Data Protection Concept for the LEOSS Public Use File

Authors: Fabian Prasser and Florian Kohlmayer

Version history

Concept version	Anonymisation pipeline version	Comment
03.04.2020	0.2	Initial version
11.04.2020	0.3	Minor fixes and polishing
21.04.2020	0.4	Adjust threshold

Introduction

The *Lean European Open Survey on SARS-CoV-2 Infected Patients* (LEOSS) is a European non-interventional prospective cohort study (see https://leoss.net and the study protocol [4]). The core concept behind LEOSS is the collection of anonymous data. For this purpose, no identifying data is stored in the registry, which requires registration, authentication and authorisation before data entry. The Public Use File covers a subset of the data collected. This document provides a detailed description of the additional measures implemented before releasing the LEOSS Public Use File. These measures include terms of use that must be accepted prior to data access as well as anonymisation measures to ensure that data about individual patients cannot be re-identified. The Public Use File comprises the variables listed in Table 1.

Variable	Description	Domain
Age at diagnosis	Age of patient at time of diagnosis	<= 25 26 - 45 46 - 65 66 - 85 > 85
Gender	Sex of patient	Male Female
Month first diagnosis	Month of first confirmed diagnosis of COVID-19	1 – 12
Year first diagnosis	Year of first confirmed diagnosis of COVID-19	4 digit year
Uncomplicated phase	Indicates whether the patient has been through the uncomplicated phase of COVID-19	Yes No
Complicated phase	Indicates whether the patient has been through the complicated phase of COVID-19	Yes No
Critical phase	Indicates whether the patient has been through the critical phase of COVID-19	Yes No
Recovery phase	Indicates whether the patient has been through the recovery phase of COVID-19	Yes No
Vasopressors in complicated phase	Indicates whether vasopressors where used in the complicated phase	Yes No Missing/unknown N/a

Vasopressors in critical phase	Indicates whether vasopressors where used in the critical phase	Yes No Missing/unknown N/a
Invasive ventilation in critical phase	Indicates whether invasive ventilation was used in the critical phase	Yes No Missing/unknown N/a
Superinfection in uncomplicated phase	Type of (if any) superinfection in uncomplicated phase	Bacterial Bacterial & fungal None Missing/unknown N/a
Superinfection in complicated phase	Type of (if any) superinfection in complicated phase	Bacterial Bacterial & fungal None Missing/unknown N/a
Superinfection in critical phase	Type of (if any) superinfection in critical phase	Bacterial Bacterial & fungal None Missing/unknown N/a
Symptoms in recovery phase	Symptoms (if any) in recovery phase	Yes No Missing/unknown N/a
Last known patient status	Last known status	Recovered Not recovered (means recovery phase not achieved) Dead from covid-19 Dead from other causes Unknown/missing

Table 1: Overview of the variables of the LEOSS Public Use File.

Qualitative risk assessment

From a qualitative perspective it can be noted that the dataset contains no directly identifying information and contains only a very small subset of variables that are typically assumed to be associated with a high risk of re-identification (“age at diagnosis”, “gender”, “month first diagnosis”, “year first diagnosis”). Notably, the dataset only features one variable, “month first diagnosis”, that would need to be removed according to the de-identification standard laid out in the Safe Harbor method of the Privacy Rule of the US HIPAA law [1] (note that “age at diagnosis” is top-coded at 85) or that is mentioned as a high-risk variable by the European Medicines Agency's Policy 007 Implementation Guideline for anonymous sharing of clinical trials data [2]. There are multiple studies indicating that the risk of re-identification of HIPAA protected data is very small, see e.g. [3]. From a qualitative perspective, we therefore conclude that the privacy risk of publishing the LEOSS Public Use File is very low, even in its original form.

In addition, the study protocol of the LEOSS registry specifies, that no data values that correspond to less than 10 individuals will be included in the Public Use File [4]. While this provides little formal guarantees, it does provide an additional layer of protection for individuals with rare characteristics regarding individual variables.

Quantitative analysis and anonymisation process

We performed additional anonymisation procedures to ensure that the dataset is protected according to the current state-of-the-art also from a formal and quantitative perspective.

For this purpose, we follow the requirements described by the Article 29 Data Protection Working Party, which was an advisory body composed of a representative of the data protection authority of each EU Member State, the European Data Protection Supervisor and the European Commission which became the European Data Protection Board with the introduction of the EU General Data Protection Regulation (GDPR) [5]. With its "Opinion on Anonymisation Methods" [6] the board formulated requirements and guidelines for effective anonymisation measures and presented an assessment of common methods. According to the opinion, the following privacy threats should be addressed by anonymisation methods [6]:

- Singling out: "the possibility to isolate some or all records which identify an individual in the dataset" [6]
- Linkability: "the ability to link, at least, two records concerning the same data subject or a group of data subjects" [6]
- Inference: "the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes" [6]

To assess which variables must be transformed to protect records from singling out and linkability, we implemented the approach proposed by Malin et al. and analysed the stability, availability and distinguishability (quantified by 1=low, 2=medium, 3=high) of the variables [7]. The results of this analysis is then used to estimate how well suited these variables are for performing successful linkage attacks (if sum of weights is > 6; we call those "key" variables). The results are shown in Table 2.

Variable	Stability	Availability	Distinguishability	Is Key
Age at diagnosis	3	3	3	Yes (9)
Gender	3	3	2	Yes (8)
Month first diagnosis	3	3	1	Yes (7)
Year first diagnosis	3	3	1	Yes (7)
Uncomplicated phase	2	2	1	No (5)
Complicated phase	2	2	2	No (6)
Critical phase	2	2	2	No (6)
Recovery phase	2	2	1	No (5)
Vasopressors in complicated phase	2	1	2	No (5)
Vasopressors in critical phase	2	1	2	No (5)
Invasive ventilation in critical phase	2	1	2	No (5)
Superinfection in uncomplicated phase	2	1	2	No (5)
Superinfection in complicated phase	2	1	2	No (5)
Superinfection in critical phase	2	1	2	No (5)
Symptoms in recovery phase	2	1	2	No (5)
Last known patient status	1	1	2	No (4)

Table 2: Assessment of the re-identification risk associated with individual variables.

To prevent singling out and linkability using the variables "age at diagnosis", "gender", "month first diagnosis", "year first diagnosis" or any arbitrary combination, we implement the k-anonymity protection model as suggested by the opinion [6]. This model ensures that each record is indistinguishable from at least k-1 other records regarding the key variables, i.e. variables that could be used for dataset linkage [8]. The Working Party recommends a value of k > 10, which is consistent with recommendations from other guidelines, including the European Medicines Agency's Policy 007 Implementation Guideline [2], which recommends a risk threshold of 0.09 (corresponding to k=11).

The LEOSS Public Use File will be released in 11-anonymous form regarding the key variables listed in Table 2.

While the opinion states that the risk of inference is also partially addressed by 11-anonymity, it still recommends additional protection. Table 3 presents the results of an analysis used to determine which variables could be used in inference attacks.

Variable	Risk of inference	Reason
Age at diagnosis	No	Basic demographics. More likely to be already known. Not sensitive.
Gender	No	Basic demographics. More likely to be already known. Not sensitive.
Month first diagnosis	No	Basic demographics. More likely to be already known. Not sensitive.
Year first diagnosis	No	Basic demographics. More likely to be already known. Not sensitive.
Vasopressors in complicated phase	Yes	Sensitive medical information
Vasopressors in critical phase	Yes	Sensitive medical information
Invasive ventilation in critical phase	Yes	Sensitive medical information
Superinfection in uncomplicated phase	Yes	Sensitive medical information
Superinfection in complicated phase	Yes	Sensitive medical information
Superinfection in critical phase	Yes	Sensitive medical information
Symptoms in recovery phase	Yes	Sensitive medical information
Last known patient status	Yes	Sensitive medical information
Uncomplicated phase	No	Perfect correlation with variables describing complications, interventions and symptoms (see Text).
Complicated phase	No	Perfect correlation with variables describing complications, interventions and symptoms (see Text).
Critical phase	No	Perfect correlation with variables describing complications, interventions and symptoms (see Text).
Recovery phase	No	Perfect correlation with variables describing complications, interventions and symptoms (see Text).

Table 3: Assessment of variables that could be used in inference attacks (Note: as a result of perfect correlation, it is not necessary to protect correlated variables from inference if the variables on which they depend have been appropriately protected.).

Some variables, in particular those describing whether patients went through a particular phase, are perfectly correlated with the variables describing complications, interventions and symptoms (i.e. their value can be derived from the fact whether information on complications, interventions or symptoms has been provided for the according phase). Hence, there is no need to protect those variables, as long as the more detailed medical variables are protected accordingly.

For the eight variables that need to be protected from inference, we implemented the well-known t-closeness model [9] with t=0.5. This approach has been recommended by the opinion [2] and the parameterisation takes into account the high level of privacy protection already achieved. By combining protection against singling out and linkage with additional protection against inference of sensitive information, the resulting dataset is strongly protected from the threats addressed by relevant guidelines and laws.

Protected continuous publishing

The LEOSS Public Use File will be updated continuously when new data is entered into the registry. To ensure that all data remains adequately protected, we implement a static data transformation scheme and withhold individual records as long as they do not meet the requirements described in this document. Moreover, the process described in this document will be re-assessed regularly and updated if necessary.

Additional safeguards

In addition to the qualitative and quantitative anonymisation procedures laid out above, users need to accept terms of use prior to downloading the LOESS Public Use File. These terms clearly state that the data must only used for research on COVID-19, that re-identification must not be attempted, that the data must be stored securely and re-redistribution is not permitted. This is very similar to the approach taken the European Medicines Agency on its Clinical Data Portal [10].

Technical implementation

The anonymisation process described has been implemented using the open source ARX Data Anonymisation Tool [11]. The code of the complete pipeline is publicly available online [12].

References

[1] Methods for De-Identification of PHI | hhs.gov. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. Accessed: 05.04.2020

[2] External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use. EMA/90915/2016 Version 1.4. 2018

[3] Sweeney, L. (2010, March). Data sharing under HIPAA: 12 years later. In Workshop on the HIPAA Privacy Rule’s de-identification standard.

[4] LEOSS Lean European Open Survey on SARS-CoV-2 Study Protocol Version 1.1 March 16th, 2020. https://leoss.net/wp-content/uploads/2020/03/LEOSS-Protocol-Submission-1-20200316.pdf. Accessed: 11.04.2020

[5] Article 29 working party archives 1997 – 2016, https://ec.europa.eu/justice/article-29/documentation/index_en.htm. Accessed: 05.04.2020

[6] ARTICLE 29 DATA PROTECTION WORKING PARTY. 0829/14/EN WP216. Opinion 05/2014 on Anonymisation Techniques. https://www.pdpjournals.com/docs/88197.pdf. Accessed: 05.04.2020

[7] Malin, B., Loukides, G., Benitez, K., & Clayton, E. W. (2011). Identifiability in biobanks: models, measures, and mitigation strategies. Human Genetics, 130(3), 383–392.

[8] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

[9] Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. 23rd International Conference on Data Engineering, 106–115.

[10] Terms of use. clinicaldata.ema.europa.eu. https://clinicaldata.ema.europa.eu/web/cdp/termsofuse. Accessed: 05.04.2020

[11] Fabian Prasser, Johanna Eicher, Helmut Spengler, Raffael Bild, Klaus A. Kuhn. Flexible Data Anonymisation Using ARX — Current Status and Challenges Ahead. J Software Pract Exper 2020;1–28 (2020).

[12] Gitlab repository of the anonymisation pipeline. https://gitlab.com/infektiologie-ukkoeln/leoss-anonymisation.git. Accessed: 05.04.2020