HORIZON-HLTH-2023-CARE-04-03
Environmentally sustainable and climate neutral health and care systems

NetZeroAICT

Digital Contrast for Computerised Tomography
-Towards Climate Neutral and Sustainable Health Systems-

Starting date of the project: 01/12/2023
Duration: 48 months

= Deliverable D2.1 =

Active data management plan produced (v1)

Due date of deliverable: 29/02/2024
Actual submission date: 29/02/2024

Work Package: 2
Responsible Work Package Leader: Anders Nordell, CMRAD
Responsible Task Leader: Jonatan Eriksson, CMRAD
Version: V1.0

Dissemination level		
PU	Public – fully open (automatically posted online on the Project Results platforms)	X
SEN	Sensitive — limited under the conditions of the Grant Agreement	
Classified R-UE/EU-R	EU RESTRICTED under the Commission Decision No2015/444	
Classified C-UE/EU-C	EU CONFIDENTIAL under the Commission Decision No2015/444	
Classified S-UE/EU-S	EU SECRET under the Commission Decision No2015/444	

Co-funded by the European Union

AUTHOR

Author	Institution	Contact (e-mail, phone)
Jonatan Eriksson	CMRAD	jonatan.eriksson@cmrad.com
Anders Nordell	CMRAD	anders@cmrad.com
Ernest Casany Pujol	CMRAD	ernest@cmrad.com

DOCUMENT CONTROL

Document version	Date	Change
V0.1	29.02.2024	First draft
V1.0	29.02.2024	Edits by Scientific Coordinator

VALIDATION PROCESS

Reviewers		Validation date
Work Package Leader	Anders Nordell	29.02.2024
Project Manager	Martina Nešverová	29.02.2024
Scientific Coordinator	Regent Lee	29.02.2024
Coordinator	Anders Nordell	29.02.2024

DOCUMENT DATA

Keywords	Data management
Point of Contact	Name: Jonatan Eriksson Partner: CMRAD Address: Phone: +46 76 535 65 98 E-mail: jonatan.eriksson@cmrad.com
Delivery date	

DISTRIBUTION LIST

Date	Version	Recipients
29.02.2024	V1.0	EC, all partners via Google Drive

DISCLAIMER

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Executive Summary

This document represents the 1st version of NetZeroAICT data management plan (DMP). The aim of this document is to describe specifications for data types, collection, curation, access, storage, transfer, security, traceability, integrity, and preservation during and after the end of the project. It addresses the FAIR principles of findability, accessibility, interoperability, and reusability and a system for unique persistent data and metadata identifiers. The first version of the DMP is delivered in M3, with an update planned to be submitted in M24. Nevertheless, DMP is a living document to which more details will be added whenever relevant.

List of Abbreviations

CE = Contrast Enhanced

CT = Computerized Tomography

CTA = CT Angiogram

DICOM = Digital Imaging and Communications in Medicine

DMP = Data Management Plan

FAIR = Findability, Accessibility, Interoperability, and Reusability (FAIR).

IV = Intravenous

PACS = picture archiving and communication system

RCM = Radiocontrast Media

Table of Contents

EXECUTIVE SUMMARY **3**

1. INTRODUCTION **5**

2. DATA SUMMARY **6**

2.1. PURPOSE OF THE DATA COLLECTION/GENERATION AND ITS RELATION TO THE OBJECTIVES OF THE PROJECT 7
2.2 TYPES AND FORMATS OF DATA THE PROJECT WILL COLLECT AND GENERATE 8
2.3 DATA PROTECTION 8
2.4 WILL YOU RE-USE ANY EXISTING DATA AND HOW (INCLUDING DATA PROTECTION ASPECTS)? 10
2.5. WHAT IS THE ORIGIN OF THE DATA? 14
2.5. WHAT IS THE EXPECTED SIZE OF THE DATA? 14
2.6. TO WHOM MIGHT IT BE USEFUL ('DATA UTILITY')? 14

3. FAIR DATA **15**

3.1. MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA 15
3.2. MAKING DATA OPENLY ACCESSIBLE 16
3.3. MAKING DATA INTEROPERABLE 17
3.4. INCREASE DATA RE-USE (THROUGH CLARIFYING LICENSES) 18

4. ALLOCATION OF RESOURCES **18**

4.1 WHAT ARE THE COSTS FOR MAKING DATA FAIR IN YOUR PROJECT? 18
4.2 HOW WILL THESE BE COVERED? NOTE THAT COSTS RELATED TO OPEN ACCESS TO RESEARCH DATA ARE ELIGIBLE AS PART OF THE HORIZON 2020 GRANT (IF COMPLIANT WITH THE GRANT AGREEMENT CONDITIONS). 18
4.3 WHO WILL BE RESPONSIBLE FOR DATA MANAGEMENT IN YOUR PROJECT? 18

5. DATA SECURITY **18**

5.1. WHAT PROVISIONS ARE IN PLACE FOR DATA SECURITY (INCLUDING DATA RECOVERY AS WELL AS SECURE STORAGE AND TRANSFER OF SENSITIVE DATA)? 18
5.2. IS THE DATA SAFELY STORED IN CERTIFIED REPOSITORIES FOR LONG TERM PRESERVATION AND CURATION? 18

6. ETHICAL ASPECTS **19**

7. CONCLUSIONS **19**

8. DEGREE OF PROGRESS **19**

9. DISSEMINATION LEVEL **19**

1. Introduction

The deliverable D2.1 *Active data management plan produced (v1)* is part of task 2.1 *Data management plan.*

As part of making research data findable, accessible, interoperable and re-usable (FAIR principles), NetZeroAICT DMP includes information on: what data will be collected/re-used, processed and/or generated; which methodology and standards will be applied; the handling of research data during and after the end of the project; whether data will be shared/made open access and how data will be curated and preserved (including after the end of the project).

NetZeroAICT Data Management Plan was developed within the first six months of the project and will be updated throughout the project, which an updated submission planned in M24. It is in line with Article 17 and Annex 5 of the *Horizon Europe Grant Agreement*. As stated in Article 17 of the Grant Agreement, the beneficiaries must manage the digital research data generated in the action ('data') responsibly, in line with the FAIR principles and by taking all of the following actions:

(a) establish data management plan ('DMP') (and regularly update it);

(b) as soon as possible and within the deadlines set out in the DMP, deposit the data in a trusted repository; if required in the call conditions, this repository must be federated in the EOSC in compliance with EOSC requirements.

(c) as soon as possible and within the deadlines set out in the DMP, ensure open access — via the repository — to the deposited data, under the latest available version of the Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC0) or a licence with equivalent rights, following the principle 'as open as possible as closed as necessary', unless providing open access would in particular:

  (i) be against the beneficiary's legitimate interests, including regarding commercial exploitation, or

  (ii) be contrary to any other constraints, in particular the EU competitive interests or the beneficiary's obligations under this Agreement; if open access is not provided (to some or all data), this must be justified in the DMP;

(d) provide information via the repository about any research output or any other tools and instruments needed to re-use or validate the data.

It is necessary to mention that the disclosure of the project's data should never jeopardize the project's main objective and the potential protection of generated intellectual property (e.g., patent, product design) and further industrial application. The confidentiality obligations, the security obligations, and the obligations to protect personal data still apply. If there would be any conflict, the data will not be made openly accessible. All partners contribute to defining the data that will be generated in the project and assess which data can be made public. In case some research data will not be made openly accessible, the Data Management Plan will provide an explanation for it (IPR issues, exploitation, etc.). Overall, the access to the data and research results will always follow the rule "as open as possible, as closed as necessary".

2. Data Summary

NetZeroAICT aims to reduce the environmental footprint created by computerized tomography (CT), a common diagnostic imaging modality, by developing state-of-the-art trustworthy AI to synthesize 'CT Digital Contrast' and reduce the global reliance on iodinated radiocontrast media (RCM). CT is one of the most common diagnostic imaging examinations in healthcare and is estimated to result in 3 mega tonnes of CO2 emission globally (9kg/scan). Iodinated RCM required for CT scans further accounts for 3% of pharmaceutical waste released into the wastewater (estimated 200,000 tonnes of iodine/year). We aim to reduce 30% of this environmental footprint in Europe by 2033 by deploying the digital contrast AI in all European countries.

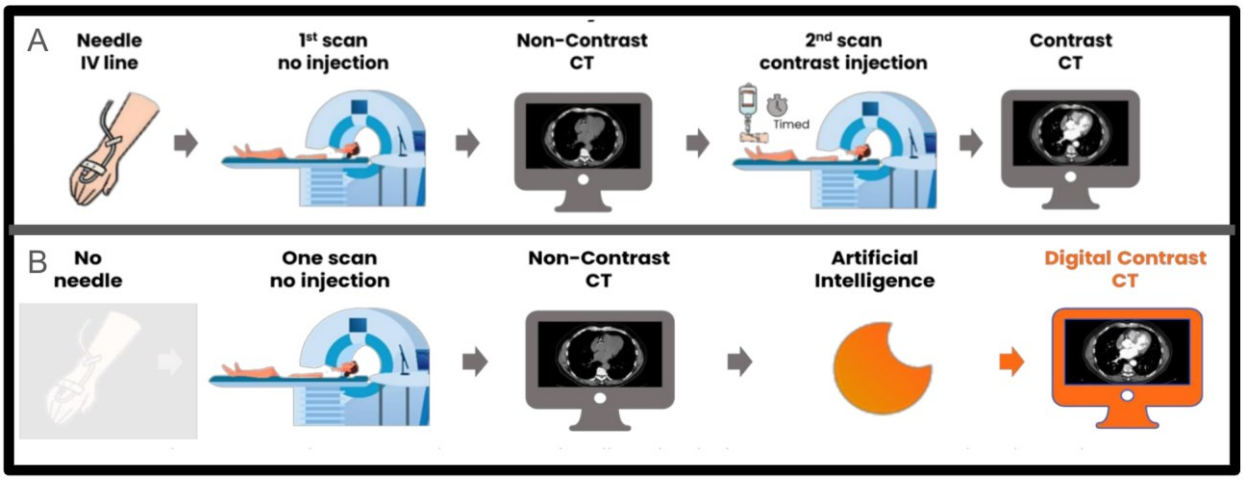

Figure 1: A) Conventional Computerized Tomography (CT) workflow. A needle and intravenous (IV) line is set on the patient, a first scan without the administration of contrast agent, rendering a Non-contrast enhanced CT, for the second scan contrast agent is injected prior to the scan rendering a contrast enhanced (CE) CT scan. B) In the workflow proposed there is no needle and IV line, and no second scan. An AI network is instead used to render the Digital Contrast CT.

Within the NetZeroAICT project, we will curate a data repository containing imaging data from around 1,000 000 patients will be created (with a mix of ~500k Contrast Enhanced (CE) and 500K non-CE imaging data). This date repository will be provided by the AICT Consortium, which is the core partnership leading to the expanded NetZeroAICT Consortium (as described in the original proposal).

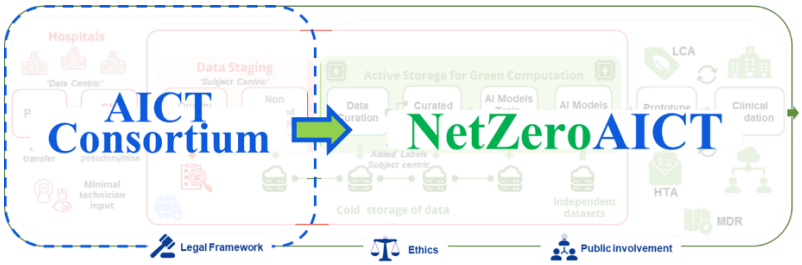

Figure 2: Transition from AICT towards NetZeroAICT consortium.

These patients have undergone a CT examination in one or more of the five clinical areas: head/neck, chest, abdomen, pelvis and limbs. Imaging data is collected from the clinical partner sites in order to develop, train and validate AI algorithms to eliminate the need for RCM and instead generate "Digital Contrast CT" from Non-CECT images.

The project has nine objectives, separated into a number of deliverables that will be delivered from seven work packages (WPs). The purpose of this document is to outline the complete set of data-related activities

of the NetZeroAICT project. The document will be updated regularly based on modifications and new details related to the NetZeroAICT data.

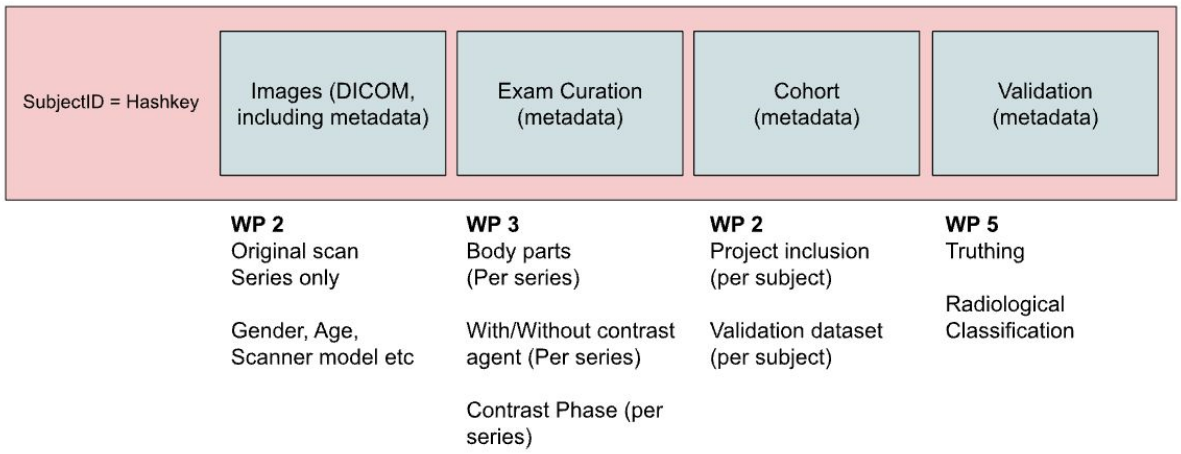

Figure 3: Schematic diagram of the image and metadata in three of the different work packages (WP).

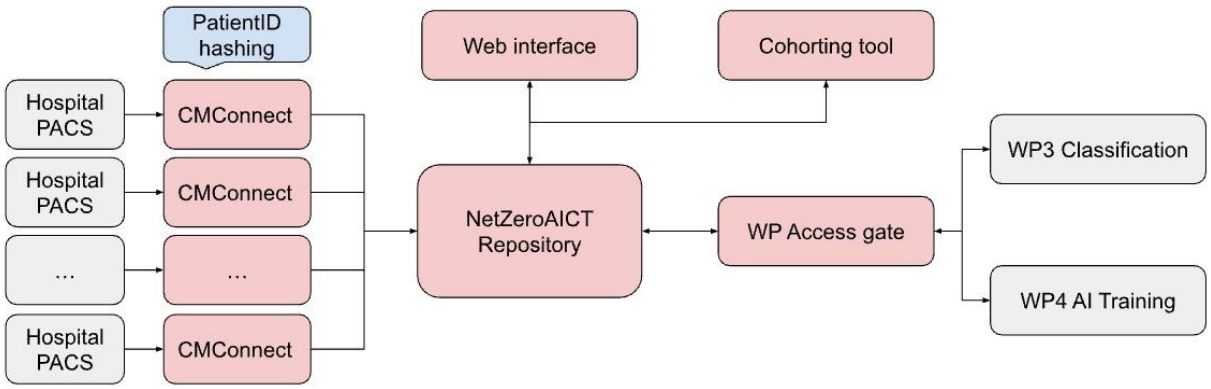

Figure 4: Schematic diagram of the data flow within the project. Each clinical partner sends their image data using the CMConnect where the pseudonymization is being performed, to the project data stage, in which the different work package projects may interact with the data.

2.1. Purpose of the data collection/generation and its relation to the objectives of the project

The main input data in this project consist of clinical imaging data acquired using the diagnostic imaging modality CT. Data has been acquired by the clinical partners in the project in order to diagnose or confirm a diagnosis in five clinical areas: head/neck, chest, abdomen, pelvis and limbs.

In order to fulfil the objective a large enough cohort with a wide spread in demographics, as well as different acquisition variables such as scanner, vendor and model (i.e. different hardware), software versions as well as acquisition parameters (e.g. tube current, slice thickness, orientation of the body, timing of scan etc – contrast phase) is needed. In order to remove the need for the contrast enhanced scan, we will train an AI network on the CE and non-CE pair.
The initial data ingestion is crucial for the success of the project, as well as the deliverables in T2.2.

- In WP 2 a repository containing 1 000 000 CT scans will be set up for AI training and validation.
- In WP 3 the data will be further classified by an AI algorithm developed within the WP: this will later on be used to classify the following data ingested in the repository. The classification performed along with the already existing metadata will facilitate the cohorting database.

- In WP 4 and 5 the data ingested in the repository will be used to train the AI network to generate the “Digital CT Contrast”, using a green compute solution.

2.2 Types and formats of data the project will collect and generate

In the NetZeroAICT project, imaging data will be collected, transferred and stored according to the **D**igital **I**maging and **Co**mmunications in **M**edicine or DICOM standard[1]. Images stored in the DICOM format, hold both the pixel values that constitute the actual image and a vast amount of metadata (“dicom tags”), along with unique IDs (UIDs). Each Patient, Study, Series and Instance hold a unique identifier. Figure 5 outlines the hierarchy of the DICOM standard.

The hierarchy of the DICOM standard is PATIENT, STUDY, SERIES and IMAGE (or INSTANCE)

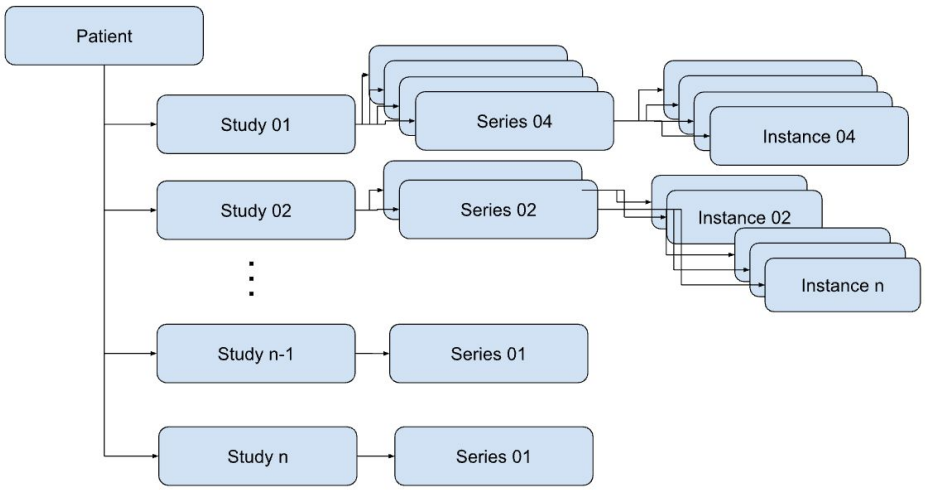

Figure 5: The hierarchy of DICOM starts with a patient, each patient can have one or several studies. Each study may consist of one or more series and each series contains one or more instances or images.

2.3 Data protection

The NetZeroAICT Project involves handling a large amount of health-related data (as defined in article 4(15) of the GDPR), which are considered special categories of data (as per article 9 of the GDPR) and is subject to multiple restrictions and requires careful consideration of the associated risks to protect the privacy rights and freedoms of the data subjects and patients. Furthermore, the project aims to process data from various sources and jurisdictions, which require compliance not only with the GDPR but also with other local regulations, such as Australia's Privacy Act 1988, Brazil's LGPD, or the UK's UK GDPR.

For the purpose of archiving the above, a system is created through that installation of a local piece of software (CM-Connect), that acts as a proxy that ensures that the local identifiable personal data remains at the local infrastructure (e.g. hospital) and all data is pseudonymized (de-identified) and minimized before being transferred to the central NetZeroAICT repository, ensuring compliance with the local laws and requirements and ensuring data protection and privacy risks minimization in the NetZeroAICT repository (depicted in Figure 6 below). A detailed description of the data privacy and IT-security assessment has been described in deliverable D1.1.

[1] https://www.dicomstandard.org/

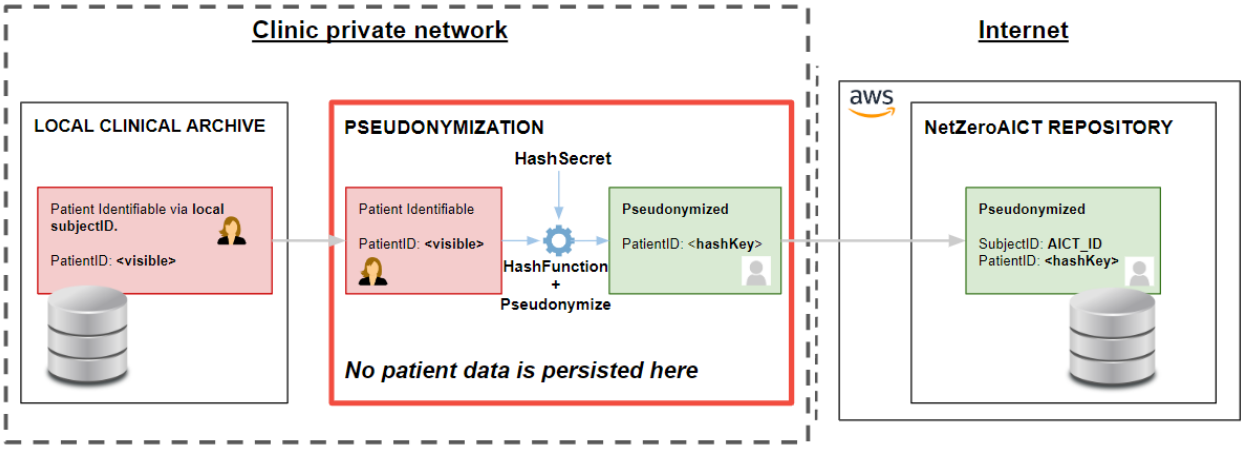

Figure 6: Schematic diagram of data flow with pseudonymization from clinical partner to NetZeroAICT repository.

For the avoidance of doubt, "Pseudonymization" commonly refers to a de-identification method that removes or replaces direct identifiers (for example, names, phone numbers, government-issued ID numbers, etc.) from a data set, but may leave in place data that could indirectly identify a person (often referred to as quasi-identifiers or indirect identifiers). Applying such a method, and nothing else, might be called "simple pseudonymization".

Frequently, security and privacy controls designed to prevent the unauthorized re-identification of data are applied on top of simple pseudonymization to create strong pseudonymization.

Pseudonymization is defined in the GDPR (art. 4.5) as: "*the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person*".

When we refer to pseudonymization, we assume some form of strong pseudonymization that would meet the GDPR definition and include the controls or "technical and organizational measures" referred to above.

In this regard, pseudonymized data leaves all of the indirect identifiers intact, and individuals could potentially still be identified using the indirect identifiers.

Notwithstanding the above, the GDPR still considers pseudonymized data as personal data, as long as other relevant Data Protection Authorities, such as the UK Information Commissioner's Office (ICO) that considers pseudonymized data to be personal information. Hence, pseudonymized data remains "personal data'' and is therefore subject to the requirements of the GDPR. But the GDPR provides some regulatory incentives to adopt pseudonymization, and there are therefore some significant benefits to employing it. Specifically, pseudonymised data helps the NetZeroAICT Consortium to meet the GDPR requirements, but it does not fully release the organization from them.

Strong pseudonymization methods are being used, and that these methods would be considered acceptable by European Data Protection Authorities (DPAs).

Furthermore, we need to take into account that the PSI establishes that "it is important to emphasize that should some such datasets nevertheless be published and made available, after a careful assessment of risks and benefits, the disclosure and any further reuse must be made in full compliance with data protection law [...]. This is because these data, despite some (sometimes very significant) measures taken to decrease the risks of their re-identification, nevertheless continue to be considered personal data".

In this regard, the NetZeroAICT consortium is aware that most of the known re-identification attacks on clinical, administrative, and survey data were done using indirect identifiers. And recently, an academic researcher gathered information from newspaper articles about vehicle accidents to re-identify individuals in a hospital discharge database, using information such as the year of birth, date of the accident, the hospital the individual went to, and where they lived.
There is evidence that basic demographics, such as the date of birth and the postal code, can uniquely identify almost all of the population. For example, these two pieces of information are unique identifiers for almost all of the population in Canada, and the Netherlands, and a high percentage of the population in the United States. These basic demographics are easy to get from public sources and can be used to re-identify individuals.

Therefore, Collective Minds applies anonymization standards that go beyond pseudonymization to cover the indirect identifiers. As data sharing initiatives take flight, there is an urgency to address the standards gap before data go out of the door.

As stated by the ARTICLE 29 DATA PROTECTION WORKING PARTY (hereinafter WP29) “Anonymisation may be a good strategy to keep the benefits and to mitigate the risks. Once a dataset is truly anonymised and individuals are no longer identifiable, European data protection law no longer applies”.

Collective Minds applies pseudonymization and data minimization techniques to further deidentify the data before it is processed/stored/shared, and in a manner that will meet multiple requirements: (a) ensure that the probability of re-identifying individual patients is small, (b) meet the regulatory and legal thresholds for what is an anonymized dataset, and (c) ensure that the anonymized data quality is sufficiently high to allow meaningful analysis.

In addition, the Consortium has its own privacy management program that guarantees compliance with the data protection and privacy assurance life cycle. This includes performing a Data Protection Impact Assessment (DPIA) and risk assessment, analyzing the legal basis of the data processing for secondary purposes, and establishing a compliant data protection framework based on the GDPR, the most stringent regulation to address security incidents, data breaches, and any data subject request or exercise of privacy rights.

2.4 Will you re-use any existing data and how (including data protection aspects)?

The clinical images that constitute the fundamentals of this project are acquired by each clinical partner for a primary purpose in order to e.g. diagnose, confirm a diagnosis or track disease or treatment progression. This data is stored within each hospital’s picture archiving and communication system (PACS). For the performance of the Project, the data initially collected by each clinical partner, and stored in each PACS, will be transferred to the project data repository and **re-used in the project, which will be processed for secondary purposes**.

In this regard the processing of data conducted on the Collective Minds Platform for the performance of the Project is considered a **secondary use of personal data**, so it is **conducted regardless of the initial legal basis** on which each controller (e.g. institution, hospital, etc.) relies on for processing and collecting the data uploaded to the Platform.

In these cases, the use of pseudonymization techniques enables data processing for secondary purposes without the need to obtain the explicit consent of the data subjects.

Furthermore, whenever necessary, Collective Minds applies further anonymization techniques on the pseudonymized datasets (in local environments) to obtain strongly pseudonymised data that will enable

safe further data processing minimizing the privacy risks of the data subjects to the maximum extent possible.

For the analysis of the risks involved with the sharing of data we are taking into account the ideas and recommendations provided in the Opinion 06/2013 on open data and public sector information (June 2013) (hereinafter 'PSI Opinion') reuse and the Opinion 05/2014 on Anonymisation Techniques (April 2014) of the ARTICLE 29 DATA PROTECTION WORKING PARTY (hereinafter WP29), along with the Statement of the WP29 on the impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the EU. This is because the recommendations for data sharing within the public sector encompass the highest standards to be taken into account for data sharing and processing for secondary purposes.

As reviewed in more detail in Opinion 3/2013 of the WP29 on purpose limitation, assessing whether further processing of personal data is incompatible with the purposes for which those data have been collected requires a multi-factor assessment. Account shall be taken in particular of:

(a) the relationship between the purposes for which the personal data have been collected and the purposes of further processing;

(b) the context in which the personal data have been collected and the reasonable expectations of the data subjects as to their further use;

(c) the nature of the personal data and the impact of the further processing on the data subjects;

(d) the safeguards applied by the controller to ensure fair processing and to prevent any undue impact on the data subjects.

As the relationship between the purposes (a) and the main technical IT security safeguards (d) are described hereinabove. To safely and lawfully share and process the data, we identify two general sources of risk: **1)** The **context** (b) of the data sharing; and **2)** the **content** and/or **data** stored on the platform, in which we need to analyze both the nature and the specific safeguards applied on the data to minimize the risk and the impact of the processing (c) and (d).

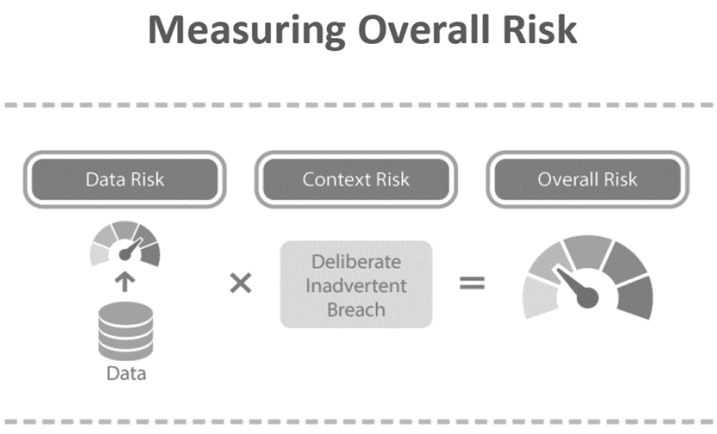

Figure 7: Diagram of the risk approach taking into account both the Data Risks and the Context Risks

The context represents the security, privacy, and contractual controls that are in place. For example, one context can be a public data release (e.g., an open data initiative) or, by contrast, a researcher who analyzes

the data in a very secure enclave. These are two very different contexts and the risk of (re)identification is different in each of these, even for the same data.

As established by the WP29 in the PSI Opinion:

"[...] it is important to carefully consider what measures - including both legal and technical measures - could be put in place to help ensure that data protection concerns [...] will be addressed. It is particularly important to consider how re-users will access the data [...]. In this respect it is crucial what additional security controls will be put in place, such as, for example, a [...] verification system to prevent automated access and minimise the risk of harvesting an entire database. Using specific technical measures could help reduce misuse of personal data and negative impacts on data subjects that otherwise could be made possible by unlimited and unconditional access of reusers to entire datasets [...]".

Furthermore, in the PSI Opinion the WP29 recommends that:

The terms of the licence to re-use [...] should include a data protection clause, whenever personal data are processed including also situations where anonymised data sets derived from personal data will be made available for reuse;

Where appropriate, public-sector bodies should ensure that personal data are anonymised, and license conditions specifically prohibit re-identification of individuals and re-use of personal data for purposes that may affect the data subjects;

As described above, Collective Minds applies legal and technical measures to the following context:

- Facilitating and enhancing collaboration between health professionals, professors (and students), physicians (medical doctors), and researchers is necessary and desirable. Furthermore, this collaboration must include the sharing of knowledge of Real-World Data that enhances the knowledge and provides real impact on those professionals' activity knowledge and training.

The main contextual risks are:

- The sharing of sensitive data, such as health information.
- Having no control over who has access to the data and what further processing activities are carried out on it (processing for secondary purposes that are unrelated to the primary purposes).
- An attempt to identify or re-identify a data subject.
- The processing of personal data for malicious, unlawful, and/or unethical purposes, such as identifying the patient or uncontrolled data processing.

As a result, the Consortium, through Collective Minds, takes the following measures:

- A closed platform, where only registered users that are part of the research Consortium can access and process the datasets: The Collective Minds Platform.
- Establishing technical and operational measures on the Collective Minds Platform to avoid unauthorized access, preventing rogue/malicious behavior or data processing outside the controlled environment of the Platform.
- Control who can register as a User of the Collective Minds Platform through a vetting procedure to verify that only trusted and relevant healthcare professionals, professors, physicians (medical doctors), and researchers can be registered as users of the Platform.

- Full traceability and access control.
- Users must sign and accept legal undertakings (Terms and Conditions) before registering on the Platform. The Terms and Conditions include, among other things:
    - A confidentiality / non-disclosure agreement regarding the data and the content of the platform, including personal data and or anonymized data and any intellectual property rights.
    - There are several data notices and a privacy policy that must be read, followed, and accepted.
    - To ensure that the user fully understands its obligations beyond the terms and conditions and general terms, information to the user is provided, generally in the form of warnings, informative messages, and guidelines.
    - The prohibition of sharing or processing data outside the Platform control and traceability.
    - The prohibition of processing data for unauthorized, unlawful or unethical purposes.
    - It is prohibited to attempt to or identify a data subject, or to conduct any activities of re-identification.
    - It is possible to legally enforce any of the above measures if they are violated.

Aside from the contextual risks described above, technical and organizational measures are implemented on the **Platform's data (content)**, before being transmitted and stored in the repository, in order to avoid a potential or actual impact of the rights, freedoms, and dignity of data subjects.

As provided by the WP29, in the Opinion 05/2014 on Anonymisation Techniques (April 2014)*: "'Open data' may provide clear benefits for society, individuals and organizations, but only if everybody's rights are respected to the protection of their personal data and private life".*

Before entering into the analysis of the measures regarding the data risks (content), it is important to clarify that a health dataset will have two types of variables that we care about from a privacy perspective: (a) direct identifiers, and (b) indirect identifiers.

Direct identifiers are details such as names, addresses, social insurance numbers, telephone numbers, email addresses, and any other unique identifiers. These direct identifiers are typically removed from the dataset when it is shared for secondary purposes. The unique identifiers, such as a medical record number, would be converted to a pseudonym so that it can still be used to relate all of the records that belong to the same patient. When you do all of these things to protect against re-identifying individuals from these types of variables, the dataset is considered pseudonymized.

As a result of the procedure described above, the Collective Minds Platform does not collect, process, store, or share any direct identifiers, as it is designed as a technical system that ensures all data uploaded to the Platform is strongly pseudonymized before transmission.

Therefore, **this make secure processing for secondary purposes possible** since, following the latest recommendations from the EU Data Processing Authorities, the data is processed through a 'proxy' function (CM Connect) that enables to perform and apply operations and measures on the data, within the safe and controlled environment of each data controller before uploading the data into the Collective Minds Platform.

2.5. What is the origin of the data?

The NetZeroAICT repository will receive CT data from and international consortium (AICT Consortium, which includes clinical sites from: France, Greece, Poland, Belgium, England, Scotland, Australia and Brazil), see Table 1 for further information. Each of these hospital groups will share their existing CT clinical archive in a 'pathology agnostic' manner. The repository will therefore capture the most diverse geographic, populational and disease context as compared to the other research imaging datasets in existence. We will have 1 million cases in the NetZeroAICT repository. The imaging data will be provided by the clinical partners of the project listed in Table 1.

Table 1. List of clinical partners providing data to the NetZeroAICT consortium

Clinical partner	Country
GZA	Belgium
Poznan	Poland
Heraklion	Greece
USP	Brazil
Glasgow	UK
Nice	France
AZ Sint-Jan	Belgium
UFF - Universidade Federal Fluminense	Brazil
Leicester	UK
Healius	Australia

2.5. What is the expected size of the data?

In this project the aim is to include a cohort of 500 000 patients.

D2.3 100 000 data sets transferred to repository (M6)
D2.4 500 000 data sets transferred to repository (M24)
D2.5 1 000 000 data sets transferred to repository (M36).

We define the CE as one data set and the non-CE as one with an average size of 500 MB or 0.5 GB.

The base data is estimated to be 1 000 000 data sets * 0.5 GB = 500 000 GB or 500 TB by M36.

2.6. To whom might it be useful ('data utility')?

The data collected within the NetZeroAICT project will be useful for:

1. The NetZeroAICT Consortium Partners: They will generate new scientific developments and generate research impact (publications), IP and innovation.

2. The current and future AICT partners. This is applicable to the extent that the NetZeroAICT is a subproject of the wider AICT project and Consortium.
3. The repository created will be a valuable asset in the future as well for other research projects in need of data to design, train and validate AI algorithms. Please see the FAIR data section for details.
4. Ultimately, patients and the public in general, whether they provided pseudonymised health information or not. If successful, this project will reduce waste, have a positive environmental impact, and generate a positive outcome for patients that will not require contrast to diagnose their images, which may be harmful and painful.

3. FAIR data

3.1. Making data findable, including provisions for metadata

3.1.1. Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

Data used in this project primarily consists of medical imaging data in accordance with the DICOM standard. There is a unique ID assigned to each subject, which is specific to the project. Site-specific hashing is applied to create the subject identifier which clinical partner sites can use to verify whether a patient is included, but the identity remains unknown for everyone else. A unique identifier is also assigned to every study which is affiliated to the subject. A unique series identifier is assigned to each series associated with a study in accordance with this schema. Moreover, the NEMA 2021b standard is followed for the minimization of DICOM metadata.

3.1.2. What naming conventions do you follow?

A globally unique 64-character identifier is assigned to each included subject. The identifier is the hash output (SHA-512/256) of the locally known patient ID in combination with a clinical site-specific secret key.

3.1.3. Will search keywords be provided that optimize possibilities for re-use?

We will extract a selection of DICOM tags and store them along with the subject ID as part of the data ingestion process. This enables the advanced cohorting tools to identify groupings via key-word matching. In future continuations, these key-word searches could be reused since they are not project-specific. To allow the identification of source and derivative materials, data generated during the project life cycle will also be saved and searchable.

3.1.4. Do you provide clear version numbers?

Yes, all data handling will be governed by a clear versioning schema. As a result, all processing can be tracked and traced.

3.1.5. What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Classification data will be generated as part of the project. This data will be required to have the necessary information to uniquely link to specific subject inclusion.

3.2. Making data openly accessible

3.2.1. How will the data be made accessible (e.g. by deposition in a repository)?

As explained in sections 1.3 and 1.4, when performing “open data” research projects originating from sensitive data, such as health data, it is necessary to find a balance between privacy and data protection of the data and make data as accessible and transparent as possible to researchers, and even to the public.

NetZeroAICT partners commit to make the research data as open as possible while taking confidentiality and IP protection into account. Relevant datasets will be uploaded to trusted repositories that will be agreed and implemented at the beginning of the project´s implementation. The exploitation opportunities, the protection of generated IP, the confidentiality obligations, the security obligations and the obligations to protect personal data will always be considered before making the data openly accessible and will be defined in the Consortium Agreement.

Generated or re-used data submitted to repositories will follow the FAIR (Findable, Accessible, Interoperable and Re-usable data) principles. This approach will enable the target user and scientific community to derive the intended benefits from the project outcomes. In addition to making the peer-reviewed scientific articles and research data openly accessible, we will also put effort into making other project results openly accessible as part of a nuanced IPR protection and public good asset mobilization plan developed via appropriate stakeholder engagement (WP1,6,7,8).

These will include designs, preprints, blue prints, data sets, detailed engineering plans (e.g. of a demo plant), systems, programs, applications, books, tools, software, protocols, algorithms, workflows, electronic notebooks, models, simulations, etc. and their access will be granted via public deliverables, project website, designated online repositories, etc. This will ensure the performance of further open data activities that may be available for external parties, such as the public and patient associations to ensure adequate transparency and accountability.

3.2.2. What methods or software tools are needed to access the data?

Through the Collective Minds Radiology research platform, NetZeroAICT data will be accessible for search and visualization. To access this feature, you will need a modern web browser and a personal login.

A role defined by the consortium will determine the scope of access to the individual person. The sub-processors will be able to handle data in accordance with the terms of the individual contracts.

As explained above, certain results based on aggregated data (not personal data) may be made available in a wider spectrum without needing any additional software or tools.

3.2.3. Is documentation about the software needed to access the data included?

Yes, documentation detailing the process for accessing the data is included.

3.2.4. Is it possible to include the relevant software (e.g. in open source code)?

The included and derived software will be made available in accordance with the exploitation plan.

However, the software to manage and store the data (repository) is proprietary and provided as a Software as a Service. Therefore, this specific part of the code is not provided in open source.

3.2.5. Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

All data will be stored within a logically segregated repository in the Collective Minds Radiology virtual private cloud. This is hosted by Amazon Web Services (AWS) in Frankfurt, Germany.

3.2.6. Have you explored appropriate arrangements with the identified repository?

As a proven and tested solution for large data storage and processing within the research community, Collective Minds Radiology provides all the necessary functionality to handle repository data.

3.2.7. If there are restrictions on use, how will access be provided?

In accordance with a consortium-defined role schema, access will be provided and granted on an individual basis. The purpose of this is to control how and to what extent data is made available for a single individual.

3.2.8. Is there a need for a data access committee?

The NetZeroAICT general assembly will create a data governance committee to take decisions on granting and revoking access to the repository.

3.2.9. Are there well described conditions for access (i.e. a machine readable license)?

There will be a detailed description of all data access before it is facilitated.

3.2.10. How will the identity of the person accessing the data be ascertained?

Access to the repository is granted by the repository administrators and is based on a personal login.

Signing up for access requires an email address, but must be verified by an existing privileged repository member.

3.3. Making data interoperable

3.3.1. Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

The classification data will be stored using standard formats and include the necessary meta information for linking to the subject ID and DICOM studies.

3.3.2. What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

DICOM is the primary standard that will be used for storing data in this project.

3.3.3. Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

DICOM is the primary standard that will be used for storing data in this project.

If project-specific ontologies are used, mappings will be provided.

3.4. Increase data re-use (through clarifying licenses)

3.4.1. How will the data be licensed to permit the widest re-use possible?

The plan is to establish a concept where consortia affiliation will be the basis to earn a data access license. The license will come with responsibilities and benefits.

Before reuse of data, the data access board of the NetZeroAICT (or AICT) consortium must approve such reuse.

One long term goal is to maximize the re-use of the data generated in the NetZeroAICT project.

3.4.2. How long is it intended that the data remains re-usable?

The value of the data collected, and the new data generated has a long term value and is expected to be valuable for at least 20 years.

3.4.3. Are data quality assurance processes described?

Data quality assurance processes are currently being developed.

4. Allocation of resources

4.1 What are the costs for making data FAIR in your project?

The costs for making data FAIR will be updated at M36.

4.2 How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

A long-term sustainability plan will be developed, where these costs will be factored in.

4.3 Who will be responsible for data management in your project?

Each consortium partner has a role in data management during the data processing life cycle, but Collective Minds Radiology will be ultimately responsible for data management on behalf of NetZeroAICT Consortium.

5. Data security

5.1. What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

As the organization responsible for maintaining the repository, Collective Minds Radiology adheres to all modern data security and redundancy standards when handling research data.
A detailed description of the data privacy and IT-security assessment has been described in deliverable D1.1.

5.2. Is the data safely stored in certified repositories for long term preservation and curation?

Yes, data is safely stored in AWS S3 and AWS S3 Glacier that follows up to date standards such as SOC 1, SOC 2, SOC 3, ISO 27001, ISO 27017, ISO 27018, and PCI DSS.

6. Ethical aspects

There is a dedicated team that represents 1) legal, privacy and data protection 2) Ethics and 3) patient engagement, that is devoted to specifically analyze, oversee and ensure compliance with the ethical standards, assess ethical risks and minimize them.

This taskforce will perform ethics and fundamental rights impact assessment during the performance of the project. Furthermore, an Ethics Committee will also be in charge of ensuring ethical aspects are duly considered.

7. Conclusions

The Data Management Plan identifies the data sets that will be collected, processed, re-used or generated by the NetZeroAICTproject. It specifies which data will be generated, which methodology and standards will be followed, whether and how the data will be exploited and/or made accessible for verification and re-use, and how the data will be curated and preserved.

The initial version of the Data Management Plan is provisional and will be updated over the course of the project. The updated version of the Data Management Plan will be submitted as D2.2 by M24.

8. Degree of progress

The deliverable is 100% fulfilled. An Updated Data management plan will be submitted as D2.2 by M24.

9. Dissemination level

This document is publicly available.