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 D2.3 =

100k CT datasets transferred to repository (V1)

Due date of deliverable: 30/11/2024
Actual submission date: 29/11/2024
Requested revision submitted: 25/11/2025

Work Package: WP2
Responsible Work Package Leader: Anders Nordell, CMRAD
Responsible Task Leader: Jonatan Eriksson, CMRAD
New responsible task leader: Carlos Santin , CMRAD

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)
Carlos Santin	CMRAD	carlos@cmrad.com, +34 661 931 386

DOCUMENT CONTROL

Document version	Date	Change
V1.0	2024-11-25	Original
V1.1	2025-10-24	Additions due to review comments. Added strategy for FAIR data, extended charts depicting Body Part and Manufacturer distribution per site.

VALIDATION PROCESS

Reviewers		Validation date
Task Leader	Carlos Santín	28/11/2024 04/11/2025
Work Package Leader	Anders Nordell	28/11/2024 04/11/2025
Project Manager	Mariana Pacheco Blanco	28/11/2024 13/11/2025
Scientific Coordinator	Regent Lee	28/11/2024 13/11/2025
Coordinator	Anders Nordell	28/11/2024 13/11/2025

DOCUMENT DATA

Keywords	Data sets, repository
Point of Contact	Name: Carlos Santín Partner: CMRAD E-mail:carlos@cmrad.com
Delivery date	04/11/2025

DISTRIBUTION LIST

Date	Version	Recipients
28/11/2024	V1.0	To validators
29/11/2024	V1.0	To partners, EC via EU portal
04/11/2025	V1.1	To coordination team
25/11/2025	V1.1	To partners, and EC via portal

Executive Summary

This document reports the status of Deliverable: D2.3 - 100,000 CT datasets transferred to repository, in work package 2, associated with task T2.2.

Table of Contents

1. Introduction

In work package 2, *"WP2 – Repository, connectivity and dataflow"* the overarching objective is to develop and maintain a large-scale CT data repository for medical AI development. Deliverable D2.3 is connected to task T2.2 (Connection, Data transfer and storage).

The objective of D2.3 is to have 100,000 datasets transferred to the repository by month 12.

A public dashboard of the data transferred to the repository can be accessed at https://netzeroaict.eu/netzeroaict-dashboard/ (feature added after the submission of this original deliverable).

2. Results and Discussion

The main objective for the NetZeroAICT repository is to obtain 1 Million CT datasets for the training of digital CT contrast AI algorithms. One dataset is described as a DICOM series of a CT scan. This definition is the foundation of the following sections.

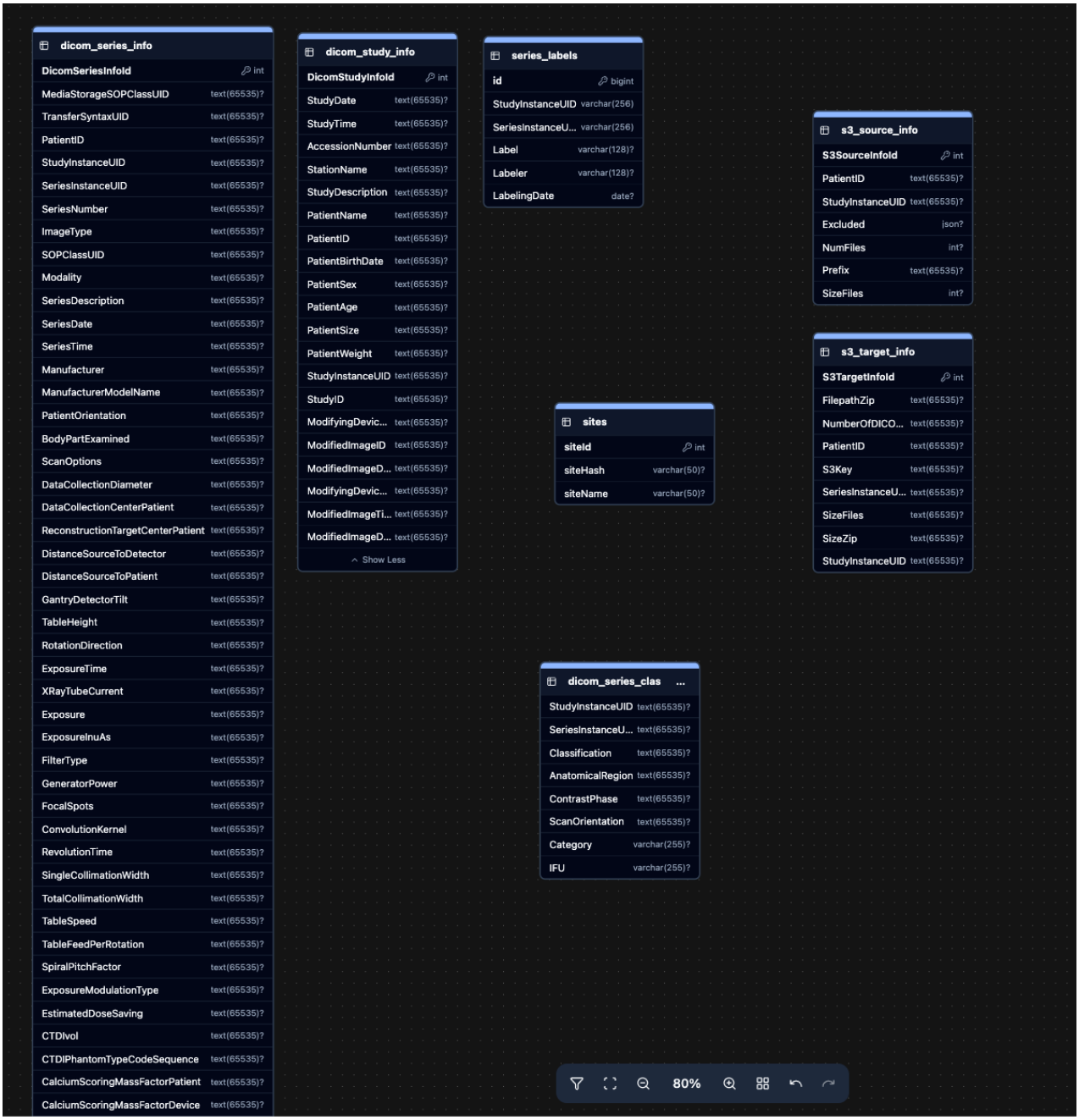
2.1. Metadata Harmonisation and Schema

The DICOM metadata information for each dataset is stored in the repository tracking database (Picture 1), including both technical and semantic information:

- **Technical metadata** (DICOM headers) for each dataset will be extracted and stored in the Database
- **Semantic metadata** will be extracted from the datasets in WP3 (Classification of data), mainly focused on Anatomical Region, Contrast Phase and Scan Orientation

Additionally, minimum provenance metadata is being stored as well, describing the clinical site data origin. Other information, like acquisition date, time, operator, etc, is wiped out in the pseudonymization step in alignment with the Data Management Plan (See D2.1)

The datasets are currently being identified based on the general DICOM study structure (Patient/Study/Series unique identifiers). All those identifiers have been pseudonymized before uploading to the repository (See D2.1)



Picture 1: Repository tracker database schema based on the relevant available DICOM tags

2.2. Adherence to FAIR principles

Adherence to FAIR principles is described in D2.1 (Data Management Plan)

The FAIR data strategy in this project ensures that all medical imaging data (primarily DICOM-formatted CT studies and related metadata) are managed according to the principles of being Findable, Accessible, Interoperable, and Reusable. The project adheres strictly to EU data protection laws (GDPR), institutional ethics approvals, and international standards like ISO 12052:2017 for DICOM. Each imaging study includes anonymized patient and acquisition data, while derived datasets may include processed images, quantitative biomarkers, and structured radiology reports. All datasets will be stored using consistent identifiers, standardized metadata schemas, and clear documentation of provenance and processing steps.

To operationalize the FAIR principles, metadata will be standardized and optionally assigned persistent identifiers to ensure findability. Access will be controlled through secure, role-based systems, allowing only anonymized or pseudonymized data to be shared. Interoperability will be supported by a repository API (and possibly a FHIR interface, standardized vocabularies such as SNOMED CT and RadLex, and export options in widely compatible formats like JSON or XML). Reusability will be promoted through detailed documentation, labeling systems for dataset management, and long-term preservation within consortium-managed infrastructure (potentially under open licenses for anonymized data).

2.3. Adoption of Standardised Ontologies

As seen above, all datasets ingested into the repository are described using a shared technical and semantic layer, which could be harmonised through existing imaging and biomedical ontologies. Specifically, and due to the intense focus on Medical Imaging, we plan to integrate mainly DICOM metadata schemas for core imaging and acquisition parameters.

In the future, and depending on the future uses of the data in the repository, the following could be considered:

- DICOM-SR (Structured Reporting)
- Specific Radiomics ontologies to provide semantic descriptors of image features, modalities, and findings.
- SNOMED CT, LOINC, and ICD-10 for clinical and diagnostic concepts associated with imaging data, if in a later phase the management of Clinical Data would be required.

This approach would enable cross-linking between datasets, imaging studies, and related clinical or omics information, and supports machine-actionable interoperability with other EHDS-compliant repositories.

2.4. Current Status of the repository

Data ingestion from the clinical partners have been ongoing and, as of the date of elaboration of this document, we have 154,732 datasets transferred to the repository.

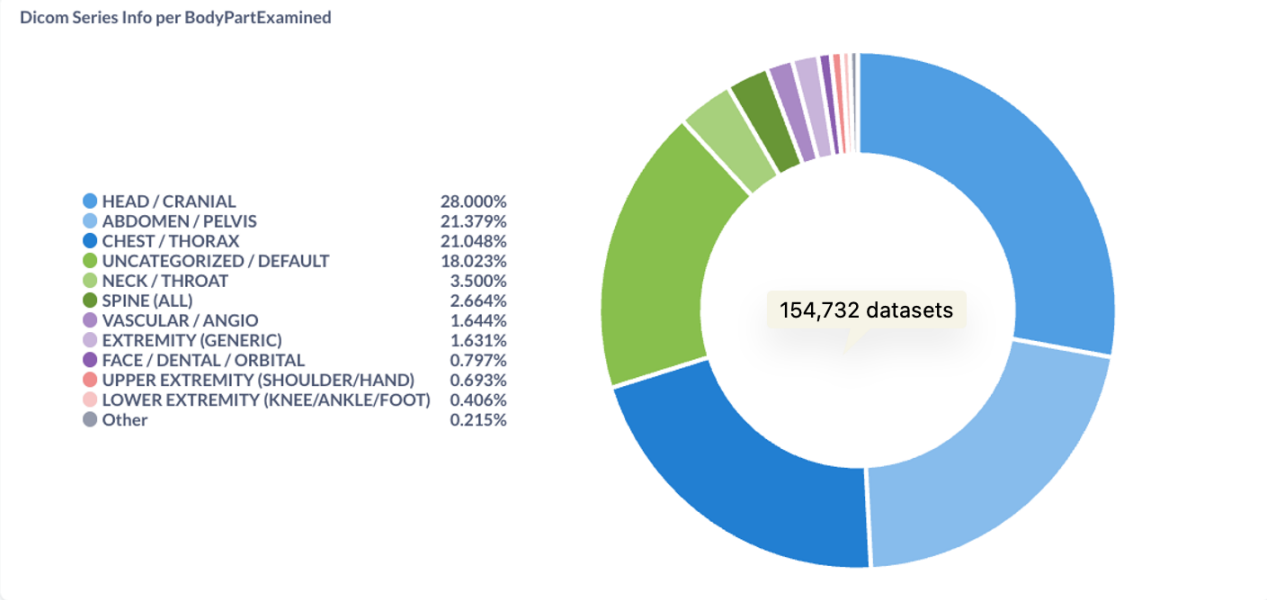


Figure 2. Pie chart showing the distribution of body parts examined in the available data sets.

Currently, there is a clear prevalence of Head, Abdominal and Chest datasets. The information above is based on the raw data received, and it is known to be relatively common to get DICOM series without categorization (this is the reason for having 18% of uncategorized data). The classification of those should be addressed in WP3. Some Indications for Use might not be sufficiently represented in the dataset, mainly IFU4 (visualization arteries in extremities). We will work with the sites to try to increase, if available, the amount of data on this IFU.

Count of SeriesInstanceUID	Column Labels								
Row Labels	AZ Sint-Jan Brugge	CHU	GZA	Healius	Heraklion	Leicester	Opole	USP	Grand Total
HEAD / CRANIAL	18.12%	1.21%	15.09%	4.88%	18.91%	40.33%	16.94%	17.81%	28.05%
ABDOMEN / PELVIS	12.76%	0.73%	37.38%	1.48%	22.13%	28.00%	29.88%	2.45%	21.42%
CHEST / THORAX	25.16%	0.81%	21.09%	14.07%	16.16%	20.29%	14.98%	31.03%	21.09%
UNCATEGORIZED / DEFAULT	23.70%	97.12%	13.42%	79.30%	35.98%	5.20%	19.49%	21.01%	18.08%
NECK / THROAT	4.68%	0.00%	4.86%	0.25%	2.98%	3.79%	0.47%	1.18%	3.51%
SPINE (ALL)	9.19%	0.02%	2.84%	0.00%	2.27%	0.16%	2.84%	0.06%	2.67%
VASCULAR / ANGIO	0.00%	0.06%	0.00%	0.00%	0.00%	0.04%	1.07%	26.10%	1.65%
EXTREMITY (GENERIC)	2.39%	0.00%	5.30%	0.00%	1.45%	0.34%	14.24%	0.00%	1.64%
FACE / DENTAL / ORBITAL	1.39%	0.00%	0.00%	0.00%	0.00%	0.95%	0.00%	0.00%	0.80%
UPPER EXTREMITY (SHOULDE	2.38%	0.05%	0.01%	0.00%	0.12%	0.23%	0.09%	0.00%	0.69%
LOWER EXTREMITY (KNEE/AN	0.24%	0.00%	0.00%	0.00%	0.00%	0.67%	0.00%	0.36%	0.41%
Grand Total	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Table 1. Distribution of body parts per clinical site in the available data sets.

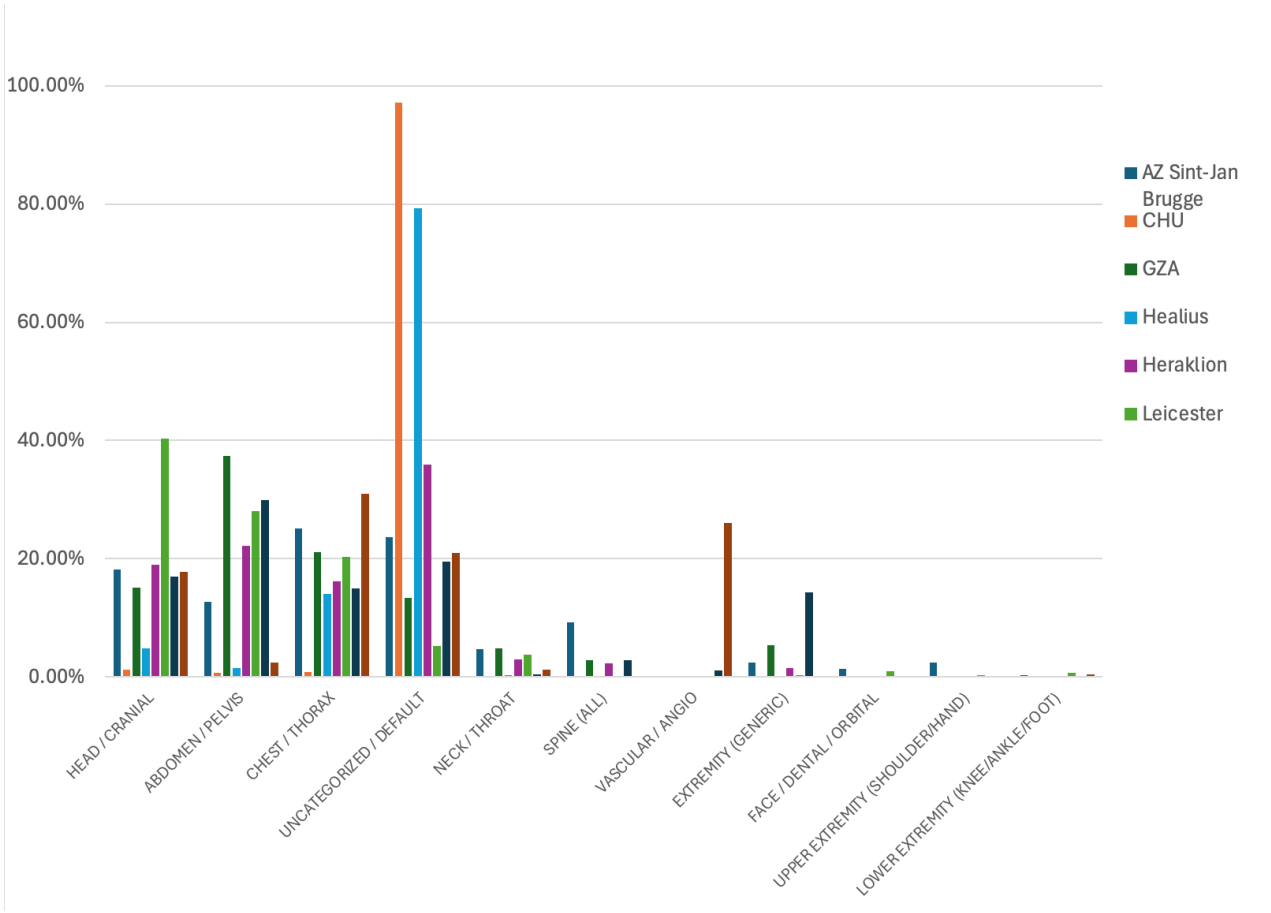


Figure 3. Distribution of body parts per clinical site in the available data sets.

Row Labels	Leicester	AZ Sint-Jan Brugge	Heraklion	USP	GZA	Healius	Opole	CHU	Grand Total
TOSHIBA	77.00%	0.01%	0.03%	45.91%	0.63%	0.03%	0.02%	0.00%	40.56%
SIEMENS	20.94%	9.85%	62.49%	0.67%	94.34%	78.27%	98.20%	0.00%	28.33%
GE	0.20%	74.84%	37.22%	0.25%	0.53%	15.65%	0.12%	0.35%	21.18%
OTHER	0.08%	13.15%	0.06%	0.04%	0.15%	0.27%	1.24%	0.04%	3.25%
CANON	1.77%	0.01%	0.15%	0.07%	0.00%	5.73%	0.00%	99.61%	2.70%
PHILIPS	0.00%	2.13%	0.05%	28.25%	4.35%	0.05%	0.43%	0.00%	2.46%
HITACHI	0.00%	0.00%	0.00%	24.81%	0.00%	0.00%	0.00%	0.00%	1.52%
Grand Total	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**

Table 2. Distribution of CT manufacturers per clinical site in the available data sets.

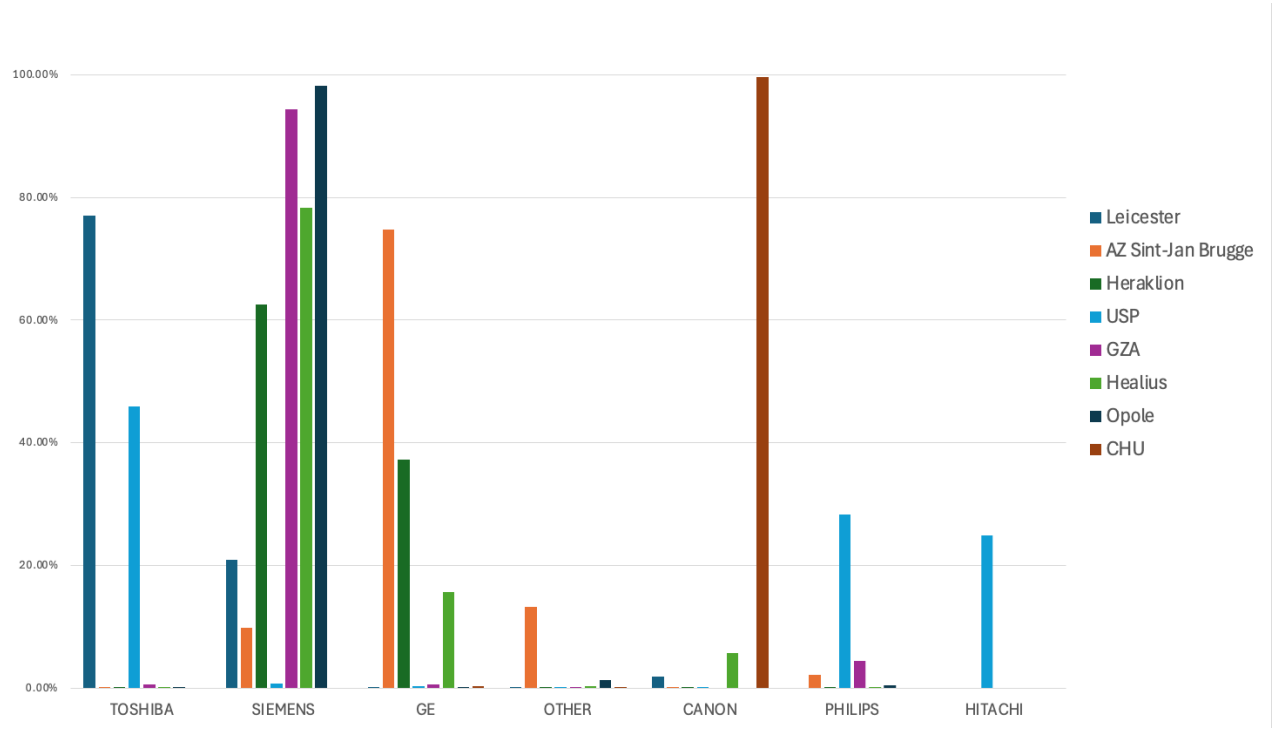


Figure 4. Distribution of manufacturers per clinical site in the available data sets.

The manufacturer data brings some interesting insights, like the fact that the repository might have a too-big prevalence from some manufacturers, and an underrepresentation of important players like Philips. This is an inherent bias due to the limited amount of sites in the consortium, that we will need to balance by trying to onboard new clinical sites covering the underrepresented modality manufacturers (Philips being the clearest example, due to the imbalance of market share against the representation in our repository)

A different angle for data analysis is the patient's dimension. By analysing the available data and timepoints per patient, it is clear that a big majority of patients (76.6%) have just one timepoint. 14.6% of the patients have 2 timepoints and 4.78% have 3 timepoints (Figure 5), which is consistent with the patient journey

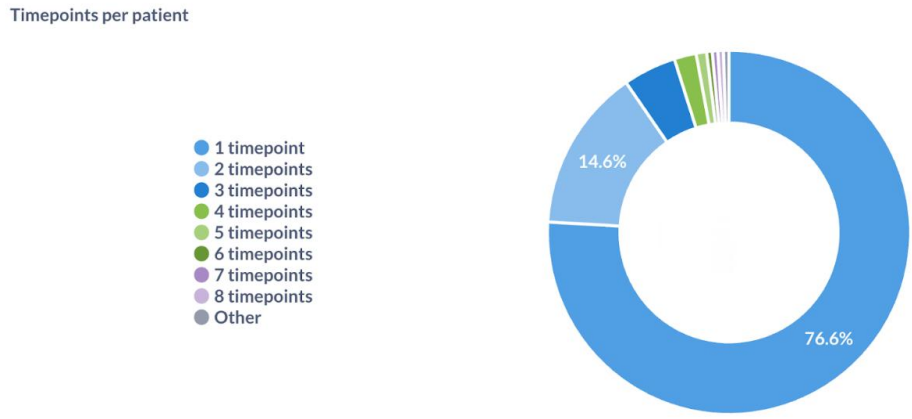


Figure 5. Distribution of the amount of timepoints per patient in the available data sets.

Additionally, the sex split in the repository data can be seen in figure 6, showing a balanced dataset in terms of sex:

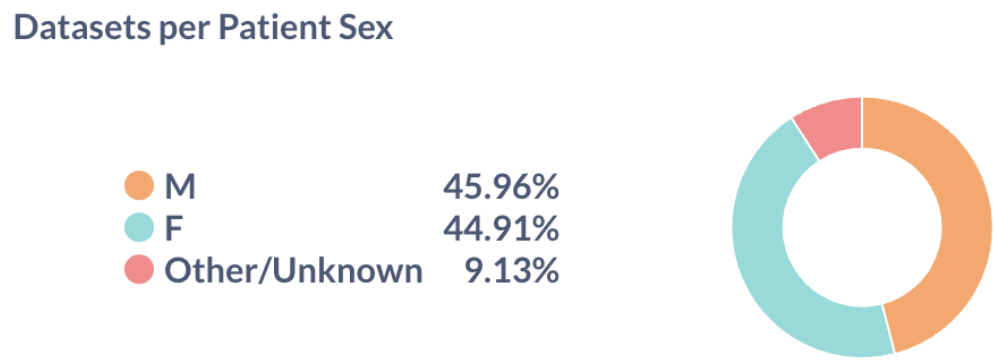


Figure 6. Distribution by patient sex in the repository data.

3. Conclusions

As more than 100k data sets are currently stored in the repository, deliverable D2.3 is considered fulfilled. There are currently some challenges with the repository data, mainly related to the limited number of sites participating in it. In order to have a bigger representation on device manufacturers more sites will have to be onboarded, covering the identified gaps. Regarding the distribution of body parts, and the potential lack of extremity scans, we will check with the sites to see if this distribution is common and if the number of scans of underrepresented body parts can be increased in the subsequent months.

1. Degree of progress

This deliverable is 100% fulfilled. Deliverable *2.4 100k, 500K, 1M CT datasets transferred to repository V2* due at M24 will follow up to update on the status of the data in the repository

2. Dissemination level

Public