

ANONYMISATION PROTOCOL FOR CLINICAL REPORTS - CARMEN-I CORPUS

July 2023

SUMMARY

One of the objectives of the collaboration between the Hospital Clínic de Barcelona (HCB) and the Barcelona Supercomputing Center (BSC) is to launch a dataset or corpus of clinical reports originated during the COVID-19 pandemic. The goal of this dataset is helping the scientific community to develop systems that improve the processing of health data.

When selecting which reports adding to the corpus, it is very important to ensure that we can guarantee the privacy of the people who appear in them. To this end, this document describes the protocol created for the data anonymisation, as well as the control mechanisms put in place for this purpose. It also includes addenda to the MEDDOCAN guidelines for the annotation of sensitive data, criteria for inclusion/exclusion of documents, and a list of indirect identifiers.

AUTHORS

Barcelona Supercomputing Center
(BSC)

Salvador Lima-López
Eulàlia Farré-Maduell
Luis Gascó Sánchez
Martin Krallinger

Hospital Clínic de Barcelona
(HCB)

Xavier Pastor
Antonio López Rueda
Santiago Frid
Artur Conesa

TABLE OF CONTENTS

1. Introduction	
2. Overview of the anonymisation protocol	
3. Phase I: Report selection	
4. Phase II: Annotation	10
5. Phase III: De-identification and validation	12
6. Conclusion	16
Annex I. Addenda to the MEDDOCAN guidelines for the annotation of sensitive data	17
Annex II. Validation criteria for anonymised documents	20
Annex III. List of indirect identifiers	22

1. Introduction

Clinical data, such as reports or structured databases, are invaluable for the development of **artificial intelligence** tools. These tools use the data to learn how to perform different tasks in order to improve healthcare and facilitate the work of healthcare workers. However, despite its usefulness, we must not forget that these sensitive data belong to individuals who have a right to **privacy**.

For this reason, the processing of **data with sensitive content** is highly regulated. At a **Spanish national level**, two main laws were applied, the Ley Orgánica de Protección de Datos de Carácter Personal (LOPD) 15/1999 [*Organic Law on Personal Data Protection 15/1999*], dated 13 December, as well as the Real Decreto 1720/2007 [*Royal Decree 1720/2007*], dated 21 December, and the Ley 34/2002 [*Law 34/2002*], dated 11 July. All these laws were repealed with the entry into force of the Ley Orgánica 3/2018 de Protección de Datos Personales y garantía de los derechos digitales [*Organic Law 3/2018 on the Protection of Personal Data and Guarantee of Digital Rights*], on 5 December 2018, which adapts the **General Data Protection Regulation** (GDPR) of the **European Union**. These laws directly impact the storage, processing, access, transfer and disclosure of data records of individuals¹.

The identification of individuals in clinical reports may have a number of unintended consequences on their rights and freedoms. We should be aware that the patient is not the only person who appears in the reports, and that identification can be carried out both by **direct identifiers** (data that point directly to a person, such as name, age or place of birth) and **indirect identifiers** (data that do not point to anyone in particular but, together with other sensitive data, can be used to recognise a person).

For all these reasons, the use of clinical data raises a number of ethical and legal issues. One possible solution that allows us to use the data in the development of

¹ Report on Personal Data Protection of the Plan for the Advancement of Language Technology (May 2019). Technical Office of Health.

tools while maintaining the privacy of individuals is **anonymisation**. Anonymisation is a technique for processing sensitive data that consists of modifying or removing information in a document that allows for identifying a person. When processing text, anonymisation consists of two main steps: (1) detection of sensitive data and (2) modification of the detected data. Being such a delicate process, although we can make use of automatic tools, human supervision is crucial at all times.

One of the problems with current data protection laws is that they point out the need to define and control the anonymisation process but do not make any specific standardised proposals. This means that we must determine ourselves what protocols and mechanisms we will use for the project. This document describes the anonymisation protocol and control mechanisms established to guarantee the privacy and anonymity of the clinical reports within the collaboration between the Barcelona Supercomputing Center (BSC) and the Hospital Clínic de Barcelona (HCB) for the creation of the CARMEN-I corpus.

2. Overview of the anonymisation protocol

The anonymisation of medical texts is a complex and specialised process that requires a detailed action plan. This action plan must consider the characteristics of the domain and the type of text (in this case, clinical reports), the direct and indirect sensitive data that may appear in them, and the validation of the final product to ensure that the objectives have been met.

For this project, we defined our own protocol with two main objectives: 1) to guarantee that we maintain the privacy of the anonymised reports (opening up the possibility of publishing them openly once they are free of sensitive data); and 2) the creation and improvement of an automatic anonymisation assistance tool adapted to the data of the hospital. Within this protocol, the work is organised in different phases (each with its own tasks, objectives, and results). Different control mechanisms are also established to ensure the quality of the anonymisation work. Figure 1 summarises graphically the anonymisation protocol of the project.

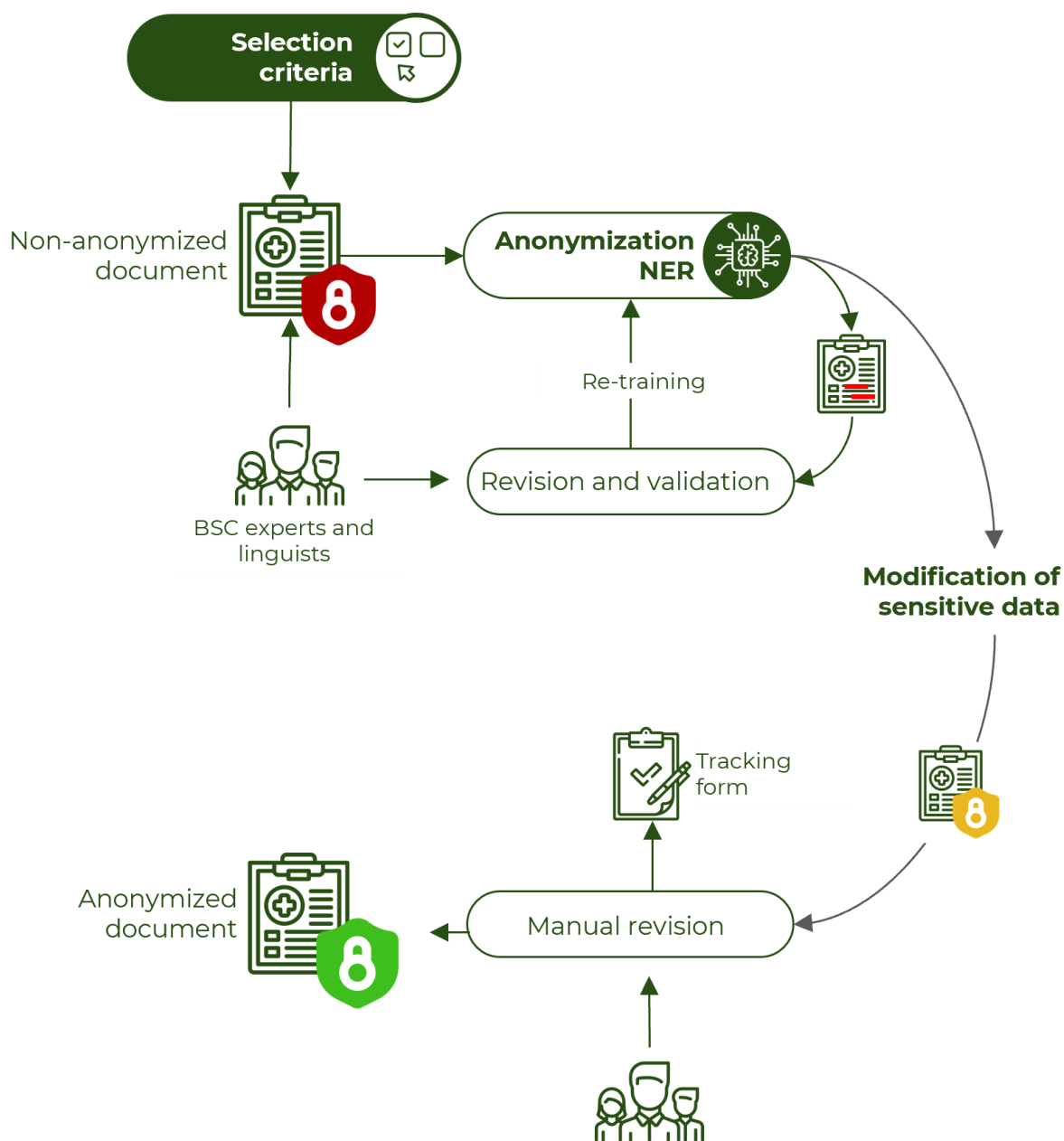


Figure 1. Summary diagram of the anonymisation protocol.

Briefly, these are the three phases into which we divide the anonymisation process:

- **Phase I: selection of reports**

In this phase, a set of reports is selected and prepared for being used in subsequent phases.

- **Phase II: manual annotation of sensitive data**

Using automatic tools as an aid, sensitive data from the reports selected are manually annotated. This annotation is used to improve the anonymisation support tool iteratively.

- **Phase III: de-identification and validation of the anonymised data**

Finally, the sensitive data annotated in Phase II are modified and the result is validated by the experts of the hospital on the basis of established criteria to ensure that it is impossible to re-identify the patient.

It is important to note that, although the three phases occur sequentially (Phase II cannot be done if Phase I has not been done before, etc.), once the process has started, the phases can occur simultaneously. This means, in particular, that the validation phase (Phase III) can start even if all the selected documents (Phase II) have not yet been annotated in order to speed up the process, or that new documents (Phase I) can be selected, if necessary, when the other phases have progressed.

The work carried out throughout the process is accompanied by control mechanisms to ensure the quality of the work and the protection of the privacy of the people mentioned in the reports in a systematic way. The main mechanisms are:

- **Continuous review and validation of the work already done**

All Phase II (annotation) and Phase III (validation) results are systematically reviewed by people from more than one institution.

- **Definition of indirect identifiers**

Through a review of the literature and the clinical reports, we extracted a list of possible indirect identifiers that require special attention during the validation process. Annex III presents a list of possible identifiers.

- **Definition of criteria to be followed for validation**

In order to make the validation process as less subjective as possible, criteria for the validation of anonymised reports are agreed upon. These criteria are detailed in Annex II.

- **Validated report tracking sheet**

Every report that has been validated is recorded individually on a tracking sheet.

The following sections explain more in-depth the phases and mechanisms mentioned during this section.

3. Phase I: Report selection

The first phase of the project focuses on the selection of the reports that will be used in subsequent phases.

The report selection is done semi-manually. In other words, first, a large number of documents that share some automatically calculated characteristics are selected, and then one or more persons select the most suitable reports by hand. For each of the two parts of the selection, different criteria are taken into account.

On the one hand, for the automatic part, the following elements can be considered, among other examples that are not included:

- **Report type**

We may want to include or exclude some specific types of reports. For example, radiology reports contain little sensitive data, as a large part of the content is taken up by a radiological description, so we may want to prioritise them. On the contrary, clinical courses have more noise and are more difficult to process. The latter may also contain more sensitive data, so they will normally be omitted.

- **Report length**

By probability, shorter reports usually contain less sensitive data than longer reports. In addition, dealing with shorter reports allows us to handle a higher number of documents. Still, it is good for both the model and the publication of the corpus to have documents of varying length.

On the other hand, for the manual selection, a reading of some reports is carried out, considering mainly two factors:

- **The existence of sensitive data**

Reports that contain large amounts of sensitive data, are very specific or do not meet any of the criteria in Annex II (e.g., more than eight

co-morbidities) are discarded directly. By doing this, we do not make extra efforts working with reports that may be problematic in the long run.

- **Variability of content**

For the construction of the corpus and the improvement of the tool, it is important that the content varies and has substantial variability. Since the number of available reports is very high and the number of working hours is limited, it is very relevant to try to prioritise reports with heterogeneous content. For example, if we select radiology reports, the image descriptions must be varied.

Reports selected during Phase I are automatically pre-annotated using the anonymisation support tool for later use in Phase II. The selection of documents can be done more than once to work on anonymisation incrementally, both manually and with the help of tools.

In addition, in the first selection of documents, reports are manually analysed by carefully reading them to understand the type of sensitive information we will be dealing with later on. The result of this analysis is embodied in the Addenda to the annotation guidelines for sensitive data (explained in Phase II) and the characterisation of indirect identifiers in clinical reports (explained in Phase III).

4. Phase II: Annotation

In the second phase, sensitive data from selected reports are manually annotated with the help of automated tools.

To anonymise a text correctly, we must first detect exactly the sensitive data contained. The way to ensure that we do this exhaustively and correctly is by manually marking or annotating these elements. The annotation must be done with defined categories and according to rules that allow the result to be consistent. These categories and rules are explained in a document called annotation guidelines.

For this task we will use the MEDDOCAN sensitive data annotation guidelines developed by the Barcelona Supercomputing Center (available at Zenodo²). These guidelines were created for the MEDDOCAN evaluation campaign³, focusing specifically on the development of annotated resources and anonymisation tools. Since European law does not list in detail what elements are considered sensitive information, these guidelines were developed based on the *US Health Insurance Portability and Accountability Act (HIPAA)*.

HIPAA defines 18 categories of sensitive data that must be removed or modified to consider a text as “anonymised”. The MEDDOCAN guidelines expand and adapt these categories to be as comprehensive as possible. In total, the guidelines define 28 possible labels for the annotation of sensitive data, including: patient or healthcare staff names, ages, dates, family members, locations, occupations, different numerical identifiers (e.g., car number plates or medical record numbers), etc.

The MEDDOCAN guidelines are very exhaustive, but for their practical application we need to consider that each hospital and type of report has its own characteristics and context. Therefore, based on the analysis made in Phase I, the

² <https://doi.org/10.5281/zenodo.4279337>

³ <https://temu.bsc.es/meddocan/>

guidelines are enriched to ensure that they also cover the idiosyncrasies of the Hospital Clínic reports and the COVID-19 pandemic situation. In our case, for example, we added specific rules for room identifiers within the hospital or for inpatient hotels that were created during the beginning of the pandemic. These new rules are described in the addenda available in Annex I.

The annotation of these texts is done by the BSC and a team of linguists from the University of Barcelona (UB). To ease the process, a series of automatic predictions are available to help the annotator work faster. The anonymisation support tool we use in this project is trained with the MEDDOCAN clinical case corpus. This tool is iteratively re-trained with the annotated reports to improve its adaptation to the texts from the hospital. The adaptation does not only lead to the improvement of the tool, but also facilitates the work of the annotators, helping them to annotate more reports with less effort.

Before moving on to Phase III, each document is reviewed a minimum of two times. Ideally, each review is done by a different person on the team to avoid possible omissions in the annotation.

5. Phase III: De-identification and validation

The third phase is dedicated to the validation by clinical experts of the anonymised reports to ensure that re-identification is not possible. This is done by de-identifying the text in two different ways: "masking" sensitive information and replacing it with synthetic equivalents.

Once the reports have been annotated and reviewed, a first automatic modification of the sensitive data is made. The modification is done using a technique known as "masking", whereby each sensitive data item is replaced by the category to which it belongs. For example, the sensitive data '3 January 2021' is replaced by the word 'DATE'.

Clinical experts will carefully read the *masked* documents to validate whether all sensitive data have been modified and whether re-identification is possible. Since this validation has a certain level of subjectivity, it must be done according to established criteria that determine when a document is anonymised or not.

For this purpose, simple criteria for the validation of reports have been defined and are developed in Annex II. These criteria include a series of aspects to which attention should be paid when validating if a document is anonymised correctly. These criteria are divided into inclusion criteria and exclusion criteria. Inclusion criteria refer to the correct fulfilment of the steps of the anonymisation protocol described in this document. Exclusion criteria refer to failures in the process of modifying sensitive data, as well as the existence of indirect identifiers that help to identify a natural person.

Indirect identifiers are data that, although they do not point directly to a person (as a name or date of birth might), can help in some way to recognise someone. Examples include physical descriptions, mentions of personal circumstances (such as family situation) or the existence of rare diseases.

By its nature, this type of data is one of the most complicated aspects to assess. In order to estimate its severity and potential impact, two fundamental aspects must be taken into account:

- **Specificity**

Identifiers can be very general or very specific. Specific identifiers pose a higher risk and may be a clear reason for exclusion even if no further identifying information appears in a report. An example would be a rare disease with less than 10 cases.

- **Combination**

The risk of re-identification increases the higher the number of identifiers that appear together. The existence of indirect identifiers in a report should be assessed globally and not separately.

In order to comprehensively account for this type of data, Annex III includes a list of possible proxy identifiers relevant to clinical reports. This list has been created through three sources: reading related literature (e.g., on social determinants of health), during the analysis of the clinical reports in Phase I, and after discussions between the team at BSC and clinicians from the hospital.

These identifiers take into consideration four main agents or actors that may appear in clinical reports:

- **Patients**

They are the most important stakeholders of clinical reports, and most of the content is about them and their health problems and treatments.

- **Family members**

They are very important to patients, both in terms of medical history and social support network, so it is common for them to be named in reports.

- **Health or care personnel**

They are in charge of caring for and guiding the patient, so they have a special role in clinical reports, often being the authors of the report.

Professionals from different healthcare centres may appear in the same report.

- **Other people who may be mentioned**

We must also consider other people who are sometimes relevant to the report, such as police officers, firefighters, translators, or co-workers.

In addition, although not a natural person, we will also consider the hospital as an entity whose privacy must be maintained, being able to veto reports with certain types of information (e.g., participation in clinical trials), as well as other healthcare facilities where the patient has received or will receive care.

It is important to stress again that this type of sensitive information alone is not sufficient to identify anyone and that each report must be judged as a whole. Care must be taken not to introduce discriminatory bias into the data set by systematically rejecting certain population groups just because they are more recognisable.

Taking all this into account, the clinical experts validate each report individually and record their decision on a tracking sheet. For each document, this sheet will mark whether the anonymised version guarantees the privacy of the persons mentioned in the report. Reports that are rejected must be justified with an explanation as to why. This justification can be based on the validation criteria (Annex II) and the list of indirect identifiers (Annex III). In our case, we also take advantage of this task to ask clinicians to categorise the language of each report into three classes: Spanish, Catalan, or bilingual (the latter class includes any report in which both languages are used, regardless of the percentage of each).

For publication, as a final step, a second modification of the sensitive data is performed. In this second step, the *masked* data is replaced by synthetic data. Thus, the result of the whole process is an anonymised report that appears to be real, which increases its usefulness for the development of language processing tools. This substitution uses a combination of different rule systems created specifically for this task, as well as different lexical resources (such as lists of

names, locations or occupations), which take the real data as input and return a synthetic one trying to maintain some resemblance to the original data.

Some of the systems are very simple (such as changing telephone numbers to another random *string*), while others become quite complex and even interact with each other. For example, the date substitution system moves all dates in a document by a random number of days, months and years to try to maintain temporal consistency where possible. This number of days, months and years is chosen again in each document, so it is not possible to re-identify any of them temporally. Another example of interaction between substitutions occurs with the different categories of places, where an attempt is made to recognise the original country, so the cities and towns used as substitutions belong to the same country in order to maintain certain geographical coherence. In general, the different substitution functions for this step are inspired by those of systems previously described in the literature such as HitzalMed⁴.

⁴ Lima-López, S., Perez, N., García-Sardiña, L., & Cuadros, M. (May 2020). HitzalMed: Anonymisation of clinical text in Spanish. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 7038-7043).

6. Conclusion

In short, this document has presented in a general way the anonymisation protocol followed for the CARMEN-I corpus, composed of anonymised medical records written during the COVID-19 pandemic (specifically between March 2020 and 2022) and resulting from the collaboration between the Barcelona Supercomputing Center and the Hospital Clínic de Barcelona.

Since one of the aims of the collaboration was the open publication of the CARMEN-I corpus for the community, this protocol is quite strict and exhaustive. Multiple people from different fields are involved, with control over specific tasks to ensure a complete anonymisation of the clinical reports. We hope that this document can serve as inspiration for future projects and incursions of healthcare institutions into the world of natural language processing, using it as is or changing aspects to suit their needs.

Annex I. Addenda to the MEDDOCAN annotation guidelines of sensitive data

This annex is intended to be a complement to the MEDDOCAN anonymisation guidelines for the anonymisation of reports from the Hospital Clínic de Barcelona. The aim is to cover types of information that were not considered in the original guidelines because they were specific, but that tend to appear in clinical reports. The proposed addenda have been drafted based on the analysis of the original guidelines as well as a number of clinical reports. The original MEDDOCAN guidelines are available on Zenodo.⁵

| MEDDOCAN_HCB- 1

inpatient hotels and residences (tag <institution>)

During the busiest times of the pandemic, specific places to treat COVID patients, such as inpatient hotels, were established. We will also find quite a few mentions of residences. This type of information will be marked with the label INSTITUTION.

Example: "During her stay at the Hotel Rosalía de Castro, the patient has had a good evolution".

Example: "Patient referred from Residencia Cuatro Caminos".

| MEDDOCAN_HCB- 2

room names and room types (tag <identif_number>)

It is common for reports to talk about the rooms where patients have been treated in the hospital. This type of information is usually a code of one or two letters plus some numbers. It does not always have to be preceded by the word "room". We will annotate these names with the tag IDENTIF_NUMBER.

Example: "PHYSICAL EXPLORATION IN ROOM E001"; "*PHYSICAL EXPLORATION HO17".

⁵ <https://doi.org/10.5281/zenodo.4279337>

In general, speciality room types and areas of the hospital are not included in this rule (e.g., "liver ICU" or "internal medicine ward" are general room types and are not annotated). However, there are some room names that are equivalent to room codes and will be annotated:

- AVI
- UVIR (i.e., ICU room)

| MEDDOCAN_HCB- 3

ids from clinical trials and vaccine batches (tag <identif_number>)

In the reports, we can also find mentions of clinical trials, for which the study ID is usually specified. This number should be anonymised with the label IDENTIF_NUMBER. In the case that the batch from which a vaccine originates is mentioned, it should also be anonymised in the same way.

Example: "starts on 04/04/2017 EC QWERTY1234 for treatment of respiratory failure in patients...".

| MEDDOCAN_HCB-4

HCB acronym (tag <hospital>)

The acronym HCB stands for "Hospital Clínic de Barcelona" and should also be marked with the label HOSPITAL.

Example: "Follow-up by Dr. Fernández (Haematology HCB)".

| MEDDOCAN_HCB-5

acronyms HC3, SAP, SIRE (tag <other_subject_care>)

These acronyms refer to computer systems specific to the Catalan health system. For the moment, we will mark them with the label OTHER_SUBJECT_CARE.

| MEDDOCAN_HCB-6

hospital's own services (tag <other_subject_care>)

There are some hospital services that are specific to employees and may disclose patient information indirectly. For example: "Occupational Health". These will be annotated as OTHER_SUBJECT_CARE.

Annex II. Validation criteria for anonymised documents

This annex describes a set of simple criteria to be followed for the final validation of anonymised documents.

- Inclusion criteria:

Every report of the corpus must meet the following conditions:

Criterion	Description
CI-1 [meddocan]	<p>Sensitive data in the report must be labelled according to the MEDDOCAN guidelines.</p> <p>If there are sensitive data that are not covered in the MEDDOCAN guidelines, they may be included after agreement between the BSC and the HCB, as long as they are easily replaceable data types (such as numerical identifiers or proper names, as opposed to long descriptions). These additions to the original guidance are reflected in the addenda (Annex I).</p>
CI-2 [validation]	<p>Reports to be included in the corpus must have been reviewed at least three times:</p> <ol style="list-style-type: none">(1) Firstly, by the team of linguists who will annotate the sensitive content.(2) Then by the BSC team to review the annotation and modify the sensitive content.(3) Finally, by the HCB team who will validate the document with the modified sensitive data.
CI-3 [registration]	<p>Every report of the corpus must be duly recorded on the control sheet for subsequent control.</p>

- Exclusion criteria:

The exclusion criteria raise a number of situations to be taken into account for rejecting anonymised reports. These reports may be rejected completely or re-annotated by the team at BSC to ensure proper anonymisation.

Criterion	Description
EC-1 [data-modified]	<p>All sensitive data that are direct identifiers must have been modified, either by obfuscation or by substitution. Reports containing direct sensitive information must be excluded when validated.</p> <p>For this purpose, the reports to be reviewed by the HCB team will include the modified sensitive data annotated with the original data as a comment to each annotation. In this way, it will be possible to check that the synthetic data and the original data differ.</p> <p>In the case of direct identifiers that have not been modified for any specific reason, the BSC team may review the annotation to correct it. Afterwards, the document should be validated again.</p>
EC-2 [identifiers-indirect]	<p>A single indirect identifier is usually not enough to identify someone, especially if it is not very specific. The risk of re-identification often comes from the combination of several of these identifiers, especially if they are linked to a direct identifier.</p> <p>Therefore, the reports of the corpus cannot include very specific indirect identifiers or identifiers that combined together allow for the re-identification of a natural person. The slightest possibility of re-identification translates into the immediate exclusion of the document. The final decision will be made by the validating clinician.</p> <p>Indirect identifiers are listed in Annex III.</p>
EC-3 [exclusion-direct]	<p>The hospital may request the exclusion of individual reports for any reason, even if the report meets the inclusion criteria.</p>

Annex III. List of indirect identifiers

Indirect identifiers are a type of sensitive data that, although they do not point directly to a person, can be used to make inferences about and re-identify that

person. This annex provides a non-exhaustive list of some of the identifiers that may appear in the clinical report, dividing them into sub-types according to who they affect and the clinical aspect to which they pertain.

A. Information about natural persons

AA. Sociodemographic indirect identifiers

Sociodemographic indirect identifiers are non-medical data that suggest information about the life of natural persons.

Identifier	Description
AA1 [physical-appearance].	<p>Reports may contain descriptions of a person's physical appearance, including possible dysmorphism.</p> <p>Affects: patients, relatives, other people.</p> <p>Examples: "he is 1.95 m tall", "he has an aquiline nose", "has a problem with the 3rd and the 4th finger of the right hand", "he lost his right arm in an accident".</p>
AA2 [life-conditions].	<p>Reports may contain descriptions of a person's living conditions, including but not limited to: living alone or in company, characteristics of the living space, approximate location, etc.</p> <p>Affects: patients, relatives, other people.</p> <p>Examples: "she lives alone since her husband died", "she has no space at home to isolate herself", "when she was a child, she lived in a tin house", "she lives in the street".</p>
EC-AA3 [disability].	<p>Particular attention should be paid to reports that refer to specific disabilities of individuals.</p> <p>Affects: patients, relatives, other people.</p> <p>Examples: "he uses a wheelchair".</p>
EC-AA4 [education]	<p>Particular attention should be paid to reports in which references are made to a person's education.</p>

	<p>Affects: patients, relatives, other people.</p> <p>Examples: "left school at the age of 8", "graduated in Philosophy".</p>
CE-AA5 [occupation].	<p>Particular attention should be paid to reports in which descriptions are made of a person's occupation or employment status that cannot be anonymised.</p> <p>Affects: patients, relatives, other people.</p> <p>Examples: "his job consists of...", "he suffers from stress due to harassment at work", "he has not been receiving any social care benefit for 6 years".</p>
CE-AA6 [violence].	<p>Particular attention should be paid to reports that refer to situations of abuse and violence.</p> <p>Affects: patients, relatives, other people.</p> <p>Examples: "she was sexually abused at home", "her father beat her until she was 15 years old".</p>
EC-AA7 [personal- circumstances]	<p>Particular attention should be paid to reports that refer to other personal situations that cannot be anonymised.</p> <p>Affects: patients, relatives, other people.</p> <p>Examples: "he has no relatives outside his country of origin", "he still does not have many friends in Barcelona", "he was expelled from home by his family due to drug use", "language barrier".</p>
CE-AA8 [orientation-ide ntity].	<p>Particular attention should be paid to reports in which references are made to people's sexual orientation and gender identity.</p> <p>Affects: patients, relatives, other people.</p> <p>Examples: "engages in bisexual relations", "transgender".</p>

AB. Indirect identifiers related to diseases and symptoms

Some diseases and symptoms, either on their own or in combination, can point to a physical person.

Identifier	Description
CE-AB1 [rare-illnesses]	<p>Reports referring to people suffering from rare diseases are almost directly excluded. In the European Union, a rare disease is considered to be a disease that affects <1 person out of 2000.</p> <p>Affects: patients, relatives.</p> <p>Examples: "patient with cystic fibrosis", "sister with Marfan syndrome".</p>
CE-AB2 [comorbidities]	<p>Reports that mention more than 8 medical records for only one person are excluded, as a high number of diagnoses may facilitate identification.</p> <p>Affects: patients, relatives.</p> <p>Examples: "patient with a history of lung, liver, kidney, spleen and heart transplantation, ...".</p>
CE-AB3 [geographical-zone]	<p>Particular attention should be paid to reports on endemic diseases, as they can provide clues and guidance on the origin of the person suffering from the disease.</p> <p>Affects: patients, relatives.</p> <p>Examples: "patient with Chagas disease".</p>

AC. Indirect identifiers related to procedures and treatments

Just as for diseases and symptoms, pharmacological procedures and treatments can also serve to re-identify a person given the right context.

Identifier	Description
------------	-------------

<p>CE-AC1 [uncommon-procedures]</p>	<p>Particular attention should be paid to reports containing unusual medical procedures. For example, face transplantation, uterus transplantation, Siamese twins' separation.</p> <p>Affects: patients, relatives.</p> <p>Examples: "in 2007 he had a transplant of the little toe on his left foot".</p>
<p>EC-AC2 [biomaterials].</p>	<p>Particular attention should be paid to reports specifying the codes, brands, models, etc. of biomaterials used in a procedure.</p> <p>Affects: patients, relatives.</p> <p>Examples: "he has three screws in his knee: one from X, one from Y and another from Z".</p>
<p>EC-AC3 [experimental-drugs]</p>	<p>Particular attention should be paid to reports referring to the use of experimental drugs.</p> <p>Affects: patients, relatives.</p> <p>Examples: "After talking to the family, it was decided to use chloroquine, which had to be suspended after 48 hours due to the presence of an arrhythmia that required defibrillation".</p>

B. Information about the healthcare centre

We may also find data referring to health institutions.

Identifier	Description
<p>CE-B1 [clinical-trials]</p>	<p>Reports that discuss the participation of the hospital, staff or patients in clinical trials are excluded, especially if the subject of this trial is described.</p> <p>Examples: "the patient is proposed to participate in the EC/XXXX trial on the effectiveness of drug X vs placebo".</p>

C. Other information

In the event that we find sensitive data that we want to take into account and that do not fall under the previous sections, they will be included in this table:

Identifier	Description
CE-C1 [commercial-br ands]	Particular attention should be paid to reports in which trademarks are mentioned. Examples: “he had a Micra™ AV pacemaker implanted”.