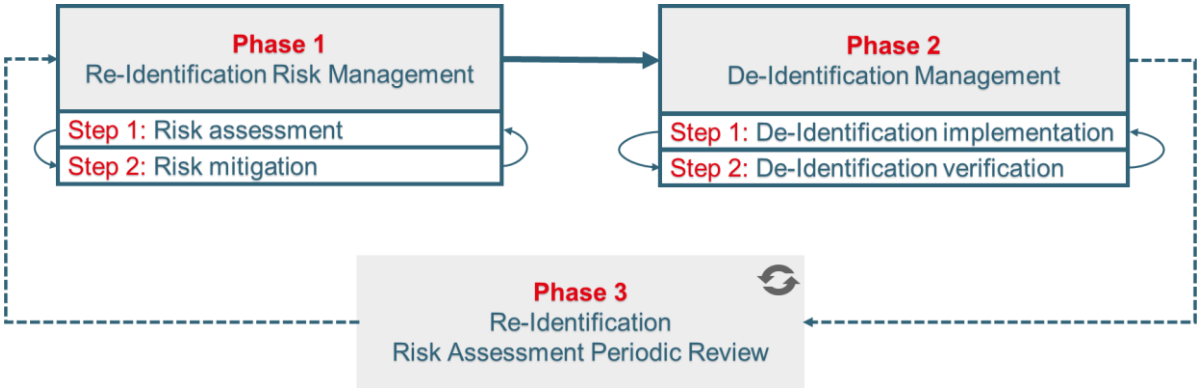

Figure 1. *De-identification of health-related data – recommended phased approach. Phase 1 comprises the re-identification risk management assessing and mitigating patients' re-identification risk. Within phase 2 risk mitigation actions specified in phase 1 are implemented and verified accordingly. Phase 3 describes the periodic review of the risk assessment performed according to project specifications.*

Responsibilities:

- It is considered that the entire de-identification workflow is under the responsibility of the project leader of the research. It is up to the project leader to perform the required activities of the workflow by calling in appropriate experts (i.e., IT, legal, scientific, etc.) of any of the concerned stakeholders (e.g., hospitals, research institutes, etc.).

3.2. PHASE 1 – Re-identification risk management

3.2.1. Introduction

The study protocol describes the project's set up as well as needs to comment on key points of the pseudonymization ('Verschlüsselung') or anonymization approach to reduce the risk of patient's re-identification. We therefore propose that the study protocol should at least include:

- the overall re-identification risk profile (low, medium, high), calculated as specified below
- controls in place to transfer and process the data
- information on high-risk identifiers (as marked in the 'Template Use case evaluation and risk assessment')

As such, the assessment purpose of the re-identification risk management is to iteratively identify the required mitigation measures (i.e.in terms of operational, organizational and technical measures) to be implemented to reduce the residual risk of data re-identification to its lowest level. The residual risk level and its associated mitigation measures should be agreed between all project stakeholders and considered by the ethical committee.

The methodology is based on an iterative 2-step approach. The first step consists of the evaluation of the data de-identification risk level of the research project. If the risk level is evaluated as 'high', the second step focuses on defining the appropriate mitigation measures. This means that several iterations of both steps may be required to reach a residual re-identification risk profile acceptable by all project stakeholders.

3.2.2. Methodology

3.2.2.1 Step 1 – Risk assessment

To manage the re-identification risk, a template for assessing this risk has been developed, termed "Template use case evaluation and risk assessment" (See Appendix 6 below). The "Template use case evaluation and risk assessment" is organized in excel format, consisting of six tabs. Tabs 3-5 collect information on the project's specific set up and de-identification strategy, thus determining its re-identification risk (that is depicted in tab 6).:

1. Template change history
2. Version history
3. Project overview
4. IT-security and contractual measures
5. Data de-identification
6. Project risk profile

7. The category "Project overview" contains information on general aspects of the project. It informs about the general data types being obtained and processed in the project. The category "IT-security and contractual measures

" aims at collecting information on the research project environment with questions grouped by topics (geographic risk, contracts and policies, data access, infrastructure and security).

The tab "Data de-identification" aims at gathering information on variables foreseen to be collected and their respective expected de-identification rules. The list of identifying and quasi-identifying variables contains

a. demographic variables,

b. multimedia variables,

c. DICOM attributes,

d. genomic variables and

e. other variables such as additional project specific quasi-identifiers.

It is recommended to agree on the types of data used in this project (structured, unstructured, multimedia, genomic data) and to complete the list of variables together with the project partners before identifying with them the most appropriate de-identification rules. An appropriate de-identification rule is a rule that reduces the re-identification risk while it ensures that the data set can still by analyzed in a useful manner in the scope of the concerned research project (i.e., no excessive distortion of data (e.g., age ranges too wide, image unusable due to high level of blurring, removal of sensitive but mandatory data, etc.).

The risk profile of the project, calculated in the category "Project risk profile", could vary from low to high based on controls in place and rules selected. The latter impact the mitigation measures that will have to be applied prior to any data transfer between the provider and the recipient.

Responsibilities:

- If the project is set up as a multi-center project with multiple (data) controllers having joint data controllership, it is in the responsibility of the project lead to consider different information security levels of the controllers' institution. The information security policy of the institution with the lowest security level should be the one taken into account as a basis, when assessing the risk. The questions in the "Template use case evaluation and risk assessment" might be answered separately by each participating institution to assess and align an appropriate security level for data processing.

Risk profile evaluation

In the "Template use case evaluation and risk assessment" document, each answer about IT-security and contractual measures and data de-identifications rules is evaluated in regard to pre-defined risk level and risk weight. The evaluation follows the subjective approach, however, it relates to expert experiences and published reports following risk based approaches[25,26].

[25] Rosner, Gilad and Rosner, Gilad, De-Identification as Public Policy (October 1, 2019). Journal of Data Protection & Privacy 3(3): 1-18 , Available at SSRN: https://ssrn.com/abstract=3639304

[26] Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc. 2010 Mar-Apr;17(2):169-77. doi: 10.1136/jamia.2009.000026. PMID: 20190059; PMCID: PMC3000773.

The risk level scales from 0 (lowest) to 3 (highest) which determine how far the answer provided or the rule selected has a risk impact in the scope of the question raised. In addition, each answer or variable de-identification rule has been associated with a risk weight which gives the relative importance of each of them within the global evaluation of the research project. The risk weight scales from 1 (lowest) to 10 (highest).

To evaluate the risk per question or variable, based on the answer(s) provided or de-identification rule foreseen, the following formula is used [risk level] * [risk weight]. The result of the calculation gives a risk value.

The sum of risk values of the answers provided and the de-identification rules selected for the different variables allows the evaluation of the overall risk profile of the project, which is calculated as follows (see Table 1 below):

1. The sum of the risk values and of the high-risk answers or high-risk de-identification rules (i.e., answers or rules with a risk level equal to 3) is calculated.
2. A relative weight (percentage) is linked to both IT-security and contractual measures and data de-identification rules, also called categories.
    a. The risk score of each of those categories is determined by comparing its sum of the risk values with the risk thresholds that have been defined in order to determine per category a risk score that could be 1, 2 or 3 as shown in Table 2.
3. Finally:
    a. The total number of high-risk answers or rules is calculated by summing the high-risk answers and rules on, respectively, IT-security and contractual measures and data de-identification rules.
    b. The evaluation of the total risk score of the project is calculated by the following formula: (Sum ([category risk score] * [category weight]) / 2)
    c. The overall project risk profile is obtained by comparing the “total risk score” with the risk score thresholds defined in the Table 2. The resulting project risk profile will be either “Low”, “Medium” or “High”.”.

Both Table 1 and Table 2 below depict an exemplary extract of the “Template use case evaluation and risk assessment” document available as Appendix A.

Table 1. *Overall risk profile evaluation table (example for a specific research project)*

Project risk profile					
Controls (geographic risk, contracts and policies, cohort size and profile, data access, infrastructure and security)			**Risk value subtotal**	**Category weight**	**Risk score**
Number of high risk answers	6		94	50%	**1**
BioMedIT usage	No				
Data (demographic and administrative, multimedia, genomic variables and DICOM attributes)					
Number of high risk rules:	5		112	50%	**2**
Risk assessment outcome					
Number of high risk	11		Total Risk Score:		**0,75**

Table 2. *Risk value and risk score thresholds*

Categorization of risk score thresholds		
Low (Risk score = 1)	**Medium** (Risk score = 2)	**High** (Risk score = 3)
< 129	129 to 258	> 258
< 105	105 to 210	> 210
Project risk score thresholds		
< 0,51	**0,51 to 1,00**	**> 1,00**

In case the risk is identified as medium or high, the calculated risk profile needs to be further evaluated to potentially reduce it (Step 2 – Risk mitigation).

3.2.2.2 Step 2 – Risk mitigation

Selected answers or rules should be reviewed if the project risk profile calculated at the end of the step 1 (i.e., risk assessment) is:

1) either high or medium
2) and, at least, one high-risk answer or rule has been selected.

This analysis of high-risk answers and rules should be dedicated to reducing the re-identification risk profile by mean of mitigation actions conceivable and acceptable in the scope of the research project (e.g., improving

legal agreement and/or IT security policy, or choosing de-identification rules having a lower risk level (and risk weight)).

The evaluation of acceptable mitigation actions is easily performable via the usage of the “Template use case evaluation and risk assessment” file. The "Template use case evaluation and risk assessment” document has been conceived to allow the project risk profile (from low to high) to be automatically recalculated based on answers provided and/or rules selected. As such it enables the project leader to evaluate the impact on the overall re-identification risk based on mitigation actions foreseen in agreement with all parties involved.

At that stage, a close collaboration between the project leader (researcher), supported by the different subject matter experts, and the data provider (e.g., data engineer in an IT department) is essential to determine the appropriate mitigation actions. Selected actions aim to reduce the re-identification risk while preserving health-related data quality for research.

3.2.3. Outcomes

At the end of phase 1, the overall project risk profile is calculated based on the risk assessment and evaluation of risk mitigation actions. Comments and project specific conditions shall be documented in the “Template use case evaluation and risk assessment” document and, where necessary, in the study protocol.

The “Template use case evaluation and risk assessment” document enables the project leader to integrate the overall outcome of this risk assessment (low, medium, high) in the study protocol, providing the ethics committee with a summary on the project-specific re-identification risk. Furthermore, if there is the need to keep the original value for certain high risk identifiers (e.g. full date of birth), the project leader is requested to disclose these high risk identifiers in the study protocol. The “Template use case evaluation and risk assessment” document is highlighting this request in the column “Condition needs explicit description for ethics approval”. The associated risk for re-identification but also the respective risk mitigation actions complete the justification for using identifying data.

Under such conditions, the project leader should ensure that any revision of the use case evaluation and risk assessment document is documented properly and that any deviation is correctly documented in the comments column associated in the tabs “IT-security and contractual measures” and ”Data de-identification”.

3.3. PHASE 2 – De-identification management

3.3.1. Introduction

The aim of this 2nd phase of the de-identification workflow is to:

- Implement the de-identification rules which have been defined and agreed during the different steps of the re-identification risk assessment (phase 1).
- Verify that the de-identified dataset produced by the data provider (data engineer):
    - respects the rules which have been defined for each of its variables and
    - is still useful in the scope of the concerned research project (i.e., no excessive distortion of data)

3.3.2. Methodology

3.3.2.1 Step 1 – De-identification implementation

The effective implementation of the de-identification rules specified in the "Template use case evaluation and risk assessment" file during 2nd phase is not described in this document. This implementation remains entirely under the control and the sole responsibility of the health-related data provider (i.e., mainly hospitals). As such, it allows those rules to be applied on the requested dataset based on methodologies in use or specifically developed by the provider institution.

The outcome of this phase 2 is a de-identified dataset compliant with the de-identification rules previously defined in the "Template use case evaluation and risk assessment" and. It provides a dataset which is finally agreed between all the parties at the end of phase 1, the de-identification management, where potential mitigation measures have been defined.

Responsibilities:

- It is the sole responsibility of the data provider to verify that the de-identification objectives set forth previously are met and that the appropriate level of anonymization[27] or pseudonymization of health-related data has been reached successfully. If this is not the case, it is up to the provider to either (1) adjust its de-identification techniques to make sure that predefined de-identification rules are applied correctly or (2) inform the project leader that the defined rules are not sufficient to acquire the expected dataset de-identification level.

3.3.2.2 Step 2 – De-identification verification

The correct de-identification of a given data set should be verified by the data provider (i.e., data engineer of IT department). Therefore, data providers need to have quality assurance and quality control measures in place, to ensure and check for the correct de-identification of the data in a given project. These measures may comprise manual checks (of complete data sets or defined samples) or the use of an independent algorithm that cross-checks the results of the used de-identification algorithm. Furthermore, study teams (researchers) shall notify the data provider if they encounter identified information in their data sets.

3.3.3. Outcomes

The outcome of this 2nd phase is a de-identified data set that was verified for correct de-identification. Furthermore, updating the tab "Version history" of the "Template use case evaluation and risk assessment", the project leader confirms that:

1. the project outline (tab "IT-security and contractual measures") has not changed compared to what has been stated during the 1st phase
2. the foreseen de-identification rules have been successfully applied on the de-identified dataset produced by the data provider.

Responsibilities:

- It is in the responsibility of the project leader to ensure during the whole life cycle of the research project that the specified IT-security and contractual measures are indeed in place or that the specification is updated accordingly (requiring a new risk assessment). It is in the responsibility of the data

[27] Anonymization only takes place when there is a direct need for it (i.e., open dataset), since it is more difficult to create and especially maintain an anonymized dataset. Especially if the project expects continues or evolving deliveries.

provider to implement and verify the correct de-identification of the data set as specified in the "Template use case evaluation and risk assessment" and according to the internal processes of the institution (hospital).

3.4. PHASE 3 – Re-identification risk assessment periodic review

3.4.1. Introduction

As a research project is by essence subject to changes during its lifecycle, it is crucial to re-evaluate, on an ad-hoc basis or at a specific frequency, the de-identification risk profile in the light of relevant changes. As mentioned previously, external conditions may also affect the risk profile of the project (e.g., technological evolutions). Most importantly, changes within the project itself (e.g., additional data; see below) should be taken into account during the periodic re-identification risk assessment review. If changes are found to increase the risk to an inacceptable level, it would be required to perform once more the whole workflow.

3.4.2. Methodology

For the re-evaluation of the de-identification use case, mainly two types of changes can be anticipated:

Type 1: Significant modifications of the original dataset or of the project context (incl. due to external factors):

- Examples (non-exhaustive list):
  - the dataset should be shared with third parties,
  - new variable(s) should be added to the original dataset,
  - a new third-party data processor should be involved in the project and its information security level is lower than those initially evaluated,
  - new rules should be selected on some variables as the original dataset was finally not usable (e.g., too drastic de-identification rules applied),
  - or a combination of multiple of those types of changes.

Under such conditions, the whole de-identification workflow should be run again (phases 1 and 2, with its re-identification risk assessment and de-identification management). Please note that if the re-evaluation yields an in-acceptable risk in a project where data has already been shared, effective data protection actions need to be defined on a case-by-case basis. It has also to be kept in mind that based. It has also to be kept in mind that based on this re-evaluation, changes to the study protocol may need to be re-submitted to the ethics committee[28].

Type 2: Additional new records should be added to the original dataset (i.e., with no impact on the list of variables):

- In this case, only phase 2, meaning the implementation of the de-identification rules previously selected, should be applied on those new records.
- Nevertheless, as increases in the cohort size may augment the risk of re-identification (see also risk scores in question C-05 of the "Template use case evaluation and risk assessment"), the risk evaluation should be re-performed to check whether the increased cohort size indeed translates into changes of the project's risk profile.

[28] Please also refer to the "Substantial amendment: YES or NO or "it depends" published by swissethics: https://swissethics.ch/assets/Meldungen/substantial_amendment_yes_no_e.pdf

Type 3: Developments of new technologies that increase the re-identification risk are not foreseen to be under the responsibility of the project but should be considered in updates of data de-identification recommendations.

3.4.3. Outcomes

The outcomes of phase 3 depend on the type of change(s) which has/have been handled. It includes an update of the "Template use case evaluation and risk assessment" and possibly an adapted specification of the de-identification strategy.

Responsibilities:

- It is the responsibility of the project leader to keep the "Template use case evaluation and risk assessment" up to date during the entire research project lifecycle and to consider project changes that impact the re-identification risk profile. If changes lead to an overall "high" risk profile or if certain new high-risk identifiers are to be included in the data set, it is an amendment of the initial study protocol and a submission of the substantial changes to the ethics committee might be recommended.

4. Glossary

This glossary lists terms or acronyms used in this document.

Acronym/Term	Description
Anonymization	According to HRO Art. 25: [1] For the anonymization of biological material and health-related personal data, all items which, when combined, would enable the data subject to be identified without disproportionate effort, must be irreversibly masked or deleted. [2] In particular, the name, address, date of birth and unique identification numbers must be masked or deleted.
Anonymized data	Data for which the de-identification is, in principle, irreversible, because no key or code exists to re-link the data to an identified or identifiable person[29]
BHI	Berlin Institute of Health
BioMedIT (node)	An Information Technology infrastructure provider, consisting of a high-performance compute and storage infrastructure, highly skilled data scientists and support personnel. There are three nodes available in Switzerland depending on researchers' affiliation.
CDW	Clinical Data Warehouse
CHUV	Centre Hospitalier Universitaire Vaudois
Coding	According to the HRO Art. 26: [1] Biological material and health-related personal data are considered to be correctly coded in accordance with Article 32 paragraph 2 and Article 33 paragraph 2 HRA if, from the perspective of a person who lacks access to the key, they are to be characterized as anonymized. [2] The key must be stored separately from the material or data collection and in accordance with the principles of Article 5 paragraph 1, by a person to be designated in the application who is not involved in the research project.
Data Controller	According to GDPR, Art. 4, Point 7: The natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data.
Data Engineer	Employee of data providers' institution supporting the project leader in the process of providing and curating data
Data Processor	According to GDPR, Art.4, Point 8: A natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller.
Data Provider	The entity or person providing data or any other service. Multiple data providers could be involved in a single research project.

[29] Handkommentar DSG-ROSENTHAL, Art. 3 n. 35.

Acronym/Term	Description
Data Recipient	According to GDPR, Art. 4, Point 9: A natural or legal person, public authority, agency or another body, to which the personal data are disclosed, whether a third party or not.
Data subject	Identified or identifiable natural person (according to GDPR, Art 4)
De-identified data	Data for which identifying attributes to an identified or identifiable person has been removed so that the person with access to de-identified data (but not to the original identifying data) is, in principle, not able to identify the person concerned. Removing the link can be achieved by supressing, replacing or modifying information. De-identified data encompasses both anonymized and pseudonymized data where the link to an identified or identifiable person has been removed so that the person with access to de-identified data (but not to the source data) is, in principle, not able to identify the person concerned. De-identified data may be anonymized or pseudonymized data
DPA	Data Protection Act
EHR	Electronic Health Record
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
HRA	Human Research Act
HRO	Human Research Ordinance
Identifiers	Information directly associated with a data subject that reliably identifies this data subject
Personal data	Data that relates to an identified or identifiable person; thus, the person with access to such personal data will be able to directly or indirectly identify the person concerned
PHI Group	Personalized Health Informatics Group of the SIB Swiss Institute of Bioinformatics
Project leader	According to HRO, Art. 3 and 4, POINT 1: The person responsible for the conduct of the research project in Switzerland and for protection of the participants at the research site. Note that in multi-center projects, there might the differentiation between the local project leader and the project leader (overall lead). The person also responsible for organizing the research project, and in particular for the initiation, management and financing of the project in Switzerland, provided that no other person or institution headquartered or represented in Switzerland takes responsibility for this (sponsor). The project leader responsible for a research project must: a. be entitled to practice independently the profession specifically qualifying him or her to conduct the research project in question; b. has the training and experience required to conduct the research project in question; c. be conversant with the legal requirements for research projects or be able to ensure compliance by calling in appropriate expertise.

Acronym/Term	Description
Pseudonym	Key or code substituted to data in order to re-link the data to an identified or identifiable person
Pseudonymization	According to GDPR, Art 4, Point 7: It 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 organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person
Pseudonymized data	Data for which the de-identification is, in principle, reversible because there is a key or code to re-link the data to an identified or identifiable person[30]. In the Human Research Act (HRA), the term "coded data" is used for pseudonymized data. Given that the German term for "coded data" (i.e., "verschlüsselte Daten") used in the HRA is misleading as it may be misinterpreted to refer to "encrypted data", the term "pseudonymized data" will be uniformly used instead of "coded data" in this document. Such term is also more frequently used in literature regarding the Data Protection Act (DPA)[31]
Quasi-identifiers	Pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with a data subject that they can be combined with other quasi-identifiers to create a unique identifier to that data subject
Re-identification	Any process by which pseudonymized data is matched with the identity of the person from which data was originally sourced.
SPHN	Swiss Personalized Health Network
USZ	Universitätsspital Zürich

[30] Article 26 HRO
[31] SHK HFG-RUDIN, Vor Art. 32–35 n. 9 ff.

5. Acknowledgments

This document was elaborated in the realm of SPHN. We thank our colleagues from SIB Swiss Institute of Bioinformatics and PHI Group members who provided insight and expertise that greatly assisted in the review process.

We thank all representatives of the Swiss University Hospitals (Markus Obreiter, Alexander Leichtle, Pierre Dethare, Solange Zoergiebel, Constantin Sluka) for assistance in terms of implementation and regulatory aspects and for comments that greatly improved the document, although they may not agree with all of the interpretations or conclusions of this paper.

We are also immensely grateful to Fabian Prasser and Bradley Malin for their expertise in the methodology part and insight on the manuscript.
Special thanks go to the core Task Force members for their continuous efforts to create and finalize this document and its accompanying template.

6. Appendix A – Template use case evaluation and risk assessment

Whenever required, additional instructions to correctly filling in the Excel file referenced below have been gathered directly in the workbook including comments on the relevant cells.

The file referenced in this document as the: “Template use case evaluation and risk assessment”” is available here:

https://sphn.ch/document/template-use-case-evaluation-and-risk-assessment/

Naming of the working Excel file should be formatted as follow: “[project acronym] - Use case evaluation and risk assessment vx.x] where version should be the version of the evaluation referenced in the “Version history” sheet.

7. Appendix B – Data de-identification rules

To de-identify directly identifying and quasi-identifying variables the “Template use case evaluation and risk assessment” document describes different de-identification rules for structured, semi-structured and unstructured data. Depending on the variables used in the project, the risk of re-identification could be mitigated by the selected de-identification rule. Variables are categorized in demographic and administrative, multimedia and genomic variables and DICOM attributes. Selected answers are associated with a risk level and a risk weight leading to a risk value per variable or attribute.

The de-identification rules are derived from the Safe Harbor Method[32] and aligned with Swiss legal requirements.

The project lead is responsible to select the rules that should be applied to produce a dataset that has as little as possible data loss or distortion and that provides the lowest possible risk of patient re-identification. It is indicated in the “Template use case evaluation and risk assessment” (see appendix 6 above) which type of de-identification rule implies a recommendation for a description in the application for ethics approval.

The general approach to de-identify structured data is to remove all non-required variables from the dataset. However, for semi- and unstructured data, the approach is to replace all variable values by a pseudonym or a surrogate in order to keep, for example, reports readable and in a consistent state. Pseudonymization of variables in semi-/unstructured data (as reports) also prevents variable values to stand out from the rest of the data as real values if they are not correctly de-identified.

Finally, the de-identification process should always generate the same pseudonym for a given variable value for a patient in a specific project. If the rule defines a date shifted by a random value (e.g., if a record is shifted by minus 5 days), then all the dates for this patient occurring in any dataset of this project are shifted by the same value, no matter if it is a laboratory, medication administration or diagnosis value.

7.1. Rules for demographic and administrative variables

7.1.1. Direct identifier

These are most of the identifiers (IDs) that identify patients and unique IDs like patient ID, sample ID, case ID, laboratory report ID, report ID, (non-exhaustive list).

- IDs will be suppressed, if the ID is not necessary. However, this is not always possible, such as when the variable is required to link different datasets together (e.g., link patient data from multiple data sets via the patient or case ID).
- By default, all IDs will be replaced with a generated pseudonym. This applies to structured and unstructured data.
- IDs will be kept if necessary. Note that this condition needs to be justified when applying for ethics approval.

7.1.2. Surname

This category includes all surnames of patients, relatives as well as medical professionals (i.e., doctors, nurses, etc.). The rule listed below apply to surnames mentioned in unstructured reports.

[32] https://aspe.hhs.gov/reports/health-insurance-portability-accountability-act-1996

- Surname can be substituted with a random surname with the same initial based on a substitution pre-defined list

7.1.3. First name

This category includes all first names of patients, relatives as well as medical professionals (i.e., doctors, nurses, etc.). The rule listed below apply to surnames mentioned in unstructured reports. Middle names are treated as first names.

- First name can be substituted with a random first name with the same initial based on a substitution pre-defined list.

7.1.4. Signature

These are the signatures from medical personnel that may correspond to their login name.

- Signature will be substituted with a random generated 6-character long sequence.

7.1.5. Date of birth

The date of birth is considered one of the key variables allowing re-identification of a patient. Instead of the date of birth, it is recommended to use the age of the patient at the event.

- Date of birth is suppressed or shifted by a random number of days (default)
- Only the year of the original birth date is kept
- Only the year and month of the original birth date are kept
- Full original date of birth is kept (dd/mm/yyyy). Note that this condition needs to be justified when applying for ethics approval.

7.1.6. Date of death

- Date of death is suppressed or shifted by a random number of days (default)
- Only the year of the original death date is kept
- Only the year and month of the original date are kept
- Full original date of death is kept (dd/mm/yyyy).

7.1.7. Age

The age of the patient will be calculated based on the admission date. Note that If one of those dates is shifted, the date of birth and date of death should be shifted in the same way.

- Age is suppressed (default)
- Age in generalized in groups of 5 or more years
- Original age is kept except for people with more than 89y old who are put in the age class "90y+"
- Original age is kept. Note that this condition needs to be justified when applying for ethics approval.

7.1.8. Other dates

- Dates are suppressed or replaced with a surrogate date (default)
- Dates are shifted by a random number of days within +/- 365 days or generalized to the year (i.e., provide year only, suppress day/month)
- Dates are shifted by a random number of days within +/- 90 days (one quarter offset to preserve seasonality) or generalized to quarter and year
- Dates are shifted by a random number of days within +/- 30 days (one quarter offset to preserve seasonality) or generalized to quarter and year

- Dates are shifted by a random number of days within +/- 7 days (default; one week offset)
- Original dates are kept. Note that this condition needs to be justified when applying for ethics approval.

7.1.9. Locations (street, zip code, city, region, country)

- Locations are suppressed
- Locations are replaced by pseudonym
- Only countries are kept
- Only regions are kept
- Only cities are kept. If cities have less than 20.000 inhabitants, cities are replaced by region
- Only the zip codes are kept. If zip codes refer to regions with less than 20.000 inhabitants, the last 2 numbers of the zip codes are suppressed
- The original locations are kept. Note that this condition needs to be justified when applying for ethics approval.

7.1.10. Organizations

- Organization name will be suppressed or replaced by surrogate organizations (default)
- Organization type will be kept (e.g., hospital, clinic, etc.)

7.1.11. Organizational Units

- Organizational unit will be suppressed (default)
- Organizational unit will be generalized (e.g., Neurology, Radiology, Urology, etc.)

7.1.12. Professions

This could be an individual profession reference (e.g., Neurologist) or a profession category (e.g., Physician).

- Profession will be suppressed (default)
- Original profession is kept, but replaced by a random profession for identifying ones

7.2. Rules for multimedia variables

There are three categories for the rules of de-identifying multimedia variables. Each category allows a selection of the different de-identification rules:

7.2.1. Audio Data

- Patient voice is kept in audio files
- Patient voice blurring/noise algorithm

7.2.2. Photographic Images & Videos

- Patient face (including frontal view) is kept in image or video files
- Patient face from the side
- Partial patient face
- Blurring of patient face
- Patient identifying body part (e.g., tattoo is kept in image or video files)
- Blurring of identifying patient body parts or additions (e.g. implants etc.)

7.2.3. Image from medical device

- Engraved patient information is kept in images and videos

- Scrubbing engraved patient information
- Unmodified head images
- Defacing head images (preventing reconstruction of face/ear/teeth)
- Not removing implants
- Removing implants
- No efforts are made to remove images with identifying characteristics from the cohort
- Algorithm is used to remove images with identifying characteristics (please explain the algorithm and what kind of characteristics it will remove.in the comment field)
- Manually remove patients with identifying characters

7.3. Rules for DICOM attributes (= meta data information provided in the DICOM tags)

The rules for de-identifying DICOM attributes follow six categories:

1. Hardware Identifying Attributes
2. Study Description
3. Series Description
4. Derivation Description
5. Contrast Bolus Agent
6. Retain original values of other DICOM attributes that would be removed by default according to the recommendations of nema.org

Each category allows a selection of the three de-identification rules:

- Original value is suppressed
- Original value is replaced by pseudonym
- Original values are kept. Note that this condition needs to be justified when applying for ethics approval.

DICOM attributes (DICOM attributes listed in the confidentiality list (http://dicom.nema.org/medical/dicom/current/output/chtml/part15/chapter_E.html) will be removed unless they are listed under DCM-06.

7.4. Rules for genomic data

The usage of genomic data and the applicable de-identification rule must be mentioned in the ethics application. Additionally, the usage of the BioMedIT[33] infrastructure or an infrastructure with the same high level of IT security is strongly recommended.

SPHN takes into account that BAM/SAM files from tumour samples can contain somatic as well as germline information, while VCF-files report the somatic variants only.

Depending on the aggregation level, sharing of (germline) genomic sequences highly influences the result of the risk-assessment. De-identified BAM/SAM-files cannot be considered with a low risk for re-identification.

[33] For more information on BioMedIT, please refer to the SPHN Glossary:
https://sphn.ch/document/sphn-glossary/
and a more global description of the BioMedIT project:
https://sphn.ch/network/projects/biomedit/

Comprehensively de-identified VCF-files (i.e. all identifiers have been removed) can be considered with a low risk for re-identification and acknowledged as “anonymized data” in the context of the de-identification process following Swiss law requirements. Such anonymized files should be made available for third use purposes.

Pseudonymization of sample ID and date-shifts follow the de-identification rules provided in the «SPHN use case evaluation and risk assessment template», provided under «Direct Identifier» and «Date».

- Re-use of existing files (produced in the healthcare setting):
    - All identifying information in the file and file tags (e.g. name, birthdate, etc.) have to be removed or replaced.
    - The original sample ID is replaced by project-specific sample ID. Note that this is only applicable for digital files and not necessary in case physical samples are shared.
    - The date-stamp is shifted according to the project specifications. Shifts apply to the general rules of de-identification: (i) date-stamp is suppressed or replaced with a surrogate data (very low risk); (ii) date-stamp is shifted by a random-number of days within +/-365 days (low risk); (iii) date-stamp is shifted by a random number +/- 7days (high risk); (iv) originals are kept (very high risk).
    - If with the de-identification process other (for the researcher valuable, but not associated with new re-identification risk) information is removed, this will be extracted and transferred to the recipient in text format.
- For prospective VCF-files (newly produced in the realm of a research project):
    - The project agrees on a harmonized/common pipeline for data generation, not providing identifying information and facilitating de-identification steps.
    - The original sample ID is replaced by project-specific sample ID before the analysis of the sample.
    - The date-stamp is shifted according to the project specifications. Shifts apply to the general rules of de-identification: (i) date-stamp is suppressed or replaced with a surrogate data (very low risk); (ii) date-stamp is shifted by a random-number of days within +/-365 days (low risk); (iii) date-stamp is shifted by a random number +/- 7days (high risk); (iv) originals are kept (very high risk).
    - If with the de-identification process other (for the researcher valuable, but not associated with new re-identification risk) information is removed, this will be extracted and transferred to the recipient in text format.

For germline genomic sequences the following rules might be considered:

- Only summary statistics (e.g., MAF, p-values, ORs) on x > 1000 individuals are released
- Only summary statistics (e.g., MAF, p-values, ORs) on 1000 >x>100 individuals are released
- Only summary statistics (e.g., MAF, p-values, ORs) on 100 >x>10 individuals are released
- Only summary statistics (e.g., MAF, p-values, ORs) on x>10 individuals are released
- Original individual-level values are released

7.5. Rules for other variables

There might be additional project specific quasi-identifiers that can be used for linkage by the data recipient (e.g., clinical variables) and which need to be de-identified accordingly.

- There are no other quasi-identifiers or they are suppressed
- Quasi-Identifiers are replaced by pseudonym
- Quasi-identifiers have been modified to reduce risks (e.g., generalization)

- Original values are kept