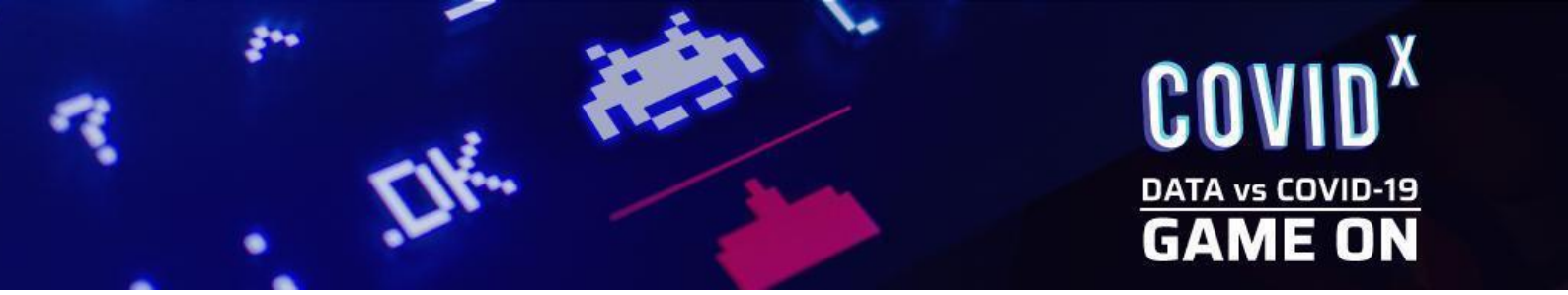

COVID X

COVID EXPONENTIAL PROGRAMME

GRANT AGREEMENT ID: 101016065

Anonymization Guide

Revision: v.0.4

Work Package	WP1
Submission date	11/01/2021
Partner	SERMAS, 8BELLS
Version	0.4
Authors	José Manuel Laperal (SERMAS), Luis Rodriguez (SERMAS), Despoina Gkatzioura (8BELLS)

DISCLAIMER

COPYRIGHT NOTICE

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

Table of Contents

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

1 Anonymization Policy

This guide is a tool to help the members of the COVID-X consortium, and it is also **the Anonymization Policy** that is also mandatory for other participating entities through the Open Calls of the project.

This document must be approved by each person responsible for the information and updated whenever necessary by the person responsible for anonymization. The objective of this document is to guarantee that each task aimed at the definitive anonymization or disassociation of personal data has a specific person in charge associated with it.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

2 Introduction

To preserve the confidentiality and privacy of the patients whose data will be part of the data sets that will be the object of the COVID-X validation pilots, it is necessary to carry out a process of anonymization of personal data, in order to eliminate the possibility of persons' identification.

This is to be achieved by locating the identifiable information included in the datasets that will be provided in the COVID-X Sandbox and applying on them the appropriate anonymization techniques, that will minimise the risk of re-identification.

Advances in technology and available information make it difficult to guarantee absolute anonymity, especially over time, but in COVID-X techniques that offer greater guarantees of privacy to people have been used in such a way that the re-identification effort of the subjects entails a sufficiently high cost so that it cannot be approached in terms of the effort-benefit ratio (Data Protection authorities consider that an anonymization process is good when the efforts to carry out identification are too high compared to the benefits obtained). That is, re-identification would imply that the benefit to be obtained may become negligible in relation to the effort used, or that said effort is not assumable by the person or entity with access to the anonymized information.

Following the principle of full functionality, from the beginning of the design of the information system, **the final usefulness of the anonymized data will be considered as a priority,** ensuring as far as possible the absence of distortion in relation to the non-anonymized data.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

3 Work Team Configuration

When roles and responsibilities are not clearly defined, access (and further processing) of personal data may be uncontrolled, resulting in unauthorized use of resources and compromising the overall security of the system. Therefore, it is crucial to have clearly defined roles and responsibilities in order to reduce risk of data breaches[1].

In the development of the anonymization process, the following segregation of functions has been carried out, according to profiles or roles (the profiles or roles of the Spanish pilot site are shown, as an example):

- **Responsible for the treatment:** Manager of the Hospital Clinico San Carlos. Its function is to decide on the purpose and objectives of the information processing.

- Data Protection Officers or Data Protection Delegate: DPO of the Biomedical Research Foundation of the Hospital Clinico San Carlos. Its function is to promote the performance of prior impact evaluations on privacy in the anonymization processes, verify the execution of the anonymization processes, ensure the independence of roles and functions, report to the person responsible for the information and the treatment on the processes of anonymization, promote audits of compliance with anonymization processes or respond to requests for information from citizens in relation to their anonymized or non-personal data, among other functions.

- Recipients or responsible for the processing of anonymized personal information: Members of the COVID-X Consortium and participating entities. They decide on the information requirements based on the final objectives for which it is intended.

- **Risk assessment team** formed by:

1. **Responsible for Security and Data Protection of the COVID-X Consortium.** In charge of carrying out the initial risk assessment, evaluating the results of the anonymization process, auditing the anonymization procedure, auditing the use of anonymized information and ultimately responsible for ensuring that the anonymized file meets the requirements related to the residual risk of re-identification.
2. **Technical team of the centre** (Information and Communications Technology Service of the Hospital Clinico San Carlos, in the example of anonymization protocol). Responsible for carrying out data extractions and guarding the anonymization keys.

- [1] "Guidelines for SMEs on the security of personal data processing", ENISA, https://www.enisa.europa.eu/publications/guidelines-for-smes-on-the-security-of-personal-data-processing.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

3. **Pre-anonymization team and anonymization team** (Innovation Unit of the Biomedical Research Foundation of the Hospital Clinico San Carlos). In charge of determining which variables will be anonymized and proposing anonymization techniques that will be selected or validated by the risk assessment team. It will also be responsible for eliminating those variables whose anonymization is not feasible or does not fit the purpose of the anonymized data, facilitating the final work of the anonymization team, and guaranteeing the value of the anonymized data. The anonymization team oversees choosing the necessary anonymization techniques and their application.
4. **Information security team:** Technical Partners of the COVID-X Consortium (8Bells, INTRA, UPM). This team oversees ensuring the necessary security measures during the life cycle of the anonymized information and during the anonymization processes, in addition to assessing the results of the Privacy Impact Assessment (PIA) and implementing measures aimed at mitigating the risks to personal information. anonymized. They oversee the security measures of the environments as well as carrying out validation tests aimed at assessing the strength of the following procedures that guarantee the irreversibility of anonymization or carry out re-identification tests. These validation tests will be performed on a sample of the anonymized data set.

The partners' security experts of the COVID-X consortium will act as an advisory body to provide technical feasibility to the clinical sites that enable the use of anonymized information and the anonymization process. Together, they will conform to the acceptable risk threshold resulting from the PIA and, if not, they must issue the corresponding reasoned opinion.

Each of the subjects and teams that fulfil these roles, act within the scope of their own competence and with total independence from the rest. Therefore, we will avoid the possibility of an error occurring at a certain level being supervised and approved at a different level by the same person.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

4 Necessary Training for the Anonymization Team

When employees are not aware of the need of applying security measures, they can accidentally pose further threats to the system[2].

One of the keys to guaranteeing the privacy of the interested parties is the training and information that is provided to the personnel involved in the anonymization process and in the exploitation of the anonymized information. During the information life cycle, all personnel with access to anonymized or non-anonymized data will be professionally trained and informed about their data protection obligations, as well as on the application of specific security measures and procedures.

The personnel involved in the anonymization process must comply with all the training and information requirements related to compliance with the regulations on the protection of personal data, especially about the security measures referred to in article 32 of the General Data Protection Regulation (GDPR).

Once the personal data has been anonymized, the personnel with access to the anonymized information will also be informed of:

- The existence and application of the anonymization policy.
- Data protection principles in the design of anonymization processes.
- Objectives set in risk management (PIA).
- Structure and responsibilities of the work team involved in the anonymization processes.
- Objectives and purpose of the anonymized information.
- Anonymization variables: identification and classification.
- Anonymization techniques used.
- Terms of use and access to anonymized information
- Specific roles and responsibilities.
- Personnel control measures with access to anonymized information (traceability).
- Obligations and duties in the event of a breach in the anonymization chain11 that makes it possible to re-identify the interested parties.

The training will be provided in such a way that it can be auditable, that is, there will be a record of the training provided and of the staff that has received the training.

[2] "Guidelines for SMEs on the security of personal data processing", ENISA, https://www.enisa.europa.eu/publications/guidelines-for-smes-on-the-security-of-personal-data-processing.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

Some of the possible risks for the re-identification of subjects with anonymized data may originate from the inadequate implementation of anonymization procedures, which could also be influenced by inadequate training or information of the personnel involved in the anonymization or in the treatment of anonymized data.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

5 Privacy Impact Assessment (PIA)

In addition to the legal requirements set out by the GDPR and other regulations on data protection, it is necessary to carry out a privacy impact assessment before the anonymization process so that, in addition to detecting privacy risks, we can avoid the use of resources that may be excessive or not necessary for the intended purpose and ensure the use of resources that are necessary to guarantee non-re-identification. This, in addition to implying unnecessary costs, can lead to re-identification risks that could be avoided.

The first step of the privacy impact assessment, or PIA, is to carry out a risk analysis of the anonymization process to subsequently manage the resulting risks with technical, contractual, organizational or any other measures.

It is necessary to remember that no anonymization technique will be able to guarantee in absolute terms the impossibility of re-identification, since there will always be an index of probability of re-identification that we must try to mitigate through the corresponding risk management.

The risk of re-identification is implicit and increases as time passes, as a consequence of the evolution and increase of indirect identifiers over time, such as, for example, the information that the interested party has contributed about itself in social networks, blogs, etc.

Aware of the risks of re-identifying the anonymized data, the data controller will promote the periodic reassessment of the existing residual risk to introduce parameters to improve the quality of the anonymization process.

5.1 Stages of the Impact Assessment

In addition to guaranteeing the privacy of the interested parties, the data controller will determine the objectives to be met by the anonymized information based on the legitimate interests of its recipient. The design of the anonymization process will be conditioned by the final objective of the anonymized information, giving rise to information of restricted use or open data.

When attempting to anonymize data belonging to the special categories referred to in article 9 of the GDPR, the existence of a team to study the feasibility of the anonymization process could be considered. The work of this team will be of special relevance and its main task would be to produce a feasibility report that will reflect in detail the reasons and specific conditions for the anonymization of specially protected data. Such a report could include, among others, for example, the ethical foundations or links of the anonymization process.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

The **Hospital Clínico San Carlos Anonymization Protocol** is shown below as an example to be used in all pilot sites.

5.1.1 Identification and categorization of assets involved in the anonymization process

5.1.1.1 Identification of personal data to be anonymized

1. As a mandatory rule, only the autonomic personal identification code (CIPA) and the patient's medical record number (NHC) are extracted. If CIPA does not exist, another identifier will be used, such as the medical record number or the provisional number that is assigned to patients who do not have CIPA.

2. The rest of personal data that allow us to identify the patients (direct or indirect identifiers), such as address, telephone, etc., are not extracted.

5.1.1.2 Anonymized information assets and associated identification variables

Those responsible for the different information systems of the Information and Communications Technology Service of the Hospital Clinico San Carlos extract the information periodically from the primary sources and after filtering it (masking/transforming direct identifiers), they store it in the intermediate transfer repository of the technical department.

The identification variables, as indicated, would be the CIPA and the NHC.

5.1.1.3 Anonymization processes and threats

These are described in the Section 8 Anonymization Protocol.

5.1.1.4 Information systems

Hardware used, limitation of the anonymization software in relation to the information assets to be anonymized:

- Hospital Information System (HP-HIS): Includes sociodemographic information on the patient, the record of their various interactions with the hospital in the form of episodes, all their discharge reports and their corresponding diagnoses coded as ICD-9 and ICD-10. They also include all the management data of the Hospital Clinico San Carlos (HCSC).
- Laboratory Information System (EOL-HIS): Includes all the results of all clinical laboratory tests performed in the HCSC Clinical Analysis Service, the Clinical Pharmacology Service and the Nuclear Medicine Laboratory.
- Microbiology Information System (GLIMMS): Includes all the results of all laboratory tests performed in the HCSC Microbiology service.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

- Haematology Information System (MODULAB): Includes all the results of all clinical laboratory tests performed in the Haematology service of the HCSC.
- Pathology Information System (PATWIN): Includes all the results of all laboratory tests performed in the Pathology Service of the HCSC.
- Radiology Information System (IMPAX): Includes all the results of all the imaging tests performed in the Radiodiagnosis and Nuclear Medicine Services of the HCSC.
- Endoscopy Information System (EndoTools): Includes all the results of all endoscopy tests performed in the Digestive System and Pulmonology Services of the HCSC.
- Cardiology Information System (XCELERA): Includes all the results of all hemodynamic and echocardiography tests performed in the Cardiology Service of the HCSC.
- Emergency Information System (SISU): Includes all the data resulting from the care of patients in the Emergency Service, including clinical history, nursing follow-up, discharge reports, etc ... generated in the HCSC Emergency Service.
- Nursing Information System (GACELA): Includes all results derived from nursing care in all HCSC inpatient units.
- Pharmacy Information System (FARMATOOLS): Includes all the data related to the management of drugs in HCSC impatiens.
- Information Aggregation System (PATIENT): Includes information from various specialized forms.
- Biobank Information System (BioEBank): Includes information regarding biological samples stored in the HCSC Biobank.
- Information System for Intensive Care (ICCA H.02)
- Reports associated with RX images (AGFA).
- Other departmental sources or individual collections may be added.

All these assets are included in a pseudo-anonymized repository called “BDClin-HCSC-IdISSC”.

5.1.1.5 Analysis of dependencies of assets involved in the anonymization process.

Does not apply.

5.1.1.6 Categorization of assets

The objective is to establish a categorization based on the criticality of each asset, considering aspects such as, for example, the degree of sensitivity of the information.

As direct personal data has been filtered, the remaining information would have the same category, so no categorization is carried out.

COVID-X has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101016065.

5.1.2 Constitution of the Work Team

In the constitution of the work team, the degree of specialization in risk analysis and data protection has been considered, as stated in the Chapter 3 "Work Team Configuration".

5.1.3 Risk Identification

The first aspect to consider in the privacy risk analysis is the initial risk classification according to three categories:

Known existing re-identification risks

- Risks of re-identification due to correlation with other data sets (inference disclosure).
- Risks of breaching the duty of secrecy due to improper access to information that has not been anonymized (e.g., new data about an individual is exposed: attribute disclosure).
- Risks of disclosure of information anonymization keys (e.g., an easy to infer algorithm or insufficient anonymization is used: identity disclosure).

Potential re-identification risks

- Risk associated with the ability to discover the keys used to anonymize the data set.

Unknown risks

- Risk of the existence of a potential attacker or adversary or what can be considered as the role of the "persecutor", when trying to identify an acquaintance, or that of a “marketer”, when trying to identify the whole dataset (deliberate attempt of re-identification).
- Risks of existence of a subject who knows the identity of a person in an information block and who seeks to obtain more information (inadvertent or intentional attempt at re-identification)

Each risk identified in the Table 1 - Risk CATEGORIES has been assigned a certain value on a quantitative or qualitative scale based on the probability of occurrence. The set of all risks will give rise to the following risk catalogue of a block of information that is intended to be anonymized.

TABLE 1 - RISK CATEGORIES

Risk Categories

COVID-X has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101016065.

General category	Sub-category
Known existing re-identification risks	Risks of re-identification due to correlation with other data sets
	Risks of breaching the duty of secrecy due to improper access to information without anonymizing
	Risks of disclosure of information anonymization keys
Potential re-identification risks	Risk associated with the ability to discover the keys used to anonymize
Unknown risk	Risk of the existence of a potential attacker or adversary or what can be considered as the role of the "persecutor"
	Risks of existence of a subject who knows the identity of a person in an information block and who seeks to obtain more information

On the other hand, 3 types of assets have been identified on which threats could occur (Figure 1):

- Pseudo Anonymised data from the HCSC repository
- Anonymized Data Lake
- Data sets extracted from the Data Lake

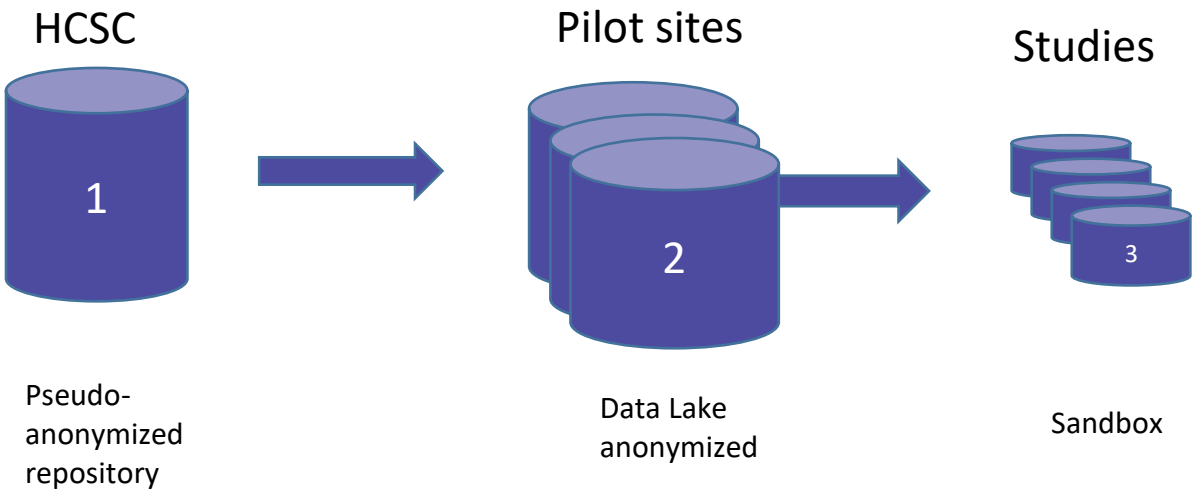

FIGURE 1: SCHEMATIC OF THE DATA PROCESSING STATIONS IN HOSPITAL CLINICO SAN CARLOS

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

5.1.4 Assessment of Existing Risks and Quantification of the Impact

According to the set of assets and the existing risk catalogue, a categorization has been made of each of the re-identification risks that have been detected (Table 2).

This categorization is considered in the phases of the anonymization process and especially when eliminating variables.

TABLE 2 - MATRIX FOR REIDENTIFICATION RISK ANALYSIS

Matrix for reídentification riks analysis (PIA)				Impact		
				Information elements		
				Data pseudo	Data lake	Sandbox
				3	2	1
Threat probability	Known existing re-identification risks	Risks of re-identification due to correlation with other data sets	2	**6**	**4**	**2**
		Risks of violation of the duty of secrecy due to improper access to information without anonymizing	2	**6**	**4**	**2**
		Risks of disclosure of information anonymization keys	2	**6**	**4**	**2**
	Potential re-identification risks	Risk associated with the ability to find keys	1	**3**	**2**	**1**
	Unknown risk	Risk of the existence of a potential attacker or adversary or what can be considered as the role of the "persecutor"	2	**6**	**4**	**2**
		Risks of existence of a subject who knows the identity of a person in an information block and who seeks to obtain more information	2	**6**	**4**	**2**

Levels of risk considered (1 to 10):

- High level (9-10)
- Medium level (6-8)
- Low level (1-5)

The assessment method of the MAGERIT risk analysis methodology has been used, in which the probability of occurrence of a risk is multiplied by the impact or damage of manifesting itself on an

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

asset. The same assessment method is also proposed by the European Network and Information Security Agency (ENISA).

It is observed that there are no risks categorized with a level 9 (high) but that the maximum risk level would be level 6 (or medium level) and only on one asset, the Pseudo Anonymised data from the HCSC repository (or pseudonymized warehouse).

5.1.5 Safeguards

For each of the risks identified with level 6, one or more safeguards are proposed to prevent a specific risk from materializing (Table 3).

TABLE 3 - RISKS & SAFEGUARDS

RISK	SAFEGUARDS
Risks of re-identification due to correlation with other data sets	The data packages that each researcher receives have a specific and unique anonymization (they could not be used to compare between them)
Risks of breaching the duty of secrecy due to improper access to information without anonymizing	Training, awareness and confidentiality documents. Segregation of functions so that keys are not known between levels
Risks of disclosure of information anonymization keys	Training, awareness and confidentiality documents. Segregation of functions so that keys are not known between levels
Risk of the existence of a potential attacker or adversary or what can be considered as the role of the "persecutor"	Rigorous anonymization measures. Security measures established in the project. Security measures of the pilot site
Risks of existence of a subject who knows the identity of a person in an information block and who seeks to obtain more information	All of the above.

Once the risks and safeguards have been identified, it is time to assess or quantify the impact of the possible materialization of a risk. It should be taken into account that the impact can be tangible (for example, material damage, possible compensation, etc.) or intangible (loss of trust, deterioration of

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

the image of the person responsible for the treatment, stigmatization of the interested parties, etc.), but in both cases we must assign a quantitative or qualitative value to the possible impact.

Finally, a catalogue of risks and a catalogue of safeguards are obtained, categorized according to the criticality of the assets that need to be protected.

5.1.6 Risk Report

The report of the resulting risks will have a summarized format and will clearly show the existing risks and their level of criticality, considering the scale that would have been used for their identification. The work team that performs the risk assessment will present to the data controller and the security team a proposal with the acceptable risk threshold for each anonymization process, for each of them to proceed to issue their opinion.

5.1.7 Determination of the Acceptable Risk Threshold

As a result of the PIA there will be a risk threshold or residual risk index of re-identification (Table 4). This risk index will be assumed by the data controller as an acceptable risk and will be taken into consideration for the design of the anonymization process. Finally, the residual risk threshold for re-identification will be known by the recipient of the anonymized information and, when the anonymized data is for public use, it will also be made public, informing the people or entities that they use information of this said risk.

To decide which risk threshold to use, we must examine the sensitivity of the data and the consent mechanism that was in place when the data was originally collected: this is *the invasion of privacy* dimension that needs to be evaluated. For example, if data is extremely sensitive, a lower threshold should be selected, to apply a more stringent anonymization process and minimize the risk as much as possible. On the other hand, if subjects gave their explicit consent to releasing the data publicly, while understanding the risks, a higher threshold can be set, but always within the acceptable range determined by the data controller.

The risk analysis must be performed periodically throughout the information life cycle and whenever there are changes in the anonymization processes or in the treatment of anonymized information. The purpose of the periodic reviews is to verify that the real state of risks coincides with the assumable risk of re-identification, verifying the effectiveness of the measures provided to mitigate the possible impact that the re-identification of individuals could have.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

For its part, the person responsible for the processing of the anonymized data must consider the initial risk catalogue prepared by the person responsible for the treatment to continue developing his/her own analysis throughout the life cycle of the anonymized information.

The person responsible for the treatment at the proposal of the risk assessment team, the anonymization team and the information security team will ultimately be the ones who decide on the acceptable risks resulting from the anonymization process.

All personnel involved in the anonymization processes will be aware of the acceptable risk threshold and the recipient of the anonymized information will be able to have access to the risk catalogue prepared by the person responsible for the treatment to adopt the appropriate measures to mitigate the possible consequences that the identification of individuals from the anonymized data would cause.

TABLE 4 - MATRIX FOR RE IDENTIFICATION RISK ANALYSIS AFTER APPLYING SAFEWARDS

Assessment of re-identification risks in the asset: Data pseudo			Risk report			
			Initial risk	Saveguard	Residual risk	Comments
Threat probability	Known existing re-identification risks	Risks of re-identification due to correlation with other data sets	6	The data packages that each researcher receives have a specific and unique anonymization (they could not be used to compare between them)	1	**There is no possibility to cross information**
		Risks of violation of the duty of secrecy due to improper access to information without anonymizing	6	- Training, awareness and confidentiality documents - Segregation of functions so that keys are not known between levels - Information about legal consequences associated with non-compliance or negligence	2	**The organization frequently reminds employees of this**
		Risks of disclosure of information anonymization keys	6	- Training, awareness and confidentiality documents - Segregation of functions so that keys are not known between levels - Information about legal consequences associated with non-compliance or negligence	2	**The organization frequently reminds employees of this**
	Unknown risk	Risk of the existence of a potential attacker or adversary or what can be considered as the role of the "persecutor"	6	Rigor in the anonymization measures, security measures established in the project and those of the pilot site	3	**The level of anonymization of the information is very good**
		Risks of existence of a subject who knows the identity of a person in an information block and who seeks to obtain more information	6	All of above	2	**Reidentification drills are carried out to detect this type of situation, whose possibilities are very low**

5.1.8 Management of Assumable Risks

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

In the design of the anonymization process, it will be necessary to foresee the consequences of an eventual re-identification of individuals that could cause damage or reduction of their rights.

Likewise, it will be necessary to foresee a hypothetical loss of information due to the negligence of the personnel involved, the lack of an adequate anonymization policy or due to an intentional disclosure of secrecy that would lead to the loss of the identification variables or identification keys of the subjects.

For each of the risks that have been determined as assumable, measures will be established to mitigate the possible impact on the privacy of the individuals that were re-identified.

5.1.9 Final Report

The possible measures that are established will be known by all the subjects involved in the anonymization processes and in the treatment of the anonymized data. All the personnel involved will have knowledge and training about the actions that they should take to mitigate the impact resulting from the materialization of any of the risks of re-identification of patients. To minimize their impact, the final report will reflect the existing risks by category and the measures or recommendations to be implemented in the anonymization processes and in the exploitation of the anonymized information.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

6 Organizational measures

The organizational security measures that have been taken in the COVID-X consortium are described below:

- An Information Security Policy has been drawn up, the content of which has been published and disseminated to all partners.
- The obligation to follow the indications of the Information Security Policy and the rest of the guidelines established by the COVID-X consortium has been established as a condition of compliance for the third parties participating in the Open Calls.
- Confidentiality agreements have been signed by all partners and participants in the Open Calls.
- Following the example at SERMAs, conditions and procedures have been established to carry out pilot studies with data from the HCSC contained in the BDClin-HCSC-IdISSC catalogue of variables, which guarantees an adequate use of the data. Researchers must send a formal request for the use of BDClin-HCSC-IdISSC for their R+D+I projects. The application must include:
  - The research project to which the data will be applied,
  - The specific data that they request to extract from BDClin-HCSC-IdISSC, and
  - The favorable opinion of the local Ethics Committee (CEIm), necessary to carry out any R+D+I project at the HCSC.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

7 Possible Techniques for Anonymization

The selection of technical anonymization measures should also be subjected to risk analysis for each anonymization process, helping to establish guidelines or principles that allow and help the subject(s) responsible for the anonymization process to choose or decide on the most appropriate techniques at each specific moment.

A combination of different techniques must be used in each pilot site, ensuring that at least two of them are used. The different techniques used in COVID-X are shown below.

7.1 Layers of Anonymization

This technique consists of adding a second level to data that has already undergone an anonymization process. The person in charge of the treatment has anonymized all the data that could be used to re-identify the subjects and sends the information to the applicant who, in order to prevent the re-identification from taking place, decides to carry out a second anonymization of the already anonymized data.

In this way, the recipient of the anonymized information ensures that their processes use their own anonymization resources, avoiding that in case of fragility of the anonymization processes of the person responsible for the treatment, the identity of the subjects could be affected.

This method will establish a difficulty proportional to the criticality or sensitivity of the anonymized information, multiplying the effort necessary for the re-identification of subjects.

In some cases, to guarantee people's privacy, it may be necessary to use geographic range distortions as in the case of people with extremely rare pathologies. In these cases, the recipient of the information will be informed of the reason why variances of geographic range or any other type of variances that were used in the anonymization process (time, length, etc.) have been used.

7.2 Data Interruption

Data disturbance is the systematic variation and suppression of data that prevents the resulting data from providing information about specific cases:

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

- Micro aggregation: A technique used to anonymize numerical data consisting of substituting specific numerical values with the mean value calculated for a certain group of data by grouping, segregating, deleting, or substituting independent records.
- Generalization: Replacing a value with a less specific but semantically consistent value.
- Random data exchange: Introducing a random distortion in a set of microdata while maintaining the detail and structure of the original information.
- Synthetic data:
  - Data distortion: Random synthetic data is generated that maintains the results of the original data set.
  - Distortion with hybrid microdata: Combining original data with synthetic data.
- Permutation of records: Exchanging of data values with key value that guarantees mean values and statistical distributions.
- Temporal permutation: Random movement of temporal ranges that does not generate distortion on the final average results.
- Rounding: substitution of variables for randomly rounded values.
- Readjustment of weights: When working with known data samples, this involves distorting the values of the original samples to avoid re-identification.
- Random noise: Injects noise while maintaining the original data structure.

7.3 Data Reduction

Using this technique, the number of original data is reduced without altering them, the level of detail of the original data is reduced, avoiding the presence of unique or atypical data without relevance to the final result:

- Elimination/Suppression of variables: Elimination of especially sensitive data that can be direct identifiers (See example in the Chapter 8 “Anonymization Protocol”.).
- Reduction of records: When after applying other measures, the subjects continue to be identifiable.
- Global re-coding: Certain categories of data are grouped into new categories, reducing the chances of re-identification.
- Upper or lower coding: For cases in which higher or lower values of a range are identifiable, it consists of expanding or reducing the higher or lower range.
- Deletion of records: deletion of data records that contain data that allows the identification of subjects. This measure will be used when it is impossible to anonymize a certain subject and will expressly indicate the deleted records and the reason why they are excluded from the final anonymization result.

COVID-X has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101016065.

Finally, in the anonymization phase, the final and irreversible disassociation of personal data is carried out. The anonymization process must be carried out as many times as necessary according to the purpose of the anonymized information and its recipient.

Each of the recipients of the studies will have anonymized data sets with different keys, anonymized with a specific objective, purpose and recipient.

In no case will a general use anonymization process be employed regardless of the recipient of the information, the type of information to be anonymized and the purpose for which the anonymized data will be used.

Below are, as a general example, some of the tasks or activities that can be performed during the anonymization phase:

- Determining the anonymization technique that is most appropriate based on the variables that have been identified in the pre-anonymization phase.
- Planning and assigning specific tasks to each member of the work team in relation to the functions assigned to each profile involved in the anonymization process.
- Determining the resources and technical equipment necessary to proceed with the anonymization of the data.
- Validating the anonymization technique by experts (unit or expert body in statistics, ethics, etc.)
- Applying the selected technique and run the anonymization process; perform tests.
- Breaking of the relational keys based on the use of the information (internal use and external use). Whenever possible, different codes will be used depending on the use to be made of the anonymized information.
- Recoding or reducing variables for residual sensitive data after the anonymization process.
- Applying data reduction techniques (suppression of fields that are not significant for later use).
- Limiting the level of disaggregation according to the geographic level affected by the file and the sensitivity of the information.
- Applying data disturbance techniques (modify quantitative data in small random quantities, exchange attributes in a controlled way between records from nearby geographic areas, respecting distributions).
- Validating and approving of the anonymized files by experts and by the evaluation team.
- Periodic reviewing of the process.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

- Auditing the anonymization process and the subsequent use of the data using metrics or scales that provide an objective interpretation of the results.
- Documenting the process that was used to anonymize the data set, as well as the results of enacting that process. The results documentation would normally include a summary of the data set that was used to perform the risk assessment, the risk thresholds that were used and their justifications, assumptions that were made, and evidence that the re-identification risk after the data has been anonymized is below the specified thresholds.

7.4 k-Anonymization

Another possible technique is k-anonymization, the objective of which is to prevent an individual from being singled out when he/she is grouped with (at least) a number "k" of subjects. To achieve this goal, attributes are generalized to the point that several subjects end up sharing an identical value. For example, instead of giving the exact figure of the salary received, or the date of birth, an interval is given (between 10,000 and 20,000 euros per year, or born between 1980 and 1990).

The privacy requirements will determine the value of$k$. A high value is related to more demanding privacy requirements, since there will need to be more subjects within the group who satisfy the same combination of identifying traits. However, a value that is too high for k can cause a loss of fidelity in the source data, thus losing its usefulness, so it must be evaluated whether this distortion is relevant to the study to be carried out and, if so, seek a reasonable balance between the result sought and the rights of the participants. On the contrary, a value that is too small will cause the weight of each of the stakeholders to be greater, which would facilitate the success of an attack by inference. Ultimately, it will have to be studied in each case, but ideally, k should be an intermediate value.

We can see this process in the following example, using a synthetic data set (Table 5):

TABLE 5 - EXAMPLE OF K-ANONYMIZATION. ORIGINAL EXAMPLE DATASET

Year of birth	Sex	Postal Code	Cause of death
1959	W	37052	Heart attack
1957	W	34806	Cancer
1959	M	43691	Cancer
1966	W	28666	Heart attack
1963	M	28574	Heart attack

In this case, we are assuming that an attacker is looking for information regarding a person who he/she knows is on the original dataset, and who was born in 1966. Having this information will allow him/her to know that said person is a woman, that she died from a heart attack, and also (if he/she knows the country where the study was carried out) that this person resided in Madrid, since the Postal Code

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

corresponds to that city. This is the main weakness of this system: it is a risk that the value of k is very low.

The most widely used k-anonymization systems are two: generalization and elimination. These also have the undoubted advantage that they do not disturb the data, since they achieve protection by replacing the values of certain data with more general ones, without introducing erroneous information.

Generalization consists of making the data less precise, for example by forming age ranges instead of using the specific year. Thus, the number of records with identical values for a set of quasi-indicators can be increased, so that it is more difficult to successfully carry out an inference attack. The generalization can be global or local, depending on whether (starting from the same value for the same type of attribute) the generalization is always carried out in the same way, or different criteria are used for each record. The table from before, applying a global generalization to it, would look like this (Table 6).

TABLE 6 - EXAMPLE OF K-ANONYMIZATION. EXAMPLE DATASET ANONYMIZED USING THE GENERALIZATION TECHNIQUE

Year of birth	Sex	Postal Code	Cause of death
1950 - 1960	M	37***	Heart attack
1950 - 1960	M	34***	Cancer
1950 - 1960	H	43***	Cancer
1960 - 1970	M	28***	Heart attack
1960 - 1970	H	28***	Heart attack

For its part, the Elimination system is shown in the following example: Let's imagine that our table contains the previous data, which can be generalized, but (in addition) we include a new record, which although it is generalized, cannot be included in any of the previous intervals. For example (Table 7):

TABLE 7 - EXAMPLE OF K-ANONYMIZATION. EXAMPLE DATASET ANONYMIZED USING THE ELIMINATION TECHNIQUE

Year of birth	Sex	Postal Code	Cause of death
1950 - 1960	M	37***	Heart attack
1950 - 1960	M	34***	Cancer
1950 - 1960	H	43***	Cancer
1960 - 1970	M	28***	Heart attack
1960 - 1970	H	28***	Heart attack
2000 - 2010	M	13***	Cancer

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

It is so far removed from the intervals in which the rest of the data are found that this interval cannot be widened enough to include it without avoiding losing such a high level of precision that the data lose its usefulness for the study in question.

The elimination system consists of eliminating these records that are "outlayer", so that they do not distort the results or become a security risk. This method is also followed when there are very unusual values, since they also constitute a risk, in this way singularization attacks are avoided.

In many cases this method is sufficiently secure, although you must assess in each case which re-identification risks are associated with each data processing, to ensure that k-anonymization sufficiently protects the information in question.

However, to choose which anonymization technique is better, it is necessary to first carry out a correct analysis of the risks (PIA) that the process will encounter, to be able to alleviate them with a technical or organizational measure.

7.5 Other Techniques

In the literature l-diversity is also described as an anonymization technique to acknowledge the imperfections of k-anonymity -such as the lack of diversity in the sensitive attributes- to overcome homogeneity assault and background knowledge. This approach revolves around the notion that the sensitive attributes in each equivalence class[3] are “well-represented”. However, risks remain. Distribution skewness and semantic similarity of the sensitive values in the equivalence class are possible attacks faced by the l-diversity technique.

t-closeness is another privacy preserving technique proposed to address the limitations in the existing k-anonymity and l-diversity methods. To do so, t-closeness limits the semantic proximity of the sensitive attributes within an equivalence class to a threshold t, thus reducing the granularity of the interpreted data.

[3] A group of records that are indistinguishable from each other is often referred to as an equivalence class.

COVID-X has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101016065.

To conclude, k-anonymity is the most commonly used technique and the one recommended in the present document. However, it could be either replaced by l-diversity or be used along with t-closeness to further enhance the privacy of published data.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

8 Anonymization Protocol

This is an example of the procedures used in the Spanish pilot site, which can be used to further understand how an anonymization procedure takes place.

The first measure to be applied in a data set that is going to be used to validate a product or any kind of study in COVID-X, is to carry out an initial classification of the data and have a scale or gradient of sensitivity of the information.

Based on a set of data that has been collected from the interested parties following the principles established in article 5 of the GDPR, the data is adequate for a specific purpose. In this data set we have:

- Microdata or direct identifiers of subjects: all those characteristics that by themselves allow the identification of a person.
- Indirect identifiers: although they do not identify a person, the crossing of several indirect identifiers could allow the identification of a person.
- Especially protected or sensitive data: those referred to in article 9 of the GDPR, in our case health data.

For example, in the HCSC a classification scheme has been developed consisting of three levels of identification of persons (microdata, indirect identification data and sensitive data), where a quantitative value is assigned to each of the identification variables. The scale is known to all the personnel involved in the anonymization process and is a fundamental key to consider in the risk analysis or Personal Data Protection Impact Assessment (PIA) of the anonymization process.

In this case, at the beginning of the anonymization process, the identification variables that are not considered necessary are eliminated, leaving only the medical record number and the CIPA code, which are considered direct identification variables (microdata).

8.1 Phase 1 - Pre-Anonymization

The pre-anonymization of the microdata is the initial part of the anonymization process, in which the possible identification variables (direct and indirect) to be taken into account in the design of the anonymization tools will be determined.

During the pre-anonymization process, the following has been considered:

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

- The determination of variables: personal data, direct and indirect identifiers, especially protected data, and other confidential data. In the HCSC, the identification variables that are not necessary are eliminated, leaving only the medical record number and the CIPA code.
- The classification and sensitivity of the variables by categories: direct identification, geographic identification, of a specially protected nature, numerical, temporal, metadata, etc.
- Identification variables that cannot be anonymized and that must be eliminated from the anonymization process.
- Anonymized variables that are essential for the purpose for which the anonymized data will be used.
- Once the variables have been categorized, the necessary protection criteria are established to guarantee people's privacy, trying to minimize the amount of personal information that will be used during the anonymization process.
- The process of anonymization of variables cannot be approached without first defining the possible identification variables that will be necessary for the purpose for which the anonymized information will be used. In this process there are variables or microdata that are tangible identification elements, but there are other indirect identification variables that allow the identification of subjects in a less tangible way, such as:
  - Clinical record number
  - Regional patient code
  - Others

The anonymization process of the data is carried out in a structured way, considering the purpose that the data is intended to give once anonymized, guaranteeing the privacy of the subjects and avoiding the distortion of the results of the anonymized information with respect to non-anonymized data.

At this stage of the process, attention is paid to the specific anonymization difficulties for certain variables, such as, for example, if it is necessary to anonymize voice records, image records or biometric and / or genetic information.

8.2 Phase 2 - First Layer of Anonymisation

The unique identifier of the patients, CIPA or Clinical record number, is transformed through a dissociation procedure (pseudonymization), from the moment it is received in the intermediate transfer repository of the technical team of the Information and Communications Technology Service of the Hospital Clinico San Carlos (hereinafter DSTI). This patient identifier code is to be used to distinguish each patient from the others in BDCLIN-HCSC-IDISSC.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

The private key of said decoupling is only known by the DSTI team. It is stored in a file that only the person in charge of the DSTI will be able to access through their user code and password. In this way, patient information is separated from their clinical data.

The possible professional codes received in the transfer files are also dissociated, generating a table that relates the professional's code with a new dissociated identifier. In this way, the absence of identifying data of the professionals collaborating with BDCLIN-HCSC-IDISSC is also guaranteed.

The person responsible for this pseudonymisation procedure is the DSTI, which may entrust the technical process to specific professionals of the Innovation Unit (IU) of the Biomedical Research Foundation of the Hospital Clinico San Carlos, under its supervision. The pseudonymised data is stored on a dedicated server at the HCSC's Data Center.

From this point on, the IU is responsible for the rest of the BDCLIN-HCSC-IDISSC management process.

8.3 Phase 3 - Elimination / Reduction of Variables

In this phase, range aggregation is used to mask subjects when there are specific microdata that allows direct identification of specific subjects or groups. For example, in the case of extremely small groups of subjects, their information should be diluted into a group with a greater numerical range, adding, if necessary, a reference to a percentage in which the existence is made clear. of the minor collective as part of a larger set.

Researchers only access data contained in those variables and records that are necessary to carry out the project and that are reflected in the protocol approved by the CEIm (Ethical Committee).

8.4 Phase 4 - Anonymisation

Finally, in the anonymization phase, the final and irreversible disassociation of personal data is carried out. The anonymization process must be carried out as many times as necessary according to the purpose of the anonymized information and its recipient.

8.4.1 k Anonymization Technique Application

Variables that may contain indirect identification data, such as the exact date of birth (including day) or death or dates of use of health resources, are subjected to a k-anonymization process of data aggregation, so that the exact date is not accessible to researchers in the data files in which they carry out the statistical analysis of the project, provided that the quality of the project results is assured. If

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

this is not possible, other obfuscation techniques will be used to guarantee non re-identification of patients.

8.4.2 Note regarding biometric data

During the anonymization process, biometric data, voice records or image records can present a specific complexity that must be addressed in the initial phases of the anonymization process. For example, about voice records, it is possible to carry out a previous transcription with their respective elimination of possible identifiers (autochthonous expressions, epideictic elements, rhetorical identifiers, etc.) to later proceed to the reproduction of the transcripts using synthesizing voice devices, should it be necessary to keep a sound record.

Image registrations present their risk of re-identification in the image, since sometimes people can be re-identified by their environment and not directly by their own generic features. The variables of re-identification of people through images can be multiple, so that sometimes the image data will require a specific treatment to prevent the re-identification of people. For example, in the case of a specific dermatological ailment in which the person has a specific tattoo or scar that reveals their identity, the image must undergo a digital treatment that makes the re-identification of the person irreversible.

Regarding biometric data, the purpose of the anonymized information may be a limitation to the anonymization of the information, giving rise to certain exceptions in which the data cannot be anonymized to avoid any critical distortion that may occur with relation to non-anonymized information. These situations will be considered in the initial phases of anonymization and especially in the PIA as an implicit risk to the process itself given the characteristics of the information. In this case, the PIA itself may raise the need to use encryption mechanisms for access to biometric data in a restricted and controlled way, mechanisms that must be agreed by the person responsible for the treatment and the person responsible for the treatment of the anonymized or encrypted information.

8.5 Phase 5 - Delivery of the Data Set

The delivery of the data set to the main researcher is associated with an acceptance by the same of certain commitments:

1. Researchers only access data contained in those variables and patient records that are necessary to carry out the proposed and approved project.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

2. The principal investigator (PI) will be responsible for ensuring that the entire research team is aware of the commitments, accepts them, and complies with them, in accordance with the document signed for the approval of the project by the CEIm.

3. Regarding the use of data and exclusivity, the research team agrees that:

- The data is to be used solely and exclusively for that project.
- The files resulting from the validation of cases or any other algorithm or procedure developed for the project that could contribute to increasing the quality of the data extracted from BDClin-HCSC-IdISSC, remain accessible without prejudice to the fact that the researcher is the author of the algorithm.
- To implement all the necessary measures so that unauthorized or inappropriate use of the data is not carried out.

4. The research team compromises to notify the Biomedical Research Foundation of the Hospital Clinico San Carlos of the publications generated as a result of this project, including the following statement: "The data for carrying out this project are part of the BDClin-HCSC-IdISSC database managed by the Innovation Unit (IU) of the Biomedical Research Foundation of the Hospital Clinico San Carlos. The results, discussion and conclusions of this project are those considered by the authors only and do not represent in any way the position of the UI regarding this issue".

Summary of techniques used in the HCSC anonymization protocol:

- anonymization by layers (pre-anonymization layer, anonymization layer)
- technique of elimination / reduction of variables
- disturbance technique

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

9 Additional guarantees of confidentiality of anonymized information

The anonymization process cannot ensure the impossibility of re-identification of subjects in absolute terms, which is why the legal guarantees necessary to preserve the rights of the interested parties must be taken into account.

Once the exploitation of the anonymized information begins, measures will continue to be taken to guarantee the privacy of the interested parties, such as those described below.

9.1 Documentary Guarantees

Since the anonymized information is intended to become information of restricted use, the privacy of personal data will be reinforced through **confidentiality agreements** that will form part of the set of legal guarantees of the anonymization process. In the case of anonymized information of restricted use, the data controller may assess the development of possible contractual clauses, codes of conduct and certification mechanisms that include the commitment by the recipient to not make any attempt to re-identify the data and guarantee the privacy of information even when re-identification breaches occur.

In this sense, some of the aspects that must be taken into account are:

- Signing confidentiality agreements involving the following actors:
  - Responsible for the treatment.
  - Responsible for the anonymization process.
  - Responsible for the treatment of anonymized data.
  - Personnel with access to anonymized information.
- Obtaining the commitment of the recipient of the information to maintain anonymization and the obligation to inform the controller of any suspicion of re-identification.
- Auditing by the person responsible for the treatment of the use of anonymized information that is made by the person responsible for the treatment of the anonymized data.
- Including the guarantees in the contract signed between the data controller and the recipient of the anonymized information.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

These and other possible guarantees that may be necessary for the treatment of anonymized information have been taken into account in the PIA as part of the safeguards aimed at minimizing the damages in the event of a possible re-identification of the involved parties.

9.2 Data Segregation

An additional guarantee that has been taken into account in order to ensure the confidentiality of the anonymized information is to have an information systems architecture that guarantees separate environments for each processing of personal data or anonymized personal information.

The anonymization process is carried out based on personal data in an independent segregated environment. In turn, the exploitation of the anonymized information is carried out in an environment outside the environments of exploitation of personal data and the environment in which the anonymization of the information is carried out.

The previously pseudonymised data is downloaded to a dedicated server of the HCSC Data Center by the DSTI to BDClin-HCSC-IdISSC checking the consistency in the types of data received, and the structure that allows its incremental update (raw data).

Periodically a new secondary use database is generated with updated information; this process is what we refer to in the document as "going to production".

Since the data is received from different information systems with different structural models, each source of information undergoes a process of normalization, standardization and harmonization, being integrated into the common data model of BDClin-HCSC-IdISSC to be exploited with R+D+I purposes (standardized data).

The segregation of environments for the processing of information also implies the segregation of the personnel that accesses the information and personal data. An additional guarantee to avoid re-identification is that those people involved in the processing of anonymized personal information do not have access to non-anonymized personal data or cannot access knowledge of the anonymization mechanisms and keys used in the anonymization processes.

9.3 Audits

The purpose of auditing the anonymization process is to ensure compliance with the anonymization policy, providing an objective opinion on the entire anonymization process. The audit can be internal or external and will be periodic.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

The quality of the audit itself is essential for maintaining the trust of the interested parties in the anonymization processes, since the lack of confidence of the interested parties in the confidentiality of the anonymization processes could cause social concern, negatively impacting the exploitation of anonymized data.

The results of the audit can be made known to the interested parties by providing them with information on the probabilities of re-identification and the best practices accredited in the anonymization process. In order to guarantee the quality of the audit, the use of internationally recognized norms, methodologies and standards is recommended.

The audit of the anonymization process will show results related to the quality objectives of the anonymization processes that were initially foreseen by the data controller.

The person responsible for the treatment or the person responsible for the treatment of the anonymized data will ensure the existence of periodic audit reports in which are stated at least the following:

- Scope and objective of the audit.
- Definition of the audit team and resources used to carry out the audit.
- Phases and planning of the audit.
- Tests and verifications carried out.
- Assessment of the results.
- Proposals to improve the anonymization process.
- Audit of the exploitation of anonymized information.

The audit will entail the necessary checks aimed at verifying the implementation of the proposals to improve the anonymization process and will allow the verification and monitoring of the effectiveness of the measures implemented.

The applicable anonymization policy must be documented and accessible to the personnel involved in the processing of anonymized data. With this objective, the following is a possible scheme of the documentary content that can be considered for the anonymization process:

- Policy of use and access to anonymized data: obligations of the personnel.
- Document of applicability of anonymization measures that will contain at least:
  - Responsible for the pre-anonymization and anonymization process.
  - Organizational measures.
  - Definition of identification variables.
  - Technical anonymization mechanisms.
  - Key policy

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

        - Confidentiality agreements
        - Rules and procedures
- Reports and opinions:
    - From the feasibility team, if it had been defined.
    - The security team.
    - Risk analysis (PIA).
    - Information audit and anonymization process.

The documentation will be updated whenever necessary due to changes in the anonymization process, in the legal requirements or due to conditions of technological evolution.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

10 Standards and References

- "Dictamen 05/2014" del Grupo de Trabajo del Artículo 29 sobre técnicas de anonimización.
- "Dictamen 06/2014" del Grupo de Trabajo del Artículo 29 sobre la noción de interés legítimo a la que se refiere el artículo 7 de la Directiva 95/46/EC.
- "Código de buenas prácticas de las estadísticas europeas para los servicios estadísticos nacionales y comunitarios", adoptado por el Comité del Sistema Estadístico Europeo el 28 de septiembre de 2011 (EUROSTAT).
- "Looking Forward: De-identification Developments – New Tools, New Challenges" (May 2013, Information & Privacy Commissioner Ontario, Canada).
- "De-identification Protocols: Essential for Protecting Privacy", (June 2014, Information & Privacy Commissioner Ontario, Canada).
- "Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy", (June 2011, Information & Privacy Commissioner Ontario, Canada).
- "Big Data and Innovation, Setting the Record Straight: De-identification Does work" (June 2014, Information & Privacy Commissioner Ontario, Canada).
- "Pan-Canadian De-Identification Guidelines for Personal Health Information", (2007, Information & Privacy Commissioner Ontario, Canada).
- "Anonymisation: managing data protection risk", (November 2012, Information Commissioner´s Office, UK).
- "CNIL – guide sécurité des données", (2010, Commission nationale de l'informatique et des libertés).
- "Lineamientos para la anonimización de microdatos", (Agosto 2014, Dirección de Regulación, Planeación, Estandarización y Normalización –DIRPEN- Colombia).
- "Norma PNE 178301, Ciudades Inteligentes. Datos abiertos (Open Data) – Versión para Información Pública".
- "A Systematic Review of Re-Identification Attacks on Health Data", (US National Library of Medicine, National Institutes of Health).
- "Perspectives on Heal Data De-identification" (Privacy Analytics, Khaled El emam, PhD).
- Guides for the Spanish Data Protection Authority.
- "Risk-level tool", European Network and Information Security Agency (ENISA), https://www.enisa.europa.eu/risk-level-tool/
- "Guidelines for SMEs on the security of personal data processing", ENISA, https://www.enisa.europa.eu/publications/guidelines-for-smes-on-the-security-of-personal-data-processing
- "Anonymizing Health Data", Khaled El Emam & Luk Arbuckle

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

- "l-Diversity: Privacy Beyond k-Anonymity", Machanavajjhala et al., 2006, IEEE.
- "A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data", Rajedran et al., 2017, IJCSNS International Journal of Computer Science and Network Security, VOL.17 No.12.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.