REal-time data monitoring for **S**hared, **A**daptive, **M**ulti-domain and **P**ersonalised prediction and decision making for **L**ong-term Pulmonary care **E**cosystems

D4.1: Representation of Multi-Modal Data and Disease Progression Monitoring Features

Dissemination level: PU
Document type: Report
Version: 1.0
Date: 31-08-2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 965315. This result reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Document Details

Reference No.	965315
Project title	**RE-SAMPLE** - **RE**al-time data monitoring for **S**hared, **A**daptive, **M**ulti-domain and **P**ersonalised prediction and decision making for **L**ong-term Pulmonary care **E**cosystems
Title of deliverable	D4.1: Representation of Multi-Modal Data and Disease Progression Monitoring Features
Due date deliverable	31-08-2022
Work Package	WP4
Document type	Report
Dissemination Level	PU: Public
Approved by	Coordinator
Authors	Alberto Acebes (ATOS), Cristina Sabater (ATOS), Yolanda Sabuco (ATOS), Gesa Wimberg (DFKI), Jakob Fabian Lehmann (DFKI), Agni Delvinioti (GEM), Miriam Cabrita (iSPRINT)
Reviewers	Serge Autexier (DFKI), Jakob Fabian Lehmann (DFKI), Gesa Wimberg (DFKI), Costas Lambrinoudakis (UPRC), Christos Kalloniatis (UPRC)
Total No. of pages	56

Partners

Participant No	Participant organisation name (country)	Participant abbreviation
1 (Coordinator)	University of Twente (NL)	UT
2	Foundation Medisch Spectrum Twente (NL)	MST
3	University of Piraeus Research Center (GR)	UPRC
4	Foundation Tartu University Hospital (EE)	TUK
5	Foundation University Polyclinic Agostino Gemelli IRCCS (IT)	GEM
6	European Hospital and Healthcare Federation (BE)	HOPE
7	German Research Center for Artificial Intelligence GMBH (DE)	DFKI
8	ATOS IT Solutions and Services Iberia SL (ES)	ATOS
9	Roessingh Research and Development BV (NL)	RRD
10	Innovation Sprint (BE)	iSPRINT

Abstract

Work Package 4 has as its fundamental objective the development of a platform, the RE-SAMPLE platform, that brings together all the data involved in the RE-SAMPLE project to feed the predictive machine learning models in a secure and privacy-preserving way and by complying to the requirements of the General Data Protection Regulation. Tasks 4.1 and 4.3 refer to the process of defining a single data model in a standardized format, based on information from multiple sources. This document collects the work developed in these two tasks.

The objective of D4.1 is to describe the accomplished work until month 18 related to the data standardization process, specifying a common data model agreed among the consortium partners. The data model is composed of data coming from different sources (Healthentia platform, Hospital Information System and Machine Learning modules). Standardization includes mappings to clinical standards, and the homogenized output data is used for further analysis and visualization. Task 4.1 continues until month 24 and task 4.3, until month 30.

This work is focused on stablishing a common methodology followed for data homogenisation and standardization requirements. The main effort is also focused on designing the RE-SAMPLE reference data model, which includes different data sources (Healthentia, Hospital Information System and Machine Learning) and data related to scores based on specific variables. Additionally, technical components in charge of data standardization and storage are developed during the period mentioned before. Clinical standards and terminologies are used within this scope, as well as a mapping between the variables of the data model and the clinical standards is performed.

Complementary documents to D4.1 are *D4.4 Multi-modal data aggregation and curation* (month 42) and *D4.9 Open clinical decision aid* (month 48), which will describe the final RE-SAMPLE data model (if any modification or refinement is made after month 18) and the process of creating an implementation guide, respectively.

Contents

List of Figures

List of Tables

Symbols, definitions, abbreviations, and acronyms

API	Application Programming Interface
CDR	Clinical Data Repository
CHF	Chronic Heart Failure
COPD	Chronic Obstructive Pulmonary Disease
CRQ	Chronic Respiratory Disease Questionnaire
EBM	Explainable Boosting Machine
EHR	Electronic Health Record
EU	European Union
FHIR	Fast Healthcare Interoperability Resources
GDPR	General Data Protection Regulation
GEM	Policlinico Universitario Fondazione Agostino Gemelli - Roma
HADS	Hospital Anxiety and Depression Scale
HDH	Health Data Hub
HIS	Hospital Information System
HL7	Health Level 7
IG	Implementation Guide
IHD	Ischaemic Heart Disease
M	Month
ML	Machine Learning
mMRC	Modified Medical Research Council Questionnaire
MST	Medisch Spectrum Twente
QoL	Quality of Life
RWD	Real World Data
SHAP	SHapley Additive exPlanations
SNOMED CT	Systematized Nomenclature of Medicine - Clinical Terms
TUK	Tartu Ülikooli Kliinikum
WP	Work Package

1. Introduction

The report D4.1 documents the work developed under the scope of tasks *4.1 Representation of multi-modal data incl. disease progression monitoring features*, and *4.3 Aggregation and curation of data from multiple sources*, and provides the process followed to define a unified patient record. I.e., this report describes how a common data model based on heterogeneous information from different sources has been created and how its adaptation to the HL7 FHIR (Health Level Seven Fast Healthcare Interoperability Resources) (Health Level 7, 2022) standard is carried out.

The first step of creating a unified data model is to identify the information providers within the consortium, that is, the medical centres and systems that will ingest data into the RE-SAMPLE platform. After that, each of the variables coming from those data sources is thoroughly analysed and included into the dataset. The variables are isolated from their source format. A common format, meaning and type are defined and agreed by the partners. The result of this analysis is what is called the reference data model. After defining the reference data model, a new data aggregation and relationships between the variables are established to determine small chunks of data consistent with the workflow of the project. Then, the transformation to the FHIR standard using clinical terminologies is implemented. A prototype of the platform designed in this document will be presented in the D4.2 [M24].

The document is structured into three major sections:

- Section 1 presents the actors that directly interact with the Clinical Data Repository (CDR) and the use cases identified for each actor, as well as the methodology followed by them.
- Section 2 contains the objectives of the document.
- Section 3 identifies the plain set of variables that each actor provides to the system.
- Section 4 presents the process of adapting the RE-SAMPLE data model to the FHIR standard, and the clinical terminologies are described. As a prelude, the main characteristics of the HL7 standard, used terminologies and components of the CDR are explained.
- Section 5 concludes the document.
- References to bibliography are included at the end.

1.1 Methodology

This deliverable is considered the reference document that collects and specifies one by one all the variables that are used in the context of the RE-SAMPLE project to be exploited by the predictive ML models. The information collected is the basis for the configuration of the platform that will determine its proper functioning in the future. Due to the iterative nature of the development of the ML modules, the data model may be modified. As indicated below, some RE-SAMPLE variables can be fed from several sources. In addition, each information provider is maintained by a different project partner. To ensure the correct documentation of the variables, all of them have been included in the dataset following the template presented in Table 1. This template will be used to collect the modifications of the data model that may occur in the next iterations, and they will be attached as an annex in *D4.4 Multi-modal data aggregation and curation* [M42]. The communication of these modifications will be made by the modifier partner, and it will be specified which actor is the provider of this new information.

Table 1: RE-SAMPLE data model template.

Variable	Type	Accepted values	Min	Max	Unit

1.2 Storing and standardization requirements

In this section, the definition of the data and the workflow is established to guide the implementation of the Health Data Hub (HDH), including the requirements of the CDR. The CDR is the core of the HDH where the standardized data will be stored.

Conceptual RE-SAMPLE **entities** are shown in Figure 1, with the *Patient* as the central object, surrounded by the rest of the entities that directly or indirectly refer to it.

The cardinality represented in the relationships between entities indicates the number of expected clinical-related procedures to be carried out depending on each conceptual grouping. Nevertheless, isolated instances of them directly related to the patient could exist.

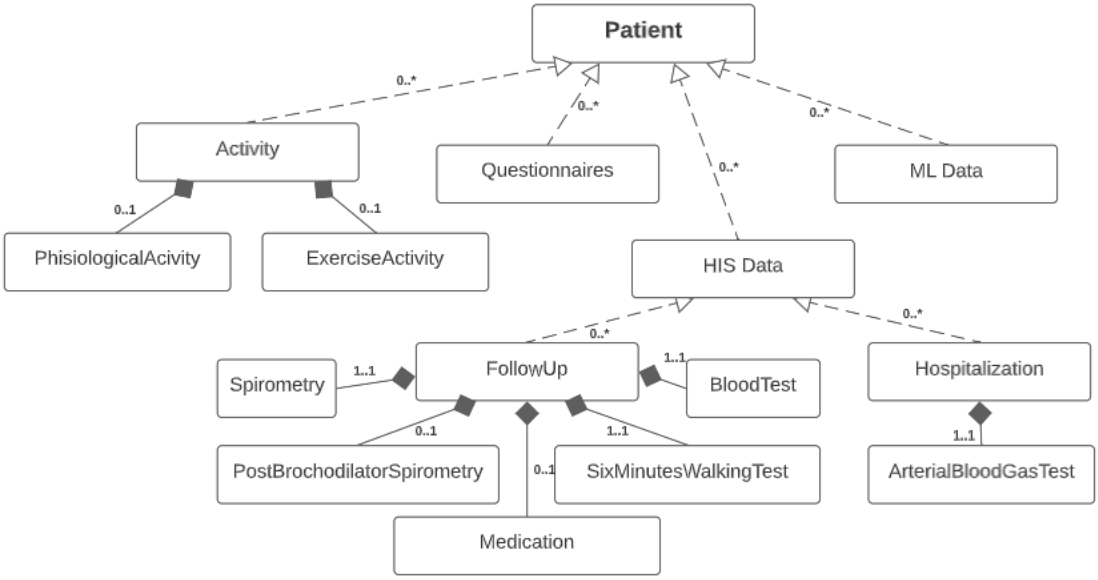

Figure 1: Clinical data repository entity relationship diagram.

The information to be gathered can be grouped into the following categories (conditioned by the different data sources that will be explained later on):

- **Activity** information (mainly from Healthentia).
    - **Physiological** (heart rate data, sleep data, daily physiological data and daily activity data) data collected.
    - **Exercise** data obtained by recording specific exercises with a wearable device.

Questionnaires data, collected through answers to previously defined questionnaires by clinicians.

- **HIS Data**, the information obtained from the Hospital Information Systems (HIS) is normally collected grouped by the following two events, although some of the procedures could also occur independently.
    - **Follow-Up** data, the information obtained by the clinicians during the baseline and the follow-up visits at the clinical sites where several procedures are carried out:
        - Spirometry (regular or after bronchodilators treatment)
        - Six-minute walking test
        - Blood test
        - Medications administered
    - **Hospitalization** data, the information retrieved by the clinicians when a patient is hospitalized. This includes the performance of a procedure:
        - Arterial blood gas test
- **Machine Learning (ML) Data** generated by the ML modules: predictions, explanations and simulations.

All this information related to the patient comes from different providers and is ingested by the CDR according to the RE-SAMPLE project requirements. Three **actors** or information providers that interact with the system have been identified:

- Healthentia application: activity data and questionnaire answers.
- HIS (through or without middleware): mainly grouped under follow-up or hospitalizations encounters, although specific isolated information can be created as well directly linked to the Patient.
- ML Module: predictions (along with explanations and simulations).

For each actor, there are several high-level generic **use cases.** They are sequences of actions (including variations) performed by the system that yield an observable result of interest to a particular actor.

Figure 2 provides an overview of all the preliminary use cases identified for each actor. During the process of enrolment, the first instance of the patient is created by Healthentia. The other actors will refer to this instance to complete it with demographics or further clinical data.

As consumers of the data, the ML modules and Healthentia are exploiting the data for its analysis and visualization.

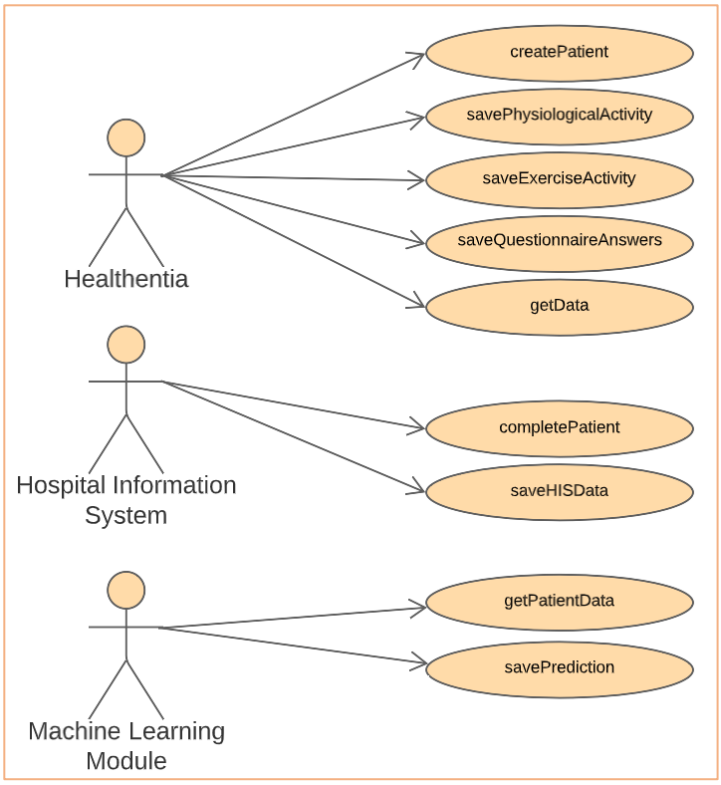

Figure 2: Clinical data repository actors and use cases.

2. Objectives

Deliverable D4.1 describes the RE-SAMPLE data model in terms of the set of variables used to represent multi-modal Chronic Obstructive Pulmonary Disease (COPD) progression monitoring related information. Moreover, the process and components to make this unified set of data compatible with clinical standards and the CDR design for its storage will be explained.

The RE-SAMPLE reference data model could have some modifications in the future, since the development of the predictive ML models is explorative from an algorithmic point of view and the identification of key features for the prediction of disease progression and exacerbations is part of task 3.2 which started in month 18 [M18]. This deliverable contains the description of the work performed during the first 18 months of RE-SAMPLE project within the scope of tasks T4.1 and T4.3 of WP4. It is planned to present the final version of the reference model in the deliverable *D4.4 Multi-modal data aggregation and curation* [M42], and a methodology to document these changes has been established.

The objectives of this document are listed as follows:

- Document and describe the work of T4.1 and T4.3 until month 18.
- Describe the design of a common data model and the methodology followed for that purpose.
- Describe the different data sources of the project and list the variables included in the data model.
- Describe the process of data standardization to clinical standards and terminologies.
- The data model will serve as a common reference for data ingestion from data providers.

3. The RE-SAMPLE reference data model

The RE-SAMPLE data model consists of the set of variables agreed by all partners collected in the project from multiple sources and necessary for the ML modules to perform their tasks. This information is provided by the different actors that have been identified in the previous section and the aggregation of all the clinical data from the different data sources takes place in CDR, which is a module of the whole solution HDH.

Given that the CDR will store the data in a standardized way, the definition in a formal way of the information available to be received is a fundamental preliminary step for its design.

The following sections detail the subsets of the RE-SAMPLE reference data model, grouped by origin. Each origin is responsible for providing such data. In addition to this, the section 3.4 Scores based on other variables has been included describing some inputs for the ML algorithms that are obtained by calculations on variables of the RE-SAMPLE reference model dataset. These scores are used for informative purposes of the algorithms and implemented directly by the data consumer.

The accepted data **types** for the ingestion in the HDH are the following:

- ***string***, is a free alphanumerical value,
- ***numerical***, a numerical value, with or without decimal component,
- ***coded***, is an alphanumerical value, free or within a list of accepted values,
- ***date***, value that represents a date, using ISO8601 format with time zone information,
- ***boolean***, a true/false value, with such strings as codes.

The expected units are mainly coded according to the UCUM standard system (Gunther Schadow, 2017), and regardless of how they are provided by other systems, like the Healthentia Application Programming Interface (API) (iSPRINT, API 2022), will be ingested in the HDH coded using the UCUM standard during the standardization process.

3.1 Healthentia dataset

Healthentia is an eClinical platform certified as a Medical Device Class I, compliant with the General Data Protection Regulation (GDPR) and ISO27001 (FAGGS BE/CA01/1-72378). It aims to facilitate clinical trial optimization, by accelerating the trial processes, reducing the failure rate, and validating drug/intervention efficacy and effectiveness with Real-World Data (RWD) insights. The Healthentia platform offers a smartphone application for patients (the Healthentia app) and a web portal (the Healthentia web portal) for investigators and clinicians, who can configure and monitor the study as well as securely access smart services and insights.

In RE-SAMPLE, the Healthentia platform is the core of the technology that interacts with the users who can be the patients, the investigators, or the healthcare professionals. In the *observational cohort* (Task 5.6), the Healthentia app is used for collection of Real-World Data while the Healthentia web portal is used by investigators for creation, configuration and management of the study. Deliverable *D5.2 RWD collection application (accompanying report)* (iSPRINT, 28th February 2021) describes the use of the Healthentia platform within the observational cohort.

In the virtual companionship programme, the Healthentia app is used not only to collect RWD, but also to provide coaching to the patient on how to self-manage their chronic condition(s) via the so-called virtual companion application (Task 5.4). The Healthentia web portal serves the purpose of monitoring the study via the clinical dashboard as part of the active support programme for healthcare professionals (Task 5.3), and clinical dashboard for the shared-care facilities (Task 5.2).

As for the moment of writing the present deliverable (Summer 2022), requirements are being defined and design decisions are being made regarding the virtual companionship programme. Therefore, this section focuses on the data collected with the Healthentia application within the observational cohort.

The Healthentia app facilitates the collection of electronic Patient Reported Outcomes (ePRO) and electronic Clinical Outcome Assessments (eCOA), as well as the collection of behavioural data through pairing with wearable devices (e.g., activity trackers) and other medical devices and sensors that can be connected through Apple HealthKit for the iOS app. Twenty-five questionnaires were created based on the

first version of the observation cohort protocol (*D5.1 First study subject approvals package RWD cohort*). This protocol was updated in August 2022. Furthermore, the Healthentia platform also gathers objective data on behavioural parameters (physical activity, sleep and heart rate) through the Garmin API (preferred option), and Apple HealthKit. For the Android app, the smartphone sensors can also be used to capture activity data.

The following data from the Healthentia platform (Table 2) is considered suitable content for the project and for inclusion in the centralized repository via the HDH CDR Synchronizer. The Synchronizer is in charge of sending data from Healthentia system to the CDR with a specific sequence. There is a clear distinction between structured data from demographic information, specific exercises and daily physical activity, and answers by patients to specific questionnaires. These answers are ingested completely (whole sets of questions and answers). Both groups of data are contained in the repository and available to the data consumers, mainly, the ML module.

Table 2: Healthentia dataset.

Variable	Type	Accepted values	Min	Max	Unit
Sex	coded	- Male - Female			
Birth Date	date	ISO8601			ISO8601 ([1])
Status	boolean	- active - inactive			
Inclusion Date	date	ISO8601			ISO8601 ([1])
Withdrawal Date	date	ISO8601			ISO8601 ([1])
Education Level	coded	- Lack of education - Received elementary school education - Educated to junior high school level - Educated to senior high school level - Received university education			
Occupational status	coded	- Employed - Unemployed - Retired			
Civil status	coded	- Married - Single - Separated - Divorced - Widowed			
Social Role	coded	- Never - Once a week or less - Three times a week - Everytime it is required without limitations - Every day, family organization is my direct business			
Do you smoke at the moment?	coded	- Yes - No, but I used to smoke			
Years smoking between 10 and 20 years old	numerical				
Smoked cigarettes per day between 10 and 20 years old	numerical				
Years smoking between 20 and 40 years old	numerical				
Smoked cigarettes per day between 20 and 40 years old	numerical				
Years smoking between 40 and 60 years old	numerical				
Smoked cigarettes per day between 40 and 60 years old	numerical				
Years smoking between 60 and 80 years old	numerical				

Variable	Type	Accepted values	Min	Max	Unit
Smoked cigarettes per day between 60 and 80 years old	numerical				
Height	numerical				m
Weight	numerical				kg
Body mass index	numerical				kg/m²
COPD Presence	boolean				
Diabetes Presence	boolean				
Chronic Heart Failure (CHF) Presence	boolean				
Ischaemic Heart Disease (IHD) Presence	boolean				
Anxiety Presence	boolean				
Depression Presence	boolean				
Exacerbation status	boolean				
Did you have more symptoms than usual during the last 24 hours?	coded	- Yes - No			
Daily Activity - Steps walked	numerical				
Daily Activity - Distance travelled	numerical				m
Daily Activity - Calories burned	numerical				cal
Daily Activity - Floors climbed	numerical				
Daily Activity - Lightly active minutes	numerical				min
Daily Activity - Moderately active minutes	numerical				min
Daily Activity - Highly active minutes	numerical				min
Heart - Resting heart rate	numerical				bpm
Heart - Min heart rate	numerical				bpm
Heart - Max heart rate	numerical				bpm
Heart - Out of range minutes	numerical				min
Heart - Fat burn minutes	numerical				min
Heart - Cardio minutes	numerical				min
Heart - Peak minutes	numerical				min
Sleep - Sleep start (hours relative to midnight)	numerical				time
Sleep - Sleep end (hours relative to midnight)	numerical				time
Sleep - REM minutes	numerical				min
Sleep - Light minutes	numerical				min
Sleep - Deep minutes	numerical				min
Sleep - Awake minutes	numerical				min
Sleep - Total minutes	numerical				min
Exercise - Start Time	numerical				time
Exercise - Duration	numerical				time
Exercise - Active Duration	numerical				time
Exercise - Calories	numerical				cal
Exercise - Steps	numerical				
Exercise - Distance	numerical				km

Variable	Type	Accepted values	Min	Max	Unit
Exercise - Average Heart Rate	numerical				bpm
Exercise - Fat Burn Minutes	numerical				min
Exercise - Cardio Minutes	numerical				min
Exercise - Peak Minutes	numerical				min
Exercise - Sedentary Minutes	numerical				min
Exercise - Lightly Active Minutes	numerical				min
Exercise - Fairly Active Minutes	numerical				min
Exercise – Very Active Minutes	numerical				min
36-Item Short Form Health Survey (RAND36) - In general, would you say your health is:	coded	- Excellent - Very good - Good - Fair - Poor			
36-Item Short Form Health Survey - In Compared to one year ago, how would you rate your health in general?	coded	- Much better now than one year ago - Somewhat better now than one year ago - About the same as one year ago - Somewhat worse than one year ago - Much worse than one year ago			
36-Item Short Form Health Survey - Does your health now limit you in: vigorous activities, such as running, lifting heavy objects, participating in strenuous sports	coded	- Yes, limited a lot - Yes, limited a little - No, not limited at all			
36-Item Short Form Health Survey - Does your health now limit you in: moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	coded	- Yes, limited a lot - Yes, limited a little - No, not limited at all			
36-Item Short Form Health Survey - Does your health now limit you in: moderate activities, lifting or carrying groceries?	coded	- Yes, limited a lot - Yes, limited a little - No, not limited at all			
36-Item Short Form Health Survey - Does your health now limit you in: climbing several flights of stairs?	coded	- Yes, limited a lot - Yes, limited a little - No, not limited at all			
36-Item Short Form Health Survey - Does your health now limit you in: climbing one flight of stais?	coded	- Yes, limited a lot - Yes, limited a little - No, not limited at all			
36-Item Short Form Health Survey - Does your health now limit you in: bending, kneeling, or stooping?	coded	- Yes, limited a lot - Yes, limited a little - No, not limited at all			
36-Item Short Form Health Survey - Does your health now limit you in: walking more than a mile?	coded	- Yes, limited a lot - Yes, limited a little - No, not limited at all			
36-Item Short Form Health Survey - Does your health now limit you in: walking several blocks?	coded	- Yes, limited a lot - Yes, limited a little - No, not limited at all			
36-Item Short Form Health Survey - Does your health now limit you in: walking one block?	coded	- Yes, limited a lot - Yes, limited a little - No, not limited at all			
36-Item Short Form Health Survey - Does your health now	coded	- Yes, limited a lot - Yes, limited a little			

Variable	Type	Accepted values	Min	Max	Unit
limit you in: bathing or dressing yourself?		- No, not limited at all			
36-Item Short Form Health Survey - During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health? Cut down the amount of time you spent on work or other activities:	coded	- Yes - No			
36-Item Short Form Health Survey - During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health? Accomplished less than you would like:	coded	- Yes - No			
36-Item Short Form Health Survey - During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health? Had difficulty performing the work or other activities (for example, it took extra effort):	coded	- Yes - No			
36-Item Short Form Health Survey - During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health? Had difficulty performing the work or other activities (for example, it took extra effort):	coded	- Yes - No			
36-Item Short Form Health Survey - During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)? Cut down the amount of time you spent on work or other activities:	coded	- Yes - No			
36-Item Short Form Health Survey - During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)? Accomplished less than you would like:	coded	- Yes - No			
36-Item Short Form Health Survey - During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)? Didn't do work or other activities as carefully as usual:	coded	- Yes - No			
36-Item Short Form Health Survey - During the past 4 weeks, to what extent has your physical health or emotional problems	coded	- Not at all - Slightly - Moderately - Quite a bit			

Variable	Type	Accepted values	Min	Max	Unit
interfered with your normal social activities with family, friends, neighbours, or groups?		- Extremely			
36-Item Short Form Health Survey - How much bodily pain have you had during the past 4 weeks?	coded	- None - Very mild - Mild - Moderate - Severe - Very severe			
36-Item Short Form Health Survey - During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?	coded	- Not at all - A little bit - Moderately - Quite a bit - Extremely			
36-Item Short Form Health Survey - How much of the time during the past 4 weeks did you feel full of pep?	coded	- All of the time - Most of the time - A good bit of the time - Some of the time - A little of the time - None of the time			
36-Item Short Form Health Survey - How much of the time during the past 4 weeks have you been a nervous person?	coded	- All of the time - Most of the time - A good bit of the time - Some of the time - A little of the time - None of the time			
36-Item Short Form Health Survey - How much of the time during the past 4 weeks have you felt so down in the dumps that nothing could cheer you up?	coded	- All of the time - Most of the time - A good bit of the time - Some of the time - A little of the time - None of the time			
36-Item Short Form Health Survey - How much of the time during the past 4 weeks have you felt calm and peaceful?	coded	- All of the time - Most of the time - A good bit of the time - Some of the time - A little of the time - None of the time			
36-Item Short Form Health Survey - How much of the time during the past 4 weeks did you have a lot of energy?	coded	- All of the time - Most of the time - A good bit of the time - Some of the time - A little of the time - None of the time			
36-Item Short Form Health Survey - How much of the time during the past 4 weeks have you felt downhearted and blue?	coded	- All of the time - Most of the time - A good bit of the time - Some of the time - A little of the time - None of the time			
36-Item Short Form Health Survey - How much of the time during the past 4 weeks did you feel worn out?	coded	- All of the time - Most of the time - A good bit of the time - Some of the time - A little of the time - None of the time			
36-Item Short Form Health Survey - How much of the time during the past 4 weeks have you been a happy person?	coded	- All of the time - Most of the time - A good bit of the time - Some of the time - A little of the time - None of the time			
36-Item Short Form Health Survey - How much of the time	coded	- All of the time - Most of the time - A good bit of the time			

Variable	Type	Accepted values	Min	Max	Unit
during the past 4 weeks did you feel tired?		- Some of the time - A little of the time - None of the time			
36-Item Short Form Health Survey - During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?	coded	- All of the time - Most of the time - A good bit of the time - Some of the time - A little of the time - None of the time			
36-Item Short Form Health Survey - Please choose the answer that best describes how true of false each one of the following statements is for you. I seem to get sick a little easier than other people.	coded	- Definitely true - Mostly true - Don't know - Mostly false - Definitely false			
36-Item Short Form Health Survey - Please choose the answer that best describes how true of false each one of the following statements is for you. I am as healthy as anybody I know.	coded	- Definitely true - Mostly true - Don't know - Mostly false - Definitely false			
36-Item Short Form Health Survey - Please choose the answer that best describes how true of false each one of the following statements is for you. I am as healthy as anybody I know.	coded	- Definitely true - Mostly true - Don't know - Mostly false - Definitely false			
36-Item Short Form Health Survey - Please choose the answer that best describes how true of false each one of the following statements is for you. My health is excellent.	coded	- Definitely true - Mostly true - Don't know - Mostly false - Definitely false			
Euro Quality of Life 5 Questions 5 Dimensions 5 Levels (5Q-5D-5L)- Your mobility TODAY	coded	- I have no problems in walking about - I have slight problems in walking about - I have moderate problems in walking about - I have severe problems in walking about - I am unable to walk about			
Euro Quality of Life 5 Questions 5 Dimensions 5 Levels - Your self-care TODAY	coded	- I have no problems washing or dressing myself - I have slight problems washing or dressing myself - I have moderate problems washing or dressing myself - I have severe problems washing or dressing myself - I am unable to wash or dress myself			
Euro Quality of Life 5 Questions 5 Dimensions 5 Levels - Your usual activities TODAY (e.g., work, study, housework, family or leisure activities)	coded	- I have no problems doing my usual activities - I have slight problems doing my usual activities - I have moderate problems doing my usual activities - I have severe problems doing my usual activities - I am unable to do my usual activities			

Variable	Type	Accepted values	Min	Max	Unit
Euro Quality of Life 5 Questions 5 Dimensions 5 Levels - Your usual pain/discomfort TODAY	coded	- I have no pain or discomfort - I have slight pain or discomfort - I have moderate pain or discomfort - I have severe pain or discomfort - I have severe pain or discomfort I have extreme pain or discomfort			
Euro Quality of Life 5 Questions 5 Dimensions 5 Levels - Please tap on the scale to indicate how your health is TODAY.	integer		1	100	
Euro Quality of Life 5 Questions 5 Dimensions 5 Levels - Anxiety / Depression TODAY	coded	- I am not anxious or depressed - I am slightly anxious or depressed - I am moderately anxious or depressed - I am severely anxious or depressed - I am extremely anxious or depressed			
Hospital Anxiety and Depression Scale- I feel tense or 'wound up':	coded	- Most of the time - A lot of the time - Time to time, occasionally - Not at all			
Hospital Anxiety and Depression Scale - I still enjoy the things I used to enjoy:	coded	- Definitely as much - Not quite so much - Only a little - Not at all			
Hospital Anxiety and Depression Scale - I get a sort of frightened feeling like something awful is about to happen:	coded	- Very definitely and quite badly - Yes, but not too badly - A little, but it doesn't worry me - Not at all			
Hospital Anxiety and Depression Scale - I can laugh and see the funny side of things:	coded	- As much as I always could - Not quite so much now - Definitely not so much - Not at all			
Hospital Anxiety and Depression Scale - Worrying thoughts go through my mind:	coded	- A great deal of the time - A lot of the time - From time to time but not too often - Only occasionally			
Hospital Anxiety and Depression Scale - I feel cheerful:	coded	- Not at all - Not often - Sometimes - Most of the time			
Hospital Anxiety and Depression Scale - I can sit at ease and feel relaxed:	coded	- Definitely - Usually - Not often - Not at all			
Hospital Anxiety and Depression Scale - I feel as if I am slowed down:	coded	- Nearly all of the time - Very often - Sometimes - Not at all			
Hospital Anxiety and Depression Scale - I get a sort of frightened feeling like 'butterflies in the stomach':	coded	- Very often - Quite often - Occasionally - Not at all			
Hospital Anxiety and Depression Scale - I have lost interest in my appearance:	coded	- Definitely - I don't take as much care as I should - I may not take quite as much care - I take just as much care as ever			

Variable	Type	Accepted values	Min	Max	Unit
Hospital Anxiety and Depression Scale - I feel restless as if I have to be on the move:	coded	- Very much indeed - Quite a lot - Not very much - Not at all			
Hospital Anxiety and Depression Scale - I look forward with enjoyment to things:	coded	- As much as I ever did - Rather less than I used to - Definitely less than I used to - Hardly at all			
Hospital Anxiety and Depression Scale - I get sudden feelings of panic:	coded	- Very often indeed - Quite often - Not very often - Not at all			
Hospital Anxiety and Depression Scale - I can enjoy a good book or radio or TV programme:	coded	- Often - Sometimes - Not often - Very seldom			
Chronic Obstructive Pulmonary Disease Assessment Test - I never cough / I cough all the time	numerical		0	5	
Chronic Obstructive Pulmonary Disease Assessment Test - I have no phlegm (mucus) in my chest at all / My chest is completely full of phlegm (mucus)	numerical		0	5	
Chronic Obstructive Pulmonary Disease Assessment Test - My chest does not feel tight at all / My chest feels very tight	numerical		0	5	
Chronic Obstructive Pulmonary Disease Assessment Test - When I walk up a hill or one flight of stairs, I am not breathless / When I walk up a hill or one flight of stairs, I am very breathless	numerical		0	5	
Chronic Obstructive Pulmonary Disease Assessment Test - I am not limited doing any activities at home / I am very limited doing activities at home	numerical		0	5	
Chronic Obstructive Pulmonary Disease Assessment Test - I am confident leaving my home despite my lung condition / I am not at all confident leaving my home because of my lung condition	numerical		0	5	
Chronic Obstructive Pulmonary Disease Assessment Test - I sleep soundly / I don’t sleep soundly because of my lung condition	numerical		0	5	
Chronic Obstructive Pulmonary Disease Assessment Test - I have lots of energy / I have no energy at all	numerical		0	5	
Clinical Chronic Obstructive Pulmonary Disease Questionnaire - On average, during the past 24 hours, how often did you feel short of breath at rest?	coded	- Never - Hardly ever - A few times - Several times - Many times - A great many times - Almost all the time			
Clinical Chronic Obstructive Pulmonary Disease Questionnaire - On average, during the past 24 hours, how often did you feel	coded	- Never - Hardly ever - A few times - Several times - Many times			

Variable	Type	Accepted values	Min	Max	Unit
short of breath doing physical activities?		- A great many times - Almost all the time			
Clinical Chronic Obstructive Pulmonary Disease Questionnaire - On average, during the past 24 hours, how often did you feel concerned about getting a cold or breathing getting worse?	coded	- Never - Hardly ever - A few times - Several times - Many times - A great many times - Almost all the time			
Clinical Chronic Obstructive Pulmonary Disease Questionnaire - On average, during the past 24 hours, how often did you feel depressed (down) because of breathing problems?	coded	- Never - Hardly ever - A few times - Several times - Many times - A great many times - Almost all the time			
Clinical Chronic Obstructive Pulmonary Disease Questionnaire - In general, during the past 24 hours, how much of the time did you cough?	coded	- Never - Hardly ever - A few times - Several times - Many times - A great many times - Almost all the time			
Clinical Chronic Obstructive Pulmonary Disease Questionnaire - In general, during the past 24 hours, how much of the time did you produce phlegm?	coded	- Never - Hardly ever - A few times - Several times - Many times - A great many times - Almost all the time			
Clinical Chronic Obstructive Pulmonary Disease Questionnaire - On average, during the past 24 hours, how limited were you in these activities because of your breathing problems: strenuous physical activities (such as climbing stairs, hurrying, doing sports)?	coded	- Never - Hardly ever - A few times - Several times - Many times - A great many times - Almost all the time			
Clinical Chronic Obstructive Pulmonary Disease Questionnaire - On average, during the past 24 hours, how limited were you in these activities because of your breathing problems: moderate physical activities (such as walking, housework, carrying things)?	coded	- Not limited at all - Very slightly limited - Slightly limited - Moderately limited - Very limited - Extremely limited - Totally limited/or unable to do			
Clinical Chronic Obstructive Pulmonary Disease Questionnaire - On average, during the past 24 hours, how limited were you in these activities because of your breathing problems: daily activities at home (such as dressing, washing yourself)?	coded	- Not limited at all - Very slightly limited - Slightly limited - Moderately limited - Very limited - Extremely limited - Totally limited/or unable to do			
Clinical Chronic Obstructive Pulmonary Disease Questionnaire - On average, during the past 24 hours, how limited were you in these activities because of your breathing problems: social activities (such as talking, being with children, visiting friends/relatives)?	coded	- Not limited at all - Very slightly limited - Slightly limited - Moderately limited - Very limited - Extremely limited - Totally limited/or unable to do			

Variable	Type	Accepted values	Min	Max	Unit
Clinical Chronic Obstructive Pulmonary Disease Questionnaire - On average, during the past 24 hours, how often did you feel short of breath doing physical activities?	coded	- Not limited at all - Very slightly limited - Slightly limited - Moderately limited - Very limited - Extremely limited - Totally limited/or unable to do			
New York Heart Association Functional Classification - Do you ever have heart problems during normal activities?	coded	- No limitation of physical activity. Ordinary physical activity does not cause undue fatigue, palpitation, dyspnoea (shortness of breath). - Slight limitation of physical activity. Comfortable at rest. Ordinary physical activity results in fatigue, palpitation, dyspnoea. - Marked limitation of physical activity. Comfortable at rest. Less than ordinary activity causes fatigue, palpitation, or dyspnoeas. - Unable to carry on any physical activity without discomfort. Symptoms of heart failure at rest. If any physical activity is undertaken, discomfort increases.			
Modified Medical Research Council Dyspnea Scale - Please choose the sentence that best describes when you experience your shortness of breath:	coded	- I only get breathless with strenuous exercise. - I get short of breath when hurrying on level ground or walking up a slight hill. - On level ground, I walk slower than people of the same age because of breathlessness or have to stop for breath when walking at my own pace. - I stop for breath after walking about 100 yards or after a few minutes on level ground. - I am too breathless to leave the house or I am breathless when dressing.			
Chronic Obstructive Pulmonary Disease Symptoms - Please indicate which symptoms have changed during the last 24 hours for: Breathlessness	coded	- Not more than usual - Slightly more than usual - Significantly more than usual			
Chronic Obstructive Pulmonary Disease Symptoms - Please indicate which symptoms have changed during the last 24 hours for: Sputum volume	coded	- Not more than usual - Slightly more than usual - Significantly more than usual			
Chronic Obstructive Pulmonary Disease Symptoms - Please indicate which symptoms have changed during the last 24 hours for: Sputum Colour	coded	- Usual for me - Different from usual - Significantly more than usual			
Chronic Obstructive Pulmonary Disease Symptoms - Did you have a fever (more than 38.5C) in the last 24 hours?	coded	- Yes - No - Not more than usual - Slightly more than usual - Significantly more than usual			
Chronic Obstructive Pulmonary Disease Symptoms - Did you	coded	- Not more than usual - Slightly more than usual			

Variable	Type	Accepted values	Min	Max	Unit
experience a significant change in coughing in the last 24 hours?		- Significantly more than usual			
Chronic Obstructive Pulmonary Disease Symptoms - Did you experience a significant change in wheezing in the last 24 hours?	coded	- Not more than usual - Slightly more than usual - Significantly more than usual			
Anxiety and Depression Symptoms - Please indicate which symptoms have changed during the last 24 hours for: Felt Depressed	coded	- Not more than usual - Slightly more than usual - Significantly more than usual			
Anxiety and Depression Symptoms - Please indicate which symptoms have changed during the last 24 hours for: Felt anxious	coded	- Not more than usual - Slightly more than usual - Significantly more than usual			
Chronic Heart Failure Symptoms - Please indicate which symptoms have changed during the last 24 hours for: Weight	coded	- Not more than usual - Slightly more than usual - Significantly more than usual			
Chronic Heart Failure Symptoms - Please indicate which symptoms have changed during the last 24 hours for: Swelling of ankles or abdomen	coded	- Not more than usual - Slightly more than usual - Significantly more than usual			
Chronic Heart Failure Symptoms - Please indicate which symptoms have changed during the last 24 hours for: Waking up at night short of breath	coded	- Not more than usual - Slightly more than usual - Significantly more than usual			
Chronic Heart Failure Symptoms - Please indicate which symptoms have changed during the last 24 hours for each symptom: Felt light headed or dizzy	coded	- More than usual - Not more than usual - Slightly more than usual - Significantly more than usual			
Ischaemic Heart Disease Symptoms - Did you experience a change in: pain, pressure, heaviness, tightness in one or more of your: chest - neck - jaw - arm(s) - back - shoulders?	coded	- Yes - No			
Ischaemic Heart Disease Symptoms - Did you experience a sudden change in your breathing resulting in severe shortness of breaths?	coded	- Yes - No			
Ischaemic Heart Disease Symptoms - Did you experience black-outs?	coded	- Yes - No			

([1]) *ISO8601 'T format' is* YYYY-MM-DDTHH:mm:ss.SSSZ.

3.2 Hospital Information System dataset

A second source of information is the HIS, which collects the variables coming from the hospital's Electronic Health Record (EHR) **(**Table 3**).** Data collected varies from patient's scheduled follow-ups (including baseline condition) to emergency visits, unexpected hospitalizations between follow-ups or even medication updates over time.

Regarding *follow-ups*, they are carried out by the clinicians in each clinical site, with special distinction between the *baseline* (month "0") and the following, according to protocol, at every six months ("6", "12", "18", "24", "30" and "36"). These are scheduled interactions planned by procedure.

Events of *hospitalization* include information collected when a patient is hospitalized due to some complication. This also has its own procedure of specific measurements and medical tests, such as arterial blood gas tests.

Additionally, other circumstances may add data, too. For example, *emergency visits* not ending in hospitalization or changes in medication prescriptions due to diverse causes (for instance, exacerbations). To deal with these circumstances, data ingestion could be done without the obligation to link them to specific grouping events (encounters follow-up or hospitalization). This could change in the future if data maintenance needs require it for better understanding or to facilitate exploitation.

Within the three pilot clinical sites, the availability and frequency of downloading these data must be agreed according to the specific characteristics of their systems, although the idea is that the sooner the data is available in aggregated form, the sooner it could be used by machine learning algorithms.

In addition, it must be considered that part of the performed tests may be done in different medical centres (e.g.: Shared Care facilities) and/or their results are obtained later in time (e.g.: blood sampling results). The final details of the aggregation and ingestion of the data is still under discussion with the clinical sites, depending on their own internal technical procedures and integration capabilities.

What is listed below in Table 4 is the result of agreements between the three pilot sites (GEM, MST and TUK) on the availability of data and feasibility and convergence in this common model, regardless of the intermediate tools and internal processes of extraction and facilitation of this data. That is, the subset of data of the RE-SAMPLE reference data model expected to be ingested from the EHR of the HIS of the pilot sites (hospitals).

It should be noted that there are some attributes that can be ingested from both Healthentia and the HIS, such as *weight* information. Therefore, they are included in the dataset specification of both data sources, although in the HDH they will end up being modelled indistinctly in the same final standard resources.

The other issue to consider is that the clinical site itself is responsible for ingesting their data, performing it incrementally or with updates on existing objects that are agreed upon, depending on the use cases considered. To extract data and make it available to the ML modules, the necessary filters will be applied. For example, although several observations of weight could be stored for the patient in a recent period, it may occur that only the last one is the useful and therefore necessary, so the others are omitted. This needs to be agreed by the pilot sites.

Regarding the medication related fields, the fact of reporting a prescription start date implies the creation of the associated resource. This is an atomic resource for each specific medication. In case of updating the parameters in the following follow-ups (or due to unexpected exacerbations or emergency visits), it would be updated on the same resource.

Table 3: Hospital Information System dataset

		Variable	Type	Accepted values	Min	Max	Unit
Patient		Country	coded	ISO-3166 (3 letters)			
		Zip code	numerical				
FollowUp	General	Date	date	ISO8601			([1])
		Number of exacerbations in the past 2 years	numerical				
		Number of hospitalizations in the past 2 years	numerical				
		Number of exacerbations in the last year	numerical				
		Number of hospitalizations in the last year	numerical				
		Mini-Mental State Exam	numerical		0	4	
		Modified Medical Research Council Dyspnea Scale	numerical		0	30	
		Smoking status	coded	- Smoker - Ex-smoker			
		Height	numerical				m
		Weight	numerical				kg
		Body mass index	numerical				kg/m²
	Spirometry	Spirometry - Forced Expired Volume in 1 second	numerical				L
		Spirometry - Predicted Percentage FEV1	numerical		0	100	%
		Spirometry - Forced Vital Capacity	numerical				L
		Spirometry - FEV1/FVC	numerical				
		Post short-acting bronchodilators spirometry - FEV1	numerical				L
		Post short-acting bronchodilators spirometry - Predicted Percentage FEV1	numerical		0	100	%
		Post short-acting bronchodilators spirometry - FVC	numerical				L
		Post short-acting bronchodilators spirometry - FEV1 /FVC	numerical				
	Six Minute Walking Test	Six-minute walking test - Medication	boolean				
		Six-minute walking test - Walking aid	boolean				
		Six-minute walking test - Oxygen use	boolean				
		Six-minute walking test - Oxygen used	numerical				L
		Six-minute walking test - Systolic pressure before test	numerical				mmHg
		Six-minute walking test - Diastolic pressure before test	numerical				mmHg
		Six-minute walking test - Walked distance	numerical				m
		Six-minute walking test - Theoretical walked distance base on BMI and Age	numerical				m
		Six-minute walking test - If the patient has stopped	boolean				
		Six-minute walking test - Oxygen saturation at baseline	numerical		0	100	%
		Six-minute walking test - Oxygen saturation in min 1	numerical		0	100	%
		Six-minute walking test - Oxygen saturation in min 2	numerical		0	100	%
		Six-minute walking test - Oxygen saturation in min 3	numerical		0	100	%
		Six-minute walking test - Oxygen saturation in min 4	numerical		0	100	%
		Six-minute walking test - Oxygen saturation in min 5	numerical		0	100	%
		Six-minute walking test - Oxygen saturation in min 6	numerical		0	100	%
		Six-minute walking test - Minimum Oxygen saturation during the test	numerical		0	100	%
		Six-minute walking test - Percentage of time that patient has SP02 below 85%	numerical		0	100	%
		Six-minute walking test - Heart rate at baseline	numerical				bpm
		Six-minute walking test - Heart rate in min 1	numerical				bpm
		Six-minute walking test - Heart rate in min 2	numerical				bpm
		Six-minute walking test - Heart rate in min 3	numerical				bpm
		Six-minute walking test - Heart rate in min 4	numerical				bpm
		Six-minute walking test - Heart rate in min 5	numerical				bpm
		Six-minute walking test - Heart rate in min 6	numerical				bpm
		Six-minute walking test - Borg score dyspnea before test	numerical				
		Six-minute walking test - Borg score dyspnea after test	numerical				
		Six-minute walking test - Borg score fatigue before test	numerical				
		Six-minute walking test - Borg score fatigue after test	numerical				
	Blood Test	Hemoglobin	numerical				mmol/L
		Hematocrit	numerical				L/L
		Thrombocytes	numerical				$\times 10^9$ / L
		Leukocytes	numerical				$\times 10^9$ / L
		Fibrinogen	numerical				g/L

		Variable	Type	Accepted values	Min	Max	Unit
		Eosinophils	numerical				$\times 10^9$ / L
		Basophils	numerical				$\times 10^9$ / L
		Neutrophils	numerical				$\times 10^9$ / L
		Lymphocytes	numerical				$\times 10^9$ / L
		Monocytes	numerical				$\times 10^9$ / L
		C-reactive protein or Hs-CRP	numerical				mg/L
		NT-proBNP	numerical				pg/mL
		HbA1c	numerical				mmol/mol
	Medication Administration	SAMA (start date)	date	ISO8601			([1])
		SAMA (end date)	date	ISO8601			([1])
		SABA (start date)	date	ISO8601			([1])
		SABA (end date)	date	ISO8601			([1])
		LABA (start date)	date	ISO8601			([1])
		LABA (end date)	date	ISO8601			([1])
		LAMA (start date)	date	ISO8601			([1])
		LAMA (end date)	date	ISO8601			([1])
		ICS (start date)	date	ISO8601			([1])
		ICS (end date)	date	ISO8601			([1])
		Antibiotics (start date)	date	ISO8601			([1])
		Antibiotics (end date)	date	ISO8601			([1])
		OCS - Oral corticosteroids (start date)	date	ISO8601			([1])
		OCS - Oral corticosteroids (end date)	date	ISO8601			([1])
		PDE4-inhibitor (start date)	date	ISO8601			([1])
		PDE4-inhibitor (end date)	date	ISO8601			([1])
		ACE-inhibitors (start date)	date	ISO8601			([1])
		ACE-inhibitors (end date)	date	ISO8601			([1])
		ARB (start date)	date	ISO8601			([1])
		ARB (end date)	date	ISO8601			([1])
		Beta blockers (start date)	date	ISO8601			([1])
		Beta blockers (end date)	date	ISO8601			([1])
		Diuretics (start date)	date	ISO8601			([1])
		Diuretics (end date)	date	ISO8601			([1])
		Digoxin (start date)	date	ISO8601			([1])
		Digoxin (end date)	date	ISO8601			([1])
		SGLT2-inhibitors (start date)	date	ISO8601			([1])
		SGLT2-inhibitors (end date)	date	ISO8601			([1])
		Ivabradine (start date)	date	ISO8601			([1])
		Ivabradine (end date)	date	ISO8601			([1])
		Neprilysin-inhibitors (start date)	date	ISO8601			([1])
		Neprilysin-inhibitors (end date)	date	ISO8601			([1])
		Nitrate (start date)	date	ISO8601			([1])
		Nitrate (end date)	date	ISO8601			([1])
		Calcium antagonists (start date)	date	ISO8601			([1])
		Calcium antagonists (end date)	date	ISO8601			([1])
		Antiplatelets (Antiagrregants) (start date)	date	ISO8601			([1])
		Antiplatelets (Antiagrregants) (end date)	date	ISO8601			([1])
		Anticoagulants (start date)	date	ISO8601			([1])
		Anticoagulants (end date)	date	ISO8601			([1])
		Statins (start date)	date	ISO8601			([1])
		Statins (end date)	date	ISO8601			([1])
		Ezetimib (start date)	date	ISO8601			([1])
		Ezetimib (end date)	date	ISO8601			([1])
		Benzodiazepines (start date)	date	ISO8601			([1])
		Benzodiazepines (end date)	date	ISO8601			([1])
		Z-Products (start date)	date	ISO8601			([1])
		Z-Products (end date)	date	ISO8601			([1])
		SSRI - Selective serotonin reuptake inhibitors (start date)	date	ISO8601			([1])
		SSRI - Selective serotonin reuptake inhibitors (end date)	date	ISO8601			([1])
		SNRI - Serotonin and norepinephrine reuptake inhibitors (start date)	date	ISO8601			([1])
		SNRI - Serotonin and norepinephrine reuptake inhibitors (end date)	date	ISO8601			([1])
		Noradrenaline and dopamine reuptake inhibitors (start date)	date	ISO8601			([1])

Variable			Type	Accepted values	Min	Max	Unit
		Noradrenaline and dopamine reuptake inhibitors (end date)	date	ISO8601			([1])
		Tricyclic antidepressants (start date)	date	ISO8601			([1])
		Tricyclic antidepressants (end date)	date	ISO8601			([1])
		MAO inhibitors (start date)	date	ISO8601			([1])
		MAO inhibitors (end date)	date	ISO8601			([1])
		Anti-epileptic drugs (start date)	date	ISO8601			([1])
		Anti-epileptic drugs (end date)	date	ISO8601			([1])
		Lithium (start date)	date	ISO8601			([1])
		Lithium (end date)	date	ISO8601			([1])
		Quetiapine (start date)	date	ISO8601			([1])
		Quetiapine (end date)	date	ISO8601			([1])
		Insulin (start date)	date	ISO8601			([1])
		Insulin (end date)	date	ISO8601			([1])
		Metformin (start date)	date	ISO8601			([1])
		Metformin (end date)	date	ISO8601			([1])
		Sulfonylureumderivates (start date)	date	ISO8601			([1])
		Sulfonylureumderivates (end date)	date	ISO8601			([1])
		Glinidines (start date)	date	ISO8601			([1])
		Glinidines (end date)	date	ISO8601			([1])
		Glitazones (start date)	date	ISO8601			([1])
		Glitazones (end date)	date	ISO8601			([1])
		GLP-1-analogs (start date)	date	ISO8601			([1])
		GLP-1-analogs (end date)	date	ISO8601			([1])
		DPP-4-inhibitors (start date)	date	ISO8601			([1])
		DPP-4-inhibitors (end date)	date	ISO8601			([1])
		Acarbose (start date)	date	ISO8601			([1])
		Acarbose (end date)	date	ISO8601			([1])
Hospitalization		Admission date	date	ISO8601			([1])
		Discharge date	date	ISO8601			([1])
		Oxygen use ([2])	boolean	true/false			
		Mechanical ventilation ([2])	boolean	true/false			
		Presence of pneumonia	boolean	true/false			
	Arterial Blood Gas Test	Blood pH level	numerical		7.0	8.0	pH
		Partial pressure of carbon dioxide (PaCO2)	numerical				mmHg
		Bicarbonate (HCO3)	numerical				mmol/L ([3])
		Base Excess	numerical				mmol/L ([3])
		Partial pressure of oxygen (PaO2)	numerical				mmHg
		Oxygen saturation (O2 Sat)	numerical		0	100	%

([1]) *ISO8601 'T format' is* YYYY-MM-DDTHH:mm:ss.SSSZ.
([2]) *At hospitalization date.*
([3]) *Aware that mEq/L is also commonly used, but the use of mmol/L has been fixed for ingestion.*

3.3 Machine Learning modules dataset

The main purpose of the RE-SAMPLE platform is to prevent COPD exacerbations as good as possible and to warn clinicians and their patients if the patient's health condition is worsening. This can be achieved using predictions generated by ML models. The ML dataset that is described in this section and summarized in Table 4 contains variables computed by the predictive ML models when applied to the patient data.

Whenever there is new patient data available, an up-to-date prediction will be calculated by the ML models developed in WP3, i.e., Task 3.1. The previous predictions will not be overridden but still saved because the prediction history is considered helpful to have an overview of the patient's progress over time.

The medical outcomes that will be predicted for every patient include quality-of-life (QoL) scores like the EQ-5D (EuroQoL, 2022), Chronic Respiratory Disease Questionnaire (CRQ) (CRQ, 2022) and Hospital Anxiety and Depression Scale (HADS) (HADS, 2022), as well as the exacerbation risk. To link the prediction with the patient and provide the prediction history, the predictions always go along with the internal RE-SAMPLE patient ID and their date and time of creation.

Moreover, every prediction (every type and for every point in time) has explanations that justify the values to help the patient and clinician understand and gain trust in the predictions. In addition to that, they provide input for the virtual companionship programme (WP5) and support the shared-decision-making (WP2). There are three types of explanations:

- SHapley Additive exPlanations (SHAP) (Molnar, 2020), based on Shapley Values, is a feature importance method that explains how each of the patients' measurements influence a baseline prediction for the observed patients. This means that for every patient, every prediction and for every predictor variable there is an associated numerical computed, as well as a baseline value,
- The explanations provided by the model Explainable Boosting Machine (EBM) that makes both the influence of each feature to every patient's individual prediction and the global behaviour of the model available, so as for SHAP, there is a numerical computed for every predictor used from the patient's data with an additional intercept,
- The counterfactual explanations (Molnar, 2020) can provide intervention suggestions for every patient because the output of this method is the minimally different patient's data that would lead to a maximally improved prediction, as well as this new prediction value.

Another list that is provided is a set of simulations. A simulation is also linked to a specific prediction. It is a set of slightly modified patient data (e.g., the smoking status changed from "smoker" to "previous smoker") that goes along with the prediction of the ML model after applying it to this slightly modified patient data. For every prediction there can be several simulations, one for each modification that is considered useful by the clinician. The variables in the simulations are the same as the patient fields that are used for a prediction.

There is no need to validate the accepted values since it is mainly the patient data that is already validated before, and the output of the models will be validated within the ML components. In addition, the mapping of this information in FHIR standard resources (section 0

Risk Assessment for ML results), aided by extensions, ensures its storage with the necessary semantic information.

Regarding the lists of features in the explanations *(n)* and simulations *(m)* sections in the expected data structure to be ingested, the data entered in the key-value lists will not be validated. It will be stored as it is handled as string in the extensions created in the standard resource that maintains the prediction information, regardless they could conceptually be of numeric type.

Table 4: Machine Learning Modules dataset (clinical data).

Variable			Type	Accepted values	Min	Max	Unit
Prediction		Patient ID	string				
		Model ID	string				
		Prediction Name	coded	- ModerateExacerbation - SevereExacerbation - EQ-5D - CRQ - HADS			
		Type	coded	- Classification - Regression			
		Date	date	ISO8601			([1])
		Timeframe	numerical	integer	1	52	weeks
		Value	numerical		Classification: 0 Regression: none	Classification: 1 Regression none	%
Explanations (3)	Shap	Baseline prediction	numerical		Classification: 0 Regression: none	Classification: 1 Regression none	%
		Feature name *(n)*	coded	- Age - Packyears - Weight - BMI - FEV1_L *Etc.* ([3])			
		Feature value *(n)*	string	Any numeric will be treated as a string	Classification: -1 Regression: none	Classification: 1 Regression none	
	EBM	Intercept	numerical	Floating point num.	none	none	
		Feature name *(n)*	coded	- Age - Packyears - Weight - BMI - FEV1_L *Etc.* ([3])	Same as feature in prospective data	Same as feature in prospective data	
		Feature value *(n)*	string	Any numeric will be treated as a string			
	Counterfactual.	New prediction value	numerical		Classification: 0 Regression: none	Classification: 1 Regression: none	%
		Feature name *(n)*	coded	- Age - Packyears - Weight - BMI - FEV1_L *Etc.* ([3])			([2])
		Feature value *(n)*	string	Any numeric will be treated as a string	Same as feature in prospective data	Same as feature in prospective data	
Simulations (m)		date	date	ISO8601			([1])
		New prediction value	numerical		Classification: 0 Regression: none	Classification: 1 Regression: none	%
		Feature name *(m)*	coded	- Age - Packyears - Weight - BMI - FEV1_L *Etc.* ([3])			([2])
		Feature value *(m)*	string	Any numeric will be treated as a string	Same as feature in prospective data	Same as feature in prospective data	

([1]) *ISO8601 'T format' is* YYYY-MM-DDTHH:mm:ss.SSSZ

([2]) *Same as the feature*

([3]) *The 'names' allowed will be maintained in the model, specifying the type: decimal/integer/string.*

As an example, the following shows a preliminary structure that represents a prediction ready for ingestion in the HDH, alongside with explanations and optional simulations.

Example

```
{
  "patientId": "0a0a0a0a-0a0a-0a0a0a0a0-0a0a0a0a0a0a",
  "name": "ModerateExacerbation",
  "type": "Classification",
  "modelId": "aaaaaaa",
  "date": "2022-05-15T12:00:00.000Z",
  "timeFrame": 26,
  "value": 0.85
  "shap": {
    "baseline": 0.12,
    "age": -0.07,
    "packyears": 0.2,
    "weight": 0.75,
    "fev1_l": -0.11
  },
  "ebm": {
    "intercept": 0.48,
    "packyears": -0.78,
    "bmi": 0.23,
    "fev1_l": -1.234
  },
  "counterfactual": {
    "value": 0.56,
    "packyears": -0.78,
    "weight": 0.23
  },
  "simulations": [{
      "date": "2022-05-16T12:00:00.000Z",
      "age": 72,
      "packyears": 24,
      "weight": 60,
      "fev1_l": 1.4,
      "value": 0.25
    },{
      "date": "2022-05-17T12:00:00.000Z",
      "age": 80,
      "packyears": 15,
      "weight": 70,
      "fev1_l": 1.5,
      "value": 0.4
    }]
}
```

(n) *(n)* *(n)* *(m)* *(m)*

In addition to the clinical data specifically aggregated in the HDH in the context of the RE-SAMPLE data model, an auxiliary parallel database maintains environmental data to be used as input in the ML algorithms. The types and units used are described in Table 5. The unit system is UCUM (*UCUM*, no date), same as for the rest of the data.

Table 5: Machine Learning Modules dataset (environmental variables).

Variable	Type	Accepted values	Min	Max	Unit
Country	coded	ISO-3166 3 letters			-
ZIP_Code	numerical				-
Air Quality Index	numerical				-
Carbon monoxide	numerical				μg/m³
Nitrogen monoxide	numerical				μg/m³
Nitrogen dioxide	numerical				μg/m³
Ozone	numerical				μg/m³
Sulfur dioxide	numerical				μg/m³
Ammonia	numerical				μg/m³
PM2,5	numerical				μg/m³
PM10	numerical				μg/m³
Temperature	numerical				K
Feels_like	numerical				K
Temp_min	numerical				K
Temp_max	numerical				K
Pressure	numerical				hPa
Humidity	numerical				%
Wind_speed	numerical				m/s

3.4 Scores based on other variables

Some input variables for the ML algorithms are scores obtained and calculated based on other variables from the RE-SAMPLE reference model dataset.

Table 6 shows the formulae for the calculations done by the ML modules consuming the clinical data from the HDH. The HDH is responsible for standardization and storage of the variables to be used for such calculations, but it is not meant to be the place to host this logic.

Table 6: Formulas of calculated variables.

Score	Formula
Packs years	Taking as a reference that a pack has 20 cigarettes. packs/year= (smoked cigarettes per day / 20 cigarettes in 1 pack) * 365 days per year * years of smoking
Global Initiative for Chronic Obstructive Lung Disease Criteria (GOLD)	I - Predicted Percentage FEV1 ≥ 80 II - Predicted Percentage FEV1 50-79 III - Predicted Percentage FEV1 30-49 IV - Predicted Percentage FEV1 < 30
Global Initiative for Chronic Obstructive Lung Disease Criteria ABCD based on mMRC and exacerbation risk	A - 0 or 1 moderate exacerbations in past 1 year, mMRC score 0-1, CAT <10 B - 0 or 1 moderate exacerbations in past 1 year, mMRC score ≥ 2, CAT ≥10 C - ≥2 moderate exacerbations or ≥1 leading to hospital admission in past 1 year, mMRC score 0-1, CAT <10 D - ≥2 moderate exacerbations or ≥1 leading to hospital admission in past 1 year, mMRC score ≥ 2, CAT ≥10
Body mass index, Airflow Obstruction, Dyspnea index (BOD)	Sum score of: ○ Predicted Percentage FEV1- max 3 points: ○ ≥ 65 0 points ○ 50 to 64 1 point ○ 36 to 49 2 points ○ ≤ 35 3 points ○ mMRC- max 3 points: ○ mMRC 0 to 1 0 points ○ mMRC 2 1 point ○ mMRC 3 2 points ○ mMRC 4 3 points ○ Body mass index - max 1 point: ○ > 21 0 points ○ ≤ 21 1 point
Body mass index, Airflow Obstruction, Dyspnea and Exercicse capacity index (BODE)	Sum score of: ○ Predicted Percentage FEV1- max 3 points: ○ ≥ 65 0 points ○ 50 to 64 1 point ○ 36 to 49 2 points ○ ≤ 35 3 points ○ mMRC - max 3 points: ○ mMRC 0 to 1 0 points ○ mMRC 2 1 point ○ mMRC 3 2 points ○ mMRC 4 3 points ○ Six-minute walking test distance in meters- max 3 points: ○ ≥ 350 0 points ○ 250 to 349 1 point ○ 150 to 249 2 points ○ ≤ 149 3 points ○ Body mass index - max 1 point: ○ > 21 0 points

Score	Formula
	○ ≤ 21 1 point
Age, Dyspnoea and airflow Obstruction score (ADO)	Sum score of: ○ Predicted Percentage FEV1- max 2 points: ○ ≥ 65 0 points ○ 36 to 64 1 point ○ ≤ 35 2 points ○ mMRC - max 3 points: ○ mMRC 0 to 1 0 points ○ mMRC 2 1 point ○ mMRC 3 2 points ○ mMRC 4 3 points ○ Age in years - max 5 points: ○ 40-49 0 points ○ 50-59 1 point ○ 60-69 2 points ○ 70-79 3 points ○ 80-89 4 points ○ ≥ 90 5 points

4. The Health Data Hub

The HDH is the component in charge of receiving and storing all the data specified in Section 3, following clinical standards, aggregating heterogeneous data sources, and according to the workflow of the RE-SAMPLE project. This means that it acts as a standardized interoperability layer.

The components of the HDH are shown in Figure 3 below. This diagram is extracted from deliverable *D2.6 Architecture and technical specifications v0.1* where the architecture is explained in depth.

The component is located within the Edge Node at the clinical site and is designed according to the security scheme proposed for RE-SAMPLE. The directive is to reject any access into the edge node coming directly from the outside internet. This also means that the communication between Healthentia and the HDH requires an intermediate component, the *synchronizer*, to perform the request from the HDH to the Healthentia API. This component will be developed by ATOS. The rest of the components that interact with the HDH will not have this problem since they are hosted within the security domain of the Edge Node.

The HDH is made up of several modules depicted in the Figure 3:

- The **FHIR Implementation Guide** module hosts the FHIR profiles for the RE-SAMPLE project. The development of this guide is a part of Task 4.6 and will be documented in detail in the deliverable D4.4 Multi-modal data aggregation and curation [M42].
- The **Clinical Data Repository API** is the gateway to the CDR and performs the curation, aggregation, and standardization tasks described in this document. It will be documented following the OpenAPI specification.
- The **Clinical Data Repository Synchronizer** will perform the task of requesting data from the Healthentia API to maintain in the centralized CDR.
- The **Clinical Data Repository** is a HL7 FHIR repository based on HAPI-FHIR (Health Level 7, 2022) and will handle the storage and validation of the standardized information. Its configuration is defined in the third module of the repository, the implementation guide (IG) module.

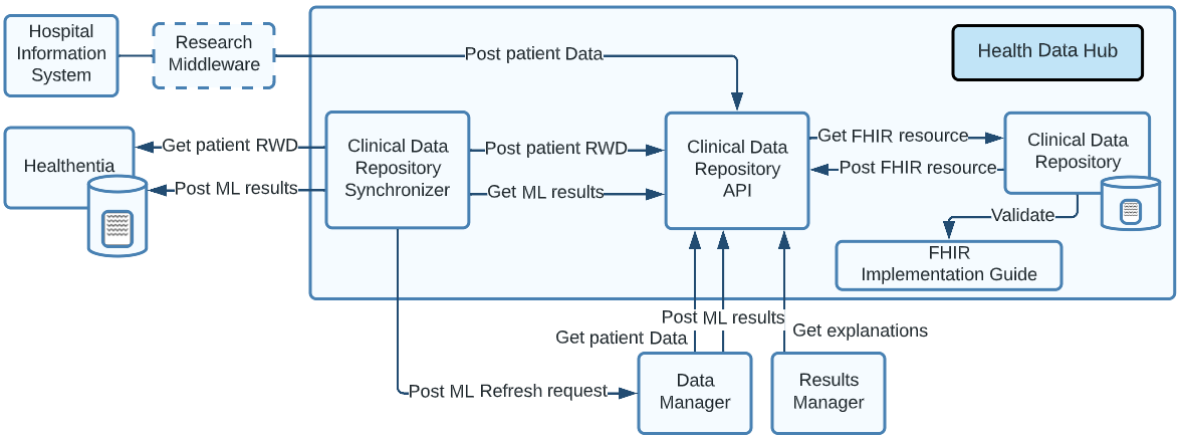

Figure 3: Health Data Hub components.

4.1 Clinical standards and terminologies

Healthcare data is characterized by its strong hierarchical structure and high complexity. The maintainability and extension of the systems that host this type of information is difficult and requires specific actions. Interoperability is defined as the ability of a product to function with another existing or future product or system without restriction of access or implementation. Interoperability differs from integration in that the latter is limited to the articulation of different components and given the rapid evolution of a system, it needs to incorporate new functionalities. The advantage of interoperable applications is the ease of scalability, maintainability, and simplicity to interact with other systems effectively without incorporating new functionalities and without losing information. One of the fundamental mechanisms for achieving interoperability is the application of standards.

After defining a common data model for the RE-SAMPLE project, the tools and actions taken to achieve interoperability are discussed below. These measures not only promote proper communication within the components of the project itself but also open the door to potential systems that want to use the platform in the future. In the RE-SAMPLE platform, the achievement of interoperability has been approached from two levels:

- Technical level: related to the relationship between systems and service. It considers the connections between these, the presentation of information and accessibility. It belongs to the **FHIR standard** scope (Health Level 7, 2022).
- Semantic level: allows information to be exchanged and interpreted unambiguously by systems that have not participated in its creation. It belongs to the scope of **clinical terminologies**. For the RE-SAMPLE project, apart from FHIR code systems, the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) (SNOMED International, 2022) has been chosen as the reference terminology.

Information models and terminology models can work together to enhance databases and EHRs. Figure 4 conveys the strengths and weaknesses of both terminology models and information models and the need to consider both when designing an overall model to store data.

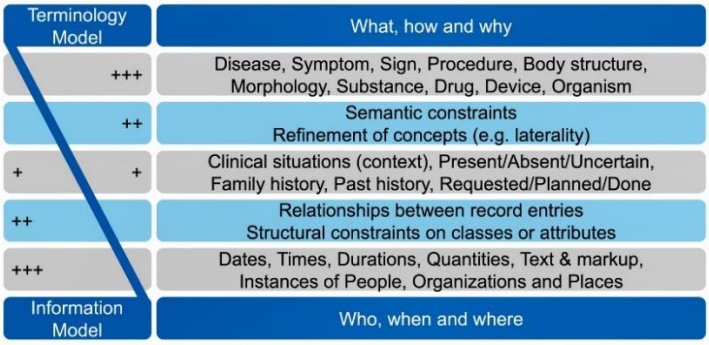

Figure 4: Spectrum of strengths of terminology and information models.

There are aspects that terminologies are best suited for, such as recording diseases, procedures, medications, and refining the records to add more specificity. In other words, this is the "what, how, and why". On the other hand, there are other issues that an information model does best, such as recording information about patients and quantities or defining the relationships between EHR data elements. This is the "who, when, and where". In the middle, there are areas that can be addressed by terminology or the information model, such as recording context specific information or family history.

Using FHIR and SNOMED CT together when designing a data model covers a greater scope since all these areas are represented and standardised by the information model or the terminology. In fact, in 2017, the SNOMED on FHIR working group was established to support FHIR implementations which use SNOMED CT as a terminology. Some new necessities coming from this team were covered by creating new FHIR content and resources such as *Profiles* and *ValueSets*, as well as a SNOMED-FHIR IG. This demonstrates that FHIR and SNOMED CT can work together as part of an EHR to facilitate quick, standard-based, and context-specific search and data entry.

4.2 Fast Healthcare Interoperability Resources (FHIR)

HL7 proposes FHIR standard (Health Level 7, 2022) in response to interoperability oriented web standards. HL7 is an organization certified by the American National Standard Institute (ANSI) and recognised as a global authority in interoperability standards of Information Technology (IT). They do not develop software but create rules to facilitate the exchange of clinical and administrative information. The main objectives of the standard are:

- Information exchange between applications developed by different providers of software.
- Improve decision support and enabling integrated HER.
- Allow connectivity between heterogeneous systems at competitive costs.
- Flexibility, allowing its implementation using various software technologies.

- Reduce resources spent on negotiating interfaces between applications.
- Reduce resources spent on programming and maintenance of proprietary interfaces.

For this project, the design is being made with the version R4, specifically v4.3.0 (most recent release in summer 2022).

4.2.1 Resource

FHIR defines the resource as the basic unit of the specification, and all interchangeable information is defined within a resource. It is a representation of healthcare concepts, a small set of main properties that has an identifier with which to register, locate and retrieve it. Each resource represents a clinical-related item and contains several attributes that can be filled to represent that item.

Resources can be used individually or grouped as messages, documents, or small services. They can be related to each other with references by unique identifiers. The philosophy of FHIR is that with the resources and their combination, called bundles, most use cases in healthcare can be represented. However, the standard also offers extension mechanisms to add new properties to the resources.

Figure 5: HL7 FHIR resources sections, Patient example.

The resources share the common characteristics represented in Figure 5:

- A common way to define and represent them through data types that define common reusable patterns of elements
- A common set of metadata
- A narrative part with extensions and attributes

4.2.2 RESTFul APIs oriented architecture

As previously mentioned, this specification is service-oriented, more specifically to Representational State Transfer (REST). FHIR uses the operations of this standard architecture to establish logical interactions at instance level, at type level and at system level, listed below in Figure 6.

Instance Level Interactions	
read	Read the current state of the resource
vread	Read the state of a specific version of the resource
update	Update an existing resource by its id (or create it if it is new)
delete	Delete a resource
history	Retrieve the update history for a particular resource
Type Level Interactions	
create	Create a new resource with a server assigned id
search	Search the resource type based on some filter criteria
history	Retrieve the update history for a particular resource type
validate	Check that the content would be acceptable as an update
Whole System Interactions	
conformance	Get a conformance statement for the system
transaction	Update, create or delete a set of resources as a single transaction
history	Retrieve the update history for all resources
search	Search across all resource types based on some filter criteria

Figure 6: FHIR REST Operations.

In the context of RE-SAMPLE, the module in charge of managing the repository will wrap this API with a simplified API adapted to the specific project use cases, thus achieving more secure and controlled access to the data by other components. Nevertheless, access to the native FHIR API could be provided, if necessary, in the future.

4.2.3 Data types

The elements of a resource can be defined by a group of types. There are four categories of data types in the R4 version (4.0.1) of HL7 FHIR (Health Level 7, 2022) used in the RE-SAMPLE project. Section 3 The RE-SAMPLE reference data model lists the data types used in the project. Four of them can be linked to HL7 FHIR simple type: string, numerical (decimal or integer), date (date or dateTime), and Boolean.

- **Simple or primitive types**: single elements with a primitive value, depicted in Figure 7.

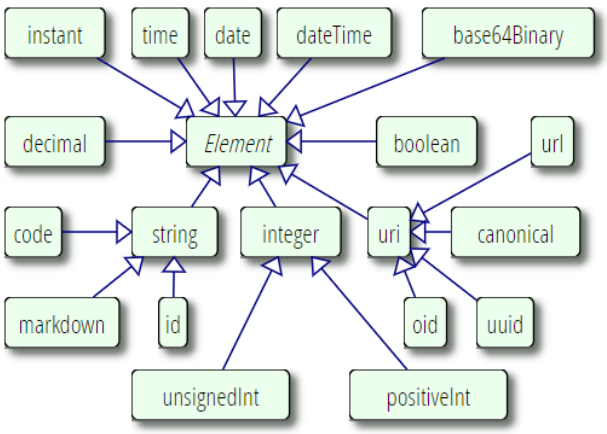

Figure 7: FHIR Primitive types.

Examples

A boolean true value:
 `<active value="true" />`
 A negative integer value:
 `<score value="-14" />`
 A high-precision decimal value:
 `<pi value="3.14159265358979323846264338327950288419716939937510" />`

A Unicode string:

```
<caption value="Noodles are called ?? in Chinese" />
```

A date of birth:

```
<date value="1951-06-04" />
```

The instant a document was created, expressed in UTC, with milliseconds:

```
<instant value="2013-06-08T09:57:34.2112Z" />
```

2:35pm in the afternoon:

```
<time value="14:35:00" />
```

- **Complex types for general purpose**: Figure 8, clusters of reusable elements. The fifth data type listed in Section 3 The RE-SAMPLE reference data model and used in RE-SAMPLE project is coded, which is linked to HL7 FHIR complex types CodeableConcept or Coding.

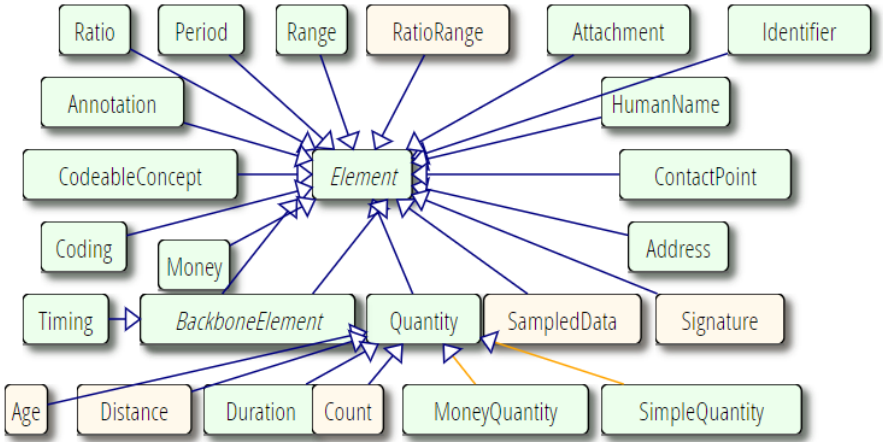

Figure 8: FHIR Complex types for general purpose.

Examples

A simple code for headache, in ICD-10:

```
<code>
  <system value="http://hl7.org/fhir/sid/icd-10" />
  <code value="G44.1" />
</code>
```

A SNOMED CT expression:

```
<problem>
  <system value="http://snomed.info/sct" />
  <code value="128045006:{363698007=56459004}" />
</problem>
```

A concentration where the value was out of range:

```
<result>
  <value value="40000" />
  <comparator value="&gt;" />
  <unit value="ug/L" />
  <system value="http://unitsofmeasure.org" />
  <code value="ug" />
</result>
```

- **Complex types for specific purpose**: cluster of elements for specific use inside the FHIR specification, depicted in Figure 9.

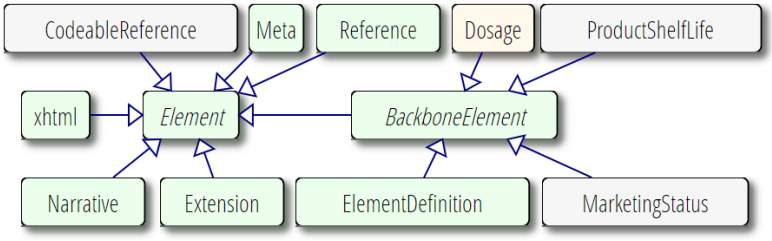

Figure 9: FHIR Complex types for specific purpose.

Examples

A relative reference to the Patient "034AB16" in an element named subject on a FHIR RESTful server:

```
<subject>
  <reference value="Patient/034AB16"/>
</subject>
```

- **Metadata**: set of types used to transmit metadata about resources, depicted in Figure 10.

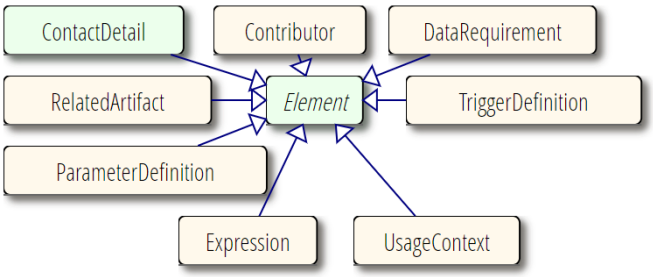

Figure 10: FHIR Metadata types.

Examples

Data requirement:

```
<dataRequirement>
  <type value="Procedure"/>
  <codeFilter>
  <path value="code"/>
  <valueSetString value="Total Colectomy Value Set"/>
  </codeFilter>
  <dateFilter>
  <path value="performedPeriod"/>
  <valuePeriod>
 <start value="2016-01-01"/>
 <end value="2016-12-31"/>
  </valuePeriod>
  </dateFilter>
</dataRequirement>
```

4.2.4 Bundles

A bundle is a container of resources with self-sufficient existence as a whole. Two resources belonging to the same bundle could depend on each other. The use of bundles aims to group resources for functional reasons (create/update all or none), or to improve performance by grouping several atomic resources in a single request. There are several use cases where bundles can be found:

- **Transaction**: bundle destined to a single operation on a server.
- **Searchset**: bundle retrieved in searches.
- **History**: bundle used in searches that retrieve results along time.
- **Messaging**: bundles for message exchange.
- **Document**: bundle as a set of resources with clinical integrity as a clinical document.

4.2.5 FHIR Implementation Guide for RE-SAMPLE

A FHIR IG is a set of rules of how a particular interoperability or standard problem is solved with the use of FHIR resources and potential extensions of them. Within the standard, an IG is considered a special type of resource, used to gather all the parts of an IG into a logical entity and to publish a computable definition of all the parts.

IGs can be referenced between each other, and thus can be used as dependencies or as a base for definition. The design and definition of the IG for the RE-SAMPLE project has the double objective of documenting and validating the use of the standard, although the final interaction on the resources of the standard is done through a transformation layer included in the HDH (*clinical-data-repository-api*). So, its implementation remains abstracted from the rest of the components. The definition process will be described in a detailed way in the deliverable *D4.4 Multi-modal data aggregation and curation* [M42].

4.3 Clinical terminologies

Medical concepts are characterized by their extremely specialization and complex. Given that there is a language barrier between the different clinical sites, there is a possibility of ambiguity in the languages and the existence of synonyms in the vocabulary, so that the exchange of information without using a standardized terminology becomes an impossible task. In addition, it must be considered that the stored information in the platform must be consumed by ML algorithms, so all the information must be discretized and encoded.

As FHIR has established alliances with organizations related to interoperability, it not only allows the use of terminologies offered by the standard itself or the creation of proprietary terminologies, but it is also compatible with external specifications.

The information coding process in RE-SAMPLE has been performed under the following premises:

- Given a variable, the information it collects is coded under the terminologies that the FHIR specification itself offers.
- If this is not possible, it is coded under the SNOMED CT ontology.
- Only if the two previous options do not offer a solution to the coding of the content of the variable, proprietary coding is used (RE-SAMPLE system).

SNOMED CT is an ontology available in multiple languages. The fact that it is an ontology instead of a terminology (relationships between concepts are permitted) allows to use synonyms that help reduce ambiguity. It includes a wide variety of clinical concepts, which allows minimizing the use and maintenance of different terminologies in which the concepts often overlap in the same EHR. It is distributed by the International Health Terminology Standards Development Organization (IHTSDO).

4.3.1 SNOMED CT Concept Model

The SNOMED CT Concept Model is a set of rules that specifies how the SNOMED CT concepts are defined, in terms of both formal logic and editorial rules. It defines the permitted set of attributes and values that may be applied to each kind of concept. The Concept Model provides value to SNOMED CT in several ways: it defines validation rules for the creation and maintenance of SNOMED CT content; and it also provides a foundation for processing the meaning of codes stored in clinical records, it facilitates effective data retrieval, and it enables clinical information to be used appropriately for decision support, analysis, aggregation, epidemiology and audit purposes.

SNOMED CT covers a wide range of clinical concepts. There are 19 main hierarchies (see Figure 11), which provide a way of organizing the concepts in SNOMED CT. Some of them are special hierarchies not intended to be recorded in a health record. For example the 'SNOMED CT Model Component' hierarchy, which is purely used to support the SNOMED CT release itself, provides the metadata that helps to document SNOMED CT concepts, relationships, descriptions, and reference sets.

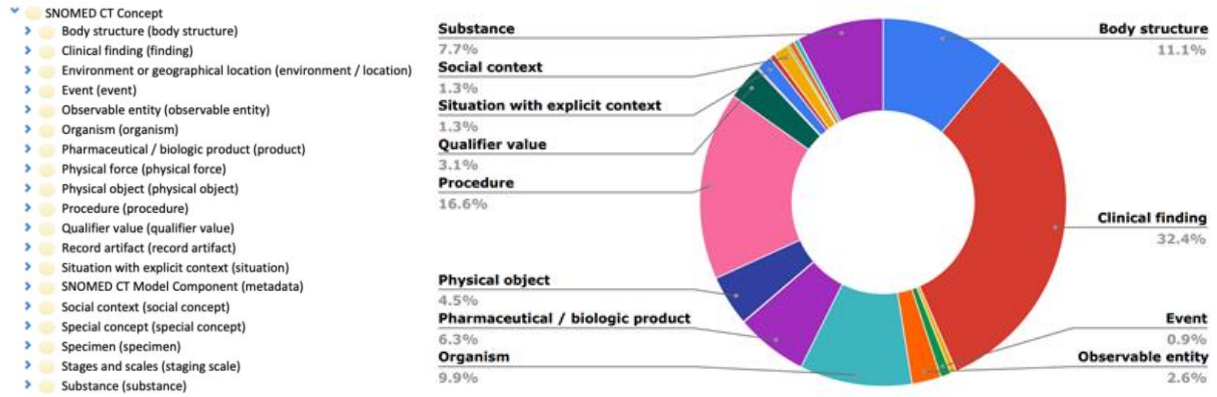

Figure 11: SNOMED CT main hierarchies and concept percentage.

A SNOMED CT concept (see Figure 12) has a unique, numerical, and machine-readable identifier and it represents a unique clinical meaning. Each concept is linked to several descriptions that represent its meaning. A description links a human-readable term to a concept. There are two types of descriptions:

- Fully Specified Name (FSN): a unique and unambiguous description of a concept.
- Synonyms: several words that point to the same FNS.

Each description also has a tag for acceptability, describing if the description is a preferred term in a language, country, region, value set, etc. or an acceptable one.

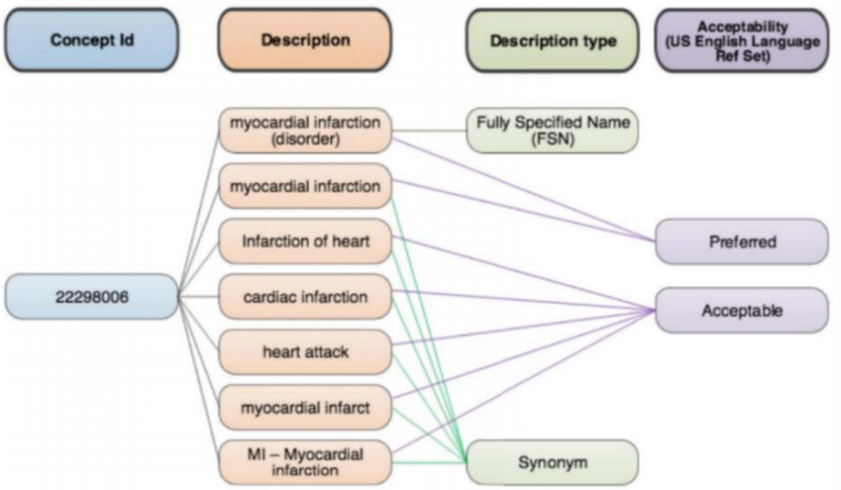

Figure 12: SNOMED CT types of descriptions.

One peculiarity of SNOMED CT is that concepts can be related to one another by relationships of hierarchy (see Figure 13) or other characteristics (see Figure 14). The concepts in SNOMED CT are placed in a poly-hierarchy (subtype definition), which means that each concept can have multiple parents. This means a concept can belong to several different groups or categories based on different aspects of its meaning.

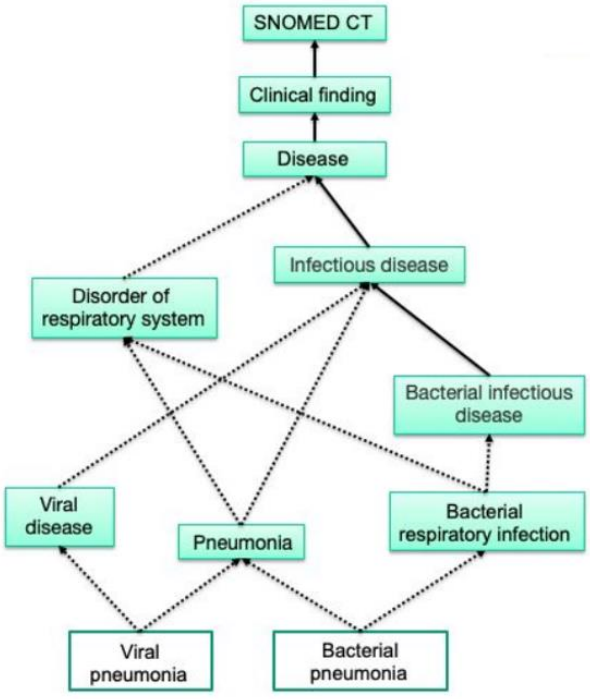

Figure 13: SNOMED CT poly-hierarchy.

Concepts are also described with additional defining characteristics. To represent these additional defining characteristics, the concepts are linked to other concepts via pre-defined semantic relationship types. These non-hierarchical relationships to other concepts are called attribute relationships, and the full definition of a SNOMED CT concept consists of both the defining subtype relationships and the defining attribute relationships. See an instance in Figure 14.

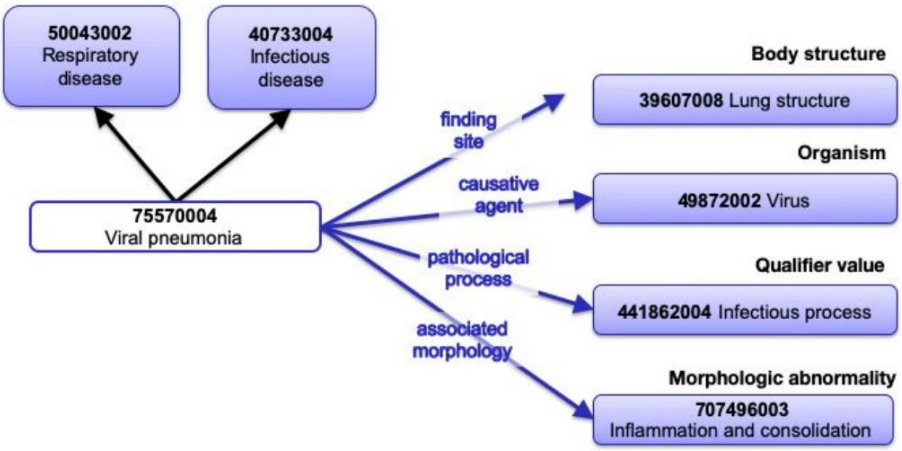

Figure 14: SNOMED CT attribute relationships.

4.3.2 Post-coordination

Post-coordination allows to represent new clinical meanings not included as pre-coordinated SNOMED CT concepts, without creating a new concept in the ontology. The idea is to group existing concepts taking advantage of the hierarchical and non-hierarchical (attribute) relationships between concepts. If this approach is not possible, another option is to use the more general concept, excluding specific information. This could alternatively be added by other means, for example by creating a textual annotation.

As an example, consider 'Pneumonia caused by SARS-CoV-2', imagining that there is not a pre-coordinated concept in SNOMED CT. The appropriate supertype in this case is the concept "Viral pneumonia". It can be linked together with the appropriate virus, represented by a particular organism. These two concepts can be linked together by a semantic link, in this case the attribute "causative agent", to form the specific type of viral pneumonia (see Figure 15).

Figure 15: SNOMED CT post-coordination example.

For the specific case of SARS-CoV-2, the post-coordinated expression will look as follows:

```
75570004 |Viral pneumonia (disorder)|: 246075003 |Causative agent (attribute)| = 840533007 |Severe
acute respiratory syndrome coronavirus 2 (organism)|
```

Post-coordination mechanism is used when codifying RE-SAMPLE variables in the data model is not possible with a pre-coordinated concept, so an alternative post-coordinated expression is sought.

4.3.3 License and Membership

Anyone who uses SNOMED CT needs to be licensed and their use of the terminology is subject to the license conditions, applicable to both organizations and individuals. The types of licenses are:

- Licenses that apply to the SNOMED CT International Release:
  - Affiliate Licenses issued by SNOMED International: an agreement that covers worldwide use of SNOMED CT International Release. It is required by all those developing, maintaining and/or distributing any type of system, application or service that incorporates or uses SNOMED CT.
  - Sublicenses issued by Affiliate License holders: sublicense.
  - issued by an Affiliate Licensee to their customers, clients or end-users. A sublicense allows users of the Affiliate's products or services to use SNOMED CT. This means that an Affiliate can include SNOMED CT as part of the system or service that they provide to their customers.

- SNOMED International Members have rights to use SNOMED CT: countries that are Members of SNOMED International also have rights to use SNOMED CT. These rights are granted as part of the Articles of Association that formally define the governance structures of SNOMED International.

- Supplementary Licenses that apply to the SNOMED CT National Extensions:
  - National licenses issued by Members to Affiliates: SNOMED International permits Members to register Affiliate Licenses on its behalf. This means that if someone who is not already a SNOMED CT Affiliate applied for a national license the member receiving that application can require the applicant to accept the SNOMED CT Affiliate License Agreement as part of the process of issuing them with a national license.

Some of the conditions of use of SNOMED CT differ according to whether that use occurs in a member country or in a non-member country. Figure 16 shows current member countries.

It is important to note that the location where SNOMED CT is used determines these conditions, not the country in which the system supplier is based nor the country from which a service is provided.

- License conditions in member Countries: SNOMED International does not charge for use of SNOMED CT. Often a key reason for a country becoming a member is to encourage the use of SNOMED CT as a terminology standard and to remove any perceived barriers to this.
- License conditions in non-member Countries: The affiliate is required to notify SNOMED International before use of SNOMED CT in a non-member country (used by the affiliate or its sublicensees); submit an annual report of usage; and pay fees for usage depending on the non-member country.

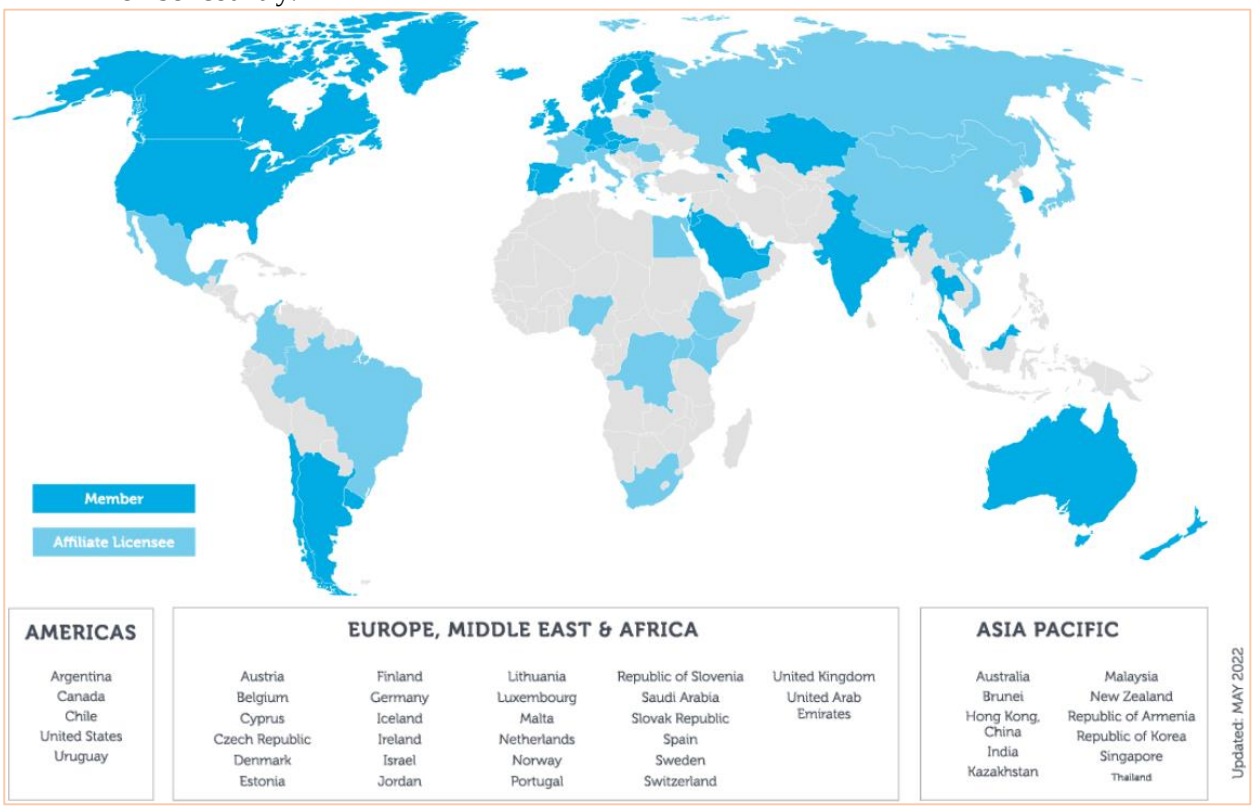

Figure 16: SNOMED International member countries.

There are some situations in which SNOMED International may agree to waive fees for use in non-member countries:

- Low-income countries categorized by the World Bank: no fees are payable.
- Qualifying research projects: discretionary and subject to SNOMED International approval.
- Humanitarian and charitable use: discretionary and subject to SNOMED International approval.

The organization was established as the International Health Terminology Standards Development Organisation in 2007. The founding 9 charter Members were Australia, Canada, Denmark, Lithuania, Sweden, the Netherlands, New Zealand, the United Kingdom and the United States. Trading as SNOMED International, the organization has grown to 42 Members and has issued Affiliate Licenses to more than 5,000 individuals and organizations.

Belgium, Estonia, Netherlands, Germany and Spain are member countries and have free use of SNOMED CT. Whereas Greece and Italy could be able to use it for free by implying its use is under a European research project scope.

4.4 Standardized RE-SAMPLE HL7 FHIR resources

This section describes the mapping performed from the different clinical data sources (subsets of the model) into the standardized set of HL7 FHIR resources (onwards, simply FHIR) and relations between them in order to obtain a formalized reference data model for the RE-SAMPLE project.

Figure 17 presents the main FHIR resources used for modelling the storage needs of the data collected in RE-SAMPLE. The use of the generic resources provides a clear and standard manner to provide a reusable and documented interoperability layout, since third systems could be ready to directly consume them assuming semantic content.

The specific needs for RE-SAMPLE not supported by the FHIR version R4 resources are provided by *extending* the resources with the needed attributes and relationships. The motivation and definition of these extensions will be documented in the FHIR IG in development for the project within the scope of deliverable *D4.9 Open clinical decision aid* [M48].

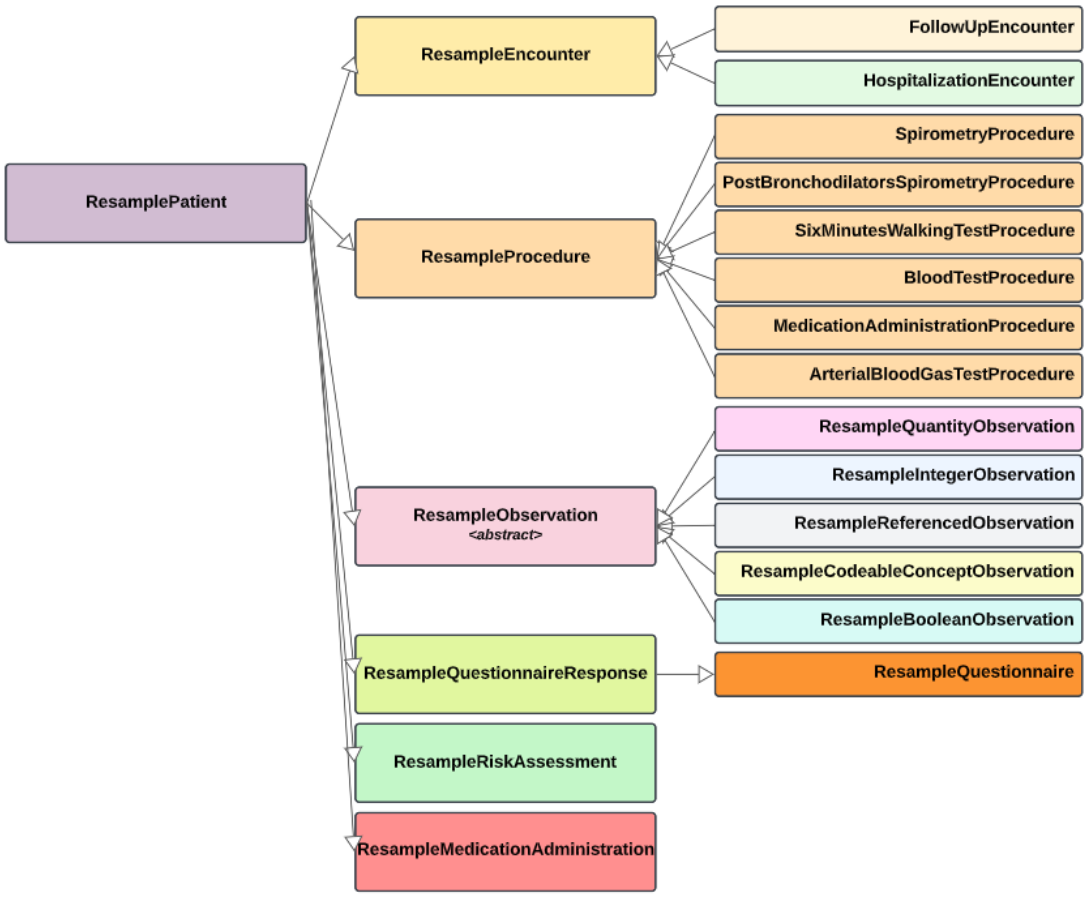

Figure 17: RE-SAMPLE FHIR resources summary.

The resources shown in Figure 17 provide a generic level of abstraction, with the common needed (and sometimes mandatory) data, later particularized in the specific suffixed resource, e.g., *ResampleObservation* > *ResampleIntegerObservation* > *Steps*. Meaning *Steps* is an *Observation* resource with the value type *Integer*.

The following subsections depict the FHIR resources used in RE-SAMPLE with a deeper detail of explanation regarding attributes and cardinality modelled to the specific scope of the project. The use of the blue colour and italic type text in the cardinalities denote variation from the HL7 FHIR standard, specifically established for validation within the scope of the project. These variations are defined by the RE-SAMPLE FHIR IG, although its definition is not described in depth in this document, as indicated in the section 4.2.5 FHIR Implementation Guide for RE-SAMPLE. It will be detailed in following deliverables.

4.4.1 Patient

The patient is the standardized resource in charge of storing the demographic data of the subject of study and serves as a reference for all other resources associated with the specific person.

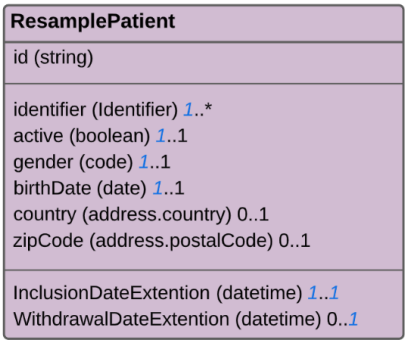

Figure 18: RE-SAMPLE FHIR resource Patient.

Conceptually, its fields should not have great variation over time. In case of the RE-SAMPLE project, as a result of the enrolment process and location of basic data, this resource (Table 7) is created first via synchronization with a third service (Healthentia) as first reference (where certain terms of use should also be accepted). Later, it is completed with data coming from the HIS, with the fields exclusively owned by it.

Table 7: Resource ResamplePatient elements detail.

Element	Source	Observations
id	-	Internal FHIR id autogenerated uuid, which is considered the "RE-SAMPLE id"
identifier	Healthentia	Secondary identifier containing Healthentia id[1]
active	Healthentia	Flag that indicates if the user is active in the study. It would be used in the process of deleting associated resources in case of withdrawal, depending on who maintains the consents
gender	Healthentia	
birthDate	Healthentia	YYYY-MM-DDThh:mm:ss+zz:zz from endpoint YYYY-MM-DDThh:mm:ss.zzzZ
country	HIS	Used for the collection of environmental data
zipCode	HIS	Used for the collection of environmental data
InclusionDateExtension	Healthentia	Date of inclusion in the clinical study
WithdrawalDateExtension	Healthentia	It would be used in the process of deleting associated resources in case of withdrawal

4.4.2 Encounter

The patient Encounter is the standardized resource in charge of storing data related to interactions between a patient and a healthcare provider with the purpose of delivering healthcare services. An Encounter encompasses the lifecycle from pre-admission, the actual encounter, stay and discharge (if applicable). During the encounter, the patient may move from practitioner to practitioner and location to location. It always references the Patient resource and can reference other type of resources or be referenced by them.

[1] This is pending on future new strategy of id mapping depending on security aspects under discussion.

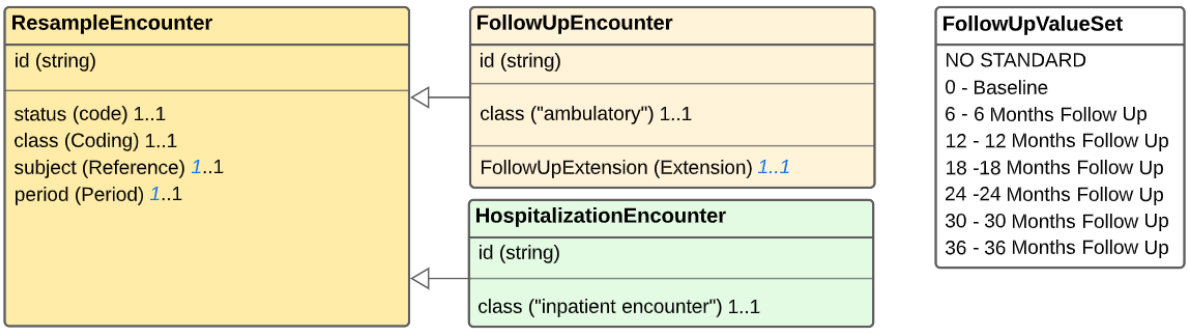

Figure 19: RE-SAMPLE FHIR resource Encounter.

Within the scope of the RE-SAMPLE project, the encounter resource (Figure 19, Table 8) is used satisfying the purpose of storing grouped data related to a patient interacting with the hospital for: a scheduled follow-up (baseline or 6-month follow-up); or unexpected hospitalizations. The follow-up encompasses different medical tests, measurements and clinical-related data of the patient which are recorded every 6 months. In the same way, during a hospitalization, standard procedures are performed on the patient. Hence, other resources reference the RE-SAMPLE *Encounter* resource, types vary from *Procedure*, *Observation* and *MedicationAdministration* resources.

Table 8: Resource ResampleEncounter elements detail.

Element	Source	Observations
id	-	Internal id autogenerated uuid
status	-	Fixed value to "finished"
class	-	Fixed value to "ambulatory" or "inpatient encounter"
subject	Healthentia HIS	Patient subject of the encounter
period	Healthentia HIS	The start/end time of the encounter
FollowUpExtension	Healthentia HIS	Extension to store the followUp encounter type (baseline or sequential)

4.4.3 Procedure

The procedure is the standardized resource in charge of storing the details of current and historical procedures performed on or for a patient as part of the provision of care. It always references the *Patient* resource and can reference other type of resources or be referenced by them.

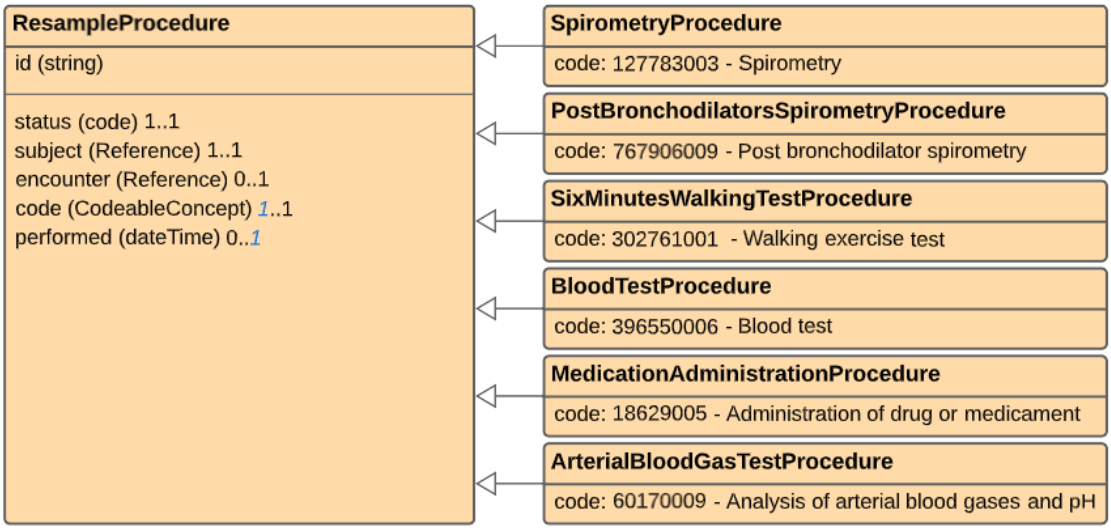

Figure 20: RE-SAMPLE FHIR resource Procedure.

In the RE-SAMPLE project, *Procedure* resources (Figure 20, Table 9) are used under the scope of patient's follow-ups (*Encounter* resource) for storing data related to procedures applied to the patient in the hospital (spirometry, spirometry after having taken bronchodilators, six-minute walking test, blood test and medications of the patient), or under hospitalization (*Encounter* resource) for storing data related to a hospitalized patient (arterial blood gas test). Some of the fields are common, but there is a distinction in the code that describes the procedure, so the *Procedures* are conceptually based on a primitive one and extended by the code. *Procedure* resources in turn are referenced by other resources (RE-SAMPLE *Observation* and *MedicationAdministration* resources).

Table 9: Resource ResampleProcedure elements detail.

Element	Source	Observations
id	-	Internal id autogenerated uuid
status	-	Always considered "completed"
subject	Healthentia HIS	Patient subject of the procedure
encounter	-	Reference to FollowUp Encounter
code	-	SNOMED CT code for the specific procedure: Spirometry, PostBronchodilatorSpirometry, SixMinuteWalkingTest, BloodTest, ArterialBloodGasTest or MedicationAdministration
performed	Healthentia HIS	To store the datetime when the Procedure was created

4.4.4 Observation

The observation is the standardized resource in charge of storing measurements and simple assertions made about a patient. Observations are a central element in healthcare, used to support diagnosis, monitor progress, determine baselines and patterns. Some examples are vital signs, laboratory data, clinical findings, device measurements or social history. It references the Patient resource and can reference other type of resources or be referenced by them.

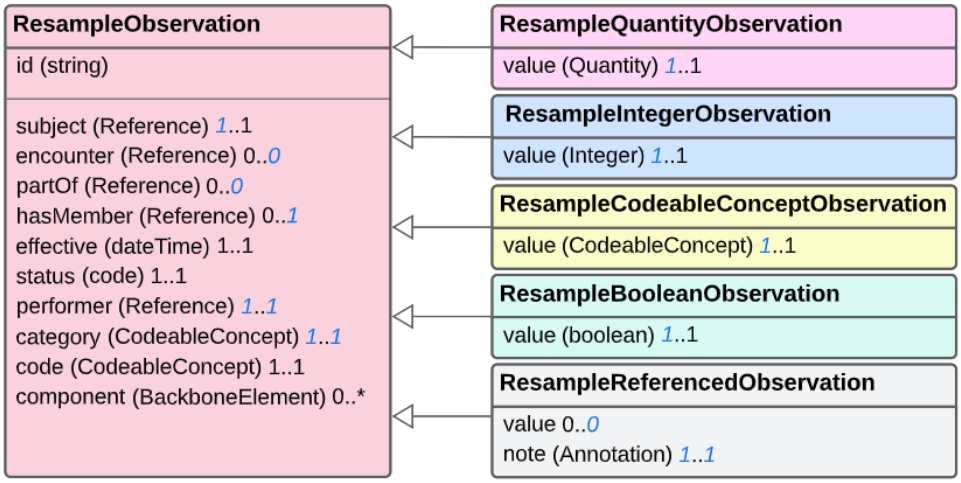

Figure 21: RE-SAMPLE FHIR resource Observation.

In the RE-SAMPLE project, most of the collected data is stored using *Observation* resources (Figure 21, Table 10) taking advantage of their great scope, versatility and ease of utility. Since most of the fields are common for all RE-SAMPLE *Observation* resources, like linking *CodeableConcepts* in the Resample *CodeableConceptObservation*, the observations are conceptually based on a primitive one and then extended based on the distinctive attributes. It mainly varies in the value field, which for RE-SAMPLE data can be of type *Quantity*, *Integer*, *CodeableConcept* and *Boolean*. Another distinction is the *ResampleReferencedObservation*, which is an observation referred by other observations with the purpose of grouping them, so it does not include a value, but it contains a note field instead.

Table 10: Resource ResampleObservation elements detail.

Element	Source	Observations
id	-	Internal id autogenerated uuid
subject	Healthentia HIS	Patient subject of the observation
encounter	-	The Encounter to which the Observation refers, in case of FollowUp or Hospitalization.
partOf	-	The Procedure to which the Observation refers
hasMember	-	Other Observation to which the Observation refers
effective	Healthentia HIS	Date when the observation took place
status	-	Fixed value to 'final'
category	-	FHIR Observation category
code	-	Value from a list of SNOMED CT ObservationCode that describes the Observation
value	Healthentia HIS	Value of the observation. It can be of the types: Quantity, Integer, CodeableConcept or Boolean
component	Healthentia	Subcomponents included in the Observation. For RE-SAMPLE Physiological Observations they are `Trend', `Long-term average' and `Short-term average'
note	Healthentia	In the case of ResampleReferenceObservation, note is used for storing the type of Exercise Observation such as walking, running, etc.

4.4.5 QuestionnaireResponse and Questionnaire

The *QuestionnaireResponse* resources are used to maintain specific answers of the patients to a particular questionnaire. The list of questions in a questionnaire is maintained in the referenced resource of type *Questionnaire*.

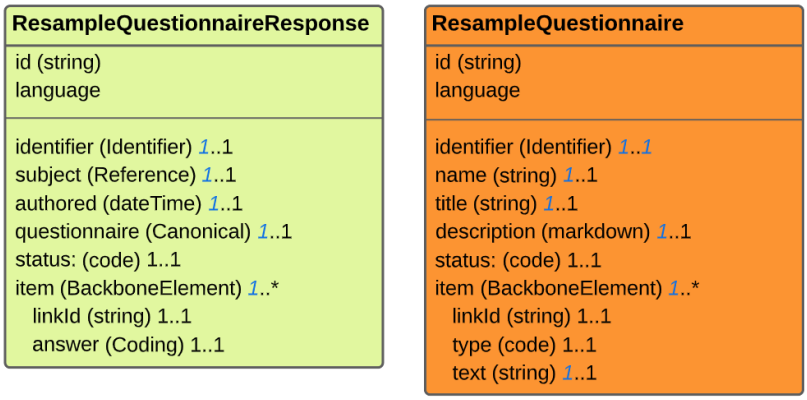

Figure 22: RE-SAMPLE FHIR resources QuestionnaireResponse and Questionnaire.

Within the RE-SAMPLE project, it is expected that all questionnaires will be designed and defined in the Healthentia platform. The idea in the project is to allow the maximum possible amount of patient data to be used by the ML algorithms (within the conditions of the GDPR). In this manner, all the answers of the questionnaires are also ingested within the HDH as they are structured, maintaining a synchronized copy, from which later scores or answers to specific questions can be obtained according to the analysis' needs. No other type of FHIR resources (like extra *Observations*) will be generated from the questionnaire's answers, just the *QuestionnaireResponse*,(Figure 22, Table 11).

Table 11: Resource ResampleQuestionnaireResponse elements detail.

Element	Source	Observations
id	-	Internal id autogenerated uuid
language	-	To distinguish the language use (English as the default)
identifier	Healthentia	Secondary identifier containing the Healthentia patientQuestionnaireId for synch.
subject	Healthentia	Patient subject of the QuestionnaireResponse
authored	Healthentia	DateTime the Responses to the questionnaire was recorded
questionnaire	Healthentia	Reference to the resource Questionnaire that maintains the original question's structure. There is no expected need of any kind of versioning, it is just to keep the questionnaires defined.

status	-	QuestionnaireResponseStatus that synchronizer always fix as “completed”
item.linkId	Healthentia	The specific question in the referenced questionnaire Unique id for item in questionnaire
item.answer	Healthentia	The answer for such question

Table 12: Resource ResampleQuestionnaire elements detail.

Element	Source	Observations
id	-	Internal id autogenerated uuid
language	-	To distinguish the language use (English as the default)
identifier	Healthentia	Secondary identifier containing
name	Healthentia	Codename from Healthentia Questionnaire
title	Healthentia	Title from Healthentia Questionnaire in English
description	-	"Questionnaire from Healthentia questionnaire of name '<name>' for the RE-SAMPLE project."
status	-	PublicationStatus: draft, active, retired, unknown
item.linkId	Healthentia	A unique id for the specific question in the referenced questionnaire
item.type	Healthentia	Type of question, from Numerical/Text to QuestionnaireItemType.INTEGER/STRING;
item.text	Healthentia	The question text in English

4.4.6 Risk Assessment for ML results

A *RiskAssessment* is a specialized type of resource, similar to the *Observation* resource, but specifically to be used as a simpler mechanism to capture risks and associate those risks with time-ranges, probabilities, etc.

The resource is associated primarily with resources of type *Patient*, *Group*, *FamilyMemberHistory*, *Procedures* and a series of *Observations*.

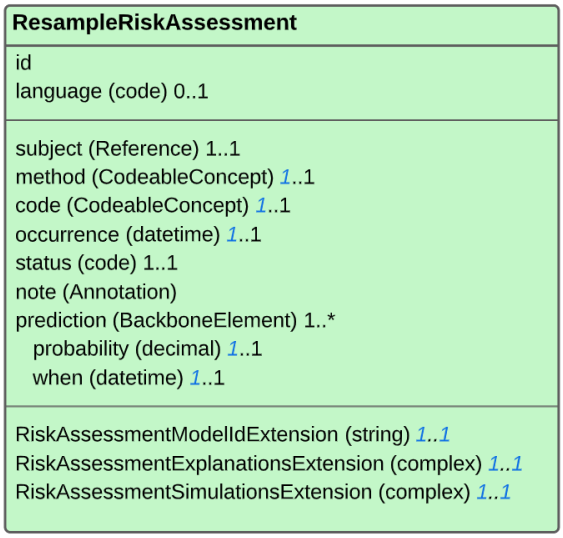

Figure 23: RE-SAMPLE FHIR resource RiskAssessment.

Within the RE-SAMPLE project, the *RiskAssessments* resource (Figure 23, Table 13) is used to store the patient data alongside the *predictions* generated by the ML modules.

Since it is also required to store *explanations* and *simulations* related to the specific prediction, to centralize all this information, extensions of the *RiskAssessment* resource have been developed to store them. In the same way that the input parsing is carried out in the ingestion process, the same data will be extracted conversely without greater logical control than the cardinality established for the explanations and simulations.

Table 13: Resource ResampleRiskAssessment elements detail.

Element	Source	Observations
id	-	Internal id autogenerated uuid
language	-	English by default. If other languages are needed consumers will translate.
subject	ML Module	Patient subject of the prediction
method	ML Module	Categorization of the prediction name: Moderate/SevereExacerbation, QoLScore, ...
code	ML Module	Categorization of the prediction type: Classification or Regression.
occurence	ML Module	Date when prediction was generated
status	-	RiskAssessmentStatus.REGISTERED
note	-	Fixed "RiskAssessment created from RE-SAMPLE ML results"
prediction.probability	ML Module	Outcome value of the Prediction value of probability
prediction.when	ML Module	Timeframe of the prediction
extension.RiskAssessmentModelIdExtension	ML Module	Reference to the specific model used to calculate the specific prediction
extension.RiskAssessmentExplanationsExtension	ML Module	Explanations (SHAP, EBM and counterfactual) related to the Prediction
extension.RiskAssessmentSimulationsExtension	ML Module	Simulations related to the Prediction

4.4.7 MedicationAdministration

The *MedicationAdministration* is the standardized resource that stores the details of the event of a patient consuming or being administered a medication, including self-administrations of oral medications, injections, intra-venous adjustments, allergy shots, device-administered insulin, etc. It always references the *Patient* resource and can reference other type of resources or be referenced by them.

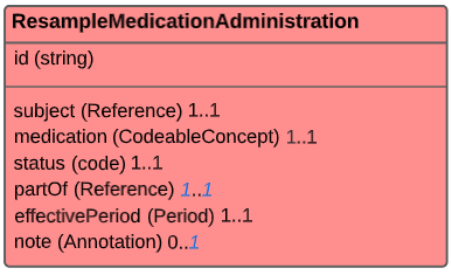

Figure 24: RE-SAMPLE FHIR resource MedicationAdministration.

In the RE-SAMPLE project, *MedicationAdministration* resources (Figure 24, Table 14) are used under the scope of the patient's medication *Procedure* included in the follow-ups (*Encounter* resource) or independently, for storing data related to medication prescribed/administered to or by the patient. Some of the fields are common, but there is a distinction in the medication code that describes the specific medication, so the *MedicationAdministrations* are conceptually based on a primitive one and extended by that code.

Table 14: Resource ResampleMedicationAdministration elements detail.

Element	Source	Observations
id	-	Internal id autogenerated uuid
status	-	Always considered "completed"
subject	Healthentia HIS	Patient subject of the procedure
partOf	-	Reference to MedicationAdministration Procedure
medication	Healthentia HIS	RE-SAMPLE code for the specific medication
effectivePeriod	Healthentia HIS	Period of dates when the medication took place (start and end)
note	-	Fixed label to "Prescribed on date x"

4.4.8 Activity

The patient activity is measured by gathering data from Healthentia's wearable system. The kind of data that the patient can monitor is divided into two potential data sources: physiological activity and exercise activity, both included in the RE-SAMPLE data model.

4.4.8.1 Physiological activity

The physiological activity data (Figure 25) includes information collected passively, that is, the patient wears the wearable device preferably during the whole day and data is measured and stored automatically. This generates a daily general picture of the different activities the patient performs during their daily activities. The information can be classified in several groups such as patient's weight, sleep data, heart-related data and activity-related data.

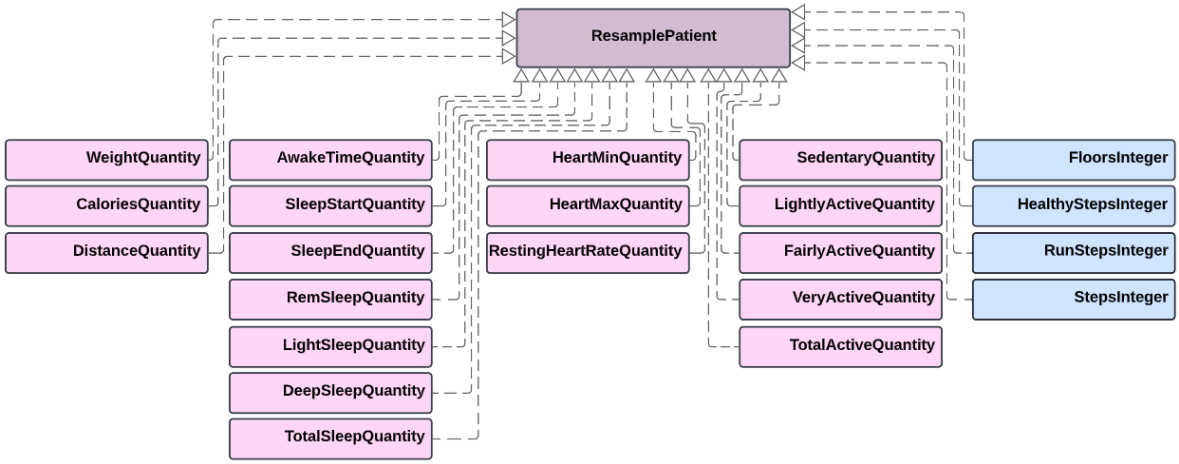

Figure 25: RE-SAMPLE FHIR resource Observation for Physiological Activity.

FHIR resources mapped to store physiological activity data are *Patient* (to which all the other resources refer to) and *Observation*. Observations in turn can be of type *Quantity* (if the stored value includes a unit of measurement) or *Integer* (if just an auto described figure is stored). Additionally, all physiological activity observations include three subcomponents to store *Trend*, *Long-term average* and *Short-term average*.

4.4.8.2 Exercise activity

The exercise activity data (Figure 26) includes information collected actively by the patient, that is, the patient initiates the recording of data via the wearable when they start a physical exercise. Data is stored and provides information regarding the daily active exercise a patient does. The collected information includes the type of exercise (e.g., walking), patient's heart-related data and activity-related data.

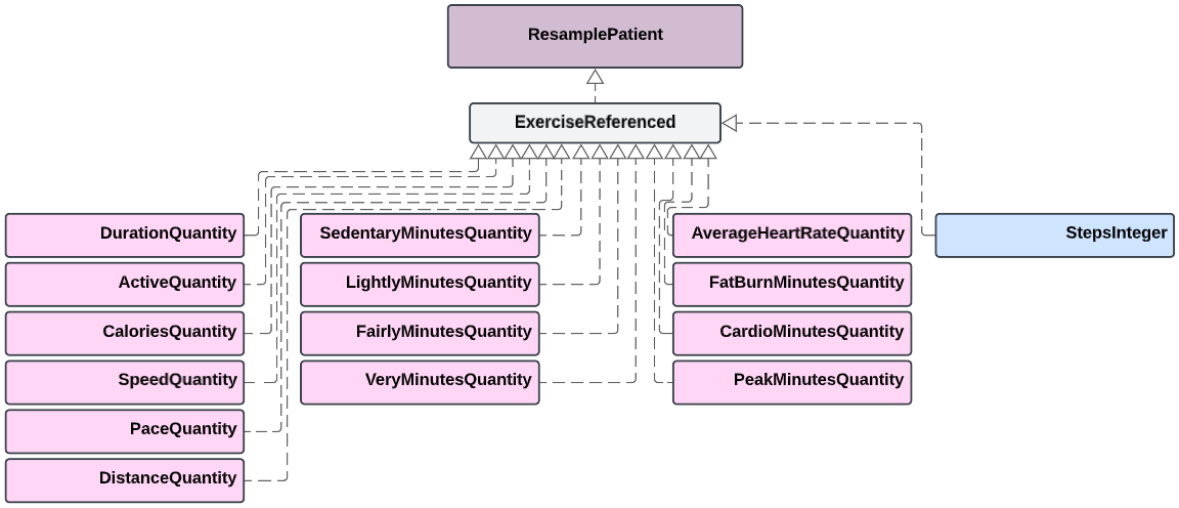

Figure 26: RE-SAMPLE FHIR resource Observation for Exercise Activity.

FHIR resources mapped to store exercise activity data are *Patient* (to which ReferencedObservation refers to), an *Observation* of the type *Referenced* (including the type of exercise and to which the rest of the observations refer to) and the rest of the observations. Observations in turn can be of type *Quantity* (if the stored value includes a unit of measurement) or *Integer* (if just an auto described figure is stored).

4.4.8.3 Follow-up Encounter

This part of the RE-SAMPLE reference data model comes from the dataset defined in section 3.2 Hospital Information System dataset, with the information coming from the EHR of the pilot sites. Follow-ups of the patient are scheduled to be performed every six months from baseline according to the clinicians procedure. In terms of effective ingestion from hospital systems, part of the data listed in this concept can also be generated under other circumstances apart from the scheduled follow-ups, such as in case of exacerbation (emergency visit) or hospitalizations.

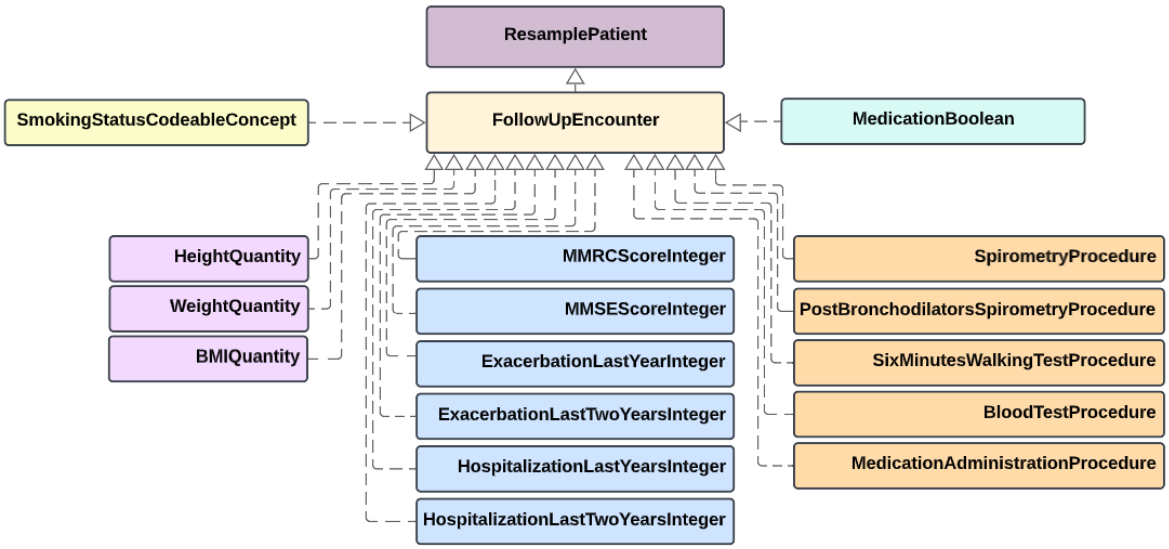

Figure 27: RE-SAMPLE FHIR resource FollowUp Encounter.

All data expected in this part of the follow-up of the patient are grouped in the concept of *FollowUpEncounter* (Figure 27) (which refers in turn to the *Patient* resource), with some atomic *Observations* and associated nested Procedures that will be extended in the following sections. The atomic observations can be of the type *CodeableConcept* (for values storing codes from a code system), *Boolean* (true/false values), *Quantity* (if the stored value includes a unit of measurement) or *Integer* (if an auto described figure is stored).

4.4.8.4 Spirometry Procedure

The data related to the performance of a spirometry within a patient's follow-up is stored with a grouping *Procedure* (Figure 28) to which four *Observations* refer to. The type of these observations is *Quantity*, since they store a value and a unit of measurement. *SpirometryProcedure* usually refers to the encounter for grouping the follow-up.

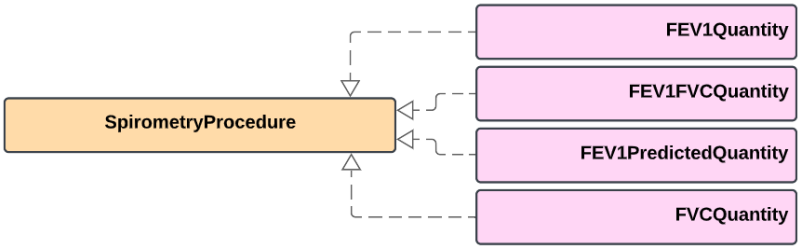

Figure 28: RE-SAMPLE FHIR resource Procedure for Spirometry.

4.4.8.5 Post Bronchodilators Spirometry Procedure

The data related to the performance of a spirometry after medicating the patient with bronchodilators during a patient's follow-up is stored with a grouping *Procedure* (Figure 29) to which four *Observations* refer to.

The type of these observations is *Quantity*, since they store a value and a unit of measurement. *PostBronchodilatorsSpirometry* procedure refers to the encounter for grouping the follow-up.

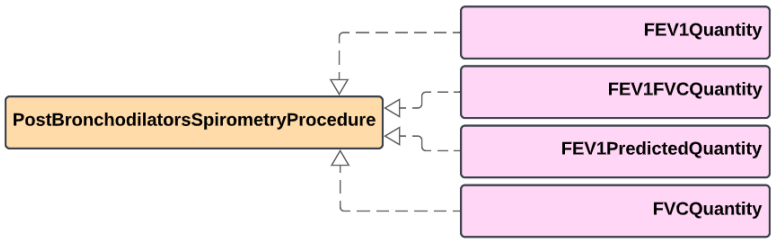

Figure 29: RE-SAMPLE FHIR resource Procedure for PostBronchodilatorsSpirometry.

4.4.8.6 Six Minute Walking Test Procedure

The data related to the performance of a six-minute walking test during a patient's follow-up is stored with a grouping *Procedure* (Figure 30) to which 25 *Quantity* observations (the stored value includes a unit of measurement) and *Boolean* observations (true/false value) refer to. The *SixMinutesWalkingTest* procedure usually refers to the *Encounter* grouping the follow-up.

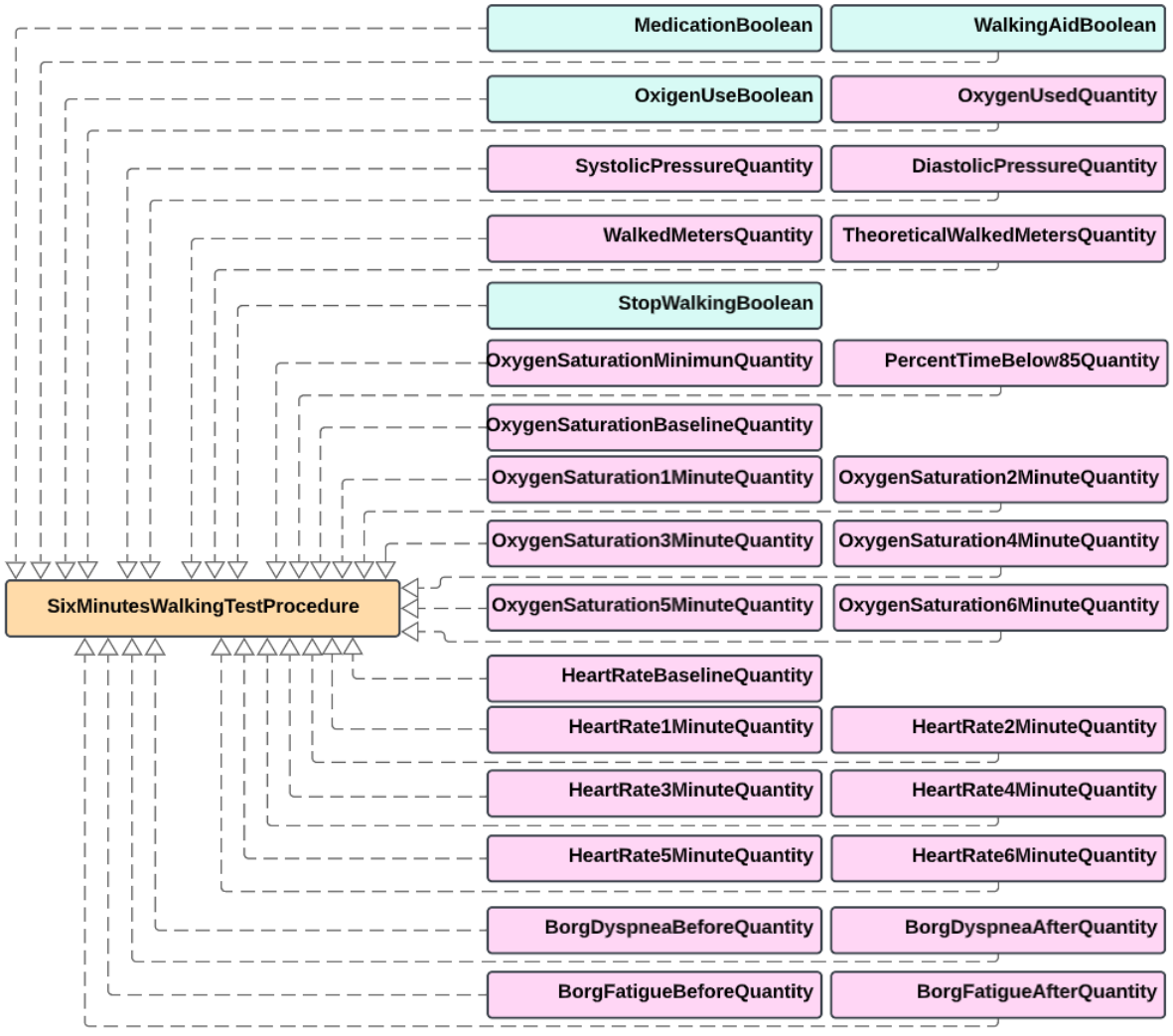

Figure 30: RE-SAMPLE FHIR resource SixMinutesWalkingTestProcedure.

4.4.8.7 Blood Test Procedure

The data related to the performance of a blood test during a patient's follow-up is stored with a grouping Procedure (Figure 31) to which 13 Observations refer to. The type of these observation is Quantity, since they store a value and a unit of measurement. The BloodTest procedure usually refers to the Encounter for grouping the follow-up.

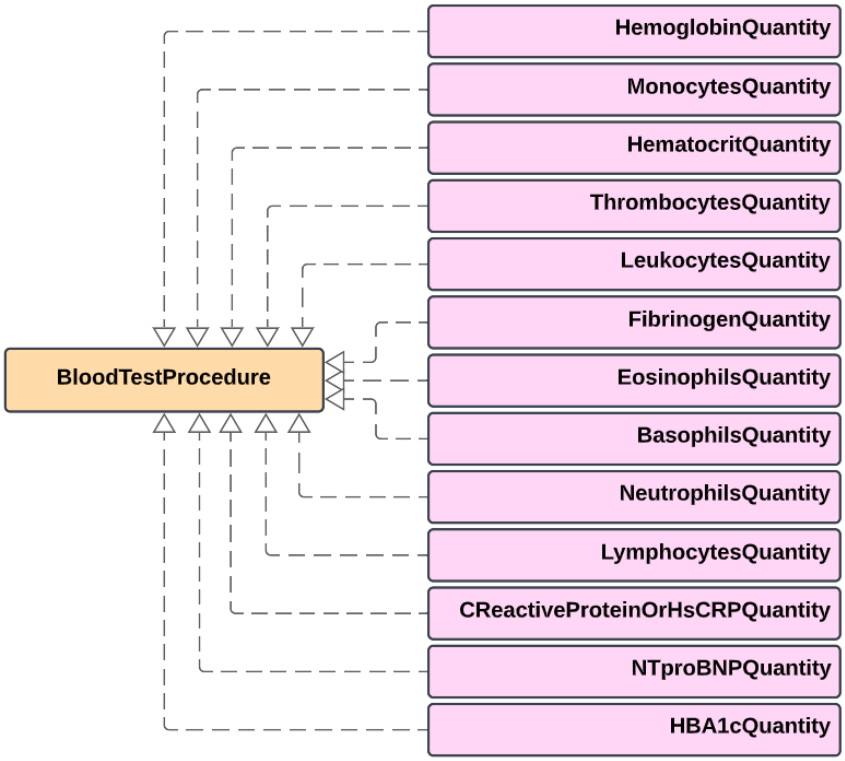

Figure 31: RE-SAMPLE FHIR resource BoodTestProcedure.

4.4.8.8 Medication Administration Procedure

The data related to the list of medications MedicationAdministrations a patient is taking is grouped in a Procedure. MedicationAdministrationProcedure (Figure 32) could optionally refer to the follow-up Encounter, although the maintenance and status of the prescriptions can be carried out independently of it.

Figure 32: RE-SAMPLE FHIR resource MedicationAdministrationProcedure.

4.4.8.9 Hospitalization Encounter

This part of the RE-SAMPLE reference data model comes from the dataset defined in section 3.2 Hospital Information System dataset, with the information coming from the EHR of the pilot sites. Hospitalizations of the patient may occur in between the six month follow-ups, so that information must also be stored and included into the RE-SAMPLE repository to be analysed. When a hospitalization occurs, there are a series of standard procedures, tests and measurements that are performed to the patient, which Figure 33 depicts.

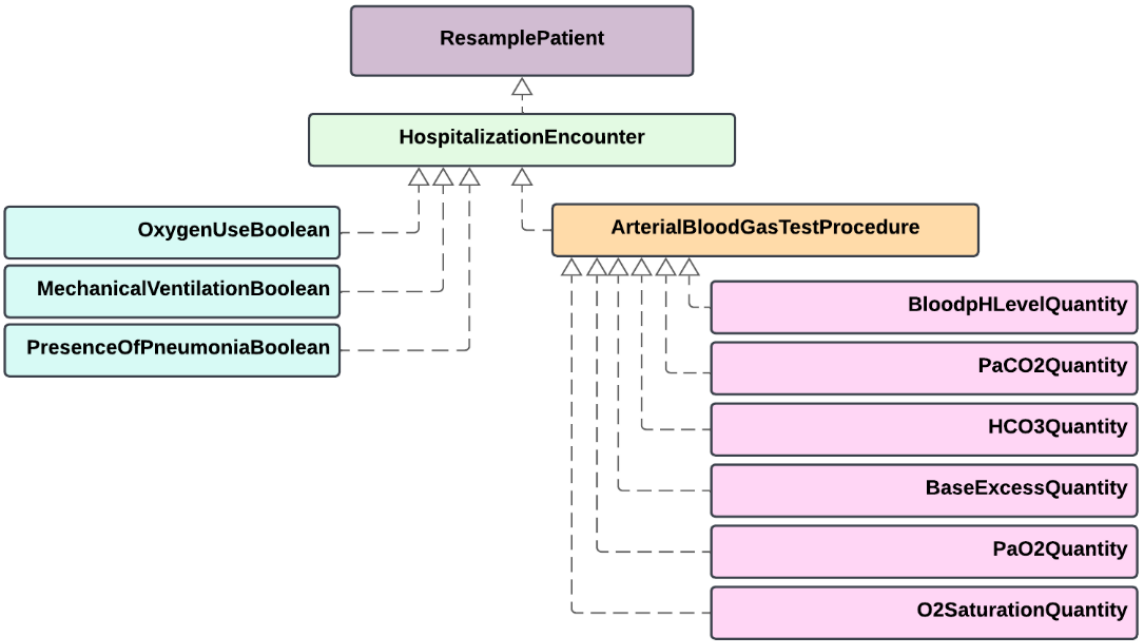

Figure 33: RE-SAMPLE FHIR resource HospitalizationEncounter.

Hospitalization data are grouped in the concept of *HospitalizationEncounter* (which refers in turn to the *Patient* resource), with some atomic observations and an associated nested *Procedure*. The atomic observations are of the type *Boolean* (true/false values). The *ArterialBloodGasTest* procedure includes observations of type *Quantity* (the stored value includes a unit of measurement).

4.4.8.10 Arterial Blood Gas Test Procedure

The data related to the performance of an arterial blood gas test within a patient hospitalization is stored with a grouping Procedure (Figure 34) to which six Observations refer to. The type of these observation is Quantity, since they store decimal values with a unit of measurement. ArterialBloodGasTest procedure usually refers to the Encounter for grouping the hospitalization.

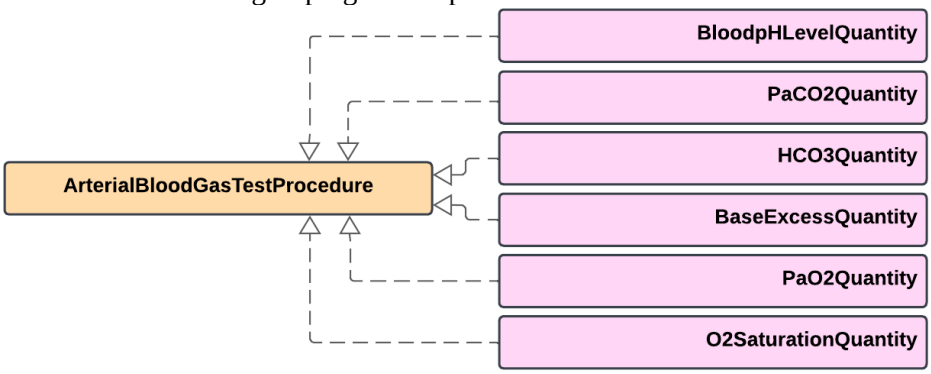

Figure 34: RE-SAMPLE FHIR resource ArterialBloodGasTestProcedure.

After the data model is defined and the data standardized, an OpenAPI is provided for the information providers to insert the data. This API preserves the structure and resource associations described in section 4.4, validating data entry and ensuring only the appropriate data is inserted into the CDR.

5. Conclusions

This deliverable has presented the work performed for homogenising and assembling the information coming from diverse data sources to a unique common place within the RE-SAMPLE project. The data sources are: HIS from three different pilots; the Healthentia platform; and the ML modules. This process included complex discussions with all the involved actors. Some of these discussions are still active at the time of this deliverable submission. Hence, some variations and fine adjustments, albeit minimal, could be expected. In such case, it will be reported in the next deliverable *D4.4 Multi-modal data aggregation and curation (M42)*.

Furthermore, the process of mapping from the different data sources to the FHIR standard will be explained along with the formal definition of the FHIR IG in *D4.9 Open clinical decision aid (M48)*, with the double objective of documenting the data maintained and validating the ingestion process.

References

CRQ. (2022). Retrieved 08 03, 2022, from Chronic Respiratory Disease Questionnaire: https://qol.thoracic.org/sections/instruments/ae/pages/crq.html

EuroQoL. (2022). *EQ-5D*. Retrieved from https://euroqol.org/

Gunther Schadow, C. J. (2017). *UCUM - Unified Code for Units of Measure*. Retrieved from https://ucum.org/trac

HADS. (2022). Retrieved from https://www.svri.org/sites/default/files/attachments/2016-01-13/HADS.pdf

Health Level 7. (2022). *Datatypes - FHIR v4.6.0*. Retrieved from https://build.fhir.org/datatypes.html

Health Level 7. (2022). *FHIR v4.0.1*. Retrieved March 7, 2022, from https://www.hl7.org/fhir/

ISO. (2022, 01 08). *ISO-3166 COUNTRY CODES*. Retrieved from iso.org: https://www.iso.org/iso-3166-country-codes.html

iSPRINT. (API 2022). *Healthentia API - Swagger UI*. Retrieved from https://demo-api.healthentia.com/swagger/index.html?urls.primaryName=v2

iSPRINT. (28th February 2021). *D5.2: RWD collection application (accompanying report).*

Molnar, C. (2020). *Interpretable Machine Learning.* Retrieved from https://christophm.github.io/interpretable-ml-book/

SNOMED International. (2022). *SNOMED CT - 5 Step Briefing*. Retrieved March 7, 2022, from https://www.snomed.org/snomed-ct/five-step-briefing