



COVID eXponential Programme

GRANT AGREEMENT ID: 101016065

D1.1 - Ethical and legal framework

Revision: v.0.2

Work Package	WP1
Due date	31/01/2021
Submission date	31/01/2021
Deliverable lead	SERMAS
Version	1.0
Authors	Jose Manuel Laperal (SERMAS), Elena Arredondo (SERMAS), Despoina Gkatzoura (8BELLS), Alexandros Karatzos (8BELLS), Carlotta Cattaneo (ICH), Silvia Alba Uribe Mayoral (UPM), Gert Helgesson (KI), Niklas Juth (KI), Sven-Ove Hansson (KI),
Reviewers	Luis Rodriguez (SERMAS), German Seara (SERMAS), Sabine Koch (KI), Carl Johan Sundberg (KI), Sokratis Nifakos (KI), Carlotta Cattaneo (ICH).

Abstract	This report contains the reference guide regarding legal and ethical issues related to personal data privacy, security and regulations, which will govern the activities that will be carried out in COVID-X. It has been divided in three main sections ("the problem", "the solution", and "how we do it"). In addition, and considering that throughout the life of the project the General Data Protection Regulation requires a constant update in privacy and related issues, this document will be periodically revised and updated, adapted to the needs and circumstances that may apper during the life of the project.
Keywords	Personal data, Legal framework, Ethical framework, anonymization, pseudonymization, security procedures, privacy, compliance, normative, GDPR, patients rights



DOCUMENT REVISION HISTORY

Version	Date	Description of change	List of contributors
V0.1	05/11/2020	1st version of the template for comments	Jose Manuel Laperal (SERMAS), Elena Arredondo (SERMAS)
V0.2	19/11/2020	Review and inclusion of contribution from ICH	Jose Manuel Laperal (SERMAS), Carrlotta Cattaneo (ICH)
V0.3	28/12/2020	Completion of several sections and contribution from KI	Jose Manuel Laperal (SERMAS), Gert Helgesson (KI), Niklas Juth (KI), Sven-Ove Hansson (KI)
V0.4	15/01/2020	Completion of several sections and contribution from technological partners	Jose Manuel Laperal (SERMAS), Despoina Gkatzoura (8BELLS), Alexandros Karatzos (8BELLS), Silvia Alba Uribe Mayoral (UPM)
v1.0	29/01/2021	Final review of the document	Luis Rodriguez (SERMAS), German Seara (SERMAS), Sabine Koch (KI), Carl Johan Sundberg (KI), Sokratis Nifakos (KI), Carlotta Cattaneo (ICH)

DISCLAIMER

The information, documentation and figures available in this deliverable are written by COVID-X project's consortium under EC grant agreement 101016065 and do not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

COPYRIGHT NOTICE

© 2020 - 2022 COVID-X Consortium Reproduction is authorised provided the source is acknowledged



Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable:		R
Dissemination Level		
PU	Public, fully open, e.g., web	X
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential project and Commission Services	

* R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

OTHER: Software, technical diagram, etc.

EXECUTIVE SUMMARY

The objective of Deliverable D1.1 “Ethical and legal framework” is to be the reference and guide for action, in legal and ethical matters, of the COVID-X project. It is a document that will be permanently updated during the life cycle of COVID-X, adapting to the needs and circumstances that may appear, creating successive versions of it.

This deliverable will complement the continuous support to the COVID-X partners with any legal and/or ethical issue that may appear during the implementation of the different components, as well as those that may appear during the exploitation phases to both the project partners and the partners that join through the Open Calls.

The document is organized in three main blocks:

Block 1 (*the problem*) contains the result of a meticulous study process of all existing regulations related to the use of personal data in clinical research, to detect all those particularities that must be taken into account and highlighting the most important. In a project in which we need to achieve QUALITY levels as high as those intended by COVID-X, it is equally necessary to achieve the same levels of PRIVACY, to the point that the project must be ready to be audited by the corresponding authorities.

In block 2 (*the solution*), the needs related to privacy and data protection are analyzed. The privacy requirements are analyzed indicating how the project should be developed to include security-by-default and by-design, how to comply at all times with the principle of accountability, minimization, and ethical aspects that are to be taken into account.

In block 3 (*how we do it*) we have developed the strategies that are been followed in the different aspects that come into play (data loading, anonymization...), to respond to the previous requirements. A set of specific techniques are collected to carry out risk analysis, impact evaluations, along with different warnings and considerations that must be taken into account in the anonymization processes of personal data (strategy adopted in the COVID- X for more efficient management of project objectives). Also included in this block are other details intrinsically linked to privacy, such as information on the security of the architecture and its components, security measures taken in data storage, additional measures to guarantee the exercise of patients' rights, and organizational measures such as data treatment agreements and system audits.

As by-products of this deliverable D1.1, a specific guide has also been developed to correctly carry out the data anonymization process. This document has been prioritized so that it would be useful for the partners from the beginning, while other by-products such as the *Security Policy* or the *Action Guide for Security Breaches*, will be developed in the following month, within the necessary permanent updating of privacy (from the beginning and throughout the life cycle) as indicated in the General Data Protection Regulation (GDPR).

TABLE OF CONTENTS

Contents

1	OVERVIEW OF LEGAL ISSUES. INTRODUCTION TO REGULATORY ASPECTS AND PROBLEMS AROUND BIOMEDICAL RESEARCH.....	9
2	APPLICABLE EUROPEAN UNION (EU) LEGISLATION AND LOCAL NORMATIVES	10
2.1	General EU regulations about data protection	10
2.2	Local Regulation in Spain	14
2.3	Local Regulation in Sweden	15
2.4	Local Regulation in Italy	15
3	GENERAL REGULATIONS ABOUT SCIENTIFIC INVESTIGATION	16
3.1	EU regulations	16
3.2	Spanish normative	18
3.3	Swedish normative	23
3.4	Italian normative	23
3.5	Regulation of Health Data records in the GDPR and local regulation in Spain, Sweden, and Italy	24
3.5.1	<i>Special categories of data</i>	24
3.5.2	<i>Treatment for archival purposes of public interest, scientific / historical and statistical research</i>	26
4	KEY ASPECTS DEALT WITH IN COVID-X	33
4.1	Key aspects about principles established by the GDPR:	36
4.1.1	<i>Privacy by-design and by-default</i>	36
4.1.2	<i>Information and transparency</i>	40
4.1.3	<i>Accountability</i>	41
4.1.4	<i>Minimization</i>	42
4.1.5	<i>Data Quality</i>	43
4.1.6	<i>Anonymization/dissociation</i>	43
4.1.7	<i>International transfers</i>	45
4.2	Key ethical aspects	45
4.2.1	<i>General</i>	45
4.2.2	<i>Consent vs anonymization</i>	47
4.2.3	<i>Unexpected information to patients</i>	49
4.2.4	<i>Mandatory ethical and legal normative from the Open Calls Countries</i>	51
4.2.5	<i>Traceability and monitoring</i>	52



4.2.6	Consent (children included)	52
4.3	Key aspects for data uploads	52
4.4	Key aspects for Storage and Securitisation	53
4.5	Personal data security inside COVIX	54
5	TREATMENT GUIDELINES (Data security protocol)	56
5.1	Data Protection Policy	56
5.2	Risk Assessment	56
5.3	Impact assessment.....	57
5.4	Anonymization guide considerations	58
5.5	Security of the architecture design.....	61
5.6	Security measures for the safe storage of the data	63
5.7	Patient rights procedure	64
5.8	Treatment agreement documents	65
5.9	Audit and reporting mechanism for the authorities	66
5.10	Data breach notification mechanism.....	67
6	Conclusions	68
	References.....	69
	Appendix A	70

LIST OF TABLES

TABLE 1 – SPECIFIC PRIVACY PROTECTION OBJECTIVES AND THE MEASURES THAT WILL BE PLACED IN ORDER TO ACHIEVE THEM.	35
TABLE 2: MAPPING BETWEEN THE CIA PRINCIPLES AND THE SECURITY COMPONENTS OF THE COVID-X SANDBOX	53

ABBREVIATIONS

Accelerate	Transfer technical, business, and ethical knowledge
AI	Artificial Intelligence
API	Application Programming Interfaces
CA	Certification authority
COVID-X	Innovation action supported by the European Commission in the framework of the EC call SC1-PHE-CORONAVIRUS-2020-2B
CPU	Central Processing Unit
CROs	Contract research organizations
DPO	Data Protection Officer
EDPB	European Data Protection Board
EU	EUROPEAN UNION
GDPR	General Data Protection Regulation
GT29	Working Group of Article 29
HCSC	Hospital Clínico San Carlos
IVO	Inspektionen för vård och omsorg
LOPD	Ley Orgánica de Protección de Datos (Organic Law of Data Protection)
LOPDGDD	Ley Orgánica de Protección de Datos y Garantía de Derechos Digitales (Organic Law on Protection of Personal Data and Guarantee of Digital Rights)
OC	Open Calls
PIA	Data protection impact assessment
PKI	Public Key Infrastructure
RBAC	Role Based Access Control
Team Solution (TS)	Open call line for clinical partners working in a team with technology providers
VPN	Virtual Private Network
WMA	World Medical Association

BLOCK 1 (THE PROBLEM)

1 OVERVIEW OF LEGAL ISSUES. INTRODUCTION TO REGULATORY ASPECTS AND PROBLEMS AROUND BIOMEDICAL RESEARCH

The personal data related to health are considered as the information concerning the past, present and future, physical and mental health of an individual. In particular, data related to people's health are those referring to their percentage of disability and their genetic information.

European legislation on data protection, **characterizes in a different way** the treatment and transfer of health data, with respect to the treatment and transfer of other types of personal data (eg name, surname, etc.) and qualifies as data specially protected or sensitive data. The General Data Protection Regulation (GDPR) determines the possibility of the treatment of health data, without the prior consent of the interested party (patient) provided that said treatment is carried out by a health professional subject to the duty of professional secrecy or by personnel also subject to an equivalent obligation of secrecy, when is necessary in any of the following cases:

- Medical prevention or diagnosis,
- Provision of health care or medical treatments,
- Health services management,
- To safeguard the vital interest of the patient or another person, in the event that he or she is physically or legally incapacitated to grant her consent.

Therefore, the data provided by the patients who come to the health center **are collected and processed to manage their health control and health care** received.

In this sense, in the treatment of personal data collected in the clinical history of patients, in cases not collected by the aforementioned, the express and written consent of the patients must be obtained. Therefore, **when they are carried out for research or teaching purposes, the express and written consent of the patients** participating in the clinical, biomedical research study or clinical trial that is going to be carried out **will be required**, or otherwise, **a dissociation mechanism in said data to safeguard its anonymity** (this has been the approach followed in COVID-X).

2 APPLICABLE EUROPEAN UNION (EU) LEGISLATION AND LOCAL NORMATIVES

2.1 General EU regulations about data protection

- Regulation (EU) 2016/679 of the European Parliament and of the Council, of April 27, 2016, regarding the protection of natural persons with regard to the processing of personal data and the free circulation of these data and by which Directive 95/46 / EC is repealed.

Directly applicable standard in all EU Member States. It regulates the protection of personal data of natural persons and is applicable both to companies or professionals responsible or in charge of data processing with registered office in Europe, who provide services to European citizens, as well as to companies or professionals with registered office outside of the EU that carry out data processing as a result of their offer of goods or services intended for European citizens.

It is the general legal framework regarding the protection of personal data in the EU, fulfilling the mandate of article 16.2 of the TFEU. As a premise, part of the obligation that the principles and rules relating to the protection of natural persons with regard to the processing of their personal data must respect their fundamental freedoms and rights, in particular, the right to data protection personal (Recital 2). However, it also declares that the right to the protection of personal data is not an absolute right, but must be considered in relation to its function in society and maintain a balance with other fundamental rights, in accordance with the principle of proportionality (Considering 4). Therefore, the purpose of the Regulation is to contribute to the full realization of an area of freedom, security and justice, but also of an economic union, to economic and social progress.

It is based on two fundamental axes or pillars. On the one hand, the principles that should govern all processing of personal data, established in Article 5 and developed throughout the Regulation: legality, loyalty, transparency, limitation of purpose, minimization of data, accuracy, integrity, confidentiality and proactive responsibility (accountability)

On the other hand, the second pillar is the focus on risk management (widely covered in the COVID-X project, in accordance with arts 24, 25 and 32 GDPR)

- Regulation (EU) 2018/1807 of the European Parliament and of the Council, of November 14, 2018, on a framework for the free circulation of non-personal data in the European Union.

It applies to the treatment in the EU of electronic data that are not of a personal nature (this is the approach followed in COVID-X, in which the data that will be managed by the consortium will no longer be considered personal data). As a general rule, it prohibits Member States from imposing requirements on where data must be located

In the case of a data set consisting of personal and non-personal data, this Regulation will apply to the non-personal data of the data set. When personal and non-personal data in a data set are inextricably linked, this Regulation will apply without prejudice to Regulation (EU) 2016/679 (GDPR). Art. 2.2

It establishes guidelines for a framework for the free circulation of non-personal data in the European Union.

It should also be taken into account that Regulation 2018/1807 will not apply if all the processing activities of "non-personal data" are carried out outside the EU, while the GDPR must be respected even when the treatment of "personal data" is carried out outside the EU, provided that the processing is carried out in the context of the activities of an establishment of the controller or processor in the Union or if the data subject is located in the EU

The Guidelines recall that the rules for the transfer of "personal data" to third countries or international organizations, under the GDPR, must be complied with in any case.

The guidelines recall that neither of the two Regulations requires companies to separate the data sets that they control or process.

- Directive 2003/98 / EC, of the European Parliament and of the Council, of November 17, regarding the reuse of Public Sector information

It regulates the reuse of existing documents kept by public sector bodies and public companies in the Member States. Its purpose is to promote the use of open data and stimulate innovation in products and services.

The new regulation extends the scope of its provisions both with regard to certain material areas of special interest and to types of data of particular relevance.

Thus, research data is considered to be included when the activity is financed with public funds (this is the case of COVID-X), although the need to protect other legal assets such as intellectual and industrial property, the confidentiality derived from the existence of legitimate commercial interests or, where appropriate, the protection of personal data. This difficult balance is formulated with a new principle to take into account in this area: the data will be "as open as possible, as closed as necessary".

To this end, the Directive establishes a set of minimum standards that regulates the reuse and practical devices intended to facilitate the reuse of research data, in accordance with the conditions established in Article 10 (the Directive itself defines "research data : as documents in digital format, other than scientific publications, compiled or prepared in the course of scientific research activities and used as evidence in the research process, or commonly accepted in the research community as necessary to validate the conclusions and results of the investigation, without differentiating if it is personal data or not).

Recital 27 indicates that the volume of research data generated is growing exponentially and can be reused by other users outside the scientific community. In order to effectively and

comprehensively address growing societal challenges, the ability to access data from different sources and from all types of sectors and disciplines, to combine and reuse it, has become crucial and urgent. Research data includes statistics, results of experiments, measurements, observations derived from field work, survey results, recordings and images of interviews. They also include metadata, specifications, and other digital objects. Research data is different from scientific articles that disclose and comment on the conclusions resulting from your scientific investigations. For many years, the free availability and reuse of scientific research data supported by public funding have been the subject of specific policy initiatives. Open access refers to the practice of offering end users free online access to research results, with no restrictions on use and reuse other than the possibility of requiring recognition of authorship (COVID-X is fully aligned with this Directive). The open access policies aim, in particular, to provide researchers and the general public, as early as possible in the disclosure process, with access to research data and to facilitate their use and reuse. Open access improves quality, reduces the need for unnecessary duplication in research, accelerates scientific progress, combats scientific fraud, and can generally support economic growth and innovation. In addition to open access, it is commendable that efforts are being made to ensure that data management planning becomes standard scientific practice and to support the dissemination of research data that is easy to find, accessible, interoperable and reusable.

Recital 28 indicates that for the reasons set out above, it is appropriate to impose on the Member States the obligation to adopt open access policies with regard to publicly funded research data and to ensure that these policies are implemented by all organizations that carry out research activities. research and organizations that fund research. Organizations conducting research activities and organizations funding research can also be organized as public sector bodies and as public companies. Open access policies normally allow a series of exceptions when making the results of scientific research available to the public. In addition, the conditions under which certain research data can be reused should be improved. For this reason, some obligations arising from this Directive should be extended to research data resulting from scientific research activities subsidized by public funding or co-financed by entities from the public and private sectors. Under national open access policies, publicly funded research data will be open by default. However, in this context, concerns regarding privacy, personal data protection, confidentiality, national security, legitimate business interests such as trade secrets, and the intellectual property rights of third parties should be duly taken into account. , according to the principle "as open as possible, as closed as necessary."

To avoid any administrative burden, the obligations arising from this Directive should only apply to research data that has already been made available to the public by researchers, organizations conducting research activities or organizations that finance research through a institutional or thematic registry, and should not impose additional costs for obtaining data sets or require additional selection of data. Member States may extend the application of this Directive to research data made available to the public through data infrastructures other than registries, through open access publications or in the form of a file attached to an article, to a

data article or an article in a journal specialized in data. Documents other than research data should continue to be excluded from the scope of this Directive.

- Article 10 on research data

Chapter 2. Without prejudice to article 1, paragraph 2, letter c), research data will be reusable for commercial or non-commercial purposes in accordance with chapters III and IV, insofar as they are financed with public funds and in that Researchers, organizations that carry out research activities, or organizations that finance research have already made them available to the public through an institutional or thematic repository. In this context, legitimate business interests, knowledge transfer activities and pre-existing intellectual property rights should be taken into account.

- **Charter of Fundamental Rights**

The Charter of Fundamental Rights during the respective legal trend dedicates a separate article to the protection of personal data. Article 8 sets out the right to the protection of personal data of an individual and thus the protection of personal data has now its own legal basis apart from the right to respect for an individual's private life and the protection of human dignity. Art. 8 of the Charter sets out the rules for the legitimate processing of personal data, notably that the processing shall be fair and for pre-specified purposes based on the consent of the data subject or other legitimate basis laid down by law. Reference is furthermore made to two rights of the data subject: the right of access to the data and the right to have it rectified. Finally, Article 8 sets out the need for an independent authority, which shall control the compliance with the data protection rules.

So, we have to be prepared for the situation that a person wants to retire his data from the Project repository

- **HELSINKI DECLARATION: ETHICAL PRINCIPLES FOR MEDICAL INVESTIGATION ON HUMAN SUBJECTS.** Approved by the General Assembly of the World Medical Association (WMA) in 1964 (amended seven times, most recently at the General Assembly in October 2013).

Adopted in the Finnish capital by the 1964 General Assembly of the WMA, it is the most important international document regulating research in humans since the 1947 Nuremberg code.

Promulgated as a body of ethical principles that should guide the medical community and others who are engaged in human experimentation. It is considered the most important document in the ethics of research with human beings, although it is not a legal instrument that binds internationally. Its authority emanates from the degree of internal codification and from the influence it has gained nationally and internationally

- Basic principles

The basic principle is respect for the individual (Article 8), their right to self-determination and the right to make informed decisions (informed consent) (Articles 20, 21 and 22) including

participation in research, both at the beginning and during the course of the investigation. The duty of the researcher is solely towards the patient (Articles 2, 3 and 10) or the volunteer (Articles 16 and 18), and while there is a need to carry out an investigation (Article 6), the well-being of the subject must always be precedent on the interests of science or society (Article 5), and ethical considerations must always come from the preceding analysis of laws and regulations (Article 9).

It is recognized that when the research participant is incompetent, physically or mentally incapable of consent, or is a minor (Articles 23 and 24) then permission must be given by a substitute who watches over the best interest of the individual. In this case, his consent is very important (Article 25).

Since 2016, the Declaration of Helsinki is supplemented by the Declaration of Taipei on ethical considerations regarding health databases and biobanks.

2.2 Local Regulation in Spain

- Spanish Constitution 1978

- Article 18

It guarantees the right to honor, to personal and family privacy and to one's own image, in addition to the secrecy of communications. It determines that the Law will limit the use of information technology to guarantee the honor and personal and family privacy of citizens and the full exercise of their rights.

- Organic Law 3/2018, of December 5, on the Protection of Personal Data and guarantee of digital rights.

Develops the content of the European Directive GDPR.

- Law 37/2007, of November 16, on the reuse of public sector information.

Develops the content of the European Directive of the same name.

- Royal Decree 1495/2011, of October 24, which develops Law 37/2007, of November 16, on the reuse of public sector information, for the state public sector, within a framework of free competition.

Develops the content of the European Directive of the same name.



2.3 Local Regulation in Sweden

- Patient Data Act (2008:355).

2.4 Local Regulation in Italy

- Code for the Protection of Personal Data: Legislative Decree 196/2003, of 30 June, as amended by Legislative Decree 101/2018, of 10 August.



3 GENERAL REGULATIONS ABOUT SCIENTIFIC INVESTIGATION

3.1 EU regulations

- Convention of the Council of Europe for the protection of human rights and the dignity of the human being with respect to the applications of biology and medicine, signed in Oviedo (April 1997).

Object of the Convention: to protect human beings in their dignity and identity and guarantee respect for their integrity and other fundamental rights and freedoms with respect to the applications of biology and medicine.

The Convention deals explicitly, in detail and extensively, with the need to recognize the rights of patients, among which stand out the right to information, informed consent and the privacy of information relating to people's health, pursuing the scope of harmonization of the laws of the various countries in these matters

Obligation of each member state to adopt in its internal legislation the necessary measures to comply with the provisions of the Convention (informs national legislation).

Consent: requires that you be free and informed. Articles 5, 6 and 7: regulate free and informed consent and the requirements in the case of people who do not have the capacity to give their consent (minors, disabled, mentally ill). Article 17, consent requirements for scientific research.

Article 10: Right of everyone to have their private life respected when it comes to information related to their health. Likewise, the right to know all information obtained regarding her health, as well as to have her wish not to be informed.

- Regulation (EU) 536/2014 of the European Parliament and of the Council of April 16, 2014 on clinical trials of medicinal products for human use, and by which Directive 2001/20 / EC is repealed

It applies to all clinical trials conducted in the European Union. It does not apply to non-interventional studies. (art.1)

Recital 29: It is appropriate that universities and other research centers, in certain circumstances that are in accordance with applicable law on data protection, may collect data from clinical trials for use in future scientific research, for example, for the purposes of research in fields such as medicine and the natural or social sciences. In order to collect data for these purposes, the subject must consent to the use of their data outside of the clinical trial protocol, and have the right to withdraw such consent at any time. It is also necessary

that research projects based on such data, before being carried out, are subject to relevant reviews in the case of research on personal data, for example, on ethical aspects.

Recital 83: Respect for fundamental rights. Expressly, the protection of personal data indicates that this Regulation must be applied in accordance with those rights and principles.

Article 28: Establishes the mandatory conditions to carry out a clinical trial, including respect for the right to privacy, as well as protecting personal data in accordance with the GDPR (art. 28.1 d)

Without prejudice to the provisions of the GDPR, at the time of granting informed consent to participate in the clinical trial, the subject may be asked to give their consent for their data to be used exclusively for scientific purposes outside of the clinical trial protocol, being able to withdraw that consent at any time. Scientific research that is making use of the data outside the clinical trial protocol will be carried out in accordance with the applicable law on data protection (art. 28.2)

The test subject can leave the clinical trial at any time without any justification, and without suffering any harm, by withdrawing informed consent himself or, if he is unable to give informed consent, his legally designated representative. Without prejudice to the provisions of the GDPR, the withdrawal of informed consent will not affect the activities that have already been carried out and the use of the data obtained based on the informed consent, before its withdrawal. (art. 28.3)

In the same sense, Recital 76: applicability to the processing of personal data of Directive 95 / 467CE (now understand the GDPR, which repeals it). Specifies that the withdrawal of informed consent should not affect the results of activities already carried out, such as the storage and use of data obtained based on the informed consent before its withdrawal

Article 56: Registration, processing, treatment and storage of information: The promoter or, as appropriate, the researcher will record, process, treat and keep all the information of the clinical trial so that it can be communicated, interpreted and verified accurately, while protecting the confidentiality of the clinical records and personal data of the test subjects, in accordance with the applicable law on the protection of personal data.

It establishes the obligation to take the appropriate technical and organizational measures so that the information and personal data that are processed are protected (COVID-x anonymizes this data as the best possible protection measure) and prevent them from being accessed, disclosed, disseminated, modified or destroyed in an unauthorized or illegal way, or accidentally lost, especially when transmitted over a network (art. 56.2)

Mandatory nature of an EU database of clinical trials that, as a general rule, should not contain personal data of the test subject (in COVID-X this type of data is eliminated in the first step of the loading processes). It will only contain personal data to the extent necessary for the purpose of the database. No personal data of the test subjects will be publicly accessible (Recital 67 and articles 81 and 82).

It does not establish any indication about the possibility of using anonymized data

3.2 Spanish normative

- Law 14/1986, of April 25, General Health

Basic norm applicable to the entire territory of the State, except for articles 31, section 1, letters b) and c), and 57 to 69, which will constitute a supplementary right in those Autonomous Communities that have issued norms applicable to the matter that in said precepts is regulated.

Article 105 bis: The processing of personal data in health research will be governed by the provisions of the seventeenth additional provision of the Organic Law on Protection of Personal Data and Guarantee of Digital Rights. (Precept introduced by the 5th final provision of the Law on Data Protection and Guarantee of Digital Rights- LOPDGDD).

- Law 33/2011, of October 4, General Public Health

Basic rule applicable to the entire territory of the State. Establishes the rights of citizens in relation to public health. Among them, the Right to privacy, confidentiality and respect for dignity:

1. All persons have the right to respect for their dignity and personal and family privacy in relation to their participation in public health actions.

2. The personal information used in public health actions will be governed by the provisions of Organic Law 15/1999, of December 13, on the Protection of Personal Data (now you have to understand all the regulations regarding data protection: GDPR and LOPDGDD) and in Law 41/2002, of November 14, basic regulating the Autonomy of the Patient and Rights and Obligations in the matter of Information and Clinical Documentation (art. 7).

- Article 41.2

It also establishes that the health administrations will not need to obtain the consent of the affected persons for the treatment of personal data, related to health, as well as their transfer to other public health administrations, when this is strictly necessary for the protection of the health of the population. (art. 41.2)

3. For the purposes indicated in the two previous sections, public or private persons will assign to the health authority, when so required, the personal data that are essential for making decisions in public health, in accordance with the established in Organic Law 15/1999, of December 13, on the Protection of Personal Data.

In any case, access to medical records for epidemiological and public health reasons will be subject to the provisions of section 3 of article 16 of Law 41/2002, of November 14, basic regulating the Autonomy of the Patient and of Rights and Obligations regarding Information and Clinical Documentation (art. 41.3)

- Security of the information.

1. At all levels of the public health information system, the necessary measures will be adopted to guarantee data security.
2. Workers in public and private centers and services and those who by reason of their activity have (art. 43)

- Law 41/2002, of November 14, basic regulating the autonomy of the patient and rights and obligations regarding information and clinical documentation.

Basic rule of application throughout the state that regulates the rights and obligations of patients, users and professionals, as well as health centers and services, public and private, in terms of patient autonomy and information and clinical documentation.

Supplementary application in clinical research projects: The rules of this Law relating to healthcare information, information for the exercise of freedom of choice of doctor and center, informed consent of the patient and clinical documentation, will be applicable supplementary in medical research projects (2nd additional provision)

It recognizes and regulates the rights of the patient, among which the right to information, informed consent and the privacy of information related to people's health stand out.

Respect for the autonomy of the patient's will and her privacy will guide all the activity aimed at obtaining, using, filing, safeguarding and transmitting the information and clinical documentation.

Any action in the field of health requires, in general, the prior consent of patients or users (art. 2)

Art.17.6. The technical security measures established by the legislation regulating the conservation of files containing personal data and, in general, by Organic Law 15/1999, on the Protection of Personal Data, are applicable to clinical documentation. (now GDPR, and LOPDGDD)

- Law 14/2007, of July 3, on Biomedical Research.

It regulates biomedical research in general and in particular, it regulates the performance of genetic analysis and the processing of personal genetic data. Article 5 indicates:

1. The protection of personal privacy and the confidential treatment of personal data resulting from biomedical research activity will be guaranteed, in accordance with the provisions of Organic Law 15/1999, of December 13, on the Protection of Character Data Personal (now understand the GDPR and LOPDGDD). The same guarantees will apply to biological samples that are a source of personal information.

2. The transfer of personal data to third parties unrelated to medical-assistance activities or biomedical research will require the express and written consent of the interested party.

In the event that the data obtained from the source subject could reveal personal information of their relatives, the transfer to third parties will require the express and written consent of all interested parties.

3. The use of data related to people's health for purposes other than those for which consent was given is prohibited.

4. Any person who, in the exercise of their functions in relation to a medical-assistance action or a biomedical research, whatever the scope of both, accesses personal data will be subject to the duty of secrecy. This duty will persist even after the investigation or action has ceased.

5. If it is not possible to publish the results of an investigation without identifying the person who participated in it or who contributed biological samples, such results may only be published when the prior and express consent of the latter has been obtained.

Regarding genetic data, it is indicated that the right to privacy and respect for the will of the subject in matters of information will be guaranteed, as well as the confidentiality of personal genetic data.

Free: the entire process of donation, transfer, storage and use of biological samples for both source subjects and depositors, must be devoid of any purpose or profit motive. Personal genetic data may not be used for commercial purposes.

Consent: the written consent of the source subject or, where appropriate, their legal representatives must be obtained in advance for the treatment of samples for research purposes or of personal genetic data.

Data quality: the data obtained from genetic analysis may not be processed or transferred for purposes other than those provided for in this Law.

- Article 50. Access to genetic data by healthcare personnel.

1. The health professionals of the center or establishment where the patient's medical history is kept will have access to the data contained in it as long as it is pertinent for the care they provide to the patient, without prejudice to the duties of reserve and confidentiality to that they are submitted.

2. Genetic personal data may only be used for epidemiological, public health, research or teaching purposes when the interested subject has expressly given their consent, or when said data has been previously anonymized (therefore if they are anonymized, consent is not necessary for their use for research purposes)

3. In exceptional cases and of general health interest, the competent authority, after a favorable report from the authority on data protection, may authorize the use of encoded genetic data, **always ensuring that they cannot be related or associated with the source subject by part of third parties only anonymizing you can achieve this**).

- Article 51. Duty of confidentiality and right to protection of genetic data.

1. The personnel who access the genetic data in the exercise of their functions will be subject to the duty of secrecy on a permanent basis. Only with the express and written consent of the person from whom they originate may personal genetic data be disclosed to third parties.

If it is not possible to publish the results of a research without identifying the source subjects, such results can only be published with their consent.

- Article 52. Conservation of data.

1. Personal genetic data will be kept for a minimum period of five years from the date they were obtained, after which the interested party may request their cancellation.

2. If there is no request from the interested party, the data will be kept for the period necessary to preserve the health of the person from whom they come or of third parties related to it.

3. **Outside of these assumptions, the data may only be kept, for research purposes, in an anonymized form, without the identification of the source subject being possible.**

The following articles refer to the use of human biological samples for biomedical research purposes:

- Article 58. Obtaining the samples.

1. Obtaining biological samples for biomedical research purposes may only be carried out when the written consent of the source subject has been previously obtained and after information on the consequences and risks that such collection may entail for their health. Said consent will be revocable.

2. The consent of the source subject will always be necessary when the intention is to use biological samples for biomedical research that have been obtained for a different purpose, whether or not they are anonymized.

Notwithstanding the foregoing, exceptionally coded or identified samples may be treated for biomedical research purposes without the consent of the source subject, **when obtaining said consent is not possible or represents an unreasonable effort** in the sense of article 3.i) of this Law. In these cases, the favorable opinion of the corresponding Research Ethics Committee will be required, which must take into

account, at least, the requirements set forth in the precept, among them, that the **confidentiality** of the data of personal character.

- Article 59. Information prior to the use of the biological sample.
 1. Without prejudice to the provisions of the legislation on protection of personal data, and in particular, in article 45 of this Law, before issuing consent for the use of a biological sample for biomedical research purposes **that does not go to be subjected to an anonymization process**, the source subject will receive the following information in writing:
 - a) Purpose of the investigation or line of investigation for which you consent.
 - b) Expected benefits.
 - c) Possible inconveniences related to donating and obtaining the sample, including the possibility of being contacted later in order to collect new data or obtain other samples.
 - d) Identity of the person responsible for the investigation.
 - e) Right to revoke consent and its effects, including the possibility of destruction or anonymization of the sample and that such effects will not be extended to data resulting from investigations that have already been carried out.
- Article 60 indicates

Specific consent may provide for the use of the sample for other lines of research related to the one initially proposed, including those carried out by third parties. If this is not the case, the source subject will be asked to grant, if it deems it appropriate, a new consent.
- Article 61. Conservation and destruction of samples.
 1. In the event that the sample is conserved, the source subject will be informed in writing of the conditions of conservation, objectives, future uses, transfer to third parties and conditions to be able to withdraw them or request their destruction. However, biological samples used in biomedical research will be kept only as long as they are necessary for the purposes that justified their collection, unless the source subject has given their explicit consent for other subsequent uses.
 2. **What is indicated in the previous section is understood to be applicable as long as the identification data of the sample have not been subjected to anonymization in accordance with the provisions of this Law (therefore anonymized samples can be used without need for consent).**

Regarding biobanks, the second transitory provision. Previously stored samples:

Biological samples obtained prior to the entry into force of this Law may be processed for biomedical research purposes when the source subject has given their consent or when the samples have been previously anonymized.

For all the above, it would be demonstrated that pseudo-anonymization would not be a sufficient technique and it is absolutely necessary that the information be totally anonymized in COVID-X.

3.3 Swedish normative

- Ethical Review Act (2003:460)
- The Law on Responsibility of Good Research Practice (2019:504): Establishes the principles and guidelines for good research practice.
- The Biobank Act (2002:297): Establishes the basic requirements for the authorization and operation of biobanks for biomedical research purposes and the treatment of biological samples of human origin.

3.4 Italian normative

- Ministerial Decree of 15 July 1997 "Implementation of the European Union guidelines of good clinical practice for the execution of clinical trials of medicines"
- Ministerial Decree of 23 November 1999 "Composition and determination of the functions of the National Ethics Committee for clinical trials of medicines"
- Executive Decree of 25 May 2000 "Electronic transmission of data relating to clinical trials of medicines"
- Ministerial Decree of 30 May 2001 "Inspection checks on compliance with the rules of good manufacturing practice and good clinical practice"
- Legislative Decree n. 211/2003, of 24 June "Implementation of Directive 2001/20 / EC on the application of good clinical practice in the execution of clinical trials of medicinal products for clinical use"
- Ministerial Decree of 17 December 2004 "Prescriptions and conditions of a general nature, relating to the execution of clinical trials of medicines, with particular reference to those for the purpose of improving clinical practice, as an integral part of health care"
- Ministerial Decree of 21 December 2007 "Procedures for forwarding the request for authorization to the competent Authority, for the communication of substantial amendments"

and the declaration of conclusion of the clinical trial and for the request for an opinion to the ethics committee"

- AIFA Resolution of 20 March 2008, containing "Guidelines for the classification and conduct of observational studies on drugs"
- Ministerial Decree of 7 November 2008 - Amendments and additions to the decrees of 19 March 1998, containing "Recognition of the suitability of centers for clinical trials of medicines"; May 8, 2003, on "Therapeutic use of medicinal products undergoing clinical trials" and May 12, 2006, on "Minimum requirements for the establishment, organization and operation of Ethics Committees for clinical trials of medicines"
- Ministerial Decree of 14 July 2009 "Minimum requirements for insurance policies for the protection of subjects participating in clinical trials of medicines"
- Ministerial Decree of 15 November 2011 "Definition of the minimum requirements for contract research organizations (CROs) in the context of clinical trials of medicines"
- AIFA Resolution of 20 September 2012 "Adoption of the CT-3 guidelines (June 2011) of the C.E. implementation of Directive 2001/20 / EC, of the ICH E2F guidelines (September 2011) and establishment of a national database relating to the monitoring of the safety of medicinal products in clinical trials"
- Ministerial Decree of 8 February 2013 "Criteria for the composition and functioning of the ethics committees"
- Ministerial Decree of 19 April 2018 "Constitution of the National Coordination Center of territorial ethics committees for clinical trials on medicinal products for human use and medical devices, pursuant to Article 2, paragraph 1, of the Law of 11 January 2018, n. 3"

3.5 Regulation of Health Data records in the GDPR and local regulation in Spain, Sweden, and Italy

3.5.1 Special categories of data

- EU

Health data and genetic data are established in the GDPR as "special categories of data". The processing of personal genetic data and data related to health (art. 9.1) is only allowed in the cases, among them, to highlight:

a) the interested party gave their explicit consent.

i) the treatment is necessary for reasons of public interest in the field of public health.

j) **the processing is necessary** for archival purposes in the public interest, **scientific** or historical research purposes or statistical purposes, in accordance with Article 89 (1), on the basis of Union or Member State law, **which must be proportional to the objective pursued**, respect essentially the right to data protection and establish adequate and specific measures to protect the interests and fundamental rights of the interested party. Therefore, the treatment

of health and genetic data is allowed without the need for consent, when necessary for scientific research purposes, requiring additional guarantees, in accordance with art. 89 and that it is also covered by specific EU regulations and the laws of the member states. **Since the proportionality with the objective pursued is difficult to assess, COVID-X has adopted the anonymization strategy as an appropriate and specific measure** to protect the interests and fundamental rights of patients.

- Considering 51 to 56

They refer to the treatment of special categories of data regulated in art. 9, without special mention of scientific research.

- Spain

- Article 9 “Ley Orgánica de Protección de Datos” (LOPD; "Special categories of data").

It determines that the data processing contemplated in letters g), h) and i) of article 9.2 of Regulation (EU) 2016/679 based on Spanish law must be covered by a norm with the force of law, which may establish additional relative requirements to your security and confidentiality.

- Seventeenth additional provision LOPD. Section 1 ("Treatment of health data")

It establishes the Spanish laws (and their development provisions) in which the treatment of health-related data and genetic data is covered, for reasons, among others, of public health or scientific research. To highlight: a) Law 14/1986, of April 25, General Health. c) Law 41/2002, of November 14, regulating basic patient autonomy and rights and obligations regarding information and clinical documentation. f) Law 14/2007, of July 3, on Biomedical Research. g) Law 33/2011, of October 4, General Public Health. i) The revised text of the Law of guarantees and rational use of 105 medicines and health products, approved by Royal Legislative Decree 1/2015, of July 24. j) The revised text of the General Law on the rights of people with disabilities and their social inclusion, approved by Royal Legislative Decree 1/2013 of November 29).

- Sweden

- Patientdatalag. (2008:355)

- Italy

- Article 2-sexies Code for the Protection of Personal Data (Processing of special categories of personal data necessary for reasons of significant public interest)
- Article 2-septies Code for the Protection of Personal Data (Guarantee measures for the processing of genetic, biometric and health data)

3.5.2 Treatment for archival purposes of public interest, scientific / historical and statistical research

- EU

- Article 89 GDPR ("Guarantees and exceptions applicable to treatment for archival purposes in the public interest, scientific or historical research purposes or statistical purposes")

1. The treatment for archival purposes in the public interest, scientific or historical research purposes or statistical purposes will be subject to the appropriate guarantees (it is an impossible concept to specify which implies a risk), in accordance with this Regulation, for the rights and the freedoms of the interested parties. These guarantees will ensure that technical and organizational measures are available, in particular to guarantee respect for the principle of minimizing personal data. Such measures may include pseudonymisation, provided that in this way said purposes can be achieved (it is risky to guarantee that pseudonymisation is sufficient guarantee in a massive data processing project). Provided that those ends can be achieved by further processing that does not allow or no longer allows the identification of the data subjects, those ends will be achieved in that way.

2. When personal data is processed for scientific or historical or statistical research purposes, the law of the Union or of the Member States may establish exceptions to the rights contemplated in articles 15, 16, 18 and 21, subject to the conditions and guarantees indicated in paragraph 1 of this article, provided that it is probable that those rights make it impossible or seriously impede the achievement of scientific purposes and when such exceptions are necessary to achieve those purposes. (see A.D. 17^a.2 e) allows by means of a norm with the force of law to establish these exceptions with certain conditions and requirements)

3. When personal data is processed for archiving purposes in the public interest, the law of the Union or of the Member States may provide exceptions to the rights contemplated in articles 15, 16, 18, 19, 20 and 21, subject to the conditions and guarantees cited in paragraph 1 of this article, provided that these rights may seriously impede or impede the achievement of scientific purposes and when such exceptions are necessary to achieve those purposes.

4. In the event that the processing referred to in sections 2 and 3 also serves another purpose at the same time, the exceptions will only apply to processing for the purposes mentioned in referred sections.

Therefore, the processing of data for scientific research is allowed, without the need for consent, provided that adequate guarantees are adopted to indicate protect the rights and freedoms of individuals, in particular, it expressly pseudonymisation, which is mandatory if the purposes of the treatment they can be achieved this way. (See NOTE at the end of this chapter)

- Considering 156 to 160, 33, and 161 to 163

(156) The processing of personal data for archival purposes in the public interest, scientific or historical research purposes or statistical purposes must be subject to **adequate guarantees**

for the rights and freedoms of the interested party in accordance with these Regulations. These guarantees must ensure that technical and organizational measures are applied so that the principle of data minimization is observed, in particular. The subsequent processing of personal data for archival purposes in the public interest, scientific or historical research purposes or statistical purposes must be carried out when the controller has evaluated the viability of fulfilling those purposes through data processing that does not allow the identification of the interested parties, or that no longer allows it, provided that there are **adequate guarantees (such as, for example, pseudonymisation of data that could be enough for massive treatment of data)**. Member States should establish adequate safeguards for the processing of personal data for archival purposes in the public interest, scientific or historical research purposes or statistical purposes. Member States should be authorized to establish, under specific conditions and subject to adequate guarantees for the interested parties, specifications and exceptions with respect to the information requirements and the rights of rectification, deletion, oblivion, limitation of treatment, portability of the data, and opposition, when personal data is processed for archival purposes in the public interest, scientific and historical research purposes or statistical purposes. The conditions and guarantees in question may entail specific procedures for the interested parties to exercise said rights if appropriate in light of the purposes pursued by the specific treatment, together with the technical and organizational measures aimed at minimizing the processing of personal data according to the principles of proportionality and necessity. The processing of personal data for scientific purposes must also observe other relevant rules, such as those relating to clinical trials (reference to the specific regulations on scientific research)

(157) By combining information from registries, researchers can gain valuable new insights into widespread medical conditions such as cardiovascular disease, cancer, and depression. Based on records, research results can be more robust as they are based on a larger population

(159) This Regulation should also apply to the processing of personal data for scientific research purposes. The processing of personal data for scientific research purposes should be interpreted, for the purposes of this Regulation, in a broad manner, including, for example, technological development and demonstration, fundamental research, applied research and research funded by the sector private. Scientific research purposes should also include studies conducted in the public interest in the field of public health. In order to comply with the specificities of the processing of personal data for scientific research purposes, specific conditions must apply, in particular with regard to the publication or otherwise communication of personal data in the context of scientific research purposes. If the result of scientific research, in particular in the field of health, justifies other measures for the benefit of the interested party, the general rules of this Regulation should be applied taking into account such measures.

(33) It is often not possible to fully determine the purpose of the processing of personal data for scientific research purposes at the time of collection. Therefore, data subjects should be allowed to consent to certain fields of scientific research that respect recognized ethical

standards for scientific research. Interested parties should have the opportunity to give their consent only for certain areas of research or parts of research projects, to the extent that the intended purpose allows.

Given the difficulty that the collection of consents can have for each study or purpose, COVID-X has adopted anonymization as the best strategy (see NOTE at the end of this chapter).

(161) In order to grant consent for the participation in scientific research activities in clinical trials, the relevant provisions of Regulation (EU) No 536/2014 of the European Parliament and of the Council (on clinical trials of medicinal products for human use).

- Spain

- Seventeenth additional provision LOPDGDD. Section 2 ("Treatment of health data - Research")

2. The treatment of data in health research will be governed by the following criteria:

a) The interested party or, where appropriate, her legal representative may grant consent for the use of their data for health research purposes and, in particular, biomedical. Such purposes may cover categories related to general areas linked to a medical or research specialty.

b) The health authorities and public institutions with powers in public health surveillance may carry out scientific studies without the consent of those affected in situations of exceptional relevance and seriousness for public health.

c) The reuse of personal data for health and biomedical research purposes will be considered lawful and compatible when, having obtained consent for a specific purpose, the data is used for research purposes or areas related to the area in which the initial study is scientifically integrated.

In such cases, those responsible must publish the information established by article 13 of Regulation (EU) 2016/679 of the European Parliament and of the Council, of April 27, 2016, regarding the protection of natural persons with regard to the processing of your personal data and the free circulation of these data, in an easily accessible place on the corporate website of the center where the research or clinical study is carried out, and, where appropriate, in that of the promoter, and notify the existence of this information by electronic means to those affected. When they lack the means to access such information, they may request its submission in another format.

For the treatments provided for in this letter, **a prior favorable report from the research ethics committee will be required.**

d) The use of pseudonymised personal data for health and, in particular, biomedical research purposes is considered lawful.

The use of pseudonymised personal data for public health and biomedical research purposes will require:

1st. A technical and functional separation between the research team and those who carry out the pseudonymization and keep the information that enables re-identification.

2nd. That the pseudonymized data are only accessible to the research team when:

i) There is an express commitment to confidentiality and not to carry out any re-identification activity.

ii) Specific security measures are adopted to prevent re-identification and access by unauthorized third parties.

The data may be re-identified at its source, when, as a result of an investigation that uses pseudonymised data, the existence of a real and specific danger to the safety or health of a person or group of people, or a serious threat is appreciated for their rights or is necessary to guarantee adequate healthcare.

e) When personal data is processed for health research purposes, and in particular biomedical, for the purposes of article 89.2 of Regulation (EU) 2016/679, the rights of those affected provided for in articles 15, 16, may be exempted. 18 and 21 of Regulation (EU) 2016/679 when:

1^o The aforementioned rights are exercised directly before researchers or research centers that use anonymized or pseudonymized data.

2^o The exercise of such rights refers to the results of the investigation.

3^o The investigation is aimed at an essential public interest related to State security, defense, public security or other important objectives of general public interest, provided that in the latter case the exception is expressly included by a rule with the rank of Law.

f) When, in accordance with the provisions of article 89 of Regulation (EU) 2016/679, a treatment is carried out for research purposes in public health and, in particular, biomedical, the following will be carried out:

1st Carry out an impact assessment that determines the risks derived from the treatment in the cases provided for in article 35 of Regulation (EU) 2016/679 or in those established by the supervisory authority. This assessment will specifically include the re-identification risks linked to the anonymization or pseudonymization of the data.

2^o Submit scientific research to quality standards and, where appropriate, to international guidelines on good clinical practice.

3^o Adopt, where appropriate, measures aimed at guaranteeing that researchers do not access data identifying the interested parties.

4th Appoint a legal representative established in the European Union, in accordance with article 74 of Regulation (EU) 536/2014, if the promoter of a clinical trial is not

established in the European Union. Said legal representative may coincide with the one provided for in article 27.1 of Regulation (EU) 2016/679.

g) The use of pseudonymised personal data for research purposes in public health and, in particular, biomedical, must be submitted to the prior report of the research ethics committee provided for in the sector regulations.

In the absence of the existence of the aforementioned Committee, the entity responsible for the investigation will require a prior report from the data protection officer (DPO) or, failing that, an expert with the prior knowledge in article 37.5 of Regulation (EU) 2016/679.

h) The research ethics committees, in the field of health, biomedical or medicine, must include among their members a data protection delegate or, failing that, an expert with the knowledge that the Regulation (EU) 2016/679 when they deal with research activities that involve the processing of personal data or pseudonymised or anonymized data.

Conclusion: it allows the processing of data for scientific research purposes without consent in two cases:

a) When consent was obtained for a specific purpose, and the data is used for purposes or research areas related to the area in which the initial study is scientifically integrated. Previous favorable report from the Ethics Committee.

b) **If they are pseudonymised, provided that specific guarantees are established** (how be sure that are enough?) and the requirements indicated are met and after a report from the Research Ethics Committee, or failing that, from the DPD of the responsible entity. SEE NOTE at the end of this chapter.

- Sixth transitory provision LOPD ("Reuse for health and biomedical research purposes of personal data collected prior to the entry into force of this law")

The reuse of personal data lawfully collected prior to the entry into force of this law will be considered lawful and compatible when any of the following circumstances occur:

- a) That said personal data be used for the specific purpose for which consent was given.
- b) That, having obtained consent for a specific purpose, such data are used for purposes or areas of research related to the medical or research specialty in which the initial study is scientifically integrated.

- Fifth final provision LOPD ("Modification of Law 14/1986, of April 25, General Health")

A new Chapter II is added to Title VI of Law 14/1986, of April 25, General Health with the following content:

"Chapter II. Treatment of health research data. Article 105 bis:

The treatment of personal data in health research will be governed by the provisions of the seventeenth additional provision of the Organic Law on Protection of Personal Data and Guarantee of Digital Rights " – LOPDGDD.

- Ninth final provision LOPD ("Modification of Law 41/2002, of November 14, basic regulator of the autonomy of the patient and of rights and obligations regarding information and clinical documentation")

Section 3 of article 16 of Law 41/2002, of November 14, basic regulating the autonomy of the patient and of rights and obligations in matters of information and clinical documentation, is modified, which happens to have the following wording:

«Article 16 [...]

3. Access to medical records for judicial, epidemiological, public health, research or teaching purposes is governed by the provisions of current legislation on the protection of personal data, and in Law 14/1986, of April 25, General Health, and other rules of application in each case. Access to the medical record for these purposes requires preserving the patient's personal identification data, separate from those of a clinical-care nature, so that, as a general rule, anonymity is ensured, unless the patient himself has given his consent to do not separate them.

The investigation cases provided for in section 2 of the seventeenth additional provision of the Organic Law on Protection of Personal Data and Guarantee of Digital Rights are excepted.

Likewise, the cases of investigation of the judicial authority in which the unification of the identifying data with the clinical care are considered essential, in which the judges and courts in the corresponding process will be subject to. Access to the data and documents of the medical record is strictly limited to the specific purposes of each case.

When this is necessary for the prevention of a serious risk or danger to the health of the population, the health administrations referred to in Law 33/2011, General Public Health, may access the identifying data of patients for reasons epidemiological or protection of public health. Access must be made, in any case, by a healthcare professional subject to professional secrecy or by another person subject, likewise, to an equivalent obligation of secrecy, with prior motivation from the Administration requesting access to the data. "

- Sweden
 - The Patient Data Act (2008:355) contains regulations on the archiving of patient records and of their use for quality registers, which may be used for research.

In Sweden, GDPR replaced the Personal Data Act (Personuppgiftslagen, PuL, 1998:204) in May 2018. The national applications of the law are described in the Swedish Act with Supplementary Provisions concerning the EU General Data Protection Regulation (2018:218), known as the Data Protection Act, and the Supplementary Provisions concerning the EU General Data Protection Regulation Ordinance (2018:219). The Data Protection Act is

subsidiary in relationship to other laws and regulations that regulate treatment of personal data, meaning that the law is not applicable if it contradicts said laws. There are many instances of such occurrences, mainly within specific sectors wherein it is described how government offices and agencies can treat personal data. Such a law or regulation is applicable only if it is in line with GDPR and treats an issue that according to GDPR can be regulated or specified at a national level. The supervisory authority for GDPR is appointed in the Supplementary Provisions concerning the EU General Data Protection Regulation Ordinance (2018:219) as the Swedish Data Protection Authority (Datainspektionen). They can decide that a company or authority that contravenes GDPR must pay an administrative fine.”

- Italy
 - Title VII Code for the Protection of Personal Data (Processing for archival purposes in the public interest, for scientific or historical research or for statistical purposes), articles from 97 to 110bis

NOTE:

Due to all these previous statements, one could question **why anonymize if it is allowed to treat health data for scientific research purposes, only with the measure of pseudo-anonymization and in this way it is easier?**

The answer is that it is true that it is allowed to process pseudonymised personal data in the field of biomedical research, but not in all situations (consent being required in some of them) and it is necessary to observe that it is continuously indicated that they must be established ADEQUATE additional guarantees (if something fails, it will be considered that the failure occurred because the guarantees were not adequate). In the case of an incident, breach or attack of singularization, it would be shown that these guarantees were not adequate, therefore all these pseudo data automatically become personal data, the entity that is treating them being the person responsible for their treatment, with all the consequences and measures that this entails (including financial penalties). **This risk is too high to assume that it can arise in the course of the project (it cannot be transferred to the Consortium)**, with the possibility of going one step further and anonymizing (it costs only a little more) and achieving a much higher and guaranteed level of privacy (**It's hard to argue that your measurements were adequate if you didn't use all possible measurements.**) For this reason, anonymizing has been the strategy followed in COVID-X.

BLOCK 2 (THE SOLUTION)

4 KEY ASPECTS DEALT WITH IN COVID-X

- **System Privacy Requirements**

Understanding how a treatment of personal data can affect the privacy of individuals is the key to designing and developing reliable systems from a data protection point of view.

The GDPR describes in its article 5 the basic principles that must be taken into account when carrying out the treatments, so that these six principles (legality, loyalty and transparency, limitation of the purpose, minimization of data, accuracy, limitation of the retention period, integrity and confidentiality) together with proactive responsibility (or accountability) become the core of the standard and the objective that every system, application, service or process must guarantee in its design, in addition to the functional requirements to satisfy the system's own.

- **Privacy and Security Objectives**

Traditionally, the design of secure and reliable systems has focused on analyzing risks and responding to threats that affect security objectives that are more privacy oriented:

- Confidentiality, avoiding unauthorized access to systems,
- Integrity, protecting them from unauthorized modifications of the information, and
- Availability, ensuring that data and systems are available when needed.

However, although unauthorized access and modification of personal data can become a critical aspect that threatens the privacy of individuals, there are other risk factors that may appear during authorized data processing and that must be identified during risk assessment for the rights and freedoms of data subjects.

Loss of autonomy in decision-making, excessive data collection, re-identification, discrimination and/or stigmatization of people, bias in automated decisions, lack of understanding by users of the scope and risks, treatment or profiling that is not legitimized, invasive or incorrect are examples of privacy risks with a clear impact on the rights and freedoms of people that cannot be managed using a traditional risk model focused on the exclusive protection of the objectives of security.

Taking into account this scenario and the possible privacy risks associated with the planned and authorized operation of the systems that collect, use and disclose personal data, it is necessary to broaden the framework of analysis so that it covers both the risks derived from its treatment not authorized as those that may arise from a planned and permitted processing of the information.

To respond to these possible risks, three new specific privacy protection objectives have been included in the COVID-X analysis, the guarantee of which becomes a safeguard of the treatment principles established by the GDPR:

- **Unlinkability:** it aims to process the information in such a way that personal data from a treatment domain cannot be linked to personal data from a different domain or that the establishment of such a link involves a disproportionate effort. This privacy objective minimizes the risk of unauthorized use of personal data and the creation of profiles by interconnecting information belonging to different data sets, establishing guarantees on the principles of purpose limitation, data minimization and limitation of the conservation period.

In COVID-X, this objective is met through the separation between the data lake and the extractions that are made from it, in which there will be a total dissociation with the source data without the recipient of that new data set being able to identify the source data (see *Anonymization guide for details*).

- **Transparency:** Clarify the treatment of data so that the collection, processing and use of the information can be understood and reproduced by any of the parties involved and at any time during the treatment. This privacy objective intends that the context of the treatment is perfectly delimited and that the information on the purposes and the applicable legal, technical and organizational conditions is available before, during and after the treatment to all the parties involved, both for the controller and for the subject whose data is processed, thus minimizing the risks that may affect the principles of loyalty and transparency.

In COVID-X, this objective is achieved in two different ways:

- On the one hand, there are data sets that are collected through informed consent in which patients are informed about the objectives of the study and how they can withdraw their participation in it at any time.
- On the other hand, covered by article 89 of the GDPR that allows the processing of data for scientific research purposes in adequate security conditions, in COVID-X an anonymization is performed in two steps: In the first one, it is carried out a pseudo-anonymization of the data, after which the treatment performed still corresponds to the purpose for which the data were collected (since the Data Lake is used to improve the management and care of the patients themselves). In the second step, a complete anonymization is

carried out, after which the result is no longer personal data and could be processed for different purposes.

- **Control (Intervenability):** guarantees the possibility that the parties involved in the processing of personal data and, mainly the subjects whose data are processed, can intervene in the processing when necessary to apply corrective measures to the processing of the information. This objective is closely related to the definition and implementation of procedures for the exercise of rights regarding data protection, the presentation of claims or the revocation of the consents given by the interested parties, as well as mechanisms to guarantee, by the responsible, the evaluation of compliance and the effectiveness of the obligations that are set by the regulations, which helps to respect the principles of accuracy and proactive responsibility set by the GDPR.

In COVID-X in the process of pseudo-anonymization of the data there are filters in which the data of those subjects who have expressed their right to revoke their consent can be eliminated. On the other hand, the security policy and the mechanisms and guides included in this document guarantee that the data is processed by the controller in accordance with the requirements of the GDPR.

These three new protection objectives, together with the existing security objectives, establish a global framework of protection in the treatment of personal data and determine, as a result of carrying out an assessment of the risks on its affectation, other types of attributes or non-functional requirements that the system must satisfy and that become the inputs to the privacy design processes (Table 1).

TABLE 1 – SPECIFIC PRIVACY PROTECTION OBJECTIVES AND THE MEASURES THAT WILL BE PLACED IN ORDER TO ACHIEVE THEM.

PRIVACY PROTECTION OBJECTIVES		
UNLINKABILITY	TRANSPARENCY	CONTROL
Data minimization Limitation of the conservation period Integrity and confidentiality	Legality, loyalty, and transparency Purpose limitation	Purpose limitation Accuracy Integrity and confidentiality Proactive accountability

Security measures and privacy measures to protect the personal data of the partners of COVIX and data from the Open Calls applicants are described in the deliverables of WP3.

The Data Protection, Ethics and Regulation dimension will provide SMEs and external consortia (of SMEs and health systems) support to ensure that the proposals submitted are compliant with the current legal and ethical regulations, in order to ensure a swift approval by the local Ethics committees.

In addition, the dimension will also support the signing of the agreements between the SMEs and the clinical sites, so the validation can take place.

This dimension will work closely with CovidProgramme, as the ethical and legal advice will be offered to the SMEs being mentored.

In order to provide this service, project will carry out a thorough review of the current European, national and local legislations, in order to be able to develop a framework for the activities that may take place during the open calls and the validation processes. Project will pay special attention to the limitations on the use of clinical data by for-profit third parties. In addition, periodic reviews of the norms and legislation will take place, so any substantial modification can be incorporate as soon as possible.

With the framework in place, Project will define the criteria that will be used to review the projects after being selected in the open calls. Our role will be to verify, before they are submitted to the local Ethical Committees for Clinical

Research, the compliance with the applicable legal and Ethical regulations. In case some part of the proposal is not feasible under the current regulations, advices will be provided regarding how it should be modified. This dimension will also guide the SMEs though the process of getting approval by the Local Ethics Committees. In case, the proposal is initially rejected because of legal or ethical issues, support will be offered to help address the Committee's requests.

After approval has been obtained, CovidProtection will carry out a monitoring of the validation process, to ensure the compliance with the regulations during the whole project.

After the completion of the validation project, CovidProtection will provide the SMEs with advices regarding futures steps to carry out more ambitious projects, such as testing the effectiveness of the proposal, or to carry out healtconomic studies.

4.1 Key aspects about principles established by the GDPR:

4.1.1 Privacy by-design and by-default

Regulation (EU) 2016/679, General Data Protection (hereinafter, GDPR), in its article 25 and under the heading 'Data protection by design and by default', incorporates into the data protection regulations the obligation to consider privacy requirements from the earliest stages of product and service design. Therefore, it gives the category of legal requirement, at the beginning of integrating the guarantees for the protection of the rights and freedoms of citizens in relation to their personal data from the first stages of the development of systems and products.

Privacy by design (hereinafter referred to as PbD) implies using a risk management-oriented approach and proactive responsibility to establish strategies that incorporate privacy protection throughout the entire life cycle of the object (either this a system, a hardware or software product, a service or a

process). The life cycle of the object is understood to be all the stages it goes through, from its conception to its withdrawal, through the phases of development, putting into production, operation, maintenance, and retirement.

Furthermore, it implies that not only the application of privacy protection measures are taken into account in the early stages of the project, but also that all the business processes and practices involved in the associated data processing are considered, thus achieving a true governance of the management of personal data by organizations.

The ultimate goal is that data protection is present from the early stages of development and is not an added layer to a product or system. Privacy must be an integral part of the nature of that product or service.

- **Prevention and anticipation approach.**

Any system, process or infrastructure that is going to use personal data must be conceived and designed from the beginning, identifying, a priori, the possible risks to the rights and freedoms of the interested parties and minimizing them so that they do not materialize in damages. A PbD policy is characterized by the adoption of proactive measures that anticipate threats, identifying weaknesses in systems to neutralize or minimize risks instead of applying corrective measures to resolve security incidents once they have occurred.

In COVID-X, this is achieved through the performance of a risk analysis that identifies the possible threats to which the data is subject, the vulnerabilities that these threats may materialize, the possible impact if these risks materialize and the mitigating measures that can be put in place to reduce these risks.

- **Privacy focus as default setting**

Privacy must be an integral part of the systems, applications, products, and services, as well as the business practices and processes of the organization. It is not an additional layer or module that is added to something pre-existing but must be integrated into the set of non-functional requirements from the moment it is conceived and designed.

To ensure that privacy is considered from the earliest stages of design, Project have to:

- Consider as a necessary requirement in the life cycle of systems and services, as well as in the design of the organization's processes.
- Carry out an analysis of the risks to the rights and freedoms of people and, where appropriate, impact assessments related to data protection, as an integral part of the design of any new treatment initiative.
- Document all the decisions that are adopted within the organization with a “privacy design thinking” approach.

In COVID-X, privacy is taken into account from the first stages, in the terms set out in WP1, and is maintained throughout the life of the project, and from the first stages a risk analysis has been carried out, with a permanent update, a privacy risk analysis.

- **Privacy assurance approach throughout the life cycle**

To integrate privacy throughout all stages of data processing, the different operations involved (collection, registration, classification, conservation, consultation, dissemination, limitation, deletion...) must be carefully analysed and implemented, in each one of these are the most appropriate measures to protect the information and among which should be considered:

- Early pseudonymisation or anonymization techniques such as k-anonymity
- Classification and organization of data and processing operations based on access profiles.
- Encryption by default so that the “natural” state of the data in case of loss or theft is “unreadable”.
- The safe and guaranteed destruction of information at the end of its life cycle.

In COVID-X these aspects are dealt with in Chapter 6 and in the specific guides, such as the Guide for data anonymization.

- **Visibility and transparency approach**

One of the keys to guaranteeing privacy is to be able to demonstrate it, verifying that the treatment is according to the information given.

Transparency in data processing is key to demonstrate diligence and proactive responsibility before the Control Authority and as a measure of confidence before the subjects whose data is processed. As established in recital 39 of the GDPR, it must be totally clear for natural persons that personal data that concerns them is being collected, used, consulted or otherwise processed, as well as the extent to which said data is or will be processed.

Promoting transparency and visibility involves adopting a series of measures such as:

- Publish the privacy and data protection policies that govern the operation of the organization. In COVID-X a Privacy Policy is created that is mandatory for all project participants (members of the consortium and participants in the open calls) and that is available for consultation by any patient through a simple and in clear language and without technicalities
- Develop and publish concise, clear and intelligible information clauses that are easily accessible and that allow interested parties to understand the scope of the processing of their data, the risks to which they may be exposed, as well as how to enforce their rights regarding data protection. All the information clauses in COVID-X have been reviewed and updated in accordance with these principles to ensure adequate, clear and simple information

- Even though it is not mandatory for all those responsible, to make public, or at least easily accessible to those interested, the list of the treatments carried out in the organization, the treatments carried out by the pilot sites are published:
 1. The list of Hospital Clínico San Carlos (HCSC) treatments published on the transparency portal of the Community of Madrid that can be consulted at the following link:
<https://www.comunidad.madrid/gobierno/informacion-juridica-legislacion/proteccion-datos>
https://www.comunidad.madrid/sites/default/files/01_10_rat_sanidad.pdf
 2. The list of Humanitas treatments published can be consulted at the following link:
<https://www.humanitas.it/covidx-trattamento-dati>
 3. The list of KI treatments published can be consulted at the following link, considering that patients should turn to IVO (Inspektionen för vård och omsorg) when having complaints and to Lof for rectification, etc.
 - <https://www.ivo.se/for-privatpersoner/informera-eller-anmala-halso-och-sjukvard/>
 - <https://lof.se/>
- Disseminate the identity and contact of the person responsible for organizing privacy matters. In the aforementioned transparency portal, the identity of the data responsible, the DPO and how to exercise the rights associated with them can be consulted.
- Establish accessible, simple, and effective communication, compensation and claim mechanisms aimed at the data subjects. There are some forms available in which users can exercise their rights, which can be accessed at the links:
 - Madrid:
<https://www.comunidad.madrid/gobierno/informacion-juridica-legislacion/proteccion-datos>
 - Italy
<https://www.humanitas.it/form-gdpr>
 - Sweden: The procedures that patients should follow to exercise their rights can be found on the webpages or by directly contacting these authorities (IVO and Lof):
 - <https://www.ivo.se/for-privatpersoner/informera-eller-anmala-halso-och-sjukvard/>
 - <https://lof.se/>

- **User-centered approach**

This approach indicates that processes, applications, products and services should be designed “with the user in mind”, anticipating their needs.

The user must have an active role in the management of their own data and in controlling the management that others do with them. User inaction should not compromise privacy, taking up one of the aforementioned principles and advocating a default privacy setting that offers the highest level of protection.

A design of processes, applications, products, and services that are focused on guaranteeing the privacy of data subjects implies:

- Implement “robust” privacy settings by default and in which users are informed of the consequences to their privacy of modifying the pre-set parameters.
- Provide complete and adequate information that leads to informed, free, specific, and unequivocal consent that must be explicit in those cases where it is required.
- Provide interested parties with access to their data and detailed information on the purposes of the treatment and the communications made.
- Implement efficient and effective mechanisms that allow interested parties to exercise their rights regarding data protection.

These aspects are dealt with in COVID-X, as explained in the previous chapters.

4.1.2 Information and transparency

The implementation of the principle of transparency established by the Regulation indicates that the interested parties are fully informed of the processing of their data in a timely manner. Whenever a treatment is carried out, the subjects whose data is processed should know what information is being processed, for what purpose and to which third parties it is communicated in addition to the rest of the information established in articles 13 and 14 of the GDPR. Transparency regarding this information becomes a basic privacy requirement since it allows interested parties to make informed decisions about the treatments carried out and to provide, where appropriate, free, specific, informed, and unequivocal consent. Any modification that occurs in the treatment with respect to the information previously provided should be communicated, including possible security gaps that may significantly affect the rights and freedoms of data subjects.

This strategy is supported by the existence of privacy clauses that facilitate the globalization of this information to the interested parties, with all the information required by the GDPR in relation to what personal data is processed, how it is processed and why by identifying the reason and purpose. Details must be provided in relation to the data retention periods, as well as the communications of these that are made to third parties. Along with all this information, which must be easily accessible and provided continuously over time to promote true transparency, it must also be indicated with whom and how data subjects can contact to raise questions regarding their privacy, as well as the rights that assist them in matters of personal data protection. In order to provide COVID-X with the maximum possible transparency and information for patients, the data collection clause has been modified, informing that the data collected will be part of the data set whose purpose is the best management of the patient’s health care. Likewise, patients are also informed that their data could be used for

research purposes always prior and suitably anonymized, emphasizing that this information does not imply express consent but is provided to respond to the principles of information and transparency of the GDPR and that these processes would never have any impact on your privacy since they would be completely anonymized, making it impossible to identify any patient. In any case, the patients have always the possibility to object and request not to be included in the research.

4.1.3 Accountability

The objective of this principle is that, in accordance with article 24 of the GDPR, the data controller can demonstrate, both to the interested parties and to the supervisory authorities, compliance with the data protection policy that is being applied, as well as of the rest of the legal requirements and obligations imposed by the Regulation. Specifically, it is about the implementation of the accountability or proactive responsibility required by the GDPR, based on a critical, continuous and traceable self-analysis of all the decisions made in the framework of the treatments and guarantee of an authentic governance of personal data within the organization. The following tactics allow you to carry out this strategy to guarantee and be able to demonstrate that the treatments are compliant with the GDPR:

- Record: document every one of the decisions made over time even when they have been contradictory, identifying who made them, when and the justification for doing so.
- Audit: To review in a systematic, independent, and documented way the degree of compliance with the data protection policy.
- Report: Documenting the results of the audits carried out and any incidents that occur in the personal data processing operations and make it available to the supervisory authority when necessary. In the case of new treatments and if the result of the impact assessment related to data protection shows that the treatment would entail a high risk for the rights and freedoms of the interested parties if the person in charge does not take measures to mitigate them, carry out prior consultation referred to in article 36 of the GDPR.

In accordance with the indications of the European Data Protection Board (EDPB; former Article 29 Working Group), the following common liability measures have been established in COVID-X, among which this principle of accountability can be materially specified:

- Internal procedure for reviewing and approving each data extraction in all pilot sites, which must also be endorsed by the Ethics Committee corresponding to each pilot site.
- Written and binding data protection policy that applies to all new data processing operations in Open Calls (e.g., compliance with data quality criteria, notification, security principles, access, etc.) , obligatory for all the agents involved.
- Exact and precise inventory of all the procedures involved in all data processing operations, including those related to access, correction, deletion, which will be published and available in accordance with the principle of transparency.
- The involvement of the DPO of each pilot site.

- Carrying out training activities on data protection for the members of the COVID-X project consortium
- The establishment of an internal mechanism for the resolution of complaints from interested parties, in which the DPO is directly involved.
- The establishment of internal procedures for effective management and notification of security breaches (security breaches).
- Conducting data protection impact assessments (PIA) at the beginning of the project and in specific circumstances.
- The application and supervision of verification procedures that guarantee that the measures are not only nominal, but that they are applied and work in practice (internal or external audits).

4.1.4 Minimization

The objective of this principle is to collect and process the minimum amount of data possible, so that, avoiding the processing of data that is not necessary for the purposes pursued in the treatment, the possible impacts on privacy are limited. This can be achieved by collecting data from fewer subjects (reducing the size of the study population) or less data from subjects (reducing the volume of information collected) for which the following tactics have been used:

- **Select:** Choose only the relevant sample of individuals and the necessary attributes following a conservative attitude when establishing the selection criteria and perform the treatment only on the data that respond to this criterion (white list). In COVID.X this strategy is contemplated using the catalogue of variables that limit exactly what type of information the researcher is looking for, preventing her from accessing the entire Data Lake.
- **Exclude:** This is the inverse approach to the previous one and consists in excluding beforehand the subjects and attributes that are irrelevant to the treatment performed (blacklist). In this case, an open attitude should be adopted, trying to exclude as many records as possible unless it can be justified that they are necessary for the intended purpose. In COVID-X, the filters that are applied at the beginning of the pseudo-anonymization process are used for this purpose.
- **Pruning:** Partially deleting personal data as soon as it is no longer necessary, which means determining in advance what the retention period is for each of the data collected and establishing automatic erasure mechanisms when said period is reached. If the data is part of a record containing more information that needs to be kept, the value of the unnecessary fields can be changed to a pre-set default value.
- **Delete:** completely delete personal data as soon as it is no longer relevant, ensuring that it is not possible to recover even from backup copies made. It is also necessary to take into account that only the strictly necessary data should be communicated and shared and that, in the case of treatments that infer new personal information, those data that are generated and are not necessary for the intended purpose should also be selected for exclusion.

4.1.5 Data Quality

The data origin is the first aspect that must be considered in the chain of treatments contemplated in a massive data processing system such as COVID-X.

An important part of the complexity of the analysis of these treatments will occur in those cases in which the Data Lake is loaded with information from multiple sources because depending on the level of reliability offered by the different data sources, the quality of the primary data may remain compromised from start-up and crawl throughout its life cycle.

This is where the classification of sources into endogenous (data sources internal to the organization itself) and exogenous (external sources) produces the need to apply filters and compensatory controls to a greater or lesser extent.

Beyond the quality of the primary data, the integration of the same from different origins, or sources of origin, is not always easy despite the application of sophisticated debugging techniques often supported by data dictionaries. This means that other supplementary data must be used to help ensure the reliability of the embedded data.

Actions related to data quality are included in the Anonymization Guide

4.1.6 Anonymization/dissociation

The GDPR considers personal data to be any information concerning identified or identifiable natural persons. Identifiability, which implies the application of the regulations, refers to the fact that a person can be identified by a piece of information or by the combination of information from different sources. More precisely, to determine whether a person is identifiable, "all means that can reasonably be used and without disproportionate effort" must be used. The analysis of identifiability must be based on two criteria, that of reasonableness in the availability of the means (technical, human and data sources) and on the proportionality of efforts to be able to directly or indirectly identify the natural person by the responsible for the treatment or by any other person.

The anonymization strategy focuses on limiting the exposure of the data, establishing the necessary measures to guarantee the protection of the objectives of confidentiality and unlinking. To respond to this strategy, the following tactics are useful:

- Restrict: restrictively manage access to personal data by limiting it through an access control policy that implements the principle of "need to know" both in space (detail and type of data accessed) and in time (processing stages). This measure is detailed in the access policies of all pilot sites.
- Obfuscate: make personal data unintelligible to those who are not authorized to consult it using encryption and hashing techniques, both in information storage and transmission operations. This measure is detailed in the Anonymization guide described in later chapters.

- Dissociate: eliminate the link between data sets that must be kept independent, as well as the identifying attributes of the data records to avoid correlations between them, with special attention to metadata. This average has a strong impact on the usefulness of the resulting data after applying it. The ability to establish associations between search variables is precisely one of the tools to obtain innovative results in scientific research, therefore the Anonymization Guide describes the techniques to apply to each study with the least possible impact.
- Add: Group the information related to several subjects using generalization and suppression techniques to avoid correlations. In COVID-X, anonymization techniques are used that contemplate this strategy, applying it as mentioned, in the way that produces the least possible impact.

Another possible approach related to the anonymization of the data in the strategy of avoiding, or at least minimizing, the risk that, during the processing, in the same entity, of different personal data belonging to the same individual and used in independent treatments, a complete outline of the subject can be made. For this, it is necessary to maintain independent treatment contexts that make it difficult to correlate groups of data that should be unlinked. The following tactics help implement the separation strategy:

- Isolate: collect and store personal data in different databases or applications that are logically independent or even run on different physical systems, adopting additional measures to guarantee such unlinking, such as the scheduled deletion of indexing tables between databases. In COVID-X the results of the studies are delivered to each researcher without there being any connection between them. Additionally, one of the anonymization processes is carried out with this objective.
- Distribute: disseminate the collection and treatment of the different subsets of personal data corresponding to different types of treatment on processing and management units that, within the organization, are physically independent and use different systems and applications trying to implement decentralized and distributed architectures with local information processing whenever possible instead of centralized solutions with unified access and that depend on the same control unit. This tactic is not applicable to research projects in which, logically, it is desired to have as much information as possible to delve into the characteristics of diseases and how different treatments affect individuals.

This content is developed in detail in the Anonymization Guide

Another strategy associated with data anonymization is to abstract as much as possible. The objective of this strategy is to limit as much as possible the detail of the personal data that is processed. Unlike the 'minimize' strategy that makes a prior selection of the collected data, this strategy focuses on the degree of detail with which the data is processed and its aggregation using three tactics:

- Summarize: generalizes attribute values using ranges or ranges of values, instead of using the specific value of the field.

- Group: add the information of a group of records into categories instead of using the detailed information of each one of the subjects that belong to the group working with the mean or general values.
- Disturb: use approximate values or modify the real data by using random noise instead of working with the exact value of the personal data.

In each treatment it is necessary to study how the degree of detail of the input data affects the result of this, and what is the precision necessary for the treatment to be effective. In particular, the time that has elapsed since the collection of the data may affect their relevance, so it is advisable to periodically review the information stored and apply this type of strategy.

In COVID-X this strategy is contemplated in the project's anonymization guide, specifically in the use of K anonymization.

4.1.7 International transfers

This assumption does not apply to the COVID-X Project, since all the data remains in their locations of origin and will be consulted from other locations through the Project tools, without any data transfer.

4.2 Key ethical aspects

Later paragraphs outline the ethical-related principles and procedures to be established and followed in the project. These principles will be combined with national policies and regulations regarding ethics to ensure that all pilots are run in an ethically appropriate manner.

Each pilot site manager will be responsible for establishing and implementing all ethical procedures according to the NATIONAL ETHICAL MANAGEMENT PRACTICES, that are specific (but aligned with the general guidelines) and relevant to the respective pilot site (ensure conditions for auditing from relevant authorities, drafting of material necessary for obtaining validations, drafting of informed consent forms if required, etc).

4.2.1 General

In the data collection process, major ethical issues will concern autonomy and informed consent, anonymization, and security. One important research issue is under what conditions informed consent is ethically required for automatic processing of patient records. In Sweden, register studies, i.e. research studies based on information from large databases, are usually exempted from the requirement of informed consent, provided that the research is performed on extracted anonymized data. Whether some forms of automatized processing of patient records should be subject to a similar

exemption needs to be investigated. For this purpose, comparisons with current practices concerning consent requirements in both research and clinical practice will be useful.

In the automatic interpretation of data, major issues will be the risk of misinterpretations and false alarms. Experiences from other uses of AI indicate that the potential for various forms of bias needs to be carefully investigated. The risk of false conclusions due to mass significance problems will have to be dealt with in part with statistical precautions, in part with human judgment based on mechanistic understanding and clinical experience.

In the phase of applying the outcomes of data processing in clinical and public health decision-making, issues of responsibility and accountability will arise. Much will depend on how the human assessment of the machine-generated information is organized and performed. If insufficiently understood correlations (“black box outcomes”) are used for critical decisions concerning individual patients, then the physician–patient relation may be affected in ways that can have ethical repercussions.

To guarantee the physical security of the user's personal data (in clinical records and others), in the process of acquiring different sources during the different phases until the total anonymity of the data in all pilots, it must be remembered that

Personal data can only be processed (e.g. collected and further used) if:

- The data subject has unambiguously given his or her consent, i.e. if he or she has agreed freely and specifically after being adequately informed.
- Data processing is necessary for the performance of a contract involving the data subject or in order to enter into a contract requested by the data subject, e.g. processing of data for billing purposes or processing of data relating to an applicant for a job or for a loan;
- Processing is required by a legal obligation (medical institutions).
- Processing of data is necessary to protect an interest that is essential for the data subject's life (medical institutions).
- Processing is necessary to perform tasks of public interests or tasks carried out by official authorities (such as the government in public health or similar situations).
- Finally, data can be processed whenever the controller or a third party has a legitimate interest in doing so. However, this interest cannot override the interests or fundamental rights of the data subject, particularly the right to privacy or ethical principles.

This provision establishes the need to strike a reasonable balance, in practice, between the legitimate interest of the data controllers and the privacy of data subjects. This balance is first evaluated by the data controllers under the supervision of the data protection authorities, although if required, the courts have the final decision. In order to guarantee privacy and anonymity of research data sets according to local and European ethical regulations, COVIDX must be prepared to pass audits from local authorities (this will be an objective from the second year).

4.2.2 Consent vs anonymization

The art. 89.2 GDPR allows, when personal data is processed for scientific research purposes, member states to establish exceptions to rights (except to suppression) “whenever it is probable that these rights make it impossible or seriously impede the achievement of scientific purposes and when those exceptions are necessary to achieve those ends.

This is a principle that it can use to process data for which it is not necessary to have an informed and express consent. This is usually enough in typical Research studies that take place in a closed environment and do not have any advertising in the media. However, this will not happen in COVID-X, because patients may want their data withdrawn from any study where they have not given their express consent. Why do we think this can happen? Criticisms of private companies that are part of the "health business" are frequent in the media. It is possible that citizens know that large pharmaceutical companies (for example) are paying money to experiment with their data and making money from the results of those studies, they decide to exercise their right to have their data removed from COVID-X.

Therefore, in COVID-X had been defined the scenarios for:

- If the data has been irreversibly anonymized, it cannot be identified for extraction. Additionally, the normative allows what has been used up to the moment of the negative, to remain in the studio.
- The data that is in the main repository, before the anonymization processes, could be necessary to extract them for the citizens who request it using the filter mechanism of the first part of the pseudo-anonymization process.

It is necessary to refer to the fact that personal data may not be used for purposes incompatible with those for which the data had been collected, which does not mean that they cannot be used for purposes other than those for which they were collected, but rather that these should not be incompatible. The Article 29 Working Group describes in detail in its Opinion on the aforementioned principle of purpose in which cases we are faced with incompatible purposes.

The non-incompatibility analysis is basic in Big Data and applies to COVID-X, given that to a large extent, it bases its analytics on subsequent treatment with purposes other than the original purpose. The Article 29 Working Group has analysed this aspect in its Opinion (WP 203). In this regard, to know if subsequent uses of personal data are compatible, the Opinion establishes the following criteria:

- There must be a relationship between the original purpose and the subsequent purpose or purposes. The purpose of COVID-X is always the scientist investigation
- Further processing must be within the reasonable expectations of the interested party. Always will be the improvement of health.
- The nature of the data being processed, and their sensitivity must be considered. The anonymization process is the proof

- The impact that this treatment will have on the interested parties should be considered. In COVID-X will be possible to apply the patterns of investigation in the original data when this could be interesting for the patients
- The protection measures established by the data controller must be considered, technical and organizational measures: encryption, pseudonymization, functional separation, transparency, opposition to treatment. All these measures are described in this document.

Article 6.4 of the GDPR has incorporated these criteria into the standard, making explicit mention of encryption and pseudonymisation within the appropriate guarantees, with which they will be directly applicable to the compatibility analysis of all subsequent treatments that are not based on consent of the interested party and, therefore, must be considered in the impact assessment carried out by the person responsible for these treatments.

Provided that some pilot sites must necessarily enable access to longitudinal retrospective datasets via the Sandbox, the Partners, by mutual consent, consider that for such pilot sites adopting a retrospective approach, the access via sandbox should be granted solely to data fully and irreversibly rendered anonymous.

Therefore, retrospective datasets, once internally selected by the project team, shall be subject to a suitable anonymisation procedure at the end of which it will be impossible to trace back, also for the project team and indirectly, to the data subjects' identities. The anonymisation procedure will be irreversible.

In addition, for such retrospective studies, the "date of the study" will be considered as the date when the database containing the data to be anonymized and further analyzed is frozen, so no new data is added, and not the date when the COVIDx study started (i.e. November 1st, 2020).

According to the opinion no. 5/2014 of the EDPB, the anonymisation is the processing of the personal data with the aim of irreversibly preventing the identification of a data subject. Different anonymisation techniques may be put in place, as there is no prescriptive rule in this regard in the EU legislation. In particular, the anonymisation techniques are broken down into three groups: randomization, generalization, and differential privacy.

However, none of the above mentioned techniques may clearly and absolutely satisfy the criteria of an actual anonymisation (i.e., it is impossible to single out a person, to link data concerning a person and to infer any information about individuals). Therefore, it is essential for Partners to assess, on a case-by-case basis, the existing risks, by totally or partially applying one or more techniques, combined with a differential privacy approach.

According to the aforementioned opinion of the EDPB, the adoption of randomisation, aggregation and noise addition techniques may reduce the risk of singling out, linkability, and inference – i.e., the three essential risks related to the anonymisation process. In fact, it seems reasonable to consider that, by adopting these techniques, the two elements of singling out and linkability are essentially absent,

while the inference risk cannot be excluded, at least abstractly (as Partners would keep original data complete in their archives, together with additional data concerning persons involved in the project). However, if there is the remote possibility of inferring an attribute value from the values of a set of other attributes, such reconstruction would be related to the features of a group of persons and would result from cross-referencing pseudonymised data with other non-encrypted data available in other systems and archives. Although this activity could be abstractly exercised by specialists inside the facility of the pilot sites, it would require a significant effort to proceed with singling out by inference. The chance of success of such inference attacks would also be reduced thanks to the random date shifting and noise addition, which in the re-identification process would lead to the occurrence of some false positives (and false negatives) that would further make it impossible to identify the data subjects. In fact, albeit not producing absolutely anonymous data, a process for the identification of data subjects would be neither immediate nor appealing.

Specifically, the Partners will assess which anonymization technique is more appropriate according to individual needs (e.g. processing of images, processing of medical reports, etc.), and will document in writing from time to time the criteria applied.

Below are some possible examples of how it may opt for different solutions, according to individual needs and the instructions of the Anonymization Guide.

Accordingly, the pilot sites adopting a retrospective approach are not required to comply with the information requirements and to collect the consent to access their patients' data gathered in the past.

4.2.3 Unexpected information to patients

If patient identity is retrievable from the data, it may be possible to provide feedback information affecting diagnosis and treatment. It will also be possible to perform systematic automatic searches for instance for combinations of symptoms or laboratory findings that have clinical significance but are sometimes missed in the clinic. (One example would be unusual but serious side effects of a drug.) On the other hand, if patient identity is not retrievable from the data, then such feedback will not be possible.

If clinically relevant information for identifiable individuals is possible to feedback from the research, the ethical question regarding whether or not it should be fed back (i.e. whether or not these individual or their doctors should be contacted with the information) arises. This question arises when the information is of relevance to research persons'/patients' health and actionable (i.e. when something that increases or reduces the risk to patients' health can be done). If so, then there are harm-related reasons to feed back the information, since the research persons'/patients' health may be worse off if not informed. On the other hand, if such information is fed back, the researchers may bring unwelcome health information to the research persons/patients. Moreover, even if such information is in principle actionable, there may be risks with false positives (which may lead to unnecessary anxiety and undertreatment) or false negatives (which may lead to false reassurance and undertreatment). Furthermore, by bringing feedback for a certain health problem to those affected is, in effect, to implement screening for that health problem. Usually, countries (including Sweden) have strict societal procedures for implementing screening. There is also the risk of blurring the line between research

and therapy for researchers and research persons alike, fueling the so called therapeutic misconception.

Therefore, the project needs strong health related reasons to decide to feedback health related information for therapeutic purposes. If this possibility is not ruled out, one also need to ponder whether this possibility should be part of an informed consent.

To respond to the exercise of rights to massive data processing, a very important aspect is to resolve the question about future uses and processing that were not foreseen at the time of obtaining the data.

In addition to solving the problem of how to be able to keep the affected person permanently informed about the new uses and purposes of the treatment and how to obtain the new consents or extension of the previous ones, with regard to the exercise of rights, the focus of attention should focus on the way of providing permanent information to those affected about the way and procedure by which they can exercise their rights, as well as the way in which, once any of them have been exercised, it must be attended by the person in charge or in charge of treatment .

Naturally, the consortium adheres to strict security policies in order to protect the medical data that is acquired, stored and processed in the framework of the Covid-X project. These security policies include the provision of anonymization guidelines, the acceptance of anonymized data only and the deployment of a Sandbox -the Covid-X Sandbox- based on a privacy-by-design principle.

Since the possibility of re-identifying anonymized data, while negligible, still exists, a privacy-by-design principle is vital when designing a system for the storing and processing of medical data. This principle entails authentication and authorization mechanisms, Role Based Access Control (RBAC), encrypted data transfer, monitoring and auditable trace of activities. Moreover, the said principle indicates the utilization of up-to-date, open source, virtualization tools, in order to enhance the system's robustness with respect to errors and malicious acts.

If, despite all these security measures, a malicious actor manages to get access to the Covid-X Sandbox, she/he will face the additional hindrance of anonymization.

The dissociation of personal data in COVID-X should not be an excuse and an obstacle to exercise rights. Therefore, in COVID-X all pilot sites are in a position to inform the affected party about the fact of anonymization, if it has occurred, and the risk of re-identification existing in the case of exercising a right of access. For the rest of the rights, there must be the possibility of re-identification by the person in charge or in charge of the treatment at the time of attending to their exercise by the affected party.

Although it is true that the data protection regulations will only be applicable if the re-identification of the interested party is possible, this does not exonerate the duty to respond to the request warning of the use of anonymized data and, where appropriate, of the residual risk of re-identification.

4.2.4 Mandatory ethical and legal normative from the Open Calls Countries

In similar way that the ones already involved in the project (Italy, Sweden, Spain), will be necessary that other healthcare centres from other EU countries than may participate in the Open Calls (Team Solution Call), will be responsible for following their national and local regulations regarding data security and ethic management of clinical information.

In addition, according to art. 32 of the GDPR, these companies must ensure that the providers of said services have taken the necessary actions to guarantee the security of the personal data they process.

The concept of data responsible and responsible for processing personal data is crucial in the context of data processing and the same happens in massive data exploitation projects, since the processing and analysis of the data are often outsourced. The legislation establishes specific norms for those cases in which the data is processed by different actors. In this sense, the GDPR establishes that the person in charge is the “natural or legal person, public authority, service or other body, which determines the purposes and means of the treatment, and also defines the person in charge as a natural or legal person, public authority, service or another body that processes personal data on behalf of the responsible for processing personal data”.

The person responsible for the treatment sets the purpose of the treatment and decides on the outsourcing of the same and to what degree he delegates the treatment activities to another organization. In addition, it will only choose a responsible for processing personal data who offers sufficient guarantees to apply appropriate technical and organizational measures so that the treatment is in accordance with the GDPR.

Opinion 1/2010 of the GT29 (Working Group of Article 29), establishes that, in order to act as data processor, two circumstances must exist: that it is an entity independent of the person in charge and, second, that the data is processed on his behalf. In addition, you can also carry out specific activities on the treatment, with autonomy to determine which technical means are the most appropriate

In COVID-X, third companies or investigating entities do not act as data controllers since the data that these entities receive is already anonymized.

However, the partners that are part of the COVID-X consortium, if they act as data managers since their tools will act in Data Lake where the data is pseudo-anonymized, therefore still being personal data

To respond to the responsibilities derived from this situation, agreements have been signed by all the partners.

4.2.5 Traceability and monitoring

Traceability is in this context the antonym of anonymity. As indicated in Section 5.3, information feedback is only possible if the data is non-anonymous or can be deanonymized. We therefore have an ethical dilemma that can be described on the individual level as a dilemma between privacy and the possibility of medically helpful feedback, and on the aggregated level as a dilemma between anonymity and traceability. We will perform a careful ethical analysis of this dilemma, including the possibility of an individual choice (either opt-in or opt-out) whether one's data will be anonymized.

4.2.6 Consent (children included)

To the extent that informed consent will be required for automatic processing of data, several ethical (and practical) problems will arise. In other contexts, an informed consent is usually considered to be retractable, which means that a person can withdraw her consent. In research ethics, such a withdrawal is considered to require that the withdrawn data be removed from already made analyses. However, such retroactive withdrawal may be difficult to implement in this case.

There are patients who for various reasons cannot provide informed consent. This applies for instance to temporarily unconscious persons, children, and people with dementia and some other mental conditions. Excluding these groups from the data analysis could be detrimental to quality improvement of the treatment of diseases affecting them. Proxy consent is a common solution to this problem. In the case of temporarily unconscious persons, a solution is available here that is usually not available in clinical decisions, namely to postpone the decision until the patient has regained her normal mental powers. There is a need for a separate discussion of informed consent issues in relation to data analysis. Its result will be informed by, but not identical to, the outcomes of similar discussions concerning informed consent for clinical interventions and research participation.

4.3 Key aspects for data uploads

The process of data uploads of the COVID-X project is the first step where anonymization techniques are applied so that from the beginning the possible risk of identification begins to be reduced when the data has not yet been anonymized.

As shown in the Anonymization Guide document, one of the techniques is the elimination of identification variables, avoiding that from the beginning they are no longer part of the data set and therefore it is not necessary to anonymize them, directly eliminating the risk associated with them. An example of this type of variable is the patient's name and surname.

Consult the Anonymization Guide for details.

4.4 Key aspects for Storage and Securitisation

The Covid-X consortium realizes the crucial aspects associated with the storage and treatment of medical data, as well as the extensive need to design an inherently secure system.

In order for the medical data ingested in the Covid-X Sandbox to be properly stored, the following core aspects of security need to be taken under consideration: confidentiality, integrity and availability.

In the following Table 2, a mapping between the CIA principles and the security components of the Covid-X Sandbox is shown, along with a short description of each component's role within the Covid-X Sandbox.

TABLE 2: MAPPING BETWEEN THE CIA PRINCIPLES AND THE SECURITY COMPONENTS OF THE COVID-X SANDBOX

Components	Description	CIA
Authentication	Verification that the user or entity is who they claim to be.	Confidentiality
Authorization	Once authenticated, what data can the user/entity access and what operations can they perform.	Confidentiality, Integrity
Log management and auditing	Trail of who accessed what and when, for compliance and detection.	Confidentiality, Integrity
Encryption	Transform data in-transit so that it is inaccessible to unauthorized users.	Confidentiality, Integrity
Clustering	Group of independent servers connected via network share computing tasks.	Availability
Guaranteed Delivery	Once a request to the cluster is made, one of the available nodes will deliver in case of failure.	Integrity, Availability
User Certificate	A User Certificate represents the user's digital identity, so any actions	Integrity

	performed by the user can be traced back.	
--	---	--

These components are used to cover every possible vulnerability of the system, enhancing its resistance against malicious attacks; authenticated access and encrypted transfer prevent a malicious act from occurring, authorization enhances security in case authentication was forged. Moreover, monitoring and audit logging increase transparency and give insight in the level of the undergoing risk and in the appropriate recovery process, alleviating the aftermath of the attack. Furthermore, encrypted communications between the end users and the application and between its internal components are used to maintain data integrity over its entire life-cycle and to restrict access to unauthorized users. Encrypted communications are achieved through encryption protocols, such as HTTPs, and secure connections via VPN (Virtual Private Network). Adding to the above, clustering technologies, on which distributed applications are based, support server applications that can be reliably utilized with a minimum amount of down-time and promote high availability for its users. In addition, they ensure fault tolerance, since requests that are sent to the cluster are not managed by only one node. Therefore, in case of a failure, it is guaranteed that the requests will be handled by any available node without any compromise. A technology used for authenticating users is a Public Key Infrastructure (PKI). Each user carries a cryptographic key that can be used as an identity in digital networks. Any request made by the user can be associated with that key which can be represented as a document. Responsible for signing that document is the System Owner that acts as a trusted party and signs that document using their own cryptographic key, serving as a Certificate Authority. The signed document is the user's digital signature and serves as a certificate that accompanies each request he/she makes and guarantees transparent communication, since no user is able to deny any request containing his/her signature.

The API (Application Programming Interfaces) security scheme for the Covid-X sandbox will be enabled and managed through Kong gateway. In this regard, different options for API authentication will be analysed and then selected according to the system requirements such as Key Authentication (through API keys) and OAuth2.0. On the other hand, a SSL client-server communication protocol via the gateway will be also enabled to secure the information exchange.

4.5 Personal data security inside COVIX

This point will be developed in WP3.



BLOCK 3 (HOW WE DO IT, MEDICAL GUIDELINES)

5 TREATMENT GUIDELINES (Data security protocol)

5.1 Data Protection Policy

In order to ensure that the processing of personal data are correct and respect the legal requirements and obligations imposed by the regulations, it is necessary to define a privacy framework and a governance structure that includes a data protection policy supported by senior management, as well as the roles and responsibilities that ensure its fulfilment. The culture of privacy is an essential part of the COVID-X project, both partners and other participants must know and abide by it.

With this objective in COVID -X, a training and awareness plan has been developed for all members of the project, aimed at producing a committed and responsible attitude as part of proactive responsibility.

The security policy is supported by procedures and guides that detail the implementation of the necessary technical and organizational measures.

Among these procedures are those related to guaranteeing the exercise of rights, the management and notification of security incidents, the adaptation of possible treatment orders to the legal requirements, and the accreditation of compliance with the obligations imposed by the current regulations.

- Defend: ensure compliance, effectiveness and efficiency of the privacy policy and of the procedures, measures and controls implemented to verify that they respond at all times to the performance of the treatment activities and the day-to-day running of the organization.

Consult the *COVID-X Information Security Policy* for details.

5.2 Risk Assessment

Carrying out a risk analysis and PIA together with the documentation of the decisions made based on the results obtained are the starting point for, to setting the privacy requirements that are

implemented in COVID- X in applications and systems as part of privacy by design, fully document how personal data processing is carried out and comply with the principle of proactive responsibility.

In order to evaluate this suitability, not only have the risks associated with destruction, loss or alteration, or unauthorized communication or access, been taken into account in particular, but following the guidelines of the National Security Scheme (Spanish regulations), they have been grouped around the Confidentiality, Integrity and Availability dimensions.

This content is developed in detail in the *Anonymization Guide* document.

5.3 Impact assessment

The Data Protection Impact Assessment (PIA) is a process that has to allow companies and administrations to determine whether the initiatives that involve the use of private information involve risks for the right to data protection and, As an added value, it allows them to measure, quantify said risks and assess the impact they have on the rights and freedoms of the people whose personal data they process.

For its part, the European Data Protection Regulation formulates a compliance model based on risk-focused management so that the PIA becomes a key tool to guarantee the privacy of products and services, since it serves to justify and Correctly evaluate the decisions that are made and that imply the performance of any kind of treatment.

Specifically, the GDPR establishes the obligation to carry out the PIA in the circumstance of the large-scale treatment of sensitive data, that is, those referred to in article 9 of the GDPR: those that reveal ethnic or racial origin, political opinions, religious or philosophical convictions, trade union membership, processing of genetic data, biometric data aimed at uniquely identifying a person and data related to the health, sexual life or sexual orientation of a person.

The PIA proves that COVID-X wants to achieve the specific and legitimate objective that the risk to privacy, once the necessary security measures have been adopted, is residual.

In accordance with the provisions of article 35 of the GDPR, the minimum content included in the Impact Assessment of the COVID-X project is as follows:

- Systematic description of the treatment / s considered. Typology of processed data, supports for processing, expected duration, the information technologies involved, and their functional aspects, the flow of data, the recipients and a detailed description of the life cycle in the processing of the data.

- Identification and assessment of risks. Identification of the sources of risks, the possible scenarios of unwanted events and the possible threats to which they are subject, the vulnerabilities that may facilitate the materialization of said threats, as well as the impacts or possible consequences on the private lives of the affected persons. The risks to be assessed are classified, depending on whether they affect the people whose data is processed or if they affect the COVID-X organization, which deals with said risks. The estimation of the risks has finally been carried out in terms of severity of the impacts and probability of occurrence of the risks, considering in this assessment the weighting of the risk by the existing security measures.
- Management of the evaluated Risks and selection of security measures that allow reducing the risks or their final impact. The PIA includes a description of both organizational and logical and physical security measures, and their influence on reducing the probability and consequences.
- Analysis of regulatory compliance. Once the risks, the planned measures and the impact of regulatory compliance have been assessed, COVID-X has ensured that the treatments meet the established legal requirements
- Final report and conclusion. After all the previous stages, the PIA describes the recommendations with the measures that have been adopted for the elimination, mitigation, transfer or acceptance of the risks to privacy.
- Implementation of the recommendations. To ensure the effective implementation of the measures identified by the impact evaluation, the necessary resources have been assigned in COVID-X and the monitoring of all the phases of the process and the correct implementation of the measures in relation to the objectives established for address privacy risks.
- Review and feedback. The organization of the COVID-X consortium has established a supervision and review plan that allows auditing the results of the impact assessment and the measures taken to apply them, including these reviews as one more element in its business management.

These reviews will be carried out periodically, and whenever new risks appear or when the treatment conditions are substantially modified, either due to the appearance of new technologies, new affected parties, new data, etc.

This content is developed in detail in the *Anonymization Guide* document.

5.4 Anonymization guide considerations

To guarantee the irreversibility of the anonymization processes of the personal data used in COVID-X, in this way, the non-application of data protection regulations, both the available information sources and the available technology have been considered, not only by the person responsible for the treatment but by any other person.

As indicated by the GT29, for there to be a true anonymization of personal data, it must be irreversible, that is, it must not reasonably allow the identification of the owner of the personal data, although it is

necessary to assess the risks derived from the techniques implemented for such anonymization, since given the current state of technology, situations may arise that, although a priori do not seem to allow reversibility, through the application of certain technologies, the interested party could be identified. COVID-X has provided a validation environment in which to simulate singularization and other attacks to verify the strength of the anonymization processes.

In this environment and as indicated by the WG29, an evaluation of the anonymization techniques and procedures used has been carried out to prove that the disassociation carried out prevents:

- A natural person can be identified within a data set
- The information of a natural person can be related or linked from the linking of two records within a data set (or between two independent data sets)
- Any information about the natural person can be inferred from a data set.

In COVID-X, anonymization is not considered an isolated exercise, but the existing risks are regularly reassessed.

The data protection authorities indicate, regarding the selection of anonymization techniques, the usefulness of encryption algorithms for this type of process, highlighting the “hash” algorithms as a formula to guarantee the confidentiality of the data because it is a one-way operation, that is, starting from a piece of data we can always generate the same fingerprint, but starting from a fingerprint we can never obtain the original data. However, it clarifies that a hashing mechanism does not by itself guarantee the irreversibility of the data, it must be combined with other measures such as the application of encryption algorithms, the use of time stamps, or the application of anonymization layers, being the latter is the strategy selected in COVID-X as the best means for this measure.

To determine which anonymization technique should be applied in COVID-X, we started from the purpose sought in relation to the anonymization process of the data, initially carrying out the corresponding risk analysis of the anonymization process to later manage the resulting risks with technical, organizational or any other measures.

It is convenient at this point to remember that anonymization techniques do not guarantee in absolute terms the impossibility of re-identification, so there will always be an index of probability of re-identification that has been tried to mitigate through the corresponding risk management and in the PIA carried out in COVID- X.

The anonymization processes have been focused from the concept of Data Protection from Design, which has implied that the privacy requirements have been taken into account from the initial stages of the COVID-X design, for the anonymization process and during entire life cycle.

The anonymization guide contains the details of the anonymization policy, which is documented and updated, so that it justifiably reflects the actions carried out to protect the privacy of the interested parties and is accessible to the personnel involved in the processing of anonymized data; as well as an action protocol for the anonymization process that has contemplated, the following elements:

- Identification of assets involved in the anonymization process.
- Assigned work team and segregation of functions, attending to profiles and roles in relation to the anonymization process, in line with the principle of professional independence.
- Carrying out a PIA (identification of risks, assessment of existing risks, safeguards aimed at preventing risks from materializing, quantification of the impact of the possible materialization of risks, report of the resulting risks, determination of the acceptable risk threshold, management of assumable risks, final report of existing risks and measures to be implemented to minimize their impact.
- Risk review for cases of changes in the anonymization processes and periodic re-evaluations of the existing residual risk to introduce parameters to improve the quality of the anonymization processes
- Training and information for the personnel involved in the anonymization processes with respect to compliance with the regulations for the protection of personal data, especially in relation to technical and organizational security measures, the existence and application of an anonymization policy, measures control of personnel with access to anonymized information, obligations and duties in the event of a break in the anonymization chain and the actions that must be taken to mitigate the impact resulting from the materialization of any of the re-identification risks.
- Elimination or reduction of variables that allow the identification of the people whose data is processed
- Audit of the anonymization process and subsequent use of the data.

The anonymization policy, the action protocol and the technological measures adopted regarding the anonymization procedures have been reinforced in COVID-X with the necessary legal guarantees to preserve the rights of the interested parties, such as:

- Confidentiality agreements and contractual clauses that guarantee the privacy of the information even when there are re-identification gaps.
- Commitments to maintain the anonymization of the information signed with the recipients of the studies about not taking any action to re-identify it.
- Audits of the use of anonymized information.

These guarantees are considered as part of the safeguards adopted to minimize the damages in the event of a possible re-identification of the interested parties.

In the implementation of all these measures, a pilot project has been selected with a small sample of test data (not real) from which conclusions regarding the feasibility of the proposed anonymization techniques and the anonymization procedure.

This content is developed in detail in the *Anonymization Guide document*.

Singularization prevention and measures to minimise the risk of their stigmatisation

Given that the project only processes anonymized data, the risk of re-identification and therefore of stigmatization cannot occur, since the person who suffers a certain pathology that could stigmatize them can never be identified.

However, additionally, in the extraction processes, a control is recommended in which when a number of cases is less than 10, the number 10 always appears so that a specific individual cannot be singled out.

5.5 Security of the architecture design

The safety of anonymized data is complemented by technologies that promote a secure architecture. In order to host such sensitive data, adopting the right framework can help mitigate the risks of data leaks or intruders and reduce the amount of weaknesses in our system. Behind the Sandbox application lie control systems, processes, layers, and structures that must work together in order for the result to be a secure application. Since all components are interconnected, in case of an emerging threat, all components are at risk. In order to avoid such circumstances, creating a controlled environment where each application is isolated is the way to go.

The Covid-X Sandbox incorporates virtualization technologies suitable for isolating its components such as the Docker containers. Being able to install a container to different machines lifts off hardware restrictions and allows fault management. There are resource constraints for both system Central Processing Unit (CPU) and memory usage, so a faulty task would be contained without affecting the rest of the application. Moreover, it promotes system availability, since the Covid-X Sandbox will not be tied to one node. Furthermore, it avoids excessive use of system resources that could otherwise affect the system uptime. Complementary to the above, a role-based access system that sits on top, increases the level of security by managing users accessing the Covid-X Sandbox and interacting with its components. Activity monitoring modules are also in place to manage traffic and events happening inside the cluster of containers, allowing for transparent interconnections within the container ecosystem.

Containerized environments are best managed under the umbrella of a hypervisor technology that is used for orchestrating many microservices with a well defined set of rules. That technology is Kubernetes and it entails an RBAC system following the same principles as containers. Access to the orchestrator API is only granted to specific users that have the right permissions and credentials. In addition, the cluster of containers has its unique set of rules, which regulate the container behaviour inside and outside the cluster. Connections between containers are achieved through safe network protocols that involve certificates, authentication and authorization tokens. A certification authority (CA) is in charge of signing these certificates in order to establish a secure playground for the cluster. In addition, network policies provide a way to control network traffic specifying how various network 'entities' communicate with each other. A network entity can be referred to as a network service that can be used to expose our application to other containers. Network policies also control the amount of load that these network services handle and promote load balancing. What is more, handling events inside the cluster is complemented by a monitoring tool which also measures containers' performance, such as the amount of resources that is being used. With respect to user experience, orchestrators are fault tolerant in the sense that they allow the system to continue functioning properly even in the event of failure of some of its components.

Concerning the communication between the Covid-X Sandbox components and its end users, authentication and authorization principles are used, as well as encryption protocols. These regulate both internal and external connections preventing malicious attacks. An additional security measure is activity monitoring which is supported by the Covid-X Sandbox audit trailing capabilities. All interactions between the Covid-X Sandbox components are audited and filtered through an alert system that indicates when an abnormal behaviour occurs in our system. A SIEM tool is responsible for handling events with a defined set of rules that can also interact with a visualization module, e.g. a dashboard.

Covid-X Sandbox consists of many applications which require it to be constantly up-to-date. Older versions of software are prone to security abuses that subsequently lower the Covid-X Sandbox defenses. It is important to address security on the delivery of Covid-X Sandbox software. One of the best practices is the CI/CD pipeline. The first concept refers to Continuous Integration and the second one is the Continuous Delivery. Likewise, only the right users should be allowed to access the process of the pipeline. Defining and separating the duties for each process is another good practise. In addition, maintaining the security of code by code reviewing and unit testing not only promotes scalability but also builds the fundamentals for the security of Covid-X Sandbox.

Users interacting with Covid-X Sandbox should also be considered as part of the system. Therefore it is required that everyone receives proper training when using the Covid-X Sandbox. In addition, each user has limited access to the Covid-X Sandbox avoiding any risks of privilege escalation within the system, which attributes to insider threats, or excessively using host resources.

Last but not least, all user information is encrypted and stored on a separate cluster. In case of a data leak, the information remains protected since the intruder will require a tremendous amount of effort to decrypt the files. What is more, a backup system adds an extra layer of protection to our sensitive data.

5.6 Security measures for the safe storage of the data

The Covid-X Sandbox tools and services adopt the three core security principles which are confidentiality, integrity and availability, allowing protection from multiple sources. The main scope of these services is the protection of the medical data that will be ingested in the Sandbox.

The first step towards privacy and security is anonymization. As mentioned above, all data will be anonymized prior to entering the Sandbox, following the Anonymization Guideline proposed in the Open Call Annex.

The next steps that will ensure that all data ingested in the Covid-X Sandbox is stored securely are: secure authentication and authorization mechanisms, RBAC, monitoring and auditable trace of activities.

More specifically, in order for the data to be securely stored, the Covid-X Sandbox will only be available to a limited and well-defined number of parties, that will be properly trained with respect to the security principles followed, the security risks and the measures to be taken in case of a data breach. Naturally, access to the Covid-X Sandbox will be controlled with authentication and authorization tools that support identity validation, along with access control lists. More precisely, the System Owner will be responsible for creating Covid-X Sandbox users. Each user can be assigned with a role that entitles him/her with specific privileges and can perform the appropriate actions, therefore avoiding misuse of Covid-X Sandbox tools and data. Such RBAC mechanisms will follow a minimum user access and functionalities policy, avoiding scenarios where users can escalate their rights inside the Covid-X Sandbox or consume excessive system resources.

Additionally, during the deployment of the Covid-X Sandbox, specific network protocols will be used to analyze its network traffic, i.e., keep track of all network connections, data exchanges, nodes used for communication, etc. By controlling what comes in or leaves the network, security in storage is enhanced. Furthermore, the recording of all activities, using audit logging tools, will assist in the traceability and accountability of user actions inside the Covid-X Sandbox. Each activity log is processed by an alert system that detects abnormal behavior. This measure focuses on mitigating any possible insider threat. Security logs that contain sensitive information are backed up. Additionally, customizable dashboards specifically used for activity monitoring for users and Covid-X Sandbox components are supported.

Moreover, secure data storage is achieved through securing the internal communications; encryption protocols such as HTTPs are used in order to enhance security when data is exchanged between

different clusters of the Covid-X Sandbox's components. The existence of different clusters ensures that the data is constantly available and enhances the system's fault tolerance.

A further step towards secure data storage is the utilization of modern open source virtualization tools (Docker containers, Kubernetes Orchestrator) which include isolation, encapsulation and partitioning properties and promote security. The use of certified resources, upon which the containers will be built, in combination with the constant update of the selected tools, add to the security in data storage, by avoiding any vulnerabilities of older versions and integrating new and more robust methods of recognizing and mitigating malicious acts.

5.7 Patient rights procedure

The GDPR guarantees a set of rights that are preserved in the COVID-X project:

- Right of access: the affected person has the right to know if their data is being processed, and among other information, they have the right to know the purposes of the treatment, the categories of data processed, the recipients or categories of recipients to whom they are communicated or These data can be communicated, data retention periods if possible, origin of the data when they have not been obtained from the interested party or the international transfers planned or carried out. The HCSC, as the data controller, has established a series of protocols that guarantee patients the exercise of these rights. The right of access is understood to have been fulfilled with the delivery of a copy of that information by the data controller to the interested party.
- Right of rectification: the right to have the data rectified without undue delay when it is wholly or partly inaccurate, as well as to complete any data that is not. Among the procedures that the HCSC has enabled, there is the procedure for correcting errors.
- Right of deletion (right to be forgotten) and opposition. In the treatment of data for health purposes, these rights are limited since, according to the provisions of the normative (in Spain Patient Autonomy Law for example), **the clinical history must be complete and therefore episodes cannot be hidden**, although it is true that they should be reserved and not expressly appear until the doctor treating the patient decides to consult them.
- Right not to be subjected to a treatment based solely on individual automated decisions (Article 22): right not to be subjected to a decision based solely on automated processing, including the creation of profiles, that produces legal effects on him or affects him significantly in a similar way, unless the treatment is covered by any of the exceptions provided in the cited article. This situation is not considered in the treatments carried out in COVID-X.
- Right of limitation of the treatment (article 18), for assessed cases such as, for example, in cases in which it is necessary to carry out checks on the accuracy or inaccuracy of the data or when it is necessary to suspend the erasure of the data because it is requested by the interested for the formulation, exercise or defence of claims. This situation is not considered in the treatments carried out in COVID-X.
- Right to data portability (article 20) understood as the right of the data subject to obtain, in a structured format, commonly used and mechanically read, the information that concerns him

and has provided to a data controller when that information is processed by automated means and on the basis of consent or for the performance of a contract. In relation to the latter right, the Working Group on Article 29 of Directive 95/46 / EC has adopted guidelines on the application of the right to portability. For example, the Group considers that the concept of data provided by the interested party includes the data actively provided by the interested party and the observed data (location data, search, heart rate, etc.) but does not include within the data subject to the right to portability the inferred data or deduced that they have been created by the data controller from the data provided by the interested party (such as algorithmic results). The application of this right in the COVID-X project is answered in the procedure for access to the medical record, authorized by the pilot sites. For example, the protocol to guarantee the rights of patients in the Spanish pilot site is contained in document *01PROT_Derechos SOPLAR_20180507_v.1.doc*.

5.8 Treatment agreement documents

The GDPR establishes that the person in charge is the “natural or legal person, public authority, service or other body, which determines the purposes and means of the treatment, and also defines the person in charge as a natural or legal person, public authority, service or other body that deals with personal data on behalf of the data responsible”.

The person responsible for the treatment sets the purpose of the treatment and decides on the outsourcing of the same and to what degree he delegates the treatment activities to another organization. In addition, it will only choose a manager who offers sufficient guarantees to apply appropriate technical and organizational measures so that the treatment is in accordance with the GDPR. In COVID-X, third companies or investigating entities do not act as managers or responsible for the treatment since the data that these entities receive are already anonymized and are therefore not personal data, being outside the legislation.

However, the partners that are part of the COVID-X consortium, do act as data managers since their tools will act in the Data Lake where the data is pseudo-anonymized, therefore still being personal data.

In order to respond to the responsibilities derived from this situation, treatment data agreements were signed by all partners of the consortium, since the GDPR establishes that the carrying out of treatments on behalf of third parties must be regulated in a written contract, or other legal act, that binds the person in charge with the person in charge and establishes the object, duration, nature and purpose of the treatment, as well as the type of personal data, categories of interested parties and the obligations and rights of the person in charge.

Furthermore, the security policy includes that on occasions when the person in charge of the treatment needs to resort to third parties, such as when the partners may need support or licenses or any support that may involve the participation of other third parties in turn, that act as sub-processors. In these cases, the person in charge who subcontracts services must notify the person in charge (the HCSC), informing about the type of service that has been subcontracted, as well as the guarantees that these

organizations offer to comply with the regulations. As for the obligations of subcontractors, these will be the same as those that apply to managers **and must also be included in a contract**.

5.9 Audit and reporting mechanism for the authorities

The GDPR requires in its article 5 that all processing of personal data complies with the principles related to the treatment, among others, the entity takes into account and establishes periodic verification protocols that allow verifying that the data processing continues to be compatible and lawful with the initial purpose.

The national control authorities play a primary role in the protection and guarantee of the fundamental right to the protection of personal data, and they must ensure this, as recognized by the European Court of Justice itself, in application of the powers that they hold under the Charter of Fundamental Rights of the European Union and the new European regulatory package in this field, in particular, contained in the GDPR.

In this sense, the GDPR expressly recognizes that the establishment in the Member States of control authorities capable of carrying out their functions and exercising their powers with full independence constitutes an essential element of the protection of natural persons with respect to the treatment of their Personal data.

For this reason, in COVID-X the mechanisms for dialogue with the corresponding control authorities of the pilot countries (Spain, Italy, and Sweden) have been considered, for the moments that were necessary.

A) Aspects of interest considered at the time of design in relation to the control authorities:

- Prior consultation with the control authority based on the results of the PIA carried out: According to art. 35 of the GDPR, it will be the competent control authority that establishes the list of operations or projects that require a PIA. In accordance with the guidelines published by the control authorities, which have directly identified a list of project typology conditions that must incorporate an impact assessment, among which the COVID-X project typology is clearly identified, therefore no a consultation was necessary. Given that the data processing, based on the results derived from the PIA, does not entail a high risk given the measures implemented to mitigate them, it has not been necessary to consult the control authority to continue with the treatments.
- Adoption of appropriate technical and organizational measures. In general, as provided for in article 24.1 of the GDPR, the COVID-X project consortium has applied the appropriate technical and organizational measures in order to guarantee and be able to demonstrate that the treatment is in accordance with the applicable regulations, and that it has been done taking into

takes into account the nature, scope, context and purposes of the treatment, as well as the risks of varying probability and severity for the rights and freedoms of natural persons, having specially considered the principles of privacy by design and by default, such as it has indeed been done in COVID-X.

B) Aspects of interest to consider in the development about authorities' relations:

- Record of treatment activities. In compliance with article 30 of the GDPR, it is mandatory that the person responsible, in the case of developing a Big Data project, has a Register of the treatment activities associated with it, which will be available to the competent control authority. The treatment registry of the San Carlos Clinical Hospital is published on the Transparency Portal of the different pilot sites
- A notification procedure has been developed by the person responsible for processing a possible violation of the security of personal data, to the corresponding control authority according to the provisions of article 33 of the GDPR.
- Coordination of the Data Protection Delegate name with the supervisory authority. In accordance with the provisions of article 37 of the GDPR, it is mandatory for the data controller to appoint a DPO when carrying out massive data analysis projects, as has been carried out in the COVID-X project to each of the pilot sites, which exercises the functions set forth in article 39 of the GDPR and, which acts as a point of contact with the control authority and cooperates with it, in addition to providing advice that is requested by the person in charge of the treatment (HCSC) about the impact assessment related to data protection carried out on the project.
- General duty of cooperation with the supervisory authority in the performance of its functions (Article 31 GDPR). Apart from the above, there is a duty of the person in charge for all types of treatment and projects on personal data treatment of collaborating with the control authority in the correct exercise of its powers, operating this duty as a general guarantee in favour of the protection of the privacy of individuals in this type of project.

5.10 Data breach notification mechanism

See document COVID-X security incident management.

6 Conclusions

As a result of the activity of the legal and ethical framework, on the one hand, the regulatory requirements to which the project is subjected have been identified, on the other hand, the different aspects and issues to be taken into account in the process of compliance with said requirements have been analysed. Finally, the mechanisms and measures to be implemented have been explained and developed to give the best response to compliance needs, in terms of privacy and protection of personal data.

Once the indications and action guides have been made available to the COVID-X partners, the next step will be, on the one hand, to advise partners and participants regarding any questions that may arise in the process of implementing the different privacy measures and security and, on the other hand, keep the legal and ethical framework permanently updated, as set out in the GDPR guidelines, having at all times the “accountability” capacity required in it.

References

- “Opinion 05/2014” of the Article 29 Working Group on anonymization techniques.
- “Opinion 06/2014” of the Article 29 Working Group on the notion of legitimate interest referred to in article 7 of Directive 95/46 / EC.
- "Code of good practice for European statistics for national and community statistical services", adopted by the Committee of the European Statistical System on September 28, 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
- “I-Diversity: Privacy Beyond k-Anonymity”, Machanavajjhala et al., 2006, IEEE.
- “A Study on k-anonymity, I-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.



Appendix A

Anonymization Guide.





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 Gkatzoura (8BELLS)

DISCLAIMER

The information, documentation and figures available in this deliverable are written by COVID-X project's consortium under EC grant agreement 101016065 and do not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

COPYRIGHT NOTICE

© 2020 - 2022 COVID-X Consortium Reproduction is authorised provided the source is acknowledged

Table of Contents

1	Anonymization Policy.....	4
2	Introduction	5
3	Work Team Configuration	6
4	Necessary Training for the Anonymization Team.....	8
5	Privacy Impact Assessment (PIA).....	10
5.1	Stages of the Impact Assessment	10
5.1.1	<i>Identification and categorization of assets involved in the anonymization process.</i>	11
5.1.2	<i>Constitution of the Work Team</i>	12
5.1.3	<i>Risk Identification</i>	13
5.1.4	<i>Assessment of Existing Risks and Quantification of the Impact</i>	14
5.1.5	<i>Safeguards</i>	16
5.1.6	<i>Risk Report</i>	16
5.1.7	<i>Determination of the Acceptable Risk Threshold</i>	17
5.1.8	<i>Management of Assumable Risks</i>	18
5.1.9	<i>Final Report</i>	19
6	Organizational measures	20
7	Possible Techniques for Anonymization.....	21
7.1	Layers of Anonymization	21
7.2	Data Interruption	21
7.3	Data Reduction	22
7.4	<i>k</i> -Anonymization	24
7.5	Other Techniques.....	26
8	Anonymization Protocol	27
8.1	Phase 1 - Pre-Anonymization.....	27
8.2	Phase 2 - First Layer of Anonymisation	28
8.3	Phase 3 - Elimination / Reduction of Variables	29
8.4	Phase 4 - Anonymisation	29
8.4.1	<i>k Anonymization Technique Application</i>	29
8.4.2	<i>Note regarding biometric data</i>	30

8.5	Phase 5 - Delivery of the Data Set	30
9	Additional guarantees of confidentiality of anonymized information	32
9.1	Documentary Guarantees.....	32
9.2	Data Segregation.....	33
9.3	Audits	33
10	Standards and References	36

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.

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.

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 breaches¹.

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.
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

• 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>.

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.

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 system².

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>.

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.

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.

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.
- 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 inpatients.
- 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.

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	
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

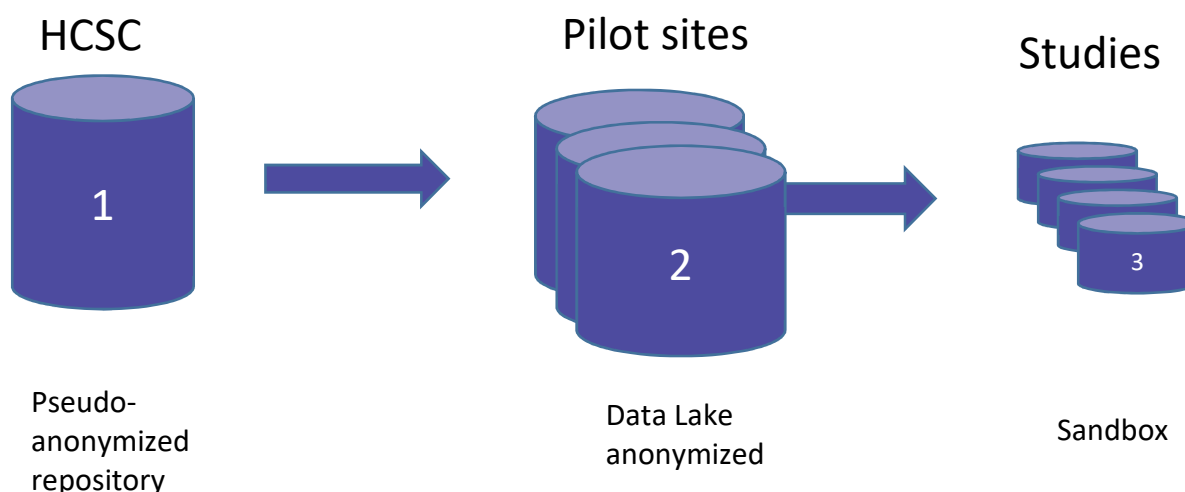


FIGURE 1: SCHEMATIC OF THE DATA PROCESSING STATIONS IN HOSPITAL CLINICO SAN CARLOS

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 reidentification risks analysis (PIA)				Impact		
				Information elements		
				Data pseudo	Data lake	Sand box
				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 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 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.

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	Safeguard	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

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.

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.

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:

- **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.

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.
- 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 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

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 class³ 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.

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.

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

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:

- 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.

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

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.
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

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.

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.

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

- 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.

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



- “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.