This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 101017331

D6.1 The data model ecosystem for AI operations and modules implementation

Deliverable No.	D6.1	Due Date	28/02/2023
Description	This deliverable provides an in-depth description and analysis of the data ecosystem in the ODIN project. It offers a detailed explanation of the data provided by the reference use cases as well as an overview of the data collection, sharing and integration in the ODIN platform.		
Type	Other	Dissemination Level	PU
Work Package No.	WP6	Work Package Title	Data results interpretation and data integration services
Version	1.0	Status	Final

Authors

Name and surname	Partner name	e-mail
Francesca Manni	PEN	francesca.manni@philips.com
Anca Bucur	PEN	anca.bucur@philips.com
Daphne Plati	FORTH	daphni.plati@gmail.com
Fotis Konstantakopoulos	FORTH	fotkonstan@gmail.com
Labrina Pliakou	FORTH	labrinapliakou@gmail.com
Fanis Kalatzis	FORTH	tkalatz@gmail.com
Vaggelis Chaniotakis	FORTH	chanvag@ics.forth.gr
Manolis Tsiknakis	FORTH	tsiknaki@ics.forth.gr
Dimitrios I. Fotiadis	FORTH	dimitris.fotiadis30@gmail.com
Sofia Granda	INETUM	sofia.granda@inetum.com
Alex Barnadas	INETUM	alex.barnadas@inetum.com

History

Date	Version	Change
06/11/2022	0.1	Initial ToC – Release of a draft version with updated TOC
23/01/2023	0.2	Version 0.2 – Release of a draft outline of the contents to WP6 members from PEN for requesting feedbacks and content based on PEN inputs.
01/02/2023	0.2	Version 0.2 – Updated version with input from FORTH, PEN, INETUM
14/02/2023	0.3	Version 0.3 – Revision and integration from FORTH, PEN, INETUM
14/02/2023	0.4	Version 0.4 – Peer review
28/02/2023	1.0	Final version

Key data

Keywords	Data ecosystem, data collection, data anonymization, data flow, AI resources
Lead Editor	Francesca Manni
Internal Reviewer(s)	SSSA: Andrea Firrincieli, Neri Niccolò Dei, THL: Mike Karamousadakis

Abstract

This document is a deliverable of the ODIN project and provides information on the work done in Tasks T6.1 (Learning Capacities and Design of the Emergency Management Modules). This task is part of the work done in WP6 on creating a high-level ecosystem for AI operations, and the report covers the period from month 7 to 24.

This deliverable will report the ODIN data ecosystem, which include the data description and the data flow implemented in the ODIN project for enabling AI operation in the context of the ODIN AI strategy. An overview of the data ecosystem is provided per Reference Use Case (RUC), namely RUC A, RUC B, RUC C. The objective of this deliverable is to provide an in-depth analysis of the data ecosystem and related injections in the ODIN platform towards the deployment and implementation of AI operations and modules.

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Acronyms

AFE	Ankle flexion/extension
AI	Artificial Intelligence
CUB	Charité-Universitätsmedizin Berlin
DL	Deep Learning
EA	Ethical Approval
FL	Federated Learning
FORTH	Foundation for Research and Technology
GDPR	General Data Protection Regulation
HCSC	Hospital Clínico San Carlos
HIS	Hospital Information System
jpg	Joint Photographic Group
KERs	Key Enabling Resources
KFE	Knee Flexion/Extension
LHS	Learning Healthcare System Approach
ML	Machine Learning
png	Portable Network Graphics
RUCs	Reference Use Cases
SAA	Shoulder Adduction/Abduction
SFE	Shoulder Flexion/Extension
UCC	Utrecht Cardiovascular Cohort

Table of contents

List of tables

List of figures

1 Introduction

This deliverable is the result of T6.1 which focused on developing learning capabilities and designing emergency management modules. This task is part of the work done in WP6 on creating a high-level ecosystem for AI operations, and the report covers the period from months 7 to 24. In Section 2, the basis and background for the ODIN data space are presented, and in Section 3 the data resources required for building and implementing the AI resources are described.
In Section 4, that data ecosystem for each ODIN use case is shown and an overview of the data injection and integration in the ODIN platform, with the focus on implementing AI operations.
Section 8 concludes this deliverable and provide an outlook for the next steps as an outcome of the activities performed within the Task T6.1 Learning Capacities and Design of the Emergency Management Modules).

In the healthcare sector, data-driven innovation is currently led by research in artificial intelligence (AI) and by the development of highly accurate machine learning (ML) and deep learning (DL) models. These models can learn from large datasets and support clinical decisions on unseen data with a good level of generalization. However, diverse and large datasets are needed to achieve a clinical-grade accuracy which is also safe, fair and equitable [1].
In order to support the adoption of big data technologies in the healthcare sector, trust, privacy and security need to be addressed to ensure the protection of health data and to comply with a highly regulated domain [2]. This poses a challenge for the adoption of AI, in terms of building trust and the ethical concerns about data use in medical AI [3].
Furthermore, it increases the complexity of data collection and poses restrictions on the data processing. Data limitation is also due to several practical challenges such as high cost of health systems to enable data collection, storage and memory costs, shortage of labels for supervised learning tasks, and dataset biases which lead to a poor generalization [3]. Thus, the collection of training datasets is a crucial step to build a robust data ecosystem for the implementation of generalizable AI models, which can be integrated and run within a digital healthcare system.
This deliverable provides an in-depth description of the ODIN data ecosystem per reference use case, as well as an overview on how the data are collected by the clinical pilots and shared with the technical partners to enable all the experiments defined so far by the pilots within the predefined ODIN use cases.
Within the ODIN project, AI/ML models are designed, developed and validated, addressing the ODIN reference use cases (RUCs). After having defined the AI-based problem, the key part is to collect and select the relevant data for training and evaluating the designed AI models. After having defined the AI-based problems per RUC in Phase 1 (M7-M17), Phase 2 has started with selecting the data for model implementations (Phase 2a) (described in detail in deliverable D6.6). This phase involves all the selection and collection of data from multiple sources, individual devices, or clinical sites, that should be collected and shared to enable model development in scaled use cases. This will form the basis to design a data ecosystem for AI operations and models in the ODIN platform.

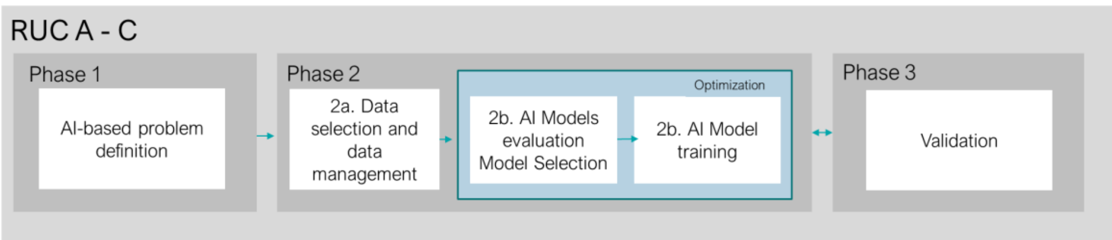


Figure 1.1: AI strategy in ODIN.

1.1 Deliverable context

Table 1.1 describes the deliverable context, detailing the planned objectives, results, related deliverables and potential risks.

Table 1.1: Deliverable context.

PROJECT ITEM IN THE DOA	RELATIONSHIP
Project Objectives	This deliverable is framed in the context of WP6 and contributes directly to the AI operations and implementation within the ODIN use cases.
Exploitable results	The results of this deliverable will be directly exploited for getting results on leveraging the data analytics and sharing within ODIN data ecosystem and interpretation as well as for the design of integration of the AI models in the ODIN platform.
Workplan	The deliverable will be constantly updated, and our partners will be encouraged to provide constant up-to-date inputs regarding the data collection and sharing activities. Progresses in data collection, descriptions and sharing will be monitored and documented. Deliverable D6.1 is the outcome of the activities performed within Task T6.1 Learning Capacities and Desi of the Emergency Management Modules.
Milestones	D6.1 is linked to the Milestone MS1 (First version of the ODIN platform, ready for components integration. All pilot specifications and KPIs defined, pilot designed and ethical approval) and MS2 (ODIN technologies catalogue defined and full version of the platform and IoT robotics and AI components).
Deliverables	The current deliverable is related to deliverable D6.2 (Data results interpretation and data integration services) that provide the data results interpretation for the AI-based models, and D6.6 (D6.6 – Development of high-level AI-based models of planning, scheduling and workflow modelling) which provides an in-depth overview of all the AI-based models developed to optimize the current pilots' workflow. Furthermore D6.1 is related with D7.1 (Pilot Studies Use Case Definition and Key Performance) where the pilots' requirements and needs are presented. D3.11 (ODIN Platform v2) and D3.5 (Privacy, Security and Trust report) are also related to D61, since the D3.11, defines the AI as a KER in the ODIN platform and its connection to HIS via the ESB and the ODIN data storage mechanisms, while D3.5 details the data anonymization and pseudo-anonymization processes. This deliverable leverages also the data sharing guidelines delineated by deliverable D8.2
Risks	Due to the data sharing agreement in definition, some of the data collection conditions outlined in this deliverable may change over the course of the ODIN project.

2 ODIN data space

Before describing the data that form the ODIN data space, we hereby provide an overview of the main steps which are performed to build the data space.

2.1 Data collection

Good and useful AI in healthcare relies on good data, in terms of quality, volume and diversity. Access to datasets that are large enough and are representative for the diversity of the target populations, to which the AI model should apply, is a significant challenge in healthcare. It is rarely the case that such a dataset, that would enable to build models which are accurate, generalizable and free of local data biases, can be collected at a single institute. Within the ODIN project, data is collected across pilots (Deliverable D7.1), to realize a data ecosystem for (1) AI model deployment and operations, (2) data sharing and usability in the ODIN platform by leveraging the federations among hospitals.

Data collection is a key process for collecting measurements, observations, and in general data to deploy the AI modules in the ODIN platform. Furthermore, it represents the first step to get insights and knowledge from the data itself. The data collection process in the ODIN data space is related to the reference use cases (RUCs) and the corresponding AI models and operations. Thereby, for each RUC, data are collected by the pilot sites (in compliance with the local ethical approval (EA)), after having defined the research objectives. The methods and procedures utilized in the ODIN data space are described in Section 4. It should be noted that the data collection process includes the definition of a data storage location, data storage period and access to the data. Section 7 provides an insight on data injections in the ODIN ecosystem to run AI operations implemented and integrated in the ODIN platform.

2.2 Data sharing

The ODIN data space aims at enriching data variability by exploiting data from different sources to generate robust AI model deployment. However, due to their sensitive nature, health data collected by different pilots and covering different populations needs to follow privacy, legal, regulatory, and ethical processes.
These processes represent the basis for developing effective analytical approaches that are generalizable and scaled within the ODIN platform.
Within ODIN, the data sharing follows all the guidelines delineated by deliverable D8.2 in terms of regulatory, privacy and ethical perspectives. This means that each partner complies to ethics and privacy as defined by the data sharing agreement. The latter has been designed and revised based on the actual data flow, described and covered in this deliverable. The data sharing strategy, defined in Phase 2 (Figure 1.1), aims at designing a data ecosystem for deploying and implementing AI modules and operations. It represents the building block to train AI models which are going to be implemented and executed in the ODIN platform.

3 Data resources for AI operations and modules

3.1 Data ecosystem requirements

The ODIN data ecosystem represents an enabling resource for the AI algorithms in ODIN, thereby it enables the specific use cases. As described in deliverable D3.10, the AI algorithms are part of Key Enabling Resources (KERs), which similarly to other resources undergo semantic/syntactic translation, and representation in the ODIN ontology. The data resources should be used by the AI algorithms within the ODIN use cases, during the training and inference phases. Thus, the data ecosystem should leverage the sharing of data resources to deploy and enable all the ODIN use cases defined in deliverable D7.1.

Hereby, we describe in detail the ODIN reference use cases (RUCs) which are enabled by the ODIN data ecosystem. A total of seven use cases have been defined and grouped in three different RUC groups, namely. The reference use cases refer to clinical activities performed within ODIN, and they are displayed and mapped to the related pilot sites in Figure 3.1 (first presented in deliverable D7.1).

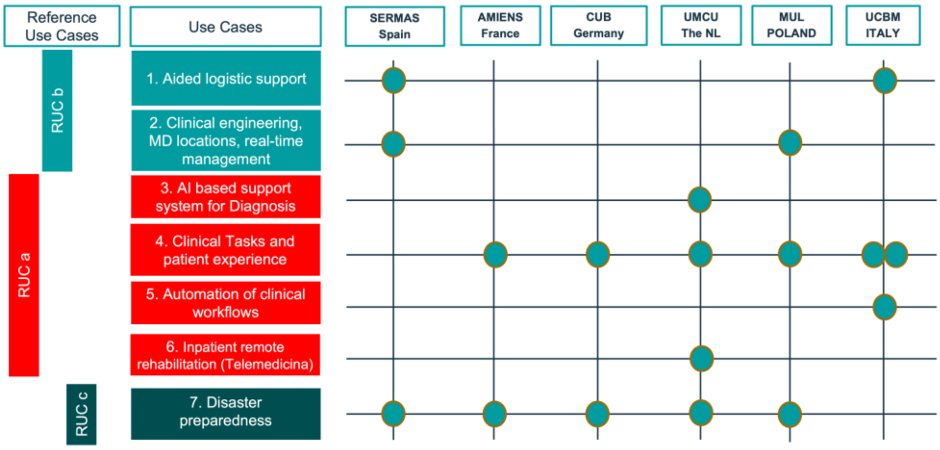


Figure 3.1: Reference Use Case (RUCs) and use case pilot distributions as indicated in deliverable D7.1 from WP7.

The ODIN data ecosystem covers the three RUCs, which include different data types. The data dictionaries, data flow and data description are provided in Sections 4, 5 and 6.

In ODIN, hospital pilots collect prospective data or provide retrospective data as resource for the AI algorithms. Nevertheless, the ODIN platform provides a data storage which is connected to the main bus (ESB) of the platform and satisfies the requirements for the ODIN data storage mechanisms (deliverable D3.10). However, this instance is designed to collect non-clinical patient data, mainly related to the hospital functionalities, which are building data, equipment database, technical documents, etc. This means that patient data are stored in the databases of the Hospital Information System (HIS), which should communicate with the ODIN platform. In deliverable D3.10, examples of possible communications between the ESB and HIS are provided for several data storage variants (CSV, API, DBMS). When a request of data is made, the specific connector is called, and the data are published by the ESB.

As mentioned above, data can exist already in the HIS or may be needed to run AI resources (e.g. an AI model). In both cases, pilot sites remain in full control and retain the ownership of data. However, as the ODIN platform provides databases to store arbitrary information, non-clinical or anonymized patient data can be also retained in these databases. Recital 26 of the GDPR defines anonymized data as “data rendered anonymous in such a way that the data subject is not or no longer identifiable” [General Data Protection Regulation 2016/679]. When patient data are anonymized so that the patient is no longer identifiable, these data can be integrated in the ODIN data bases to further improve or deploy AI operations or modules. Furthermore, to facilitate a collaborative AI model training, by preserving privacy of health data, the ODIN platform will integrate the Federated Learning (FL). Federated Learning enables to collaboratively train an AI module without sharing the data. This means that data security, privacy, custodianship remains at pilot sites, while AI model updates (e.g., weight and bias of a neural network) can be shared to train more robust and generalizable models. In this way, FL will breakdown the current barrier in sharing health-data among different institutions where privacy, legal, ethics remain big constraints. In ODIN, AI modules can leverage FL by training models for the same use cases (same data type, features) without sharing the data.

4 Data ecosystem for RUC A

In this section, we provide the data description and the data flow to be integrated in the ODIN platform. This will form the data ecosystem for RUC A which involves the following UCs as defined in D7.1.

UC3 - AI-based support system for diagnosis. UC3 aims at optimizing the diagnostic workflow by personalizing the diagnostic trajectory of patients and offer diagnostic tools to healthcare professionals to facilitate decision-making processes.

UC4 - Clinical Tasks and patient experience. In UC4, ODIN technology will improve the execution of these tasks by assisting the clinical staff and supporting the patients within the hospital ecosystem. Also, the daily work of clinicians is expected to be improved and facilitated by applying the technologies offered by ODIN. UC4 is the use case with more pilots involved within the RUC A.

UC5 - Automation of clinical workflows. Some patients require guidance and physical support during motion (walking, sitting/standing), and sometimes complementary assistance while executing certain tasks. Clinical workflows can be complex, expensive, demanding, intensive and error prone. Therefore, the automation of clinical workflows aims to automatize certain clinical processes and take the work burden of clinicians, while reducing costs and possible errors.

UC6 - Inpatient remote rehabilitation. UC6 focuses on improving the treatment of the patients at home and offering remote patient monitoring. In this way, patients' quality of life will be improved and also useful data regarding their health will be visible by clinicians.

4.1 Sleep disorder management

The data ecosystem for the sleep disorder management use case is an activity led by PEN in collaboration with the CUB pilot. Sleep disorder is a symptom of many diseases that may significantly affect the quality of daily life [4]. Current methods used for assessing the disorder are time-consuming and involve the manual scoring of Polysomnography (PSG). PSG involves monitoring various physiological processes during sleep to diagnose sleep disorders. It typically involves the use of several sensors that monitor a variety of physiological processes, including brain activity, eye movements, muscle tone, heart rate, and breathing patterns. The data collected during PSG is analyzed by a sleep expert to identify patterns of sleep disruption or disorder, which can be obstructive sleep apnea (OSA), insomnia, periodic limb movement disorder, among others.

PSG is often performed in a sleep laboratory, where the patient is monitored overnight by trained technicians. In this context, AI-based solutions such us deep learning approaches (convolutional neural networks -CNNs- and Long Short-Term Memory and Bidirectional Long Short - Term Memory -LSTM/BLSTM- networks) are raising attention for the evaluation of biomedical signals such as electroencephalography, electrocardiogram, electromyography and electrooculogram (EEG, ECG, EMG, and EOG) [6]. However, there are only a few studies in the literature where deep learning models are used for the sleep stage classification [4,6,7], especially when different co-morbidities are taken into account (e.g., insomnia and Obstructive Sleep Apnea (OSA)).
Furthermore, existing AI models for sleep disorders lack the ability to be transportable between clinical centers where data populations may show differences. Significant differences in different data populations have been observed and a challenge was raised on whether is it possible to unlock the potential clinical utility of prediction models [5]. In the ODIN platform, Philips will deploy and integrate an AI enabler for automatic detection of sleep disorders. This component will also exploit different data populations aiming at a generalizable model.
In particular, two datasets from Charité-Universitätsmedizin Berlin hospital pilot (CUB) will be used to design a robust, explainable and trustworthy model which will be one of the AI modules of ODIN platform.

4.1.1 Data description

To address the above-mentioned research problem, we will start working with retrospective data, collected by the Sleep Medicine Center at CUB.
The first dataset contains around 200 up to 230 sleep apnea patients that underwent PG and/or PSG recordings between 2007 and 2021. Generally, multiple recordings, usually performed during nights, use the same devices for acquiring PSG data (e.g., ECG, EMG or EOG devices). This dataset is the contribution of the Charité towards the ESADA project [5], and contains adult patients (18+), including both male and female patients and excludes healthy volunteers.

Inclusion Criteria:
- Adult subjects (18+) with suspicion of sleep apnea (obstructive/central).

Exclusion Criteria:
- History of any sleep disorder, or any Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV) axis I disorder other than sleep apnea.
    - PG devices: Embletta, Nox, Miniscreen, Somnotouch.
    - PSG devices: Embla, Alice V, Alice Light Edition, Somnoscreen.

The second dataset contains 64 insomnia patients that underwent PSG recording between 2008 and 2010. The patient data collected from 2008 – 2009 were used in the clinical trial "A polysomnography study to evaluate the effect, safety and tolerability of oral administration of almorexant (ACT 078573, Midnight Pharma, LLC) [8,9] in adult subjects with primary insomnia". The remaining patients were recorded under the same criteria.
This dataset contains adult patients (ages: 18 years to 64 years), including both male and female patients and excluding healthy volunteers.
Inclusion Criteria:
- Adult subjects (18-64 years) with a diagnosis of primary insomnia.

Exclusion Criteria:
- History of any sleep disorder, or any Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV) axis I disorder other than primary insomnia.
- Sleep apnea, or restless legs syndrome.
- Daytime napping of more than 1 hour per day.
- Important caffeine consumption, heavy tobacco use, alcohol or drug abuse within 2 years prior to the screening visit.
- Unwillingness to refrain from drugs, over-the-counter or herbal medication having an effect on sleep or behavior.

4.1.2 Data dictionary

A data dictionary has been created for both datasets. Table 4.1 depicts the data dictionary for the before-mentioned datasets, by reporting the data table and field names, the data field description and the data type.

Table 4.1: Data dictionary for the sleep disorder management use case.

Data Table	Data field name	Data field description	Data type
Demographics	Gender	Gender of the patient	String
Demographics	Age	Age of the patient	int
Demographics	Height	Height of the patient	int
Demographics	Weight	Weight of the patient	int
Demographics	BMI: > 18.5 and < 32 kg/m2	BMI of the patient must be between: > 18.5 and < 32 kg/m2	Boolean
Demographics	Primary Insomnia defined by DSM IV-TR	Definition of Primary Insomnia must match DSM IV-TR criteria	int
Demographics	Self-reported sleep problems for at least 3 months prior	sleep onset latency more than 30 mins, wake after sleep onset at least 30 mins, total sleep time is less than 6.5 hours, this occurs regularly (5 nights a week)	int
Demographics	Normal bedtime	Bedtime between 20:30 – 00:30	Boolean
Demographics	Sleep medication	how often per week (Monday – Friday 0-4x a day)	int
Demographics	Smoking history	since when (years), how many a day	Date, int
Demographics	Smoke at night	Does the patient smoke at night	Boolean
Demographics	Ex smoker	Is the patient an ex smoker	Boolean
Demographics	Never smoked	Has the patient never smoked	Boolean
Demographics	Alcohol abuse the past 12 months	Has the patient abused alcohol in the past 12 months	Boolean
Demographics	Alcohol consumption per day	How much alcohol does the patient consume per day	int
Demographics	Postmenopausal	Is the patient postmenopausal	Boolean
Demographics	Sterilized	Is the patient sterilized	Boolean
Demographics	Childbearing	Is the patient childbearing	Boolean
Demographics	Concomitant medication in the past 3 months	Is the patient on concomitant medication in the past 3 months: drug, diagnosis, dosage, start date, possible end date	Boolean, Boolean, Float, Date, Date
Demographics	Medical history	Medical history of the patient (diagnosis, start date, possible end date, treatment)	Boolean, Date, Date, Boolean
Demographics	Physical examination	check body parts to see if normal or abnormal e.g. sight, lungs, skin, etc.	Boolean

Sleep parameters (PSG)	Sleep Period	Sleep period of the patient	int
Sleep parameters (PSG)	Waking time in the sleeping period	How many times the patient woke up during the sleep period	int
Sleep parameters (PSG)	Total sleep time	total sleep time of the patient	int
Sleep parameters (PSG)	Sleep start	time that sleep start	int
Sleep parameters (PSG)	Sleep efficiency	the efficiency of the patients sleep (total sleep time divided by time in bed)	int
Sleep parameters (PSG)	Number of waking periods	How many times the patient woke up during the night	int
Sleep parameters (PSG)	Number of movement intervals	Number of intervals in which the patient was moving at night	int
Sleep parameters (PSG)	Sleep staging	The % of time the patient spent in each sleep stage	int
Sleep parameters (PSG)	Arousals	Number of arousals recorded at night	int
Sleep parameters (PSG)	Apnea	Number of apnea recorded at night	int
Sleep parameters (PSG)	Breathing stats	Breathing stats of the patient such as respiratory effort	int
Sleep parameters (PSG)	Oxygen saturation stats	% of oxygen saturation throughout the night	int
Sleep parameters (PSG)	Body positioning	% of body position	int
Sleep parameters (PSG)	Heart rate	average, SD, min, max bpm	int
Sleep parameters (PSG)	Snoring	Snoring of the patient recorded in minutes and %	int

4.1.3 Data flow

The retrospective dataset described in Section 4.1.2 is anonymized so that data sharing can be smoothly leveraged withing the ODIN data ecosystem. In this way, data will be available for training and testing by the relevant technical partners within ODIN, and a first exploration can be performed which will serve as proof-of-concept study for further validation on prospective data (deliverable D7.1 describes the prospective data acquisition from Charité University Hospital).
Data anonymization is performed by applying an extended version of the *expert determination* privacy rule's de-identification standard, defined with HIPAA (Methods for De-identification of PHI | HHS.gov), so that the de-id methodology can be compliant not only with HIPAA, but with other

relevant regulations like General Data Protection Regulation (GDPR). The method consists of an extended version of the expert determination approach in order to be compliant with GDPR.
The (GDPR) defines anonymous information as '*information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable*'. Anonymization is a procedure whereby personally identifiable information (PII) is transformed in a way that renders it irreversibly altered, making it impossible for a PII principal to be directly or indirectly identified by the PII controller alone or in partnership with any other party [10]. In this UC, data anonymization is applied in order to easily share data and develop the first version of AI model to automatically detect sleep disorders.
The anonymized data will be the basis for deploying AI models in the context of sleep disorder detection. Furthermore, as described in Section 3.1 and depicted in Figure 4.1, they can be stored in the ODIN databases as well as leveraged to deploy AI models (e.g., the Federated Learning module) or enabling the ODIN KERs.

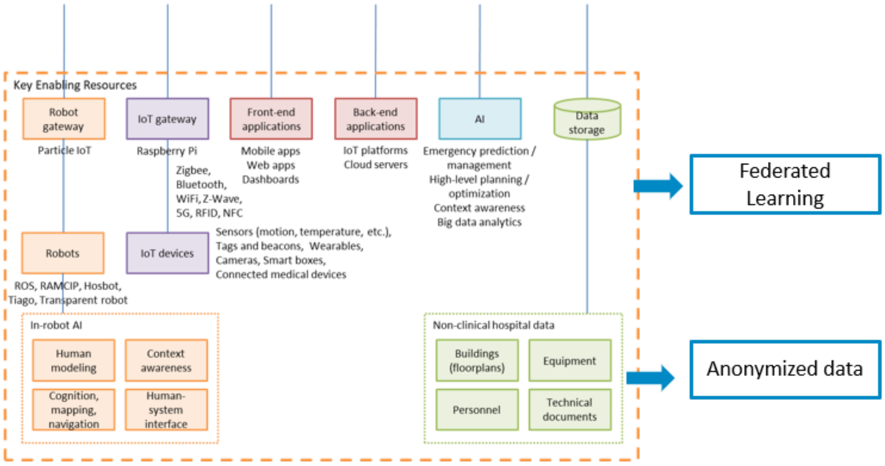


Figure 4.1: Anonymized data storage in the ODIN databases.

4.2 Automated patient inclusion system in the UCC

4.2.1 Data description

The data ecosystem for the automated patient inclusion system in the UCC use case is an activity led by PEN in collaboration with the UMCU pilot. The Utrecht Cardiovascular Cohort (UCC) is a prospective cohort study aiming at a uniform data collection for all cardiovascular patients, with the main objective of optimizing care, control quality and research in a learning healthcare system approach (LHS). In this way, the data provide an insight about cardiovascular disease risk (in approximately 10 years) and can be used for research purpose as well. By the end of 2019, +/- 7000 patients were invited to participate by performing a manual inclusion. The manual inclusion aims at including any patient with cardiovascular precondition and condition. It is based on a discussion between the patient and the nurse, which is followed by a questionnaire and an informed consent.

However, this manual process is limited by several burdens. First, the process is not sustainable due to the manual inner nature of the procedure. Second, it is time consuming and can lead to missing eligible patients and/or cardiovascular diseases. Third, an important limitation is also represented by those patients that do not provide consent. Thus, an automatic patient inclusion is needed to include all eligible patients with potential cardiovascular risk. The data employed for the automated patient inclusion method in the UCC includes age, medication, disease, measurements and other patient data acquired in the outpatient clinic.

In the following section, the data dictionary and the data flow are described.

4.2.2 Data dictionary

The data dictionary contains variables names, descriptions, and variable types from multiple components: laboratory measurements, medication, measurements (height/weight, blood pressure, etc.), medical diagnosis, medical procedures, billing codes and the agenda. It should be noted that data acquisition is going according with the protocol described in deliverable D7.1. Table 4.2-4.8 depict a scheme for the data dictionary by reporting the data table and field names, the data field description, and the data type for the agenda, lab, medication, diagnosis, measurements, operations and financial/diagnosis billing code data, respectively.

Table 4.2: Agenda dictionary.

filename	varnum	name	label	typ	length	format
AGENDA	1	PseudoID	PseudoID	Str	15	$
AGENDA	6	age_dt		Num	8	
AGENDA	7	age_dep	KOSTENPL_SUBAGENDA	Str	5	$
AGENDA	8	age_spec_agenda	SPECCODE	Str	5	$
AGENDA	9	age_spec_description_agenda	OMSCHR_LOCATIE_SA	Str	50	$
AGENDA	10	age_spec_requestor	PHYSICIAN REQUESTOR	Str	255	$
AGENDA	11	age_spec_executor	PHYSICIAN_EXECUTOR	Str	6	$
AGENDA	12	age_consulttype_code	CONSTYPE	Str	1	$
AGENDA	13	age_consulttype_description	Str	35		
AGENDA	14	age_appointmenttype	APPTYPE	Str	1	$
AGENDA	15	age_date_type_description	Str	35		
AGENDA	16	age_appointmentdescription	DESCR_appointment	Str	80	$
AGENDA	17	age_ClinPoli	CLINPOLI	Str	1	$
AGENDA	18	age_appointmentcode	CODE_appointment	Str	8	$
AGENDA	19	age_filled	filled	Str	1	$
AGENDA	20	age_lapse	lapse	Num	8	

Table 4.3: Lab dictionary.

filename	varnum	name	label	typ	length	format
LAB_SELECTION	1	lab_testcode	testcode	Str	40	
LAB_SELECTION	2	lab_dt	sample date-time	Num	8	DATETIME
LAB_SELECTION	3	lab_date	sample date	Num	8	DDMMYY
LAB_SELECTION	4	lab_time	sample time	Num	8	HHMM

LAB_SELECTION	5	lab_result_ok	testresult numeric (=1) text(=2)	Num	8	
LAB_SELECTION	6	lab_req		Str	12	
LAB_SELECTION	7	lab_dept	department test order	Str	8	
LAB_SELECTION	8	lab_dep		Str	12	
LAB_SELECTION	2	lab_dep_description		Str	120	
LAB_SELECTION	10	lab_id	sample ID	Str	30	
LAB_SELECTION	11	lab_spec	specialism test order	Str	10	$
LAB_SELECTION	12	lab_result	testresult	Num	8	
LAB_SELECTION	13	lab_resulttxt	testresult (text)	Str	32	
LAB_SELECTION	14	lab_testunit		Str	30	
LAB_SELECTION	15	pseudoid	PseudoID	Str	15	$
LAB_SELECTION	16	birthdat	year of birth	Num	8	
LAB__SELECTION	17	sex	sex (0=male,1=female)	Num	8	

Table 4.4: Medication dictionary.

filename	varnum	name	label	typ	length	format
MEDICATION	1	pseudoid	PseudoID	Str	15	$
MEDICATION	3	MO_nr	MO numBER	Str	10	
MEDICATION	4	med_start_dt	start MO datetime	Num	8	DATETIME
MEDICATION	5	med_stop_dt	stop MO datetime	Num	8	DATETIME
MEDICATION	6	med_dur_days		Num	8	
MEDICATION	7	med_dur_hours		Num	8	
MEDICATION	8	med_dur_minutes		Num	8	
MEDICATION	9	med_genName	GroepDescription	Str	40	
MEDICATION	10	med_Zlatc	5th level, chemical substance	Str	8	$
MEDICATION	11	med_label	label drug (lablename)	Str	60	$
MEDICATION	12	med_toed_txt	method of administration from knmp	Str	3	$

MEDICATION	13	med_schema_txt	frequency of dosage per day in text	Str	80	$
MEDICATION	14	med_adm_freq	frequency of administration per day (rythm)	Num	8	
MEDICATION	15	med_adm_freq_unit		Str	15	
MEDICATION	16	med_adm_dos	dosage of administration (turndose)	Num	8	
MEDICATION	17	med_adm_dos_day	dosage of administration (dailydose)	Num	8	
MEDICATION	18	med_totdosis	total dosage in period	Num	8	
MEDICATION	19	med_adm_dos_unit	unit of dosage	Str	10	$
MEDICATION	20	med_uneeded		Num	8	
MEDICATION	21	med_applicant_user	ApplicationCode	Str	6	
MEDICATION	22	med_dep	department MO ordered	Str	4	$
MEDICATION	23	med_spec	specialism MO ordered	Str	3	$
MEDICATION	24	med_ZIgpk		Num	8	
MEDICATION	25	med_ZIhpk		Num	8	
MEDICATION	26	med_ZIprk		Num	8	
MEDICATION	27	med_ZIknmp	KNMP number	Num	8	
MEDICATION	28	med_month		Num	8	
MEDICATION	29	med_year		Num	8	
MEDICATION	30	origin	Extraction file	Str	1	
MEDICATION	31	source	Clinical vs policlinical	Str	1	
MEDICATION	32	dep_name	Prescriptive Department	Str	120	
MEDICATION	33	med_start_date		Num	8	DATE
MEDICATION	34	med_start_time		Num	8	TIME
MEDICATION	35	med_stop_date		Num	8	DATE
MEDICATION	36	med_stop_time		Num	8	TIME

Table 4.5: Measurement dictionary.

filename	varnum	name	label	typ	length	format
MEASUREMENTS	1	PseudoID	PseudoID	Str	15	$
MEASUREMENTS	3	meas_dt		Num	8	DATETIME
MEASUREMENTS	4	label	Label	Str	10	$
MEASUREMENTS	5	data1	Data1	Str	100	$
MEASUREMENTS	6	data2	Data2	Str	100	$
MEASUREMENTS	7	data3	Data3	Str	100	$

Table 4.6: Diagnosis dictionary.

filename	varnum	name	label	typ	length	format
DIAGNOSIS	1	PseudoID	PseudoID	Str	15	$
DIAGNOSIS	3	diag_code	DIAGNOSIS	Str	10	$
DIAGNOSIS	4	diag_description		Str	350	
DIAGNOSIS	5	diag_text		Str	350	
DIAGNOSIS	6	diagRegistrationType		Str	30	
DIAGNOSIS	7	codelistname		Str	25	
DIAGNOSIS	8	diag_lapse	LAPSE	Num	8	
DIAGNOSIS	9	diag_start_dt		Num	8	DATETIME

Table 4.7: Operation dictionary.

filename	varnum	name	label	typ	length	format
OPERATION	1	PseudoID	PseudoID	Str	15	$
OPERATION	3	oper_description	DESCRIPTION	Str	60	$
OPERATION	4	oper _code	CODE	Str	8	$
OPERATION	5	oper _date		Num	8	DATE
OPERATION	6	oper _location	LOCATION	Str	3	$
OPERATION	7	oper_department	DEPARTMENT	Str	4	$
OPERATION	8	hos_id		Str	12	

Table 4.8: Financial/diagnosis billing code dictionary.

filename	varnum	name	label	typ	length	format
financial/diagnosis billing code	1	PseudoID	PseudoID	Str	15	$
financial/diagnosis billing code	3	number	DBCNUMBER	Str	13	$
financial/diagnosis billing code	4	startdate	Startdate	Num	8	DATE
financial/diagnosis billing code	5	enddate	Enddate	Num	8	DATE
financial/diagnosis billing code	6	status	Status	Str	1	$
financial/diagnosis billing code	7	diagnosis_ID	DiagnosisID	Str	18	$
financial/diagnosis billing code	8	diag_startdate	Startdate_Diagnose	Num	8	DATE

4.3 Early identification of patient risk malnutrition

The data ecosystem for the early identification of patient risk malnutrition use case is an activity led by FORTH in collaboration with the UCBM pilot. Usually, older hospitalized patients suffer from undernutrition or malnutrition. In the Geriatrics Unit of UCBM the meal is delivered by nurses and healthcare providers and respects the nutritionist and speech therapist indications in terms of consistency and calories. Likewise, they overview whether and to which extent patients are feeding and are quantifying the food assumption. By adding AI to the current workflow, the patients' food consumption will be automatically calculated. An AI-based system will calculate the patients' energy intake and macronutrients consumption at the main meal of the day (lunch). This will ease the burden of the staff and offer them more time for other more demanding procedures that require their attention. The goal is to offer the clinicians the opportunity for an early intervention to diagnose and treat undernutrition and minimize the risk of a further hospitalization. Also, a second goal is for elderly patients to reach the required energy intake to avoid weight and muscle mass loss, minimize the risk of aspiration and treat dysphagia.

In order to develop this system, the RGB-D camera of the TIAGO robot will take an image of the patient's food tray:

- before the patient eats his meal
- after the patient finishes his meal

The system will recognize the different types of food dishes in the tray and estimate its volume to calculate the energy intake and the macronutrients consumed. This automated system will be developed and implemented by FORTH.

4.3.1 Data description

The food image database is the key to create a highly accurate model for a dietary assessment system or a malnutrition detection system and is the first step towards the nutrition dietary system pipeline. These databases can be characterized by the number of food classes and the total number of images they include, the type of cuisine, the quality of the images, the source of the images and by their use (for training or evaluation of the food segmentation and classification system or used only for the evaluation of the food volume estimation system). For the AI model training existing dataset will be leveraged, while real food-image data acquired by the Geriatrics Unit of UCBM is collected to perform the model validation. In this phase, data augmentation is applied to increase the amount of data and consequentially the model performance and generalization.

In the context of the ODIN project for this UC, a food image database will be created containing food images from the Geriatrics Unit of UCBM. Those images must be taken following the image capture protocol that is described in section 4.3.2.

4.3.1.1 State of the art

The process of collecting food images, which will be used in the food classification model, is crucial and it directly affects the performance of the segmentation, classification and volume estimation models. A comprehensive collection of food images is the key for the performance of the AI models in dietary assessment systems. Large food image databases, such as Food-101 [11], UEC-Food100 [12], VIREO Food-172 [13], and UEC-Food256 [14], are typically used to evaluate machine learning models. Existing databases are distinguished by the different characteristics they have, such as cuisine type, the number of images, the number of food classes, the food categories, the image acquisition technique, as well as how many different food items are included in each food image. For instance, Food50 [15] contains 50 classes of food and 5000 images downloaded from the web, while Diabetes [16] has 11 classes with a total of 5420 pictures out of which 3800 images are downloaded from the web and 1620 are captured in a controlled environment. Only a few food image databases have been created by compiling images of existing food databases. For instance, the database Food85 [17] was created from the Food50 and images downloaded from the web.
Moreover, there are several databases that collected food images from specific types of cuisines. For example, the UNIMIB 2016 [18] consists of Italian food images, while MedGRFood [19] contains images of Mediterranean cuisine and includes two sub-datasets. The first sub-dataset is suitable for food image classification systems and consists of 51,840 food images which belong to 160 food classes. The second sub-dataset is suitable for food volume estimation systems and consists of 5,000 food images of known weight, belonging to 190 food classes. Also, an important element for the classifier is the way the pictures were acquired, i.e., whether they were taken in a controlled environment (in terms of lighting conditions and the food's image background) or in a free environment. In addition, with the increasing use of deep learning methods for image classification, the food image databases must contain a large number of images per class to support the training of a deep learning model. Figure 4.2 presents sample images from four food image databases.

Figure 4.2: Food images from UEC-Food100, UEC-Food256, Food-101 and MedGRFood databases.

The techniques used in the later stages of food image-based analysis nutrition systems, emphasize the need to create databases that contain a large number of images for each food class. It may be easier nowadays to collect the images for a large food image database, due to the tendency to capture food images using smartphones and to the existence of many images in social networks. Figure 4.9 summarizes the most representative food image databases and their most significant features.

Table 4.9: Food Image Databases.

Study	Database Name	Food Category	Use of database	# classes / # images	Image Source
Matsuda et al. [12]	UEC-Food100	Japanese	Classification	100/9,132	Captured by authors
Anthimopoulos et al. [16]	Diabetes	European	Classification	11/5,420	Downloaded from the web, controlled environment
Bossard et al. [11]	Food-101	USA	Classification	101/101,000	Downloaded from the web
Kawano and Yanai [14]	UEC-Food256	Japanese	Classification	256/31,397	Captured by authors
Chen et al. [13]	Vireo Food-172	Chinese	Classification	172/110,241	Downloaded from the web
Ciocca et al. [18]	UNIMIB 2016	Italian	Classification	73/1,027	Captured by authors
Chen et al. [20]	ChineseFoodNet	Chinese	Classification	208/180,000	Captured by users
Mezgec et al. [21]	NutriNet	Generic	Classification	520*/225,953	Downloaded from the web
Hou et al. [22]	VegFru	Fruit and Vegetables	Classification	292/160,731	Downloaded from the web
Donadello et al. [23]	FFoCat	Mediterranean	Classification	156/58,962	Downloaded from the web
Kaur et al. [24]	FoodX-251	Generic	Classification	251/158,000	Food-101, downloaded from the web
Gao et al. [25]	SUEC Food	Japanese	Segmentation	256/31,395	Acquired from other databases
Konstantakopoulos et al. [26]	MedGRFood	Mediterranean	Classification	160/51,840 & 190/5,000	Downloaded from the web, controlled environment

4.3.2 Data dictionary

It’s very important to take images according to the image taking protocol. This stage depends on the food segmentation, the food recognition, and the volume estimation steps of the malnutrition system.

4.3.2.1 Image capture protocol

The food should be placed on a plate (Figure 4.3) or a hospital tray (Figure 4.4). There is no restriction on the shape of the plate, it can be elliptical, circular, or even rectangular. Furthermore, the size and the material of the plate do not affect the procedure, as long as the food is clearly visible and the distinction between the tray, the plate and the food are perspicuous, so that the calculation of the amount of food is not affected. Each hospital tray could contain more than one type of foods in its different compartments, to avoid overlaps (Figure 4.5).

Figure 4.3: Food on tray.

Figure 4.4: Food on circular dish.

A reference object must be placed next to each plate. We use either a 2-euro coin or a credit card shaped reference card so that the dimension of reference is stable. There must be a contrast in the color of the reference with the background of the plate. The reference object must be placed distinctly next to the plate and must not overlap with any food item. The reference object is needed to calculate the quantity of food, so that it can be used as an object of known dimensions in the photograph, in order to enable the volume estimation step and the calculation of the quantities of each type of food contained in it.

The color of the background is defined on the basis of the color of the plate or tray accordingly. There should be a clear distinction so that the dish could be identified by the image processing techniques used in the dish recognition stage of the photograph.
To create a high-quality dataset 100 photos are required for each type of food. Ten different dishes require 10 photos taken from different angles according to the following steps:

- Photographs should be taken directly above the dish or tray (vertical to it) at a distance of 30 - 40. It is necessary that the whole plate is clearly visible in the photograph with the reference 2-euro coin placed next to it.
- Photographs should be taken at an angle of 20° - 30° from the vertical axis intersecting the center of the plate. The distance between the plate and the camera will similarly be approximately 30 - 40 cm and should follow the rules mentioned for the previous photographs.
- The remaining photos should be taken from different angles by rotating the camera around the plate, so that the food with the shortest distance to the camera changes, following the same rules as the previous photos. The horizontal distance between photos from different angles should be relatively short. Finally, images should be taken from as close to the ground as possible.

Figure 4.5 shows 3 vertical shots of the same plate, having placed the coin.

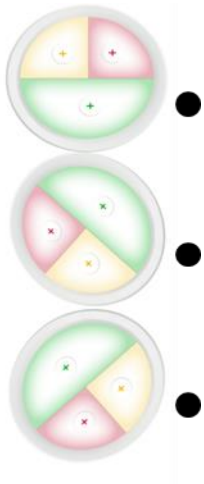


Figure 4.5: Vertical shots of the plate with the reference coin.

The resolution of the photos should be at least 500x500 pixels. All photos follow the RGB (Red, Green and Blue) color model. Also, the type of photos can be .jpg (Joint Photographic Group) or .png (Portable Network Graphics).
Considering that the photographs have been taken for each food dish, a check should take place to ensure that certain criteria have been taken cognizance of, during the images capturing process. These criteria, based on which it should be decided whether the photograph should be accepted or rejected, are as follows:

i. take at least 100 photographs of each type of food.
ii. The distance between the camera and the plate must be approximately 40 cm.
iii. Vertical shot of the food plate.
iv. Take the picture at an angle approximately 30 to 35 degrees from the vertical axis of the plate.
v. Take a photograph of the plate from different angles.
vi. When taking the photograph, the entire circumference of the plate must be visible.
vii. No overlap between food categories, in the case of more than one type of food present on the plate.
viii. There must be a reference coin with a value of 2 euros placed next to the plate.
ix. The reference coin must not overlap the plate.
x. Controlled lighting conditions to avoid creating shadows (either of the plate or the food) in the photograph.
xi. The plate should be matt in color to avoid any reflections.
xii. The resolution of the photograph must be at least 500 x 500 pixels.
xiii. Correct focus on the food plate to avoid blurring of the image.

During the process of taking images some issues had to be addressed. There were unfocused and blurry images, so that the boundaries of the food and the reference object were not clearly visible and thus the annotation could not be done correctly and accurately. Another serious issue that had to be addressed, was the overlapping between the food items or lack of the reference object. The reference object is necessary to be present and clearly visible, so that the food volume can be calculated. In addition, it is essential to represent entirely each portion of food without any overlapping between neither different nor same category of food, in order to correctly calculate the weight of the dish, as well as its nutrition facts and calories. Figures 4.6, 4.7, 4.8 and 4.9 show images with the above issues.

Figure 4.6: Food items overlapping.

Figure 4.7: The food is not visible.

Figure 4.8: Blurry image without the right depiction of the reference coin.

Figure 4.9: Overlapping between same category of food.

4.3.2.2 Description of the Food Image Database

4.3.2.2.1 Food categorization

The UCBM categorization is provided by the internal hospital Dietetics Department. Physicians, when a patient is admitted, should enter the patient's prescribed diet on the UCBM electronic health record. In particular, the system offers the dietitian a set of diets to choose from, including:

- Basic Diet
- Basic Diet for Diabetics
- Light Basic Diet
- Basic Diet Without Slags
- Basic Gastro Diet
- Soft Basic Diet
- Basic Diet for Gastrectomized Patients

- Basic Diabetic Diet
- Basic Diet for Diabetics Choir
- Basic Diet for PTCA
- Basic pre-operative Diet
- Low-Protein Basic Diet (40 g)
- Low-Protein Basic Diet (50 g)
- Low-Protein Basic Diet (60 g)
- Low-Calorie diet
- Low Calorie Diet for Diabetics
- Semi-Liquid diet
- Semi-Liquid Diet Without Slags
- Cold Semi-Liquid Diet
- Liquid Diet
- Cold Liquid Diet
- High-calorie liquid diet without waste
- Water Diet
- Diet for Dysphagia
- Celiac diet

For each diet, the categories of allowed food (shown in the Table 4.10) and the quantities adopted for the patient's meal during the day are already pre-set.

Table 4.10: Food Categorization.

Category ID	Food Category Name
1	Milk
2	Meat
3	Cheese
4	Parmesan
5	Fish
6	Egg
7	Pasta
8	Bread
9	Potatoes
10	Vegetables
11	Fruit
12	Jam
13	Sugar
14	Oil
15	Butter

All the food images that will be taken in the Geriatrics Unit of UCBM will belong in one of the abovementioned food categories.

4.3.2.2.2 Characterization and nutritional analysis of food images

For each food item on the plate/tray, the following information must be specified (Table 4.11):

Table 4.11: Nutritional Analysis of the food images.

Food Category	Based on Table 4.10	
Food Name	Food name	
Food Weight	Expressed in grams. For each type of food, a very important step is to accurately calculate the amount of each type of food contained in the dish. We emphasize that the greatest possible accuracy in calculating the amount of food will significantly contribute to the correct and accurate assessment of their nutritional composition.	
Analysis of Nutrient Composition	**Required fields** Energy (kcal) Energy (kJ) Protein (g) Total lipids (g) SFA (g) MUFA (g) Cholesterol (mg) Carbohydrates (g) Dietary Fiber (g) Water (g)	**Optional fields** Sodium (mg) Iron (mg) Calcium (mg) Zinc (mg) Copper (mg) Vitamin B6 (mg) Vitamin E (mg) Vitamin C (mg)

4.3.2.2.3 Distribution and placement of food on the dish

The meal is delivered by an operator to the patient on a tray. The room is equipped with a table and chair for having the meal. The dishes are two or three, placed on the same tray (Figure 4.10) and, in some cases, the meal can be supplied in a single disposable container with three single compartments (Figure 4.11).

Figure 4.10: Example of two/three different dishes in a tray with fruit, bread and water for UCBM food delivery.

Figure 4.11: Example of single disposable container in a tray with three single compartments for UCBM food delivery.

4.3.2.3 Content of the MedGRFood Database

In the current phase of developing the system's malnutrition models, we used 50 foods from the MedGRFood database to build the AI-based models. The MedGRFood database includes two sub-datasets, consisting of Mediterranean cuisine food images, which are divided into 11 food categories, based on the Greek Food Composition Dataset by the Hellenic Health Foundation. The food groups are the following: (i) Milk, dairy products or milk substitute, (ii) Egg or egg products, (iii) Meat or meat products, (iv) Seafood or related products, (v) Fat or oil, (vi) Grain or grain products, (vii) Nut, seed or kernel products, (viii) Vegetable or vegetable products, (ix) Fruit or fruit products, (x) Sugar or sugar products, and (xi) Miscellaneous food products.
The first sub-dataset is appropriate for food image classification systems and consists of 51,840 food images which belong to 160 food classes. Most of the images have been collected from the web, while the rest have been taken under specific conditions, completing the required number of images per food class for a balanced dataset.
The second sub-dataset is appropriate for food volume estimation systems and consists of 20,000 food images which belong to 190 food classes. All images have been collected under specific conditions, with known weight. Table 4.12 below presents some characteristics of the foods we used to develop the AI models.

Table 4.12: Content of the MedGRFood Database for the AI Models.

Food Name	Number of images for classification task	Number of images for segmentation and volume estimation tasks
Almond cream cake	325	103
Almond snowballs	325	103
Baked Anchovies	324	100
Baked potatoes	325	103
Baked stuffed eggplants	325	104
Baklava	325	104
Beef burger	324	103
Beef in tomato sauce	325	103

Beef stew with onions	325	103
Beet salad	325	103
Boiled beef with vegetables	324	103
Boiled greens	324	103
Boiled pork	324	103
Breaded chicken	324	103
Cabbage rolls	324	103
Calamari with spinach	325	104
Carbonara	324	104
Cheese pie	325	104
Chicken in tomato sauce	326	104
Chicken souvlaki	324	103
Chocolate bundt cake	324	104
Chocolate cake	327	103
Christmas honey cookies	324	104
Codfish plaki	328	104
Cookies	324	103
Croissant with cheese and ham	324	103
Cuttlefish in red sauce	324	104
Cuttlefish with spinach	324	104
Dolmades	324	104
Eggplant salad	326	104
Eggplants in red sauce	324	103
Eggs with tomato and feta cheese	326	103
Fish roe dip	326	103
French croissant	324	105
Fried calamari	324	103
Fried eggs	328	104
Fried peppers	324	104
Fried potatoes	326	103
Roasted chicken	324	104
Roasted vegetable medley	324	104
Russian salad	325	104
Salad with tomato and cucumber	324	104
Sausage and peppers	328	104
Sausage with leek	324	104
Semolina halvah	326	105
Shrimp saganaki	328	104
Spinach and rice	324	104
Tiramisu	324	104
Tuna salad	328	104
Zucchini fritters	324	104

4.4 Rehabilitation monitoring to prevent loss of mobility

The data ecosystem for the rehabilitation monitoring to prevent loss of mobility use case is an activity led by FORTH in collaboration with the UCBM pilot. Rehabilitation aims at recovering patient motor skills following an injury, trauma and/or pathology and can involve both upper and lower limbs. In case of elderly, rehabilitation aims recovering the highest possible level of self-sufficiency (particularly to carry out activities such as eating, dressing, washing, moving from bed to chair, going to the bathroom, checking the function of the bladder and intestine) and avoid loss of mobility.

In the context of the ODIN project, the aim of this UC is to monitor the patient with the RGB-d camera of the TIAGO robot while performing the prescribed rehabilitation exercises in the Geriatrics Unit of UCBM and:

- verify their correctness
- verify the compliance with the prescription
- provide information regarding the correct execution of each exercise and the number of repetitions
- give feedback to the patient

This automated system will be developed and implemented by FORTH. In the context of the ODIN project two different approaches are examined (D6.6): i) sensor-based and ii) vision-based approach to conclude which is the best suited for our research problem.

4.4.1 Data description

In the context of the ODIN project, UCBM partners have provided FORTH the appropriate videos of participants executing the rehabilitation exercises for the vision-based approach. For the sensor-based approach a physiotherapist from the Geriatrics Unit of UCBM has performed each rehabilitation exercise twice while having attached 5 IMU sensors. The collected data stream was delivered to FORTH. Both videos and IMUs data streams will be utilized by FORTH team to find the suitable solutions.

4.4.1.1 Rehabilitation exercises description

To guarantee the targeted treatment outcome, the rehabilitation exercises that are defined by physiotherapists' team in UCBM include both upper (shoulder and elbow) and lower (hip and knee) limb exercises and are explained below.

<u>Upper limps exercises</u>

Shoulder Flexion/Extension (F/E): The patient is asked to start with the palms against the side of the body, move the straight arms in front of him to the highest point he can raise his arms over the head (Figure 4.12 (a)). From there, he moves the straight arms towards his back to the highest point he can (Figure 4.12 (b)).

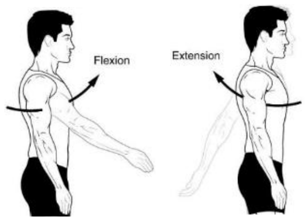


Figure 4.12: a) Flexion of the shoulder, b) Extension of the shoulder[1]

Shoulder Adduction/Abduction (AD/AB): The patient is asked to start with the palms against the side of the body, move them laterally, away from the body and reach the highest point he can (Figure 4.13 (a)). From there, he moves them towards the middle of the body and reaches the sides of the body (Figure 4.13 (b)).

Figure 4.13: a) Adduction of the shoulder, b) Abduction of the shoulder[2]

Elbow Flexion/Extension (F/E): The patient is asked to sit with the arm by his side, bend the elbow as far as possible (flexion) and then straighten the elbow as far as he can (extension) (*Figure 4.14*).

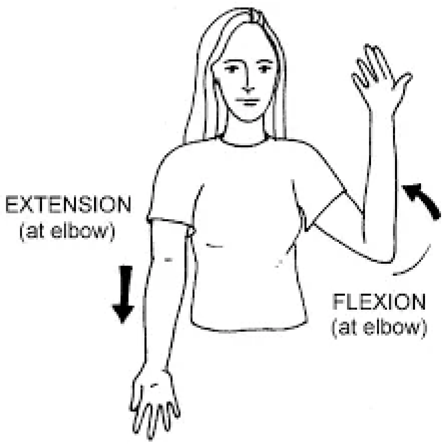
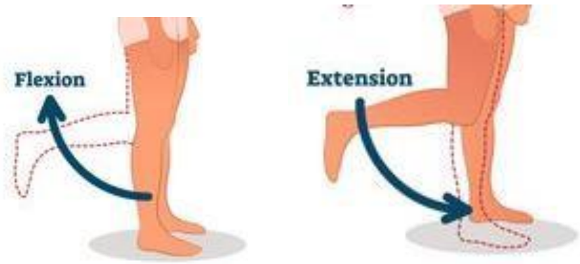


Figure 4.14: Elbow Flexion and Extension[3].

[1] https://teachmeanatomy.info/the-basics/anatomical-terminology/terms-of-movement/
[2] https://www.crossfit.com/essentials/movement-about-joints-part-1-shoulder
[3] https://mobilephysiotherapyclinic.in/elbow-exercise/

Lower limps exercises

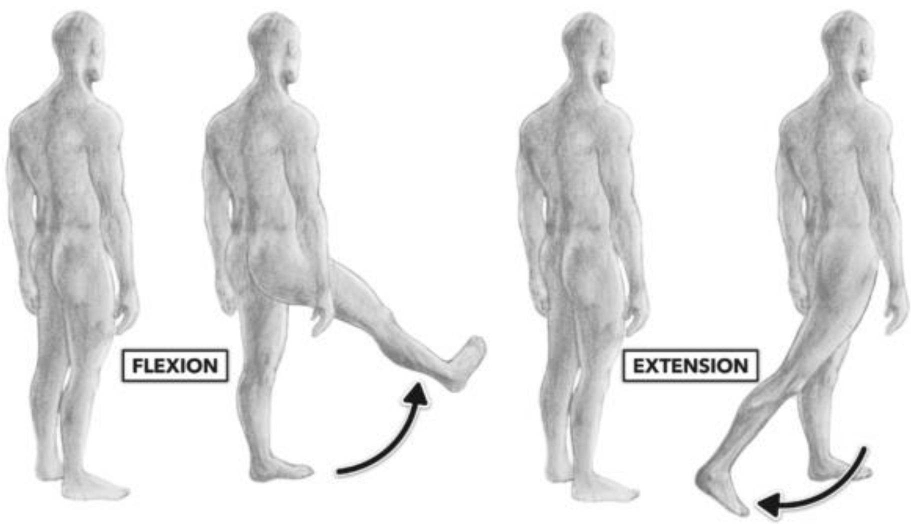
Knee Flexion/Extension (F/E): The patient is asked to start from a sitting position on the side of the bed. He keeps the ankle straight and fixed and starts lifting the foot in front of him until his leg is straightened (Figure *4.15* (a)). From there, bends his knee and brings his foot to the starting position, keeping his ankle fixed (Figure *4.15* (b)).



(*a*) (*b*)

Figure 4.15: a) Flexion of the knee, b) Extension of the knee[4].

Hip Flexion/Extension (F/E): The structure of the hip allows a wide range of motion to (and between) the extreme ranges of anterior, posterior, medial, and lateral movement. The patient is asked to raising the leg toward the front (flexion) (Figure 4.16 a), and then to push the leg toward the back (extension) (Figure 4.16 b).



(*a*) (*b*)

Figure 4.16: a) Flexion of the hip, b) Extension of the hip[5].

Ankle Flexion/Extension (F/E): Flexion and extension at the ankle are referred to as dorsiflexion and plantarflexion. The patient is asked to lift the forward portion of the foot up (rocking back on

[4] https://www.dreamstime.com/flexion-extension-vector-illustration-anatomical-movement-description-educational-arm-leg-exercise-to-bend-straighten-image187222698
[5] https://www.crossfit.com/essentials/movement-about-joints-part-5-the-hip

the heels with the balls of the feet elevated) and then push the forward portion of the foot down (raising the heels off the ground as you go up on the balls of the feet) (Figure 4.17).

Figure 4.17: Ankle Flexion/Extension[6].

4.4.1.2 Sensor-based approach dataset

The scope of this section is to collect the necessary data streams and provide evidence about the capacity of a system comprised by IMU sensors accurately and on-time. For this, a proof-of-concept study has been conducted with the collaboration of the Geriatrics Unit of UCBM and Forth.

According to this study, a physiotherapist from the Geriatrics Unit of UCBM has performed each rehabilitation exercise twice while having attached a set of IMU sensors. The next sections detail this process. FORTH team was provided the IMUs data for the following rehabilitation exercises:

- Shoulder flexion/extension (SFE)
- Shoulder adduction/abduction (SAA)
- Knee flexion/extension (KFE)
- Ankle flexion/extension (AFE)

A physiotherapist of UCBM attached on his body a set of 5 IMU sensors, as presented in Figure 4.18. For this specific proof-of-concept study, not all IMUs were required. Yet, it was chosen to connect them anyway in order to assess the capacity of the data collector module to correctly gather all the data streams that might be required for more complex exercises in a rehabilitation portfolio.

[6] https://www.crossfit.com/essentials/movement-about-joints-part-7-the-ankle

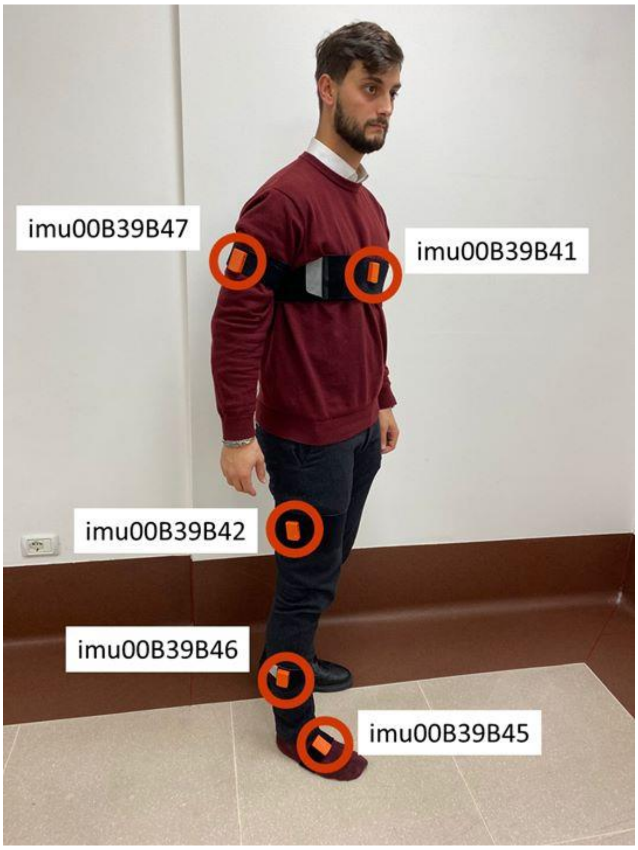
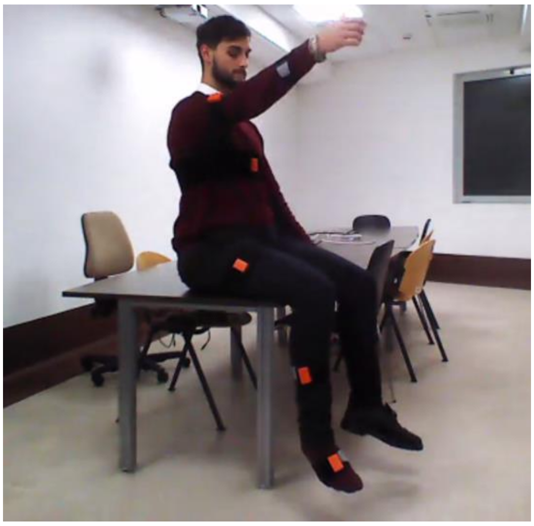


Figure 4.18: IMU placement.

Four different rehabilitation exercises have been conducted. More specifically, as presented in the Figure 4.19, exercises SFE, SAA, KFE, AFE have been presented by the physiotherapist. In parallel, the execution of the exercises was video recorded by a still RBG camera, in order to be able to evaluate the results of the analysis to some extent.



(a)

(b)

(c) (d)

Figure 4.19: Rehabilitation exercises of the proof-of-concept study. (a), (b), (c), (d).

The IMUs that were utilized were the Xsens Awinda MTw inertial-magnetic motion tracker (Xsens Technologies B.V, Netherlands)[7]. The MTw enables real-time 3D kinematic applications with multiple motion trackers by providing highly accurate orientation through an unobtrusive setup.

The MTw is a miniature IMU with a package size of 47 mm × 30 mm × 13 mm and a weight of 16 g. To sense the motion, the MTw contains inertial sensor components, namely a 3D rate gyroscope and a 3D accelerometer. On board of the sensor, the SDI algorithm is applied to the calibrated readings of the gyrosope and accelerometer. The output of the SDI, along with the calibrated magnetometer and barometer data, is then transmitted wirelessly using the Awinda Protocol to the Awinda Master. Data from the accelerometer and gyroscope are captured at a sampling frequency fS of 1000 Hz and low-pass filtered at a bandwidth of 184 Hz.

For this particular study, each IMU sensor delivered data at 100Hz. Each datapoint contained information about quaternion orientation $\overrightarrow{Q(w,x,y,z)}$, angular velocity $\overrightarrow{AV(\,x,y,z)}$ and linear acceleration $\overrightarrow{LA(x,y,z)}$.

4.4.1.3 Vision-based approach dataset

For the human pose estimation model that is based on vision-based techniques, the videos of participants executing the rehabilitation exercises were utilized.

4.4.2 Data flow

UCBM collected the appropriate videos of participants executing all the above mentioned rehabilitation exercises. The participants were all members of the UCBM team. Also, the UCBM team collected the data streams of the IMUs that were placed on a physiotherapist of UCBM executing the rehabilitation exercises twice. UCBM team uploaded the collected data (videos and IMUs data stream) in the CBMLbox in a predefined folder and the relative analysis took place.

[7] https://www.movella.com/products/wearables/xsens-mtw-awinda

4.5 Monitoring of oxygen therapy to prevent hypoxia

The data ecosystem for monitoring of oxygen therapy to prevent hypoxia use case is an activity led by FORTH in collaboration with the UCBM pilot. In the context of the ODIN project, an automated system will be developed and implemented by FORTH in order to monitor the correctness of the oxygen therapy and patient's compliance to prevent hypoxia complications in the Geriatrics Unit of UCBM pilot.

Specifically, the proposed solution will be able to monitor the correctness of the therapy in terms of the correct positioning of the oxygen mask i.e., if the oxygen mask is worn correctly.

Moreover, the system will provide vocal warnings to patients to support them in correct positioning of the oxygen mask both nasal prongs (Figure 4.20) and venturi mask (Figure 4.21) and generally in meeting the therapy rules. Also, will provide alerts to nurses/doctors in case of incorrect use of the oxygen mask. The TIAGO robot will be deployed as a robotic moving platform capable of monitoring and interacting with the patient.

Figure 4.20: Nasal prongs mask (image provided by UCBM).

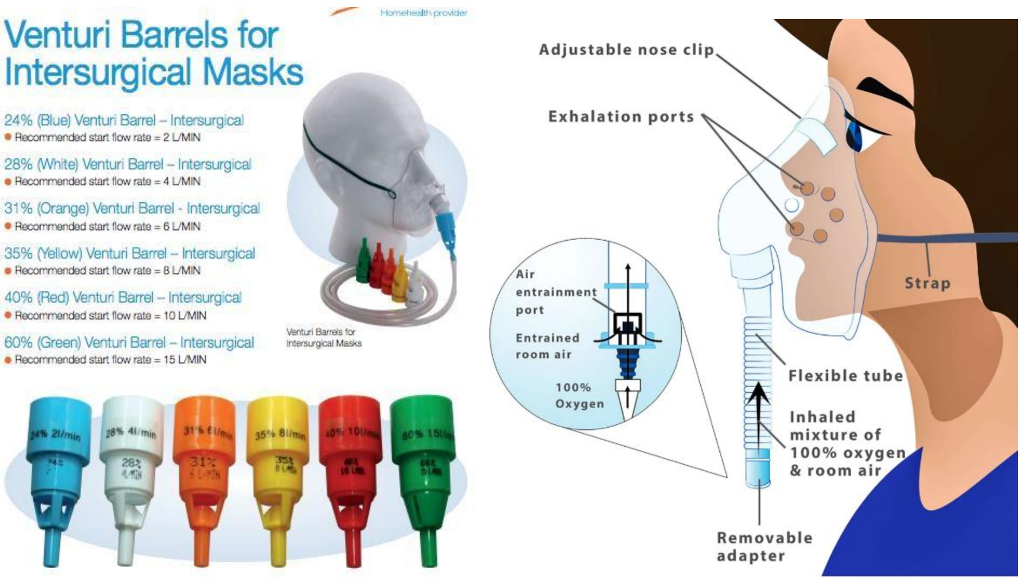


Figure 4.21: Venturi mask[8,9].

[8] https://twitter.com/AldrinAU/status/1324837215638310912
[9] https://endureind.com/products/venturi-mask

4.5.1 Data description

In this deliverable, we will present the data that will be used to detect if the patient if wearing the venturi oxygen mask correctly.

In order to build the model that will be based on deep learning techniques, it is crucial to create a large dataset that will be used for training. The dataset that will be used consists of photos of individuals wearing the venturi oxygen mask correctly, not wearing the oxygen mask and wearing the oxygen mask wrongly. Since an oxygen mask provides a method to transfer breathing oxygen gas from a storage tank to the lungs, the correct positioning of the venture mask is accomplished when it covers only the nose and mouth. If a patient is not wearing the venture mask without following these instructions, then we consider that is weaning the mask wrongly.

The process of taking photos is very important to be carried out as prescribed in the photo capture protocol since the next stages of the models' development directly depend on these steps. The developed models will be validated on photos of patients from the Geriatrics Unit of UCBM that will be collect in the purpose of the ODIN project.

4.5.1.1 Video capture protocol

To capture all the possible angles of the participants' head, we record 3 videos of each participant:

- 1st video - without wearing the oxygen mask (no mask)
- 2nd video - wearing the oxygen mask correctly (correct mask)
- 3rd video - wearing the oxygen mask wrongly (wrong mask)

In those 3 videos, the participant is recorded while performing the following head movements starting each time from the neutral position:

- Extension/flexion of the head (Figure 4.22).
- Right/left rotation of the head (Figure 4.23).
- Right lateral flexion/left lateral flexion of the head (Figure 4.24).

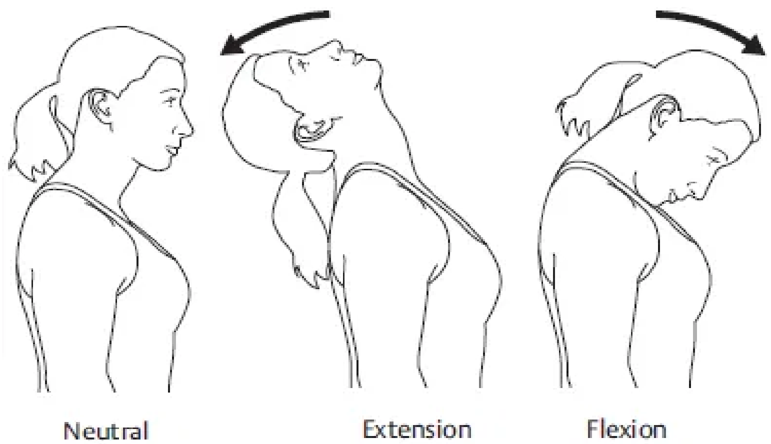


Figure 4.22: Extension/flexion of the head[10].

[10] https://musculoskeletalkey.com/neck-assessment/

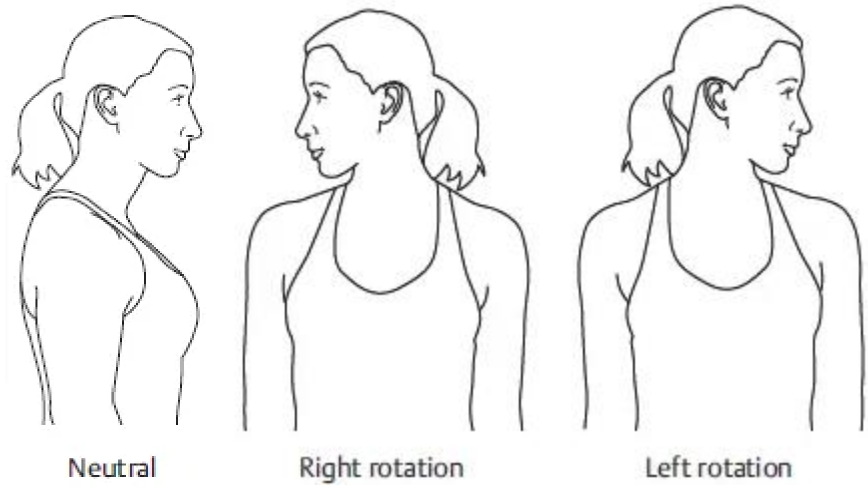


Figure 4.23: Right/left rotation of the head[10].

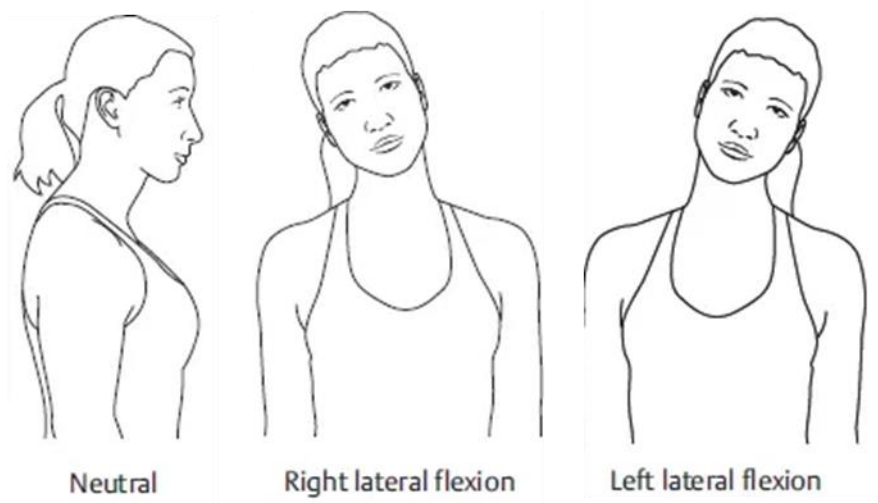


Figure 4.24: Right lateral flexion/left lateral flexion of the head[10].

Then, those 3 videos are divided in frames by using the VLC program. The derived images are used as the training dataset.

We have recorded 26 participants performing the requested head movements. The individuals that participated were all members of FORTH (Foundation for Research and Technology). The training dataset contains 4777 photos of participants. More specifically, the training dataset consists of:

- 1096 images of participants without wearing the oxygen mask (no mask)
- 1455 images of participants wearing the oxygen mask correctly (correct mask)
- 2226 images of participants wearing the oxygen mask wrongly (wrong mask)

Furthermore, a validation dataset was also created, being composed of 352 individual photos. The photos were categorized as follows per class:

- 158 images of participants without wearing the oxygen mask (no mask)
- 101 images of participants wearing the oxygen mask correctly (correct mask)
- 93 images of participants wearing the oxygen mask wrongly (wrong mask)

Figure 4.25 shows a member from FORTH team performing a left/right head rotation, while in Figure 4.26 and Figure 4.27 left and right rotation and lateral flexion are shown.

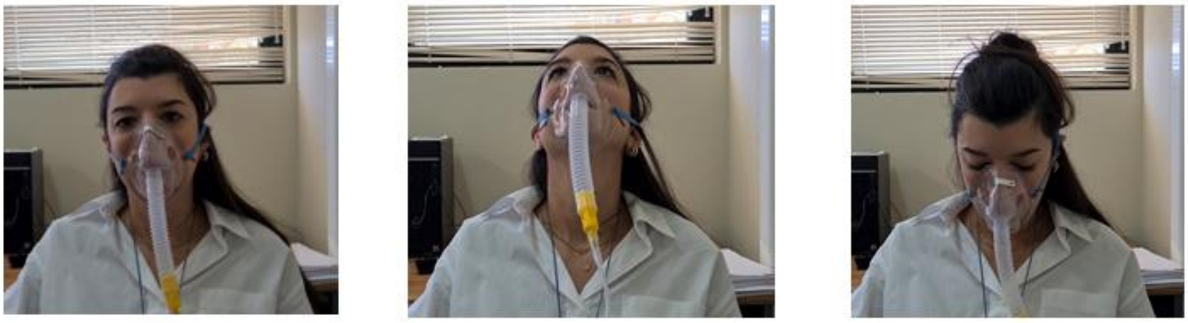


Figure 4.25: Forth team member performing extension/flexion of the head.

Figure 4.26: Forth team member performing left/right rotation of the head.

Figure 4.27: Forth team member performing left/right lateral flexion of the head.

4.5.2 Data flow

The images that are used as the training dataset came from individuals that were all members of FORTH.

5 Data ecosystem for RUC B

RUC B involves the UC1 and UC2 as already explained and defined in D7.1.

UC1 - Aided logistic support. This UC covers all the aspects about the hospital logistics. ODIN will deploy technologies to improve the design, scheduling and execution of activities such as procurement, storage, distribution of the different materials (medicines, medical supplies, meals, linens). Currently, logistics management requires redundant activities and a lot of effort from the staff dedicated to hospital logistics procedures. The vision of ODIN is to optimize these procedures, to improve the current workflow and finally minimize the work burden of the hospital workers for that they can focus on other more demanding operations.

UC2 - Clinical Engineering, MD locations, real-time management. The ODIN technology will allow real-time management of medical devices and medical locations by combining IoT, robots, and AI optimization of routine activities.

5.1 Consumables management

The data ecosystem for the consumable management use case is an activity led by FORTH, in collaboration with the Hospital Clínico San Carlos pilot. The Procurement Department of Hospital Clínico San Carlos oversees the procurement and logistics distribution of the medical equipment and consumable materials inside the hospital, being a key player in the smooth development of all clinical processes and procedures. This department has a transversal action, so the problems it may encounter affect the performance of the entire hospital. The general objective is to embed AI into the process of the hospital logistic management in relation with medical equipment and consumables, optimizing the purchase, storage and delivery, using data and real-time data when possible. More specifically, we will optimize the consumption of a consumable associated with a specific clinical process: i.e., the stent used for coronary angioplasty.

5.1.1 Data description

The UC1-Aided Logistics Support dataset contains the variables that will be used to develop algorithms for predicting the stent purchases.

The dataset involves information related to the hospital's stent acquisitions and patients' episodes during the years 2021 and 2022. No other requirement has been considered for the inclusion/exclusion of patients. No informed consent has been asked from the patients, as 1) the UC1 data set will be anonymized and its content will not be considered personal information according to the EU's General Data Protection Regulation (GDPR), and 2) the data gathered from the patients is retrospective and already registered in the hospital databases.

The historical data that are shown at the Tables below will be used to predict the number of stents that will be needed. First, the data will be cleaned and pre-processed, and then will be analyzed to predict the values. The necessary correlations between the data are established by identifying the primary keys that link the tables together.

The type of personal data gathered includes sociodemographic variables and administrative dates:

- *Patient/stent/episode relationship* table contains information about the patient, the hospital episode in which the stent was implanted in the patient, and about the stent itself (Table 5.1).
- *Consumptions and Purchases* tables contain information about the stents used and bought (Table 5.2 and 5.3).
- *Hospitalization/emergency/outpatient episodes* tables contain information about the episode in which the stent was implanted (Table 5.4).

Table 5.1: Patient/Stent/Episode Table.

Variable	Type	Description
Patient_ID	string	Patient ID unique identifier
Stent_date	date	Date the stent is implanted
Stent_ID	string	Stent ID unique identifier
Article_code	string	Article ID
Provider_code	integer	Provider ID (It is a code identifying the stent manufacturer)
Dimensions	string	Stent diameter x longitude
Hospitalization_episode	string	Hospitalization episode ID
Emergency_episode	string	Emergency episode ID
Outpatient_episode	string	Outpatient episode ID
Gender	integer	Patient gender (1 for male, 2 for female)
Age	string	Patient age in years

Table 5.2: Consumptions Table.

Variable	Type	Description
Storable	integer	If the item is storable or not
Budget_code	integer	Budget ID
Group_code	integer	Article group ID
Group_description	string	Article group description
Subgroup_code	integer	Article subgroup ID
Subgroup_description	string	Article subgroup description
Family_code	integer	Product family ID
Cost_binding_unit_code	string	Internal cost unit id
Cost_binding_unit_description	string	Internal cost unit description
Article_code	string	Article ID
Price	integer	Article cost in euros
Units_consumed	integer	Number of article units consumed
Functional_group	string	Functional group of the consumable
Hospital_institute	string	Hospital institute to which the article is assigned
Date	date	Article consumption date

Table 5.3: Purchases Table.

Variable	Type	Description
Receipt_ID	string	ID of the receipt (When an article is purchased, the Procurement Department receives a receipt. This is the receipt code)
Storable	integer	If the item is storable or not (1 it can be kept in a storage room)
Budget_code	integer	Budget ID (The identifier of the budget line with which the purchase of the stent was paid)
Group_code	integer	Article group ID
Subgroup_code	integer	Article subgroup ID

Group_description	string	Article group description
Family_code	integer	Product family ID
Cost_binding_unit_code	string	Internal cost unit id
Cost_binding_unit_description	string	Internal cost unit description
Article_code	string	Article ID
Price	integer	Article cost in euros
Date	date	Article purchase date
Order_code	integer	Purchase order ID
File_number	string	ID of the public contest. It is the identifier of the public contest that was won by the manufacturer who is selling the article to the hospital. Since it is a public hospital, purchases above a certain monetary amount need to undergo a public contest process.
Units_Purchased	integer	Number of article units purchased
Measurement_unit	string	Unit of measurement
Public_contest_date	date	Date of the public contest

Table 5.4: Hospitalization/Emergency/Outpatient Episodes Tables.

Variable	Type	Description
Patient_ID	string	Patient unique identifier
Gender	integer	Patient gender (1 for male, 2 for female)
Age	string	Age in years
Date_in	date	Date of admission
Date_out	date	Date of discharge
Service_in	string	Hospital service that receives the patient
Service_out	string	Hospital service that discharges the patient
Doctor	integer	ID of the doctor responsible for the patient discharge
Death	integer	If the patient died during the hospital stay
Diagnosis	integer	ICD-10 encoded diagnosis
Procedure	string	Procedure performed on the patient
Episode_ID	string	Episode ID

5.1.2 Data relations

Below, some tables and figures are listed to make the information contained in the data easy to understand and clear.
The patient/stent/episode Table is related to the Consumptions and Purchases Tables through Article_code (Figure 5.1).

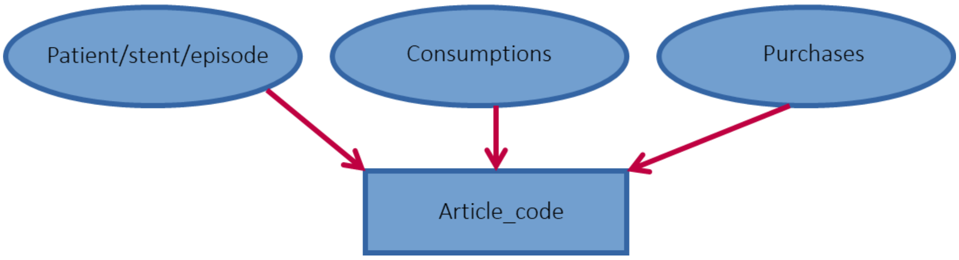


Figure 5.1: Relation of patient/stent/episode.

The patient/stent/episode Table is related to Hospitalization_episodes Table through Hospitalization_episode and Patient_ID (Figure 5.2).

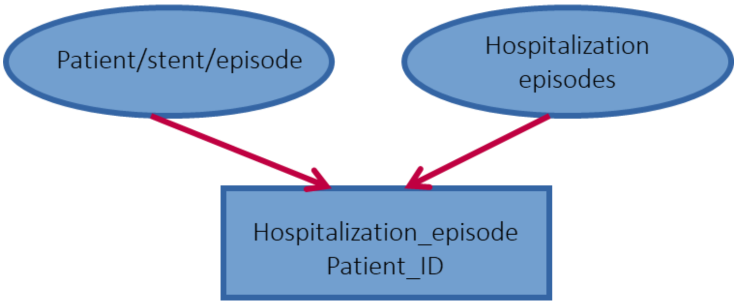


Figure 5.2: Relation of patient/stent/episode.

The patient/stent/episode Table is related to Outpatient_episodes Table through Outpatient_episode and Patient_ID (Figure 5.3).

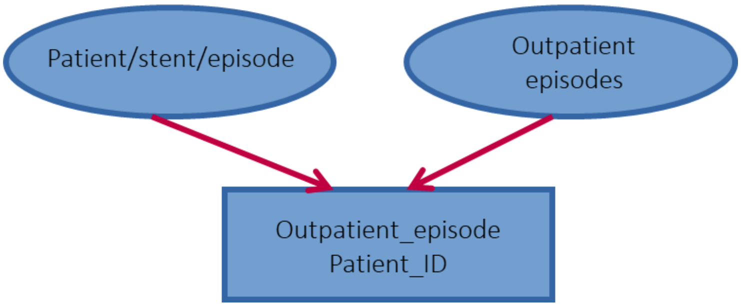


Figure 5.3: Relation of patient/stent/episode.

The patient/stent/episode Table is related to Emergency_episodes Table through Emergency_Episode and Patient_ID (Figure 5.4).

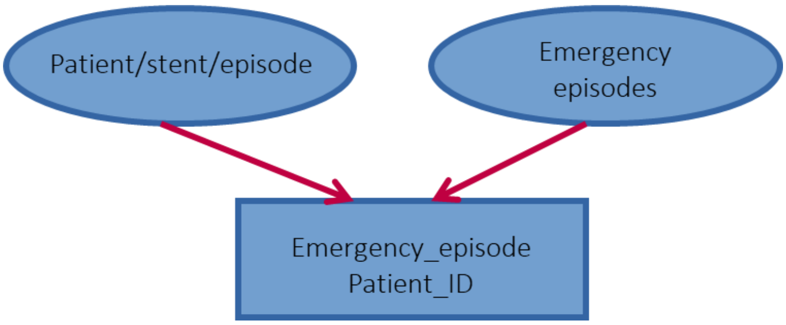


Figure 5.4: Relation of patient/stent/episode.

The Consumptions Table is related to Purchases Table through Article_code (Figure 5.5).

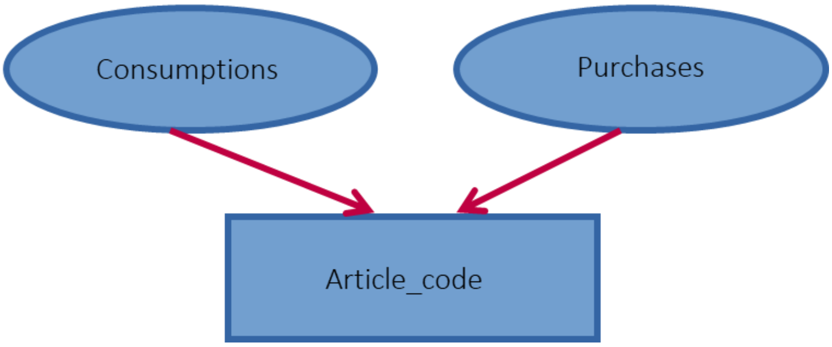


Figure 5.5: Relations of consumptions with purchases.

Based on the above, at Figure 5.6 the class diagram of the SERMAS dataset, including all the values it contains, mapped to their type, is presented. This diagram provides all the appropriate information on how the Tables are connected, the keys, and the foreign keys. Based on this, the Tables and the relations between Tables become clearer and the way in which objects may interact is defined.

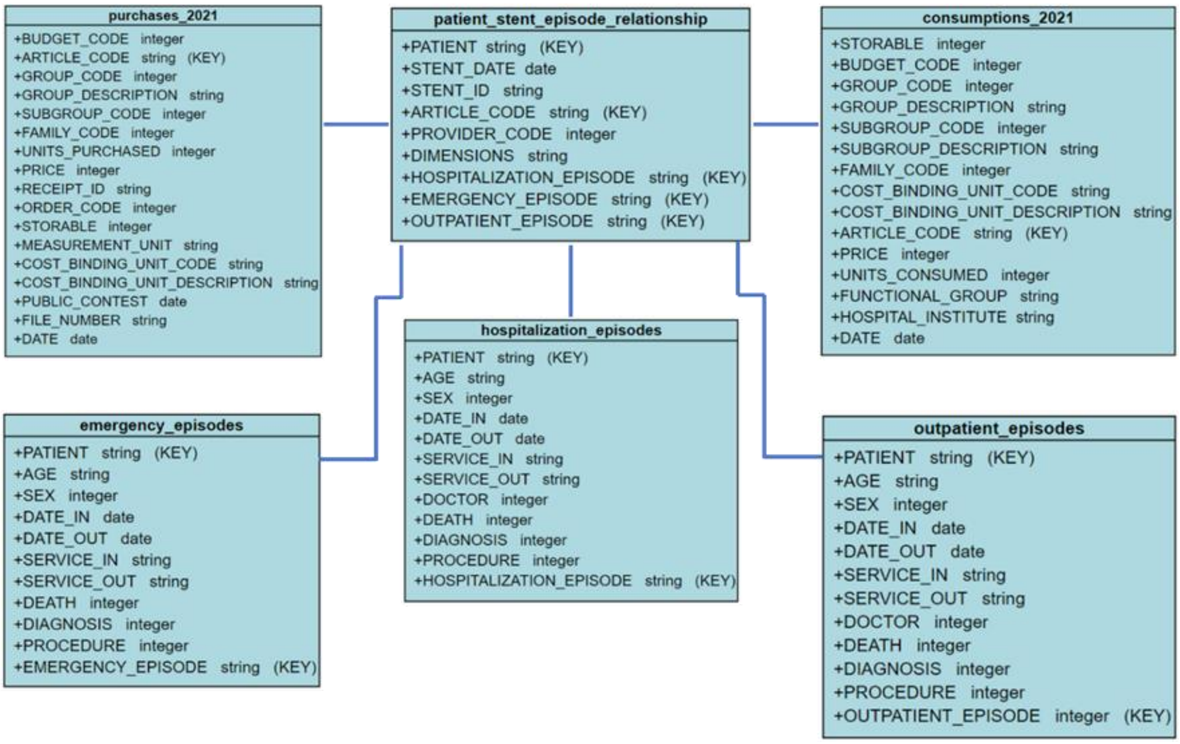


Figure 5.6: Class Diagram of the SERMAS dataset.

5.1.3 Data flow

The dataset results from the integration of information from different data sources of the Hospital Clínico San Carlos (HCSC) in Madrid. These data sources are accessed through the BDClin repository, physically located in the Data Processing Centre of the HCSC, and data sets provided by the Procurement Department. The Data Processing Centre facilities have controlled and restricted access to personnel from the Information Systems and Technology Department. Access to the servers requires two-step authentication and authorized persons must sign a confidentiality document. Once the UC1 dataset was ready, it was uploaded to the ODIN CBMLBox platform in a predefined folder. Relevant ODIN partners will have access to the data, which will not be shared with anyone outside the project. The retention period of this data will cover the entire duration of the project.

6 Data ecosystem for RUC C

6.1 Crowd management

The data ecosystem for the crowd management use case is an activity led by INETUM in collaboration with the Hospital Clínico San Carlos pilot. The aim of the RUC C is mainly to deal with the UC7 – Disaster preparedness; but also, to improve the UC5 – Automation of clinical workflows leveraging clinical care workflows and AI technologies. To do that, a machine vision system is defined and its pipeline can be found in deliverable D6.4. The objective is to process the CCTV images to improve the crowd flows of the building and be prepared against overcrowded environments giving the information for the staff to act when required

6.1.1 Data description

The pipeline of the RUC C starts with the processing of the images provided by our pilot hospitals, MUL, UMCU and SERMAS. After a processing of those images, several statistical data are generated. These data describe the information on the occupation of the hospital and allow the hospital staff to be aware when critical levels of occupation are reached. Also, several other possible applications are described, such as COVID mask detection, fall detection, fire (or cigarette) detection, etc., for which the model will be retrained on request.

The models applied to the images taken are basically two, an object detection model that detects persons, and action recognition model that detects several actions. These models are ran in real time so the data generation will be constant and must be described in a defined interval of time. Also, it has to be taken into account that the model run parallelly in several cameras located in several locations of the hospital although all the data are gathered at one central point.

6.1.2 Data dictionary

For the object detection model the relevant variables are described below.

Table 6.1: Object detection data.

Variable	Type	Description
Class	Int	Class assigned to the detection. , the class number is extracted and it is then linked to the action name with the last line.
position	[Int, Int]	Screen position (x and y) of the lower left corner of the detection (in 2D pixels).
Width	Int	Width of the object detected (in pixels).
Height	Int	Height of the object detected (in pixels).
Camera	Int	Id of the camera in which the detection has been made.
Time	Float	Time since first detection on screen (in seconds)
Detection ID	Int	ID of the detection

In the case of the class person, a further analysis is made in which the action recognition model takes place. This model outputs the action class detected.

Table 6.2: Action recognition Table.

Variable	Type	Description
Action class	Int	Action classification tells if the person detected is STANDING, WALKING, SITTING, LYING, FALLEN, STANDUP and SIT DOWN. It is a 7 x 1 array.
Fall confirmed	Bool	In the case of the detection of a fell a step to avoid false positives is added.

6.1.3 Data enrichment

The primary models utilized in the project are the pretrained YOLOv7 models that have been trained with the MS COCO image dataset [27], which consists of a vast database of 79 different object classes, including a person, handbag, backpack, bed, knife, among others. Pre-trained models such as the YOLOv5 model can be helpful because they allow you to leverage the knowledge and insights learned from a large dataset to improve the accuracy and speed of your model on your own dataset. The pre-trained YOLOv5 has already learned to recognize a wide range of objects and features from a large dataset during its training phase. This can include things like shapes, colors, textures, and other visual features that are commonly found in images. However, certain relevant object classes in the public health context, such as COVID masks, crutches, wheelchairs, and fires, are not present in this dataset. Although the detection of these objects is not the primary objective of RUC C, the model can be retrained using open-source databases to enable the detection of these objects.
By fine-tuning a pre-trained model, accuracy can be improved while reducing the computation resources needed to train it.

7 AI modules implementation in the ODIN ecosystem

7.1 Data injections in the ODIN ecosystem

In this section, the data injection in the ODIN platform for implementing AI modules is detailed. To enable usability and re-usability of the AI components in ODIN, the user might be able to choose the datasets needed for training. Furthermore, operations such as data pre-processing (including remapping) can be enabled at this phase. To do so, ODIN system administration will use the orchestrator to run services like AI modules and data loading and connect these services in a proper way.

The orchestrator allows custom design of workflows, specifically local workflows for each hospital. For AI module integration, this means that each AI model implements a general model of AI modules; in its simplest terms, this model is a service to train the AI model, using as input a data set together with a second service to execute the AI model, having a data point as input and a result as output. With these simple services integrated in the ODIN resource directory, the Orchestrator can then identify them and offer them to system administrators to build the workflows.

The local system administrator, or local data expert, or clinical staff, depending on the specific roles to access and edit the training workflow, can edit the dataset configuring all the required details. This can happen when workflow is triggered through a periodic job, when it is triggered manually by interacting with a UI, when a particular condition is met, or when an event is received. In the same way the workflow must include the necessary pre-processing, this could be the inclusion of statistical automatic modules; or modules which handle the data for manual pre-processing (e.g., presenting a UI for an expert to normalize data); or even privacy aspects such as pseudo-anonymization of the data set using automatic methods described in D3.5. All these aspects could be as complex as needed, setting multiple triggers, or conditions, or workflow paths depending on the conditions.

A similar process is used to configured for each AI module execution service in each hospital. i.e., selecting the input and a pre-processing step; however, actions to be performed could be codified depending on the results. For example, if the AI module detects the plate is empty, then the ingest statistics for the patient are updated and the robot is instructed to pick up the plates. Thus, the AI module can be reused in different hospitals, for different use cases, only by adjusting the workflows they are used in.

7.2 Data sharing and storage in the ODIN ecosystem

As described in Sections 4, 5 and 6, data sharing is a key part for deploying an AI model to address the RUCs in the ODIN framework. To do so, a data ecosystem has been created in order to share and store data for AI model training purposes. The ecosystem is also the basis to store relevant data for further processing, re-use and validation within the development of the ODIN platform. It should be noted that the integration services related to the data shared and stored in the data ecosystem will be provided in deliverable D6.2.

Overall, the data sharing and storage can be divided into two categories: personal data and anonymized data sharing and storage in the ODIN ecosystem.

When concerning personal data, the data sharing is done in compliance with the local ethical approval (EA)), the GDPR and leverage the ODIN data sharing agreement. These data will not be stored in the ODIN platform, unless data is anonymized.

As indicated in Section 3.1, anonymized data can be stored in the ODIN ecosystem and shared in order to execute available resources (e.g., AI module implementation). In deliverable D3.5, data anonymization and related differences with pseudo anonymization are described and should be used as reference for the ODIN data injection and Resource Descriptor (in case of anonymized or pseudo anonymized data). It should be noted that the anonymization process is not generally applicable for every data type and/or use cases. It's an ad hoc procedure that ensures data encryption and deletion of identifiable data so that a person cannot be identified (directly or indirectly). In Figure 10 of deliverable D3.5, a list of anonymization techniques for certain common personal attributes in data set is provided. In Section 3.3.1, guidelines for data pseudo anonymization are given. The latter consists of mapping that identifier to pseudonyms which are not linked to a specific subject. However, according to the GDPR, pseudo anonymized data remain personal and sensible data. Thus, the corresponding data sharing involving anonymized, or pseudo anonymized data may differ. In the first case, data can be shared and accessible to execute ODIN resource components. In the second case, the data flow for personal data should be applied.

8 Conclusion and next steps

This deliverable aims at presenting the ODIN data space and related aspects such as data collection and sharing, with the objective of enabling and implementing AI operations in the ODIN platform. As enabling resource, the data ecosystem should be leveraged by the AI algorithms in ODIN, thus the reference use cases.

Per each reference use case, this document covers:

- The data description which provides the required data information for AI model deployment with background understanding based on the studied use case.
- The data dictionary for each data set.
- The data flow that describes the data lifecycle for AI model training and validation, as well as potential re-usage within the ODIN platform.

The main goal of building a data ecosystem is to implement AI operations. During implementation and execution phase, the data can be injected in the ODIN platform. Thereby, after the model deployment, data should be loaded and connected with data and AI operations, according with the information (data dictionary, data flow) provided by this deliverable.
On the other hand, the data collected to deploy the relevant models within the ODIN RUCs are shared and stored to enable data re-usability. In particular, anonymized data can be stored in the ODIN ecosystem, while personal data should leverage the ODIN data sharing ecosystem and be compliant to GDPR. In both cases, this document is a reference overview for executing the AI operations for training or execution purposes. In addition, to facilitate reuse of data and AI models across hospital, AI models will be aligned with the ontology defined by task T3.2. Thus, this deliverable is used as input for the ontology described in deliverable D3.3.

Figure 8.1 represents the roadmap for the AI model development in ODIN. This deliverable has represented the basis for Phase 2a, which will focus on data selection and management (M17-M30). At the current point of the project (M25), data has been selected and sharing and collection have been initiated. In particular, the data description, storage, and usage for actual implementation in the ODIN platform is provided, with an in-depth focus on reference use cases (RUCs). Data management is a continuous process that runs in parallel with model training and development. However, during the current stage of the project, the data selection provided by this deliverable plays a critical role for the AI algorithm selection. The latter will be provided in deliverable D6.6, based on the data description and analysis provided by deliverable D6.1, and deliverable D6.2.

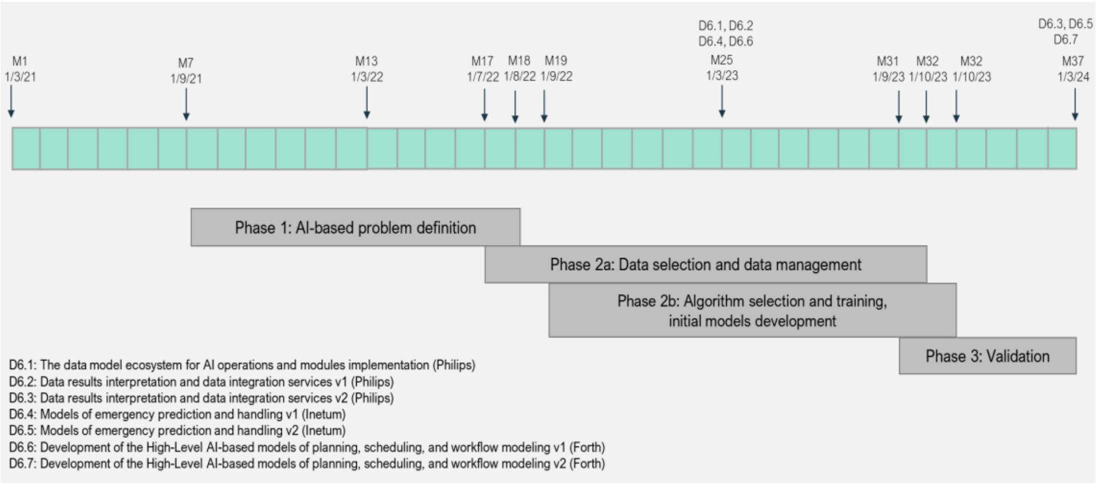


Figure 8.1: Timeline of the ODIN AI models development.

During the third year of the project, M25 – M30, the AI models will be trained and deployed in the ODIN platform, with focus on the integration in the ODIN platform. Within this year, federated scenarios will be also leveraged to enable collaborative model training without sharing the data, as privacy-preserving techniques proposed in the ODIN project. Thus, robustness and a better generalization will be achieved by the models deployed with the data described in this deliverable.

References

[1] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ digital medicine, 3(1), 1-7.

[2] Albrecht, Jan Philipp. "How the GDPR will change the world." Eur. Data Prot. L. Rev. 2 (2016): 287.

[3] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31-38.

[4] Yildirim, Ozal, Ulas Baran Baloglu, and U. Rajendra Acharya. "A deep learning model for automated sleep stages classification using PSG signals." International journal of environmental research and public health 16.4 (2019): 599.

[5] Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., & Acharya, U. R. (2018). Deep learning for healthcare applications based on physiological signals: A review. Computer methods and programs in biomedicine, 161, 1-13.

[6] Goldstein, Cathy A., et al. "Artificial intelligence in sleep medicine: an American Academy of Sleep Medicine position statement." Journal of Clinical Sleep Medicine 16.4 (2020): 605-607.

[7] Meng, C., Trinh, L., Xu, N., Enouen, J., & Liu, Y. (2022). Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. Scientific Reports, 12(1), 1-28.

[8] https://clinicaltrials.gov/ct2/show/results/NCT00608985

[9] Black, J., Pillar, G., Hedner, J., Polo, O., Berkani, O., Mangialaio, S., Hmissi, A., Zammit, G., & Hajak, G. (2017). Efficacy and safety of almorexant in adult chronic insomnia: a randomized placebo-controlled trial with an active reference. Sleep medicine, 36, 86–94. https://doi.org/10.1016/j.sleep.2017.05.009

[10] Tomashchuk, O., Van Landuyt, D., Pletea, D., Wuyts, K., Joosen, W. (2019). A Data Utility-Driven Benchmark for De-identification Methods. In: Gritzalis, S., Weippl, E., Katsikas, S., Anderst-Kotsis, G., Tjoa, A., Khalil, I. (eds) Trust, Privacy and Security in Digital Business. TrustBus 2019. Lecture Notes in Computer Science(), vol 11711. Springer, Cham. https://doi.org/10.1007/978-3-030-27813-7_5

[11] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in European conference on computer vision, 2014, pp. 446-461: Springer.

[12] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in 2012 IEEE International Conference on Multimedia and Expo, 2012, pp. 25-30: IEEE.

[13] Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 32-41.

[14] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in European Conference on Computer Vision, 2014, pp. 3-17: Springer.

[15] Joutou and K. Yanai, "A food image recognition system with multiple kernel learning," in 2009 16th IEEE International Conference on Image Processing (ICIP), 2009, pp. 285-288: IEEE.

[16] M. M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. G. Mougiakakou, "A food recognition system for diabetic patients based on an optimized bag-of-features model," IEEE journal of biomedical and health informatics, vol. 18, no. 4, pp. 1261-1271, 2014.

[17] H. Hoashi, T. Joutou, and K. Yanai, "Image recognition of 85 food categories by feature fusion," in 2010 IEEE International Symposium on Multimedia, 2010, pp. 296-301: IEEE.

[18] G. Ciocca, P. Napoletano, and R. Schettini, "Food recognition: a new dataset, experiments, and results," IEEE journal of biomedical and health informatics, vol. 21, no. 3, pp. 588-598, 2016.

[19] F. Konstantakopoulos, E. I. Georga, and D. I. Fotiadis, "3D Reconstruction and Volume Estimation of Food using Stereo Vision Techniques," in 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE), 2021, pp. 1-4: IEEE.

[20] Xin Chen, Yu Zhu, Hua Zhou, Liang Diao, Dongyan Wang, "ChineseFoodNet: A large-scale Image Dataset for Chinese Food Recognition, " arXiv preprint arXiv:1705.02743, 2017.

[21] Mezgec S, Koroušić Seljak B. NutriNet, "A deep learning food and drink image recognition system for dietary assessment, " Nutrients, 9(7), pp. 657, 2017.

[22] S. Hou, Y. Feng and Z. Wang, "Vegfru: A domain-specific dataset for fine-grained visual categorization", Computer Vision (ICCV) 2017 IEEE International Conference on., pp. 541-549, 2017.

[23] Donadello, I., Dragoni, M. (2019). Ontology-Driven Food Category Classification in Images. In: Ricci, E., Rota Bulò, S., Snoek, C., Lanz, O., Messelodi, S., Sebe, N. (eds) Image Analysis and Processing – ICIAP 2019. ICIAP 2019. Lecture Notes in Computer Science(), vol 11752. Springer.

[24] Parneet Kaur, Karan Sikka, Weijun Wang, Serge Belongie, Ajay Divakaran, "FoodX-251: A Dataset for Fine-grained Food Classification," arXiv preprint arXiv:1907.06167, 2019.

[25] Gao, J.; Tan, W.; Ma, L.; Wang, Y.; Tang, W. MUSEFood: Multi-sensor-based food volume estimation on smartphones. In Proceedings of the 2019 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Internet of People and Smart City Innovation, SmartWorld/UIC/ATC/SCALCOM/IOP/SCI, Leicester, UK, 19–23 August 2019; pp. 899–906.

[26] F. S. Konstantakopoulos et al., "Mediterranean Food Image Recognition Using Deep Convolutional Networks," in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1740-1743, 2021.

[27] https://cocodataset.org/#home