HORIZON 2020
The EU Framework Programme for Research and Innovation

CAPABLE

Cancer Patients Better Life Experience

Grant Agreement No. 875052
Start Date: 01/01/2020 (48 Months)

Deliverable No. 3.1
Information Architecture

Due Date: 31/03/2021
Submitted On: 19/07/2021

Coordinator	University of Pavia (UNIPV)
Deliverable Lead Partner	Academic Medical Center (AMC)
Contributing Partners	UNIPV, UoH, BIOM, IBM, PUT, ICSM, NKI, DEON
Contact	Prof. Silvana Quaglini
Email	silvana.quaglini@unipv.it
Website	www.capable-project.eu

Deliverable Type		
R	Document, report	**[X]**
DEM	Demonstrator, pilot, prototype	
DEC	Websites, patent fillings, videos etc.	
OTHER		
Dissemination Level		
PU	Public	**[X]**
CO	Confidential (Consortium members including the Commission Services)	
CI	Classified Information (Commission Decision 2015/444/EC)	

This deliverable is a part of a project receiving funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875052.

Table of Contents

List of Figures

List of Tables

1.Versions History

Version	Date	Author	Comments
0.5	2021-06-08	Ronald Cornet Francesca Polce Rowdy de Groot	Merging separate versions
0.7	2021-07-08	Ronald Cornet Francesca Polce Rowdy de Groot	Version for internal review
0.9	2021-07-15	Ronald Cornet Francesca Polce Rowdy de Groot	Version for final review by coordinator

2. Executive Summary

The CAPABLE project has a strong commitment to making the data collected during the project follow the FAIR Principles, addressing the Findability, Accessibility, Interoperability and Re-usability of the data.

This deliverable addresses the relevant aspects of such FAIR data, including what data the project will generate, what metadata adequately describes the data, and which standards are used so that data can be made accessible for verification and re-use.

To maximally support the Interoperability Principle, data management and metadata management are based as much as possible on standard data models and terminologies. This deliverable focuses on the information architecture. Relevant data models, terminologies, and metadata models are introduced, which will form the basis for the representation of data and metadata that are exposed to researchers external to CAPABLE. As such it complements deliverables that focus on the internal representation of data.

For data models, both **HL7 FHIR**, focusing on data exchange, and **OMOP CDM**, focusing on data storage, are supported. This enables close integration with both data exchange in a clinical context, in which HL7 FHIR plays an increasingly important role, and in the context of observational research based on real-world-data, where the OHDSI (Observational Health Data Sciences and Informatics) infrastructure, which includes OMOP, is broadly adopted. A collaboration among those two announced early 2021 will contribute to a close alignment between these two approaches.

For metadata models, Dublin Core, Data Catalog (DCAT), DataCite, and Generic Dataset Metadata Template (GDMT) will be adhered to, which establishes adherence to domain-relevant community standards for describing datasets and their contents.

Based on these data and metadata models, the actual data and metadata elements in the CAPABLE FAIR infrastructure are described, after which they are aligned with the selected models and related terminologies. Metadata will be provided via a FAIR datapoint. This will contribute to establishing the CAPABLE FAIR infrastructure.

3. Introduction

The FAIR Guiding Principles [1], which specify criteria for Findability, Accessibility, Interoperability, and Reusability of data, as shown in Table 1, have gained increasing attention of the research community. To both contribute to and benefit from an expanding FAIR ecosystem, the CAPABLE project too is committed to delivering a data management and software infrastructure that follows the FAIR Principles, in order to increase the benefit of the data, information, and knowledge captured throughout the project.

The FAIR Principles specify criteria for data as well as metadata. Metadata, as elaborated on in Section 5, is "data about data". It concerns all kinds of descriptions of data, or more broadly, of digital objects, such as authorship, access conditions, or usage statistics.

Table 1. FAIR Guiding Principles for Scientific Data Management and Stewardship

FINDABILITY	
F	F1. (Meta)data are assigned a globally unique and persistent identifier F2. Data are described with rich metadata (defined by R1 below) F3. Metadata clearly and explicitly include the identifier of the data they describe F4. (Meta)data are registered or indexed in a searchable resource
ACCESSIBILITY	
A	A1. (Meta)data are retrievable by their identifier using a standardized communications protocol A1.1 The protocol is open, free, and universally implementable A1.2 The protocol allows for an authentication and authorization procedure, where necessary A2. Metadata are accessible, even when the data are no longer available
INTEROPERABILITY	
I	I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (Meta)data use vocabularies that follow FAIR principles I3. (Meta)data include qualified references to other (meta)data
REUSABILITY	
R	R1. (Meta)data are richly described with a plurality of accurate and relevant attributes R1.1. (Meta)data are released with a clear and accessible data usage license R1.2. (Meta)data are associated with detailed provenance R1.3. (Meta)data meet domain-relevant community standards

As the FAIR principles are highly generic, they need to be made concrete for the CAPABLE infrastructure. In this Deliverable, we do so by making explicit what data and metadata standards to use, and what data and metadata will be represented using these standards.

In Chapter 4, the various components of the CAPABLE infrastructure are summarized, to describe the various data sources and the data they provide.

Chapter 5 provides an overview of the various types of metadata that constitute "rich" metadata to be provided in the FAIR infrastructure, including metadata of data sets, and metadata about data.

Chapter 6 addresses the domain-relevant data models and vocabularies to be used, which are key to the FAIR Interoperability and Reusability Principles.

Chapter 7 complements Chapter 6 by specifying relevant models and vocabularies for meta data, inspired by the inventory in Chapter 5.

Based on these, Chapter 8 briefly describes the architecture that will provide the data and metadata, as well as the current status thereof.

Finally, Chapter 9 describes an assessment of the adherence to the FAIR Principles established by this infrastructure.

4.Overview of components and data involved

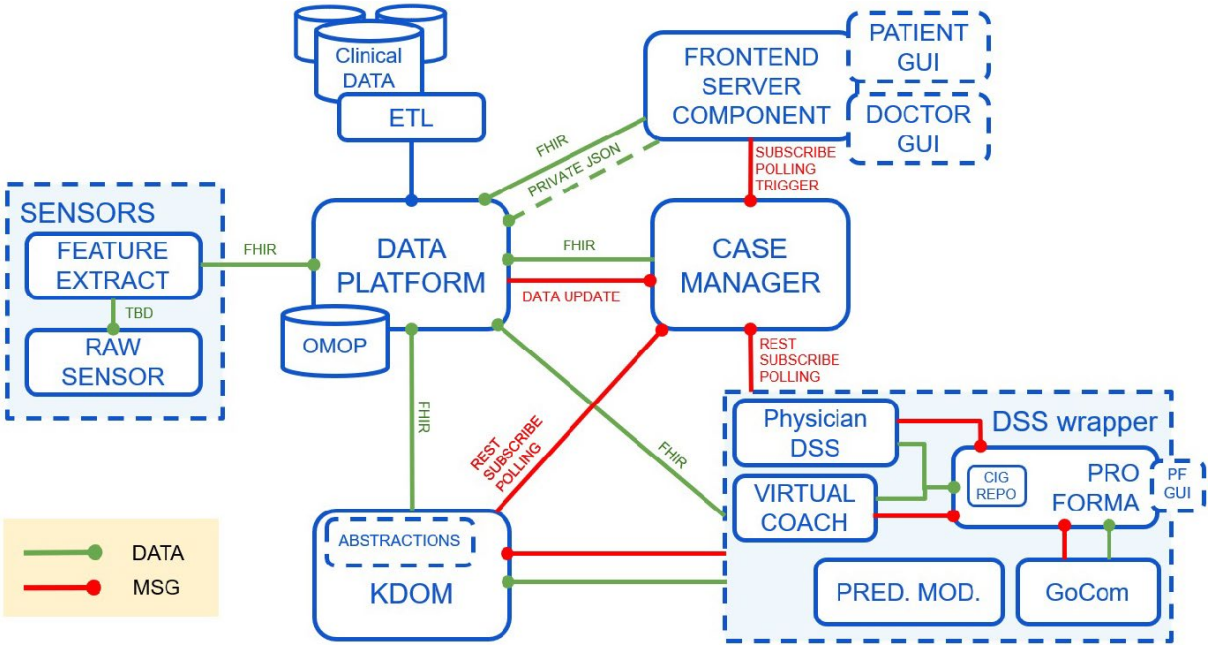


Figure 1. The CAPABLE architecture with the components involved

CAPABLE is conceived as a modular distributed system consisting of several independent and collaborative components, each of which encapsulates specific knowledge for the management of cancer patients represented through appropriate formalisms and exploiting Artificial Intelligence (AI) techniques. Figure 1 illustrates the main modules that make up the CAPABLE architecture. The system allows to remotely acquire data from patients or their home caregivers, both actively via Graphical User Interfaces (GUI) on smartphones or personal computers (through self-reporting of symptoms or the use of questionnaires) and passively (via wearable sensors). All the data collected, defined as PROs (Patient Reported Outcomes) and PREs (Patient Reported Experiences) are integrated with those available in the Electronic Health Record (EHR) made available by the hospital where the patient is being treated. All the healthcare personnel who is in charge for the patient may access and integrate those data through a web-based interface. All this clinical information is stored in a shared centralized database called Data Platform which makes it available to the entire system.

The AI components continuously and automatically process patient data, incrementing them with new information in the form of forecasts, warnings, reports of any dangerous situations or partial interpretations that are also stored into the Data Platform. These components are the following:

- Virtual Coach, with the aim of providing coaching, recommendations and reminders to the patient to motivate him/her in adhering to therapies.
- Multimorbidity Controller (a.k.a. "GoCom"), in charge of detecting and resolving conflicts between multiple drugs taken simultaneously or between diseases and prescribed drugs that may occur for patients with comorbidities. Indeed, these interactions can sometimes cause adverse events if they are not properly managed.
- Physician Decision Support System (DSS), which provides recommendations and decision support based on clinical practice guidelines to clinicians in patient management. In this sense, it provides complementary functions to those given to the patient by the Virtual Coach.

- Deontics Engine, which provides a Computer-Interpretable Guidelines (CIGs) execution service for use by the CAPABLE components that require an executable clinical knowledge representation (i.e. the Virtual Coach, the Physician DSS and GoCom). The CIGs are modeled using the PROforma language.
- Knowledge-Data Ontology Mapper (KDOM), responsible for performing clinical abstractions, that are structured concepts regarding the patient's clinical status. The KDOM carries out an ontological mapping between the abstract medical concepts coming from the CIGs (used by the other AI components for the reasoning process) and the patient raw data, bridging this semantic gap.
- Predictive Models that address the evolution of cancer by providing ad hoc predictions based on the current state of the patient and his clinical history.

All these components are decoupled and can interact with each other only through the Data Platform, seen as a blackboard where any relevant information regarding a patient is published by a component and subsequently read by any other interested. This uniform protocol prevents the onset of an unmanageable network of links among the components. To allow components to become aware of new data stored in the Data Platform without the need to continually scan it, CAPABLE provides an additional component separate from the rest, called Case Manager, designed specifically to coordinate the components and activate the right ones only when needed. The Case Manager guides the reasoning process through an event activation strategy, where events are combinations of facts that represent the main updates regarding the patient's clinical status. Each component must instruct the Case Manager on their own events of interest. The Case Manager monitors the Data Platform to detect the occurrence of such events, so that the right component can be promptly notified, reactivating the reasoning process that was previously "paused" pending the arrival of new information. Therefore, the Case Manager is the only component authorized to send messages directly to anyone else.

A more detailed description of each CAPABLE component is given in [2].

4.1. Data Platform and Patient Data Sources

The system will rely on both data already available to partners at the beginning of the project and on data that will be collected during the clinical study, which will last the entire fourth year of the project. The clinical study, that will take place at the two clinical partner organizations Istituti Clinici Scientifici Maugeri (ICSM), based in Pavia, and Netherlands Cancer Institute (NKI), based in Amsterdam, will enroll kidney cancer patients, from ICSM, and melanoma patients, from NKI. Thus, the data collected by the project pilot will be focused on these two cancer patient populations, but many of the findings intend to be generalizable to other cancer domains.

All data relating to patients managed within CAPABLE are stored persistently within the Data Platform. To guarantee a state-of-the-art level component, we selected as standards the Observational Medical Outcomes Partnership – Common Data Model (OMOP-CDM) for health care data storage and Health Level Seven – Fast Healthcare Interoperability Resources (HL7 FHIR) for health care data exchange. This exchange includes both reading data from the Data Platform by data consumers, as well as writing data into the Data Platform, by data producers. The use of FHIR resources using the Representational State Transfer (REST) Application Programming Interface (API) is the only way to access the data. For example, when the Data Platform receives a request to write a new FHIR resource from a certain source, the API is exploited to convert data in the resource into elements corresponding to OMOP CDM on the fly, so as to store it into the OMOP database. In response, the Data Platform returns the input resource with the addition of the unique identifier generated in CAPABLE and the creation timestamp, and subsequently notifies the Case Manager of the addition of this new resource. This sequence of actions is shown in Figure 2.

In order to make the database traceable and verifiable and to have a reliable record of the entire data flow over time, we decided to downgrade the CRUD (Create, Read, Update and Delete) model to the CR (Create and Read) model. Therefore, the data transaction is limited to write and read operations only, choosing not to allow deletions or updates.

Patient data collected during the pilot study and stored within the Data Platform can be divided into two groups:

- raw data reported directly by patients at home, by doctors at the hospital through their GUIs, acquired by sensors or made available by extraction from the EHR;
- data generated by the system following the interpretation of the raw data available or derived as a result of the application of knowledge of any AI component invoked.

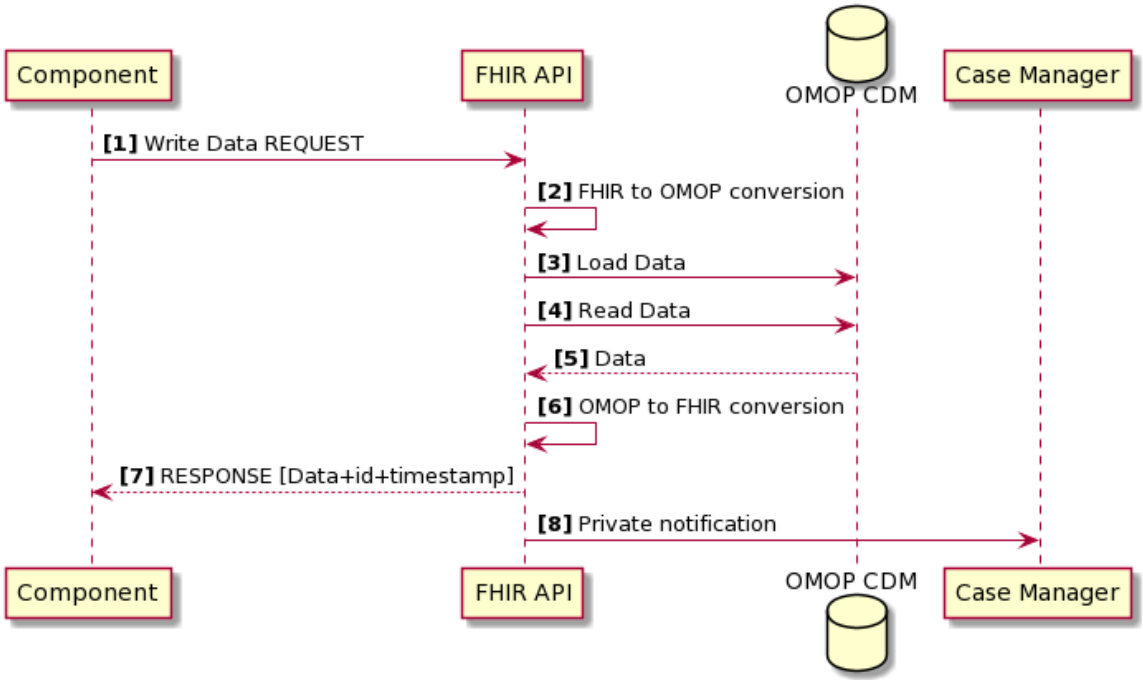


Figure 2. Sequence diagram describing the actions that occur when a generic component writes new data into the Data Platform

All these data generated and collected by CAPABLE (i.e. prospective data) and the corresponding sources from which they come will be reviewed below.

4.1.1. Raw Data

The raw input data that will be collected within the CAPABLE system during the clinical study are the following:

- patient characteristics and lifestyle (e.g., age, sex, BMI, WHO score, alcohol consumption, smoking);
- background data (e.g., presence of a home caregiver, hobbies, sports activities, awake/meals/bed usual time schedule);
- tumor characteristics (e.g., TNM staging, BRAF mutation, possible metastases and their sites);
- risk factors (e.g., diabetes, hypertension, autoimmune diseases or other comorbidities that can act as risk factors for developing treatment toxicity);
- clinical history (cancer-related surgeries with date, other surgeries, hospitalizations and other relevant events);

- cancer treatment during the pilot study (type of drugs for each treatment line, start and, end date, reason for stopping treatment, response to treatment including possible progression);
- signs and symptoms (e.g., diarrhea, nausea, vomiting, abdominal pain, dizziness, cramps, fever, bloody stools, itching, rash) together with their CTCAE (Common Terminology Criteria for Adverse Events) grade, start and end date, and actions taken to manage their impact on the current cancer treatment;
- other treatments, ongoing treatments and drugs in use (also not cancer-related);
- exams and visits;
- laboratory tests (glucose level, absolute neutrophil count, hemoglobin, serum corrected calcium, LDH level, platelet lymphocytes, serum sodium, creatinine, s100b, TSH, etc.);
- physiological data and lifestyle monitoring data (weight, temperature, oxygen saturation, blood pressure, heart rate, sleep monitoring, physical activity indicators);
- quality-of-life, nutritional, psychological or fatigue assessment questionnaires (EORTC QLQ-C30 for quality of life, Insomnia Sleep Index, MST for nutritional assessment, PHQ-9 for depression, GAD7 for anxiety, emotional thermometers, NRS2002 nutritional risk screening, caregiver burden inventory, HADS - hospital anxiety and depression scale, BFI – Brief Fatigue Inventory);
- environmental data (air quality in the patient's geographical area).

Table 2 provides a summary of the raw health and care data that CAPABLE will handle and, for each category of data, the source from which it can be collected.

Table 2. Raw data sources

Data type description	EHR	Patient App	Doctor GUI	Sensors
Patient characteristics	x		x	
Background data		x	x	
Symptoms	x	x	x	
Tumor characteristics	x		x	
Risk factors	x		x	
Clinical history	x		x	
Cancer treatment	x		x	
Other treatments	x		x	
Exams and visits	x		x	
Laboratory tests	x		x	
Physiological and lifestyle data		x	x	x
Questionnaires		x	x	
Environmental data				x

4.1.1.1. EHR

Part of the clinical data relating to the enrolled patient cohorts and generated outside the CAPABLE system is available in the EHR of the two clinical partners where the pilot trials are planned (ICSM and NKI). The relevant EHR data are regularly imported into CAPABLE through ETL (Extract, Transform and Load) procedures and stored into the Data Platform. The scheduling of this task is to be defined. These data include previous medical history,

laboratory test results, scheduled visits, based on the applications available in each hospital. Table 3, based on Deliverable 5.2 [3], shows a summary of the clinical data items to be acquired and the current feasibility assessments of extracting such data from the NKI and ICSM EHR. In those cases that some of the data are not available in a structured format (indicated as Uncertain in Table 3 because they are probably missing from the EHR), when the user is interacting with CAPABLE, the clinical staff will be asked to enter the data via the CAPABLE user interface (i.e. the Doctor GUI).

- The prospective data source available for NKI is an existing reporting/research extract that can be used. NKI EHR is currently updated every night but there are plans to update it with near-live data (in the order of minutes) in the future. It will provide data about demographic information, medication prescribed in NKI, procedures performed in NKI, blood tests (if performed in NKI), scheduled visits in NKI and possible hospitalizations.
- ICSM, in addition to the EHR of the hospital, provides a dedicated pathological register managed by REDCap (Research Electronic Data Capture), a software for electronic data collection. Their EHR will provide data on demographic information, blood tests (if done in ICSM), scheduled visits and possible hospitalizations. Scheduled visits for patients currently enrolled in pharmaceutical clinical trials are not managed by their EHR, so information is not available for this subset. Data related to the next scheduled appointments will be entered by the clinicians in the CAPABLE interface. The pathology registry is handled manually by the ICSM clinical staff and contains detailed information about the treatment lines and outcomes. The pathology register does not contain information about the scheduled administration of drugs, these data will be entered by the clinicians in the CAPABLE interface.

Table 3. Patient data items that can be extracted from the NKI and ICSM EHR

		NKI	ICSM
Data type description		**From EHR / Uncertain**	
Patient profile			
Patient demographics	name, birthdate, height, education, etc.	From EHR	From EHR
Caregiver information	Name, contact information	From EHR	Uncertain
Treatments and tests			
Medication	name, code, start date, end date, frequency, dose, route, indication	From EHR	From EHR
Home medications	name, code, start date, end date, frequency, dose, route, indication	Uncertain	Not available
Procedures (in hospital)	name, code, start date, end date	From EHR	From EHR (only if hospitalized patient)
Past procedures	name, code, start date, end date	Uncertain	From EHR (only if hospitalized patient)
Confirmed diagnoses list / problem list	name, code, date confirmed	Uncertain	From EHR (only if hospitalized patient)
Laboratory tests (blood/fluid tests)	name, code, date ordered/performed, +/-result	From EHR	From EHR
Other laboratory tests/exams	name, code, date ordered/performed, +/-result	Uncertain	Uncertain
Referrals within hospital	name, code, date	From EHR	From EHR
Questionnaires			
Questionnaires administered within the hospital and by CAPABLE	EORTC QLQ-INFO25, EORTC QLQ-Q30, EQ-5D-5L, FACT-G, FACT-M, FACT-R, GAD7, MST, PHQ9	Uncertain	From EHR

4.1.1.2. Patient App and Doctor GUI

Most of the data related to symptoms, PROs and PREs are entered directly by patients through self-reporting on their mobile app (the so-called Patient App). Patients will also fill in a set of questionnaires that will be used for outcome evaluation. More details on the questionnaires exploited in CAPABLE are available in [4]. The system also aims to monitor how the patient responds to the interventions proposed by the CAPABLE system or the clinician, if he has agreed to perform them (a patient may be asked to select one of the possible candidates or to approve or reject something) and if he has actually done them. In fact, the Patient App has a dedicated section where the patient can report if he has performed a certain activity suggested (e.g., a sports activity to improve physical well-being) and satisfaction, so as to check patient compliance and receive feedback to improve future suggestions. Similarly, the recommendations suggested to the clinician by the system can be approved by him via the web browser (i.e. the Doctor GUI). All these decisions made by users (both patients and doctors) are always saved into the Data Platform, for further compliance analysis.

Further patient data, including diagnoses and prescriptions of therapies, are collected during regular visits with the clinician, who can access each patient profile from the Doctor GUI and check treatment reports for its analysis and possible modification or addition. All data can be entered via the Doctor GUI, so that the clinician can enter all missing data items of the EHR.

4.1.1.3. Sensors

- Patient vital signs and activity information are collected automatically using wearable sensors such as smartwatches. Currently, the best candidate device seems to be the ASUS VivoWatch BP smartwatch (HC-A05 model) [5], equipped with multiple sensors able to detect and monitor various personal health measurements during the patient's daily life in a repeated and continuous way. Among these, the relevant information for CAPABLE are blood pressure, heart rate, sleep quality (i.e. duration of sleep) and the number of steps taken by the user. Table 4 shows, for each vital sign collected, the unit of measurement and the corresponding built-in sensor of the smartwatch responsible for its collection.
- The Photo-plethysmography (PPG) sensor detects changes in blood volume using an optical light source (usually red or green diode) and a photodetector. This provides information on the volumetric changes of blood circulation, from which the heart rate or the blood oxygen saturation can be extracted.
- The electrocardiogram (ECG) sensor records electrical signals in a patient's heart. It is a non-invasive way of detecting common heart problems. The ECG waveform is made up of five peaks and valleys, labeled P, Q, R, S, and T. Heart rate can be easily extracted by measuring the R-peaks of the ECG waveform, which are commonly the highest peaks.
- The G-sensor (accelerometer) is responsible for measuring the acceleration (or vibration) of the wearable device, mainly used to detect the amount of steps taken by users.

Table 4. Data collected by the ASUS Smartwatch and sensors used to detect them

Vital Sign	Unit	Sensor
Daily Steps	Steps	G sensor (Accelerometer)
Systolic Blood Pressure	Millimeter mercury column (mm[Hg])	PPG (Photo-plethysmography) sensor + electrocardiogram (ECG) sensor
Diastolic Blood Pressure	Millimeter mercury column (mm[Hg])	PPG sensor + ECG sensor
Heart Rate	Beats per minute (bpm)	PPG sensor + ECG sensor
Light sleep duration	Hours (h)	PPG + G sensor
Deep sleep duration	Hours (h)	PPG + G sensor

ASUS smartwatches synchronize their measurements with a third-party cloud platform (ASUS Omnicare API) [6], from which the CAPABLE sensor module takes the data of interest regularly and stores them (raw data or data abstractions) into the Data Platform.

4.1.2. System-Generated Data

- All raw data described above are processed by the AI components of CAPABLE during the normal functioning of the project. Each component is responsible for recognizing the conditions under which it becomes able to take part in the support process, providing new information and direct contributions to the interpretation of the patient's case. This processing step requires the AI components to apply their knowledge and augment the acquired patient data with the data they produce, which represents partial interpretations of the case, warnings, recommendations or predictions. These system-generated data are then stored into the Data Platform in order to be accessed by the patient and the physician via the Patient App and the Doctor GUI. Furthermore, these data may also provide input to other AI components. Below is the list of data generated by the system along with the AI components responsible for producing them.
- Recommendations and clinical interventions acquired from CIGs both for the patient (sent by the Virtual Coach) and for the clinician (generated by the Physician DSS), including decision support on which treatments/drugs are most suitable for the patient based on the current diagnosis, advice on the management of symptoms and side effects caused by ongoing therapies and actions suggested to improve the patient health.
- Well-being interventions represented by the Virtual Capsules, which contain instructions for evidence-based non-pharmacological interventions that aim to improve the mental and social well-being of patients and promote physical activity (e.g., Tai Chi, walking in nature, sports, breathing exercises). The Virtual Coach

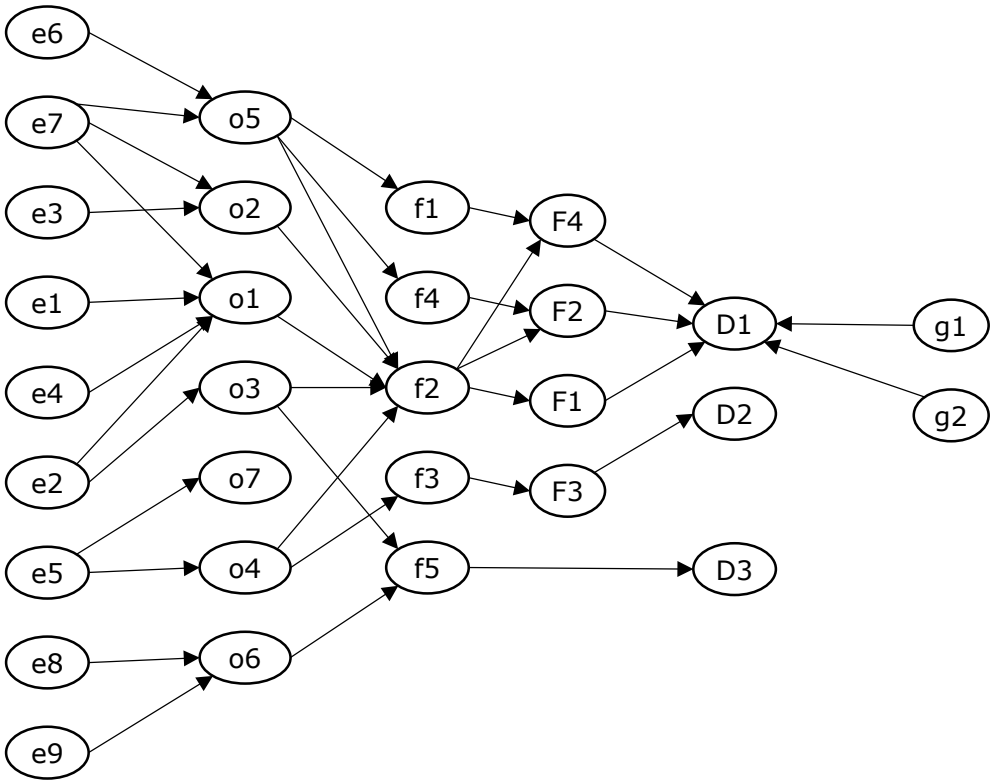
selects the most appropriate Capsules based on the patient's goals specified during a visit and personalizes the interventions according to the patient characteristics and feedback.

- Reminder to patients (generated daily/weekly by the Virtual Coach) about the prescribed treatments to follow, the drugs to be taken, the scheduled visits and the active Capsules, as well as feedback on the activities carried out by the patient.
- Educational interventions (e.g., nutritional, sleep and adverse events education) captured by validated educational material, including the AIMAC (Associazione Italiana Malati di Cancro) publications, to improve patient awareness of his conditions.
- Option sets of recommendations produced by GoCom in case it detects the presence of drug-drug interactions, drug-disease conflicts (e.g., treatments causing adverse events) or general treatment interactions (e.g., Capsules, procedures). These interactions can occur when several CIGs are applied simultaneously to a single comorbid patient as clinical practice guidelines often focus on a particular disease. If GoCom finds interactions, it will provide a series of alternative solutions to mitigate the conflicts, modeled as an option set, with explanations, priorities for each solution and goals appropriate for the patient. In this way, the clinician can have all the information he needs to make an informed decision, also considering the patient's preferences.
- Clinical abstractions that represent structured concepts regarding the patient's clinical status and that involve the combination of multiple point raw data or interpretations of the PROs reported by the patient. These abstract medical concepts, coming from the CIGs, are computed by the KDOM at the request of the other AI components. Examples of clinical abstractions are persistent diarrhea, verified when the patient has diarrhea episodes for at least 3 days in a row, and elevated level of LDH (Lactate Dehydrogenase), verified when the LDH value measured in a blood test is greater than 250 U/L.
- Predictions for patients with metastatic renal cell carcinoma and metastatic melanoma, customized based on patient data, known medical history and details of prescribed cancer treatment. Prediction results include overall survival, response to treatment, adverse events prediction for toxicities, reason for treatment termination (disease progression, toxicity, planned treatment completed, etc.), progression-free survival, time elapsed between the diagnosis of primary melanoma and the onset of metastatic disease. The predictions can be useful to clinicians in their future decisions regarding the treatment of patients. The initial knowledge used by the Predictive Models is produced by exploiting a series of training data concerning real previous patients made available by the two clinical centers participating in the consortium (ICSM and NKI), as well as by the AIMAC patients' organization. A full description of this retrospective dataset is given in [7].

4.2. Data Representation frameworks

As described in the previous section, data can be represented at various levels of granularity. There will be observations, such as raw data coming from sensors, findings as reported by patients or clinicians, such as "diarrhea", and patterns of findings, such as "persistent diarrhea", if diarrhea is reported on multiple consecutive days. Such facets may in combination with patients' background information lead to a diagnosis. For example, persistent diarrhea in the context of use of certain medication, may lead to the conclusion

that persistent diarrhea is an adverse event of medication. This is shown in Figure 3, which is adapted from [8].



Empirium → Observations → Findings → Facets → Diagnoses ← Global Complexes

Figure 3. Ontological model for clinical problem solving

Various components in the CAPABLE system, such as GoCom and KDOM, contribute to generating the higher-level information, such as facets, findings and diagnoses. Representation of the data in OMOP CDM and HL7 FHIR will need to cater for these varying levels of abstraction. For example, it should enable storing and exchanging these types of information: a finding "pruritus", a facet "pruritus, ongoing" or "pruritus, ongoing for x days", or a diagnosis "drug-induced pruritus".

Section 6.3 will further describe how this representation is realized.

5. Metadata

In the previous section we addressed the types of data to be stored and exchanged in the CAPABLE system. In this section we focus on metadata, i.e., data about the data. As metadata can cover a broad range of aspects of data, we first distinguish the various types of metadata, and then address metadata for the dataset as a whole, and metadata for the data elements in the CAPABLE system.

5.1. Types of metadata

The National Information Standards Organization (NISO) describes four types of metadata [9]. The four types of metadata are descriptive metadata, administrative metadata, structural metadata and markup languages. Descriptive metadata aids in making the resource findable and understandable. Administrative metadata describes information that is necessary to manage the resource or is related to its creation. Administrative metadata is an umbrella term and can be further divided into technical metadata, preservation metadata and rights metadata. Technical metadata describes information about digital files that are necessary and how to decode and render these. Preservation metadata helps to support the long-term management and future migration or emulation of digital files. Rights metadata describes the details of the intellectual property rights that are attached to the content. Structural metadata describes the relationship of parts of the resource to another. Markup language mix metadata and content together. Markup language is not always used and only applied on other types of metadata. Table 5 shows all of metadata types, description, primary uses and examples from NISO [9].

Table 5. The types of metadata, description, primary uses and examples, from NISO

Metadata type	Description	Primary use	Examples
Descriptive metadata	For finding or understanding a resource	Discovery Display Interoperability	Title Author Subject Genre Publication date
Technical metadata	For decoding and rendering files	Interoperability Digital object management Preservation	File type File size Creation date/time Compression scheme
Preservation metadata	Long-term management of files	Interoperability Digital object management Preservation	Checksum Preservation event
Rights metadata	Intellectual property rights attached to one another	Interoperability Digital object management	Copyright status License terms Rights holder
Structural metadata	Relationship of parts of resources to another	Navigation	Sequence Place in hierarchy
Markup languages	Integrates metadata and flags for other structural or semantic features within the content	Navigation Interoperability	Paragraph Heading List Name Date

5.2. Metadata for the datasets

We need to specify those attributes that make the datasets as much as possible follow the FAIR Principles. As no authoritative set of such attributes has been specified so far, rather

than aiming at providing an exhaustive set of attributes, we describe initial use cases to inspire a first set of attributes, which will be reviewed throughout the duration of the project. For the datasets, this will mainly be descriptive, technical, and rights metadata distinguished in Table 5. Further, we rely on Deliverable 5.1 [7], of which section 3 provides a description of data sources.

Findability

The datasets will be most relevant for people interested in subject-level data of cancer patients, melanoma patients, renal cancer patients, patients on immunotherapy, patients on targeted therapy. So, metadata should describe the inclusion criteria and possibly exclusion criteria, and indicate that it is subject-level data, as opposed to aggregate data, which could actually be provided as metadata as well. In other words, metadata should describe that the subject of records is for example "patient" or "homo sapiens".

Further, it is useful to include metadata as proposed in [10], such as study type and study status.

Accessibility

The datasets must be provided with metadata that describes access conditions. This includes possible ways of authentication and authorization, and ways in which the dataset can be accessed. This can be either as a certain type of file, or via some access service, such as a RESTful API.

Interoperability

The datasets must provide a description of the data models and vocabularies that are used, also called data dictionary or code book.

Reusability

For reusability, the dataset needs to provide a data sharing statement [10], and data quality information. For example, metadata needs to be provided about anonymization strategies that have been applied on the dataset, such as pseudonymization or data perturbation, as described in Deliverable 5.1 [7], section 3.1.2.

5.2.1. Data Set Details CAPABLE platform

Based on section 3.1.1. from Deliverable 5.1 [7] we recommend to at least include:

- Source
- Title
- Description
- Inclusion and exclusion criteria
- Geolocation
- Size
- Time period
- Format
- Identifier
- Language
- Rights
- Main data elements

5.2.2. Data Set Details ICSM Hospital - Renal Cell Carcinoma

Based on section 3.1.1. from Deliverable 5.1 [7] we recommend to at least include:

- Source: IMDC (International Metastatic Database Consortium) data set for Renal Cell Carcinoma (RCC) in which ICSM participates
- Size: 343 patients, 917 treatment lines
- Time period of reported treatments: 1998-2020
- Description: follow-up on treatments of RCC patients
- Main data elements:
  - Demographic information
  - Surgeries
  - Tumor characteristics
  - Basal Renal Functions, weight and all the hematological parameters at each treatment line
  - Metastases and brain metastases characteristics
  - Treatment details (drugs, adverse effects, dose change) and outcomes

5.2.3. Data Set Details AVL-NKI Hospital – Melanoma

Based on section 3.2.1. from Deliverable 5.1 [7] we recommend to at least include:

- Source: DMTR (Dutch Melanoma Treatment Registry) data subset from patients with melanoma treated in AVL-NKI
- Description: follow-up on patients with stage III, unresectable stage III and stage IV melanoma treated with (adjuvant) immune checkpoint inhibitors
- Size: 500 patients, one treatment line for each patient
- Time period of reported treatments: 2015-2020
- Main data elements
  - Patient Characteristics: Demographic and Social information, Comorbidities, Primary Melanoma details
  - Staging: Pathology Details, Metastases, Blood Tests, Mutations
  - Additional Surgeries
  - Treatment Details: type of treatment, dose, administrations
  - Follow-up during the treatments: toxicity, metastases, blood tests, additional treatments (radiotherapy)
  - Treatments outcomes

5.2.4. Data Set Details AIMAC

Based on section 3.3.1. from Deliverable 5.1 [7] we recommend to at least include:

- AIMAC, the cancer patients' association, provided three different kinds of data to the project:
  - data from its discussion forum;

- data from questionnaires that are filled-in when patients contact AIMAC volunteers either by telephone or by accessing the AIMAC headquarters;
- data from a questionnaire that has been recently put online during the COVID-19 pandemic.

5.3. Metadata for the data

While the previous section described the metadata for the dataset as a whole, this section focuses on describing the metadata on the data level. Here, we will focus on the models and vocabularies used for storing information and for exchanging information, and the way in which these relate. This concerns mainly technical and structural metadata as mentioned in Table 5, and is elaborated on further in Section 6.

CAPABLE makes use of OMOP CDM 5.3.1 to store the data [11]. OMOP CDM allows data to be stored in a common format and representation. This facilitates the possibility to make use of OHDSI tools and participate in OHDSI network studies. The data needs to adhere to the OMOP vocabulary to be OMOP compliant. The OMOP vocabulary contains many standardized vocabularies. CAPABLE makes use of SNOMED CT for condition occurrences, measurements and observations. LOINC for observations and measurements. RxNorm for drugs and UCUM (Unified Code for Units of Measure) for units of measurements. CAPABLE uses FHIR resources to make communication between the CAPABLE components possible.

Whereas OMOP CDM is used for data storage it doesn't provide a data exchange specification. Hence, for data exchange CAPABLE makes use of the FHIR (Fast Healthcare Interoperability Resources) standard (release 4 - v4.0.1) [12] published by HL7 (Health Level Seven). This ensures semantic interoperability and standardize the representation of healthcare data exchanged within the system. FHIR is an emerging standard developed in response to the growing use of EHRs. The goal of FHIR is to standardize the electronic exchange of clinical data, enabling healthcare providers and administrators to easily share patient information even when using different software systems. This standard is based on a compositional approach where clinical entities are represented as basic building blocks called FHIR resources. From a clinical point of view, resources are structured objects that capture different types of clinical information. FHIR allows to represent complex information as the combination of these modular resources. In practice, referral mechanisms are used to create links that relate different resources to each other.

FHIR also suggests several communication protocols for retrieving and exchanging data between healthcare applications. The most widely used protocol is based on RESTful (Representational State Transfer) web services [13]. All project data will in fact be exchanged between the Data Platform and the system components (and vice versa) as FHIR resources using the REST API. This API level is intended to receive requests (read, search and write) for standard FHIR resources, fetch/write the OMOP CDM accordingly, and create FHIR-compliant responses. The data access, thanks to the fact that it is not direct on the OMOP database but mediated by the software layer of API FHIR, allows to limit the access to particular data for specific components.

Data Platform users are expected to write new valid FHIR resources (in the agreed format and with valid attributes) among those selected by the project. A more detailed discussion of these topics is given in Section 6.2.

5.3.1. Data Mapping

Below, an outline of the mappings between OMOP and FHIR is described. It can be seen that not all elements have a one-to-one mapping, which might be expected, as OMOP has a focus on observational data, where FHIR is aimed at exchange of data in a healthcare context. The OMOP-approach to mapping is to describe the required transformations, and then develop extract-transform-load (ETL) scripts that implement the transformations.

Elements that don't exist in OMOP will either be stored in tables in a proprietary extension, so that they can be exchanged, or they will always be empty in FHIR resources.

A detailed description of the columns in the OMOP tables can be found in [14], an overview of the tables is provided in Section 6.1.

A description of HL7 FHIR resources is provided in Section 6.2, and some examples are shown in Figure 7, Figure 8, and Figure 9.

5.3.1.1. Patient resource

OMOP PERSON column	FHIR field	Logic / comment
Person_id	identifier	
	Active	Not available in OMOP. Taken from f_person table in proprietary extension
	Name	Not available in OMOP. Taken from f_person table in proprietary extension
Gender_concept_id	Gender	8507 in OMOP is "male" in FHIR 8532 in OMOP is "female" in FHIR
Year_of_birth Month_of_birth Day_of_birth	birthDate	Combine the three fields from OMOP to fill birthdate in the FHIR resource.
	deceasedBoolean	If person_id is in Death table then TRUE, else FALSE

5.3.1.2. Observation resource

*The OMOP tables OBSERVATION and MEASUREMENT map to the FHIR observation resource. CONDITION_OCCURRENCE maps to the same resource if the condition is considered a symptom in CAPABLE.

**If the condition/symptom is absent. Observation_concept_id = 3454241 (Absence of) and then the concept id of the missing condition/symptom as value_as_concept_id.

OMOP column*	FHIR field	Logic / comment
Observation_id / measurement_id / condition_occurrence_id	Identifier	
	Status	Always "final"
Observation_concept_id / measurement_concept_id / condition_concept_id	Code • System • Code • Display	Look up the concept id in CONCEPT and SOURCE_TO_CONCEPT_MAP as target_concept_id. • Vocabulary_id as system • concept_code as code • concept_name as display
person_id	Subject	Use person_id when referencing a patient
Observation_datetime / condition_start_datetime	effectiveDateTime	
Value_as_number	valueQuantity **	If CONDITION_OCCURRENCE is used then 1 for TRUE and 0 for FALSE in the FHIR resource.**

5.3.1.3. MedicationRequest resource

OMOP column	FHIR field	Logic / comment
Drug_exposure_id	Identifier	
	Status	Not present in OMOP
	Intent	Not present in OMOP
Drug_concept_id	MedicationCodeable Concept • system • code • display	Look up the concept id in CONCEPT and SOURCE_TO_CONCEPT_MAP as target_concept_id. • Vocabulary_id as system • concept_code as code • concept_name as display
Person_id	Subject	When referencing a patient
Drug_exposure_start_ datetime	authoredOn	
Provider_id	recorder	Use provide_id in PROVIDER to find specialty_concept_id.
Drug_exposure_start_ date Drug_exposure_end_ date	dosageInstruction • text • timing • frequency • period • periodUnit	Start and end date are used for timing. The others are not present in OMOP.
From table DOSE_ERA Dose_value Unit_concept_id Drug_concept_id	dosageAndRate • value • unit • system • code	Use drug_concept_id to find vocabulary_id for the system field and concept_code for the code
	maxDosePerPeriod • value • unit • system • code	Not present in OMOP
	dispenseRequest • value • unit • system • code	Not present in OMOP

5.3.1.4. Communication resource

For now it is agreed on to use Observation for the communication resource. However, the message is often an intervention and interventions belong in the PROCEDURE table according to OMOP.

OMOP column	FHIR field	Logic / comment
Observation_id / observation_source_value	Identifier	
	Status	Not present in OMOP
Observation_concept_id	Category: coding • system • code • display	Look up the concept id in CONCEPT and SOURCE_TO_CONCEPT_MAP as target_concept_id. • Vocabulary_id as system • concept_code as code • concept_name as display
Person_id	subject	When referencing a patient
	Sent	Generated for FHIR resource
	Received	Generated for FHIR resource
Person_id	recipient	When referencing a patient
	payload	Not present in OMOP

5.3.1.5. Procedure resource

OMOP column	FHIR field	Logic / comment
Procedure_occurrence_id	Identifier	
	Status	Not present in OMOP
Procedure_concept_id	Code:coding • system • code • display	Look up the concept id in CONCEPT and SOURCE_TO_CONCEPT_MAP as target_concept_id. • Vocabulary_id as system • concept_code as code concept_name as display
	Subject	In the FHIR resource it only states “0”
Provider_id	Performer	
Procedure_datetime	performedPeriod	

5.3.1.6. Task resource

A similar approach is used for the task resource as for negative findings.

OMOP column	FHIR field	Logic / comment
Observation_id	Identifier	
Value_as_string	Status	
	Intent	Not present in OMOP
Observation_concept_id		1018355 LP268497-7 Task description (LOINC concept) OR 1018098 LP267591-8 Task information (LOINC concept)
value_as_concept_id	Code	Concept id that describes the task

5.3.1.7. Goal resource

“Goals” do not really fit in OMOP as OMOP does not provide concepts for goals. A separate table for the Goal resource seems to be the best option.

6. Data Models and Vocabularies

The data captured in the CAPABLE system is represented as much as possible using domain-relevant community standards for the information model and vocabularies used. In this section we describe the models in which data are stored, OMOP CDM, and exchanged, HL7 FHIR, and the vocabularies that are used to within these models.

6.1. OMOP CDM

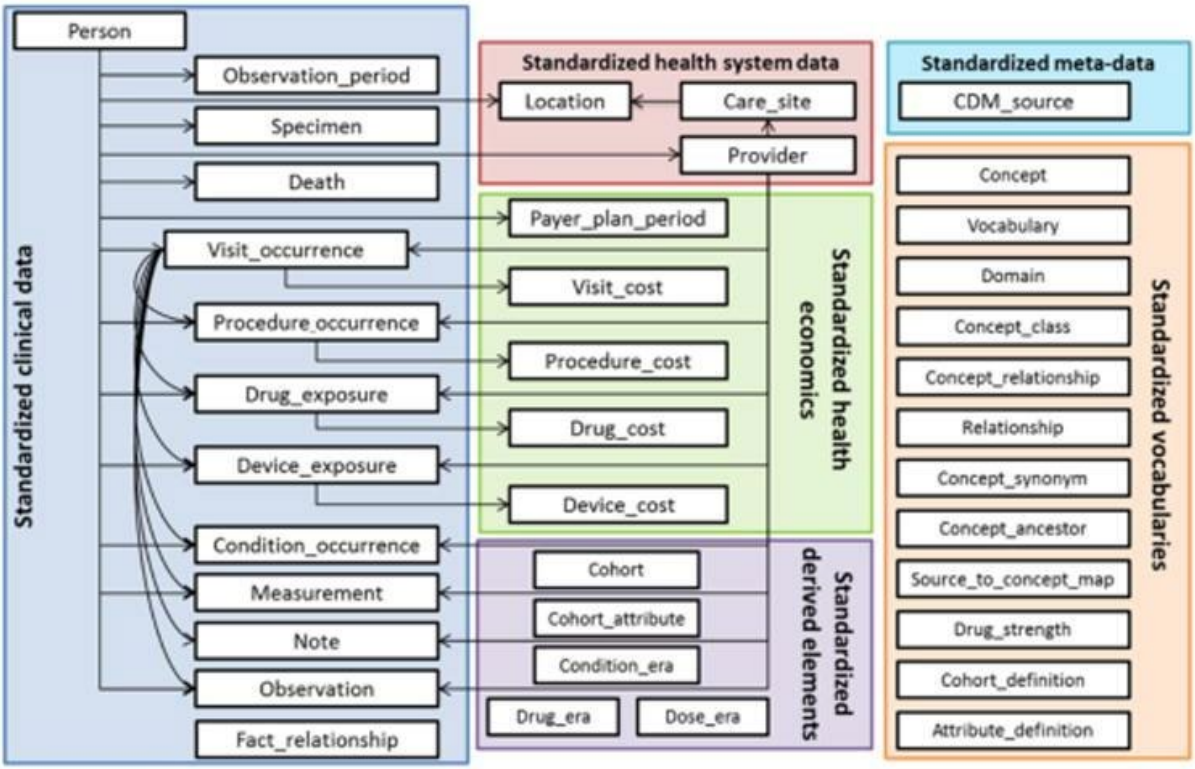


Figure 4. OMOP CDM structure, containing the classes, tables within the classes, and the relationship between tables

Figure 4, from [15], shows the classes (as labels and background colors), tables (as boxes) within those classes and relationships between tables (as arrows) from OMOP CDM version 5.3.1.

- Not all of OMOP CDM is used in CAPABLE, below we describe the tables used.
- CAPABLE makes use of all tables in the "standardized clinical data" class except for *specimen*, *device exposure*, *note* and *fact relationship*.
- CAPABLE does not use the "standardized health system data" and "standardized health economics" classes.
- From the "standardized derived elements" class the condition_era and drug_era tables are used to store derived data from the condition_occurrence and drug_exposure tables from the "standardized clinical data" class. These data will be derived with an algorithm provided by OHDSI [16].
- The "standardized meta-data" class only contains the CDM source table. This table is used to store (meta)data about the CDM instance, source dataset and vocabulary.
- All tables within the "standardized vocabularies" class are used to store the vocabulary used by OMOP. The vocabulary is provided by OHDSI [17]. CAPABLE

only makes use of standard concepts from the vocabulary that originate from SNOMED CT, LOINC, RxNorm and UCUM. Standard concepts make the standardization of methods possible. SNOMED CT concepts were used for condition occurrences, measurements, observations and procedure occurrences. LOINC concepts were used for measurements and observations. This choice was driven by the broader coverage of LOINC over SNOMED CT for measurement and observation concepts. RxNorm concepts were used for drugs and UCUM concepts were used for the units of the measurements.

Figure 5 summarizes this by showing the classes and tables that are used in the CAPABLE project.

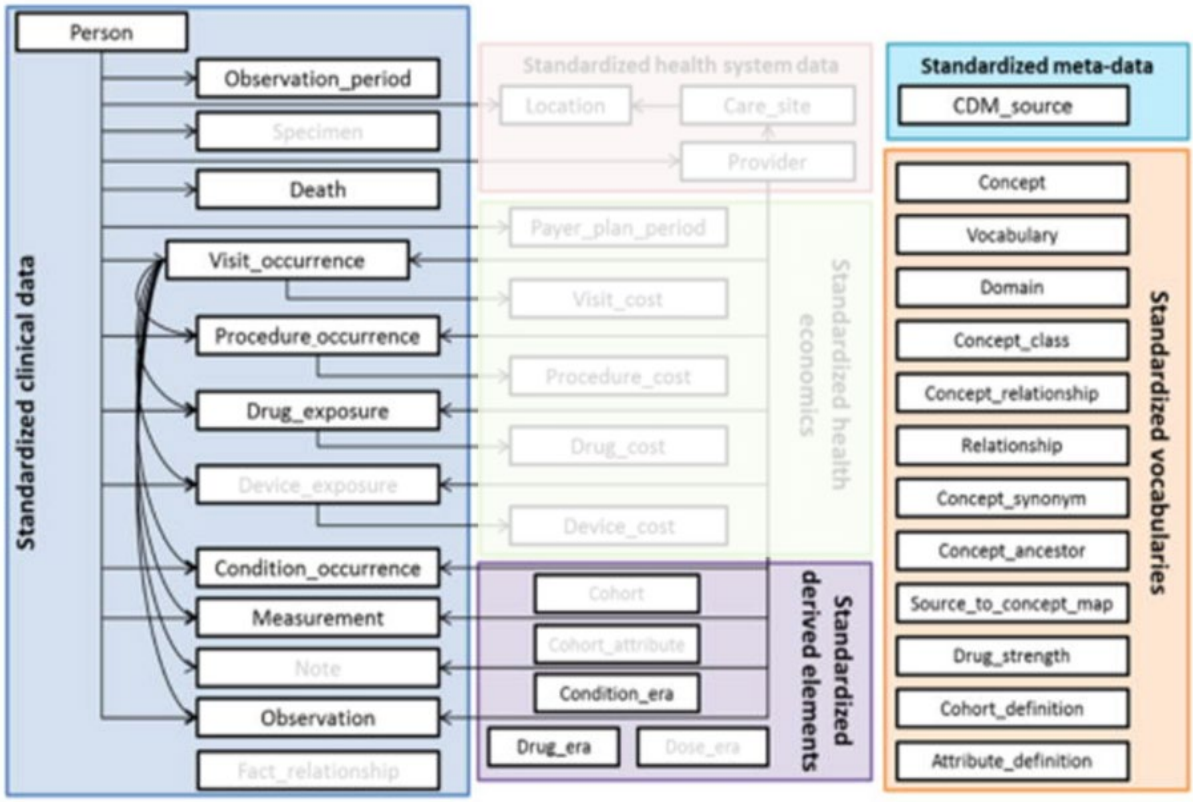


Figure 5. Classes and tables from OMOP CDM used in the CAPABLE project

6.1.1. OHDSI network studies

OHDSI network studies function as the FAIR component with the OMOP infrastructure. CAPABLE can participate in OHDSI network studies, due to the data being in the standardized OMOP CDM format [18]. The study code needs to be optimized for a range of SQL dialects as the code needs to be run in a database layer that each OHDSI network site can individually choose. OHDSI provides ATLAS, DatabaseConnector and SqlRender to help generalizing the study code.

For each study, site start-up activities may include [18]:

- Registering the study with the Institutional Review Board (or equivalent), if required
- Receiving Institutional Review Board approval to execute the study, if required
- Receiving database level permissions to read/write a schema to the approved CDM

- Ensuring configuration of a functional RStudio environment to execute the study package
- Reviewing the study code for any technical anomalies
- Working with a local IT team to permit and install any dependent R packages needed to execute the package within technical constraints

Each site has a local data analyst that will execute the study package and check the results to make certain no sensitive data is shown. OHDSI network studies never share patient-level-data. Studies only share aggregate results. Studies are announced on the weekly OHDSI community calls and on the OHDSI forum.

The current study processes are manual. However, OHDSI is working on automating this process. This platform for automated study processes is called ARACHNE [18]. ARACHNE standardizes the communication protocol and brings participating organizations into a collaborative study team. ARACHNE can facilitate the end-to-end study coordination.

While ARACHNE is currently still in active development and has not been employed at larger scale, we will not use this as a part of our FAIR infrastructure yet, but this may be reconsidered if favorable results are established within the OHDSI community during the CAPABLE project period.

6.2. HL7 FHIR Standard

The components of a modular distributed system such as CAPABLE must be technically able to exchange information and know how to decode it in order to use its contents appropriately, thus guaranteeing semantic interoperability. One of the ways to easily connect different components is to use appropriate standards to manage and harmonize the data exchanged. Various standards have been proposed in the literature for specifying models in the clinical setting. Among these, we decided to adopt the HL7 FHIR standard in CAPABLE. FHIR leverages existing logical and theoretical models to provide a simple, consistent, easy to implement and rigorous mechanism for exchanging data between healthcare applications in the form of resources. All FHIR resources share a common way of being technically defined and represented. They are built starting from data types that define models of generic and reusable elements. The resources are modeled on a wide range of concepts related to health, both clinical (such as demographics, health conditions and treatments related to patients) and administrative (such as professionals, organizations and places).

The HL7 FHIR R4 specification is organized into five levels shown in Figure 6, each of which represents a different functional area. The most relevant ones in the context of CAPABLE are levels 3 and 4. Level 3 includes a link to real-world concepts in the healthcare system explained through the *Administration* module, which deals with the base data of the people (patients and doctors), services and organizations involved. Level 4 is aimed at record-keeping and exchanging data for the healthcare process. The salient modules of level 4 are the *Clinical* category (relating to the fundamental information of a patient documented by healthcare providers during the course of clinical care), *Diagnostics* (which deals with the reporting of clinical diagnostics, including laboratory tests and symptoms) and *Medications* (for prescribing and administering drugs).

FHIR supports three standard formats for data exchange: XML (eXtensible Markup Language), JSON (JavaScript Object Notation) and RDF (Resource Description Framework). These types are among the most widespread in the computer science field and can be interpreted by any system, regardless of the tools used to develop it. FHIR therefore meets the needs of both end users and developers, allowing them to search and access information quickly. To represent the resources within CAPABLE, we chose the JSON format as it appears to be the most used for its characteristics that combine compactness

and ease of processing at a computational level, together with a discrete simplicity of reading by the human expert.

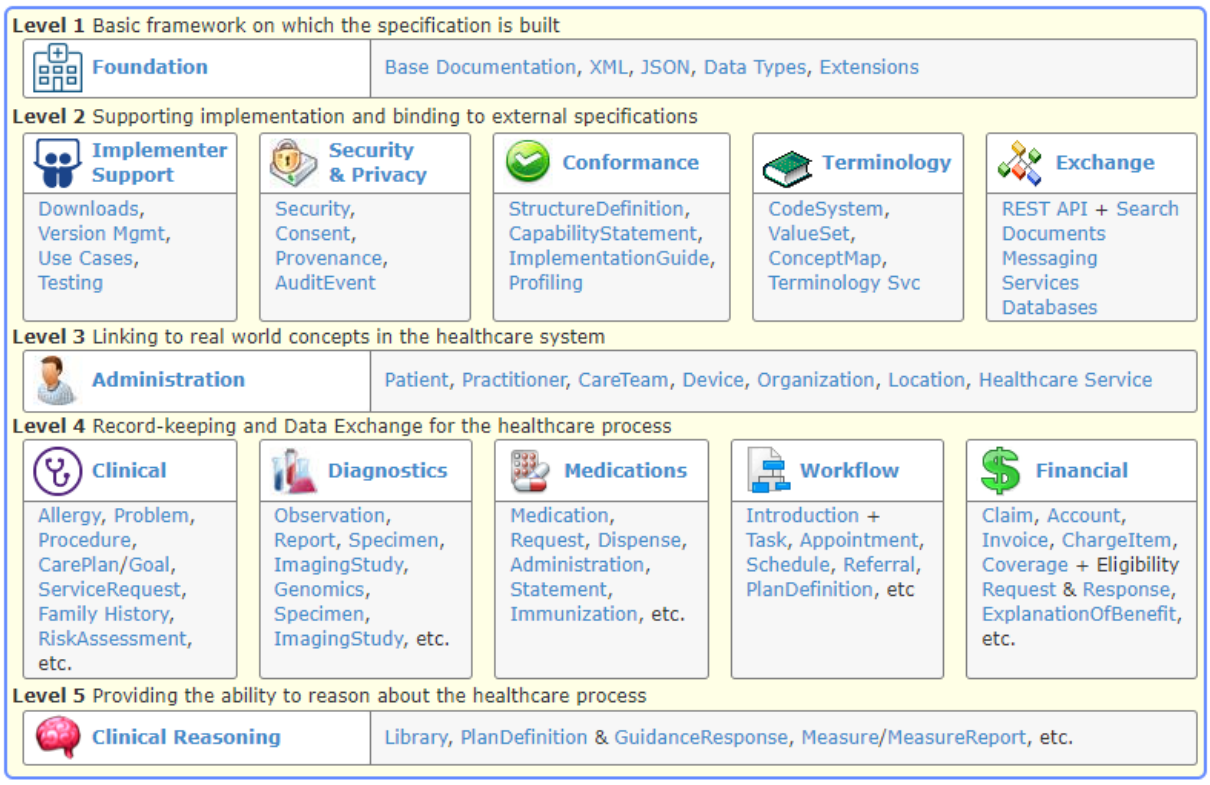


Figure 6. The five functional levels in which FHIR resources are grouped

The decision to use the FHIR standard can facilitate the implementation of the FAIR principles in the management of the exchanged data. Adherence to the FAIR principles allows to make CAPABLE data accessible and reusable by third parties outside the project. In addition, all data items collected are also made available within the project itself for the training of AI components that provide decision support to patients and clinicians. The FHIR standard, in fact, in addition to specifying the content of the data exchanged between healthcare applications, defines the possible methods to manage and implement the exchange. In particular, FHIR is described as a RESTful specification. The RESTful API is a generic interface that can be used to push and pull data between systems. This is a client/server API designed to follow RESTful design principles for CRUD operations, along with search support, using the HTTP request/response model. Client applications can then use the FHIR resources to represent the patient's clinical information and send it to the server, communicating via REST.

The specification also defines the way to package all the resources exchanged between different applications, enclosing them within a single resource called *Bundle*, that is a kind of "container". More details on the FHIR RESTful API to ensure the FAIR principles in CAPABLE is given in *Deliverable No. 3.2 - "Data-related functionality to realize a FAIR infrastructure"* [19].

6.2.1. The mapping of CAPABLE data into FHIR resources

CAPABLE will have to manage heterogeneous data from disparate sources, including EHR, questionnaires, wearable sensors, patient mobile devices for receiving PROs/PREs and medical staff web browsers for therapeutic updates, as described in Section 4.1. Indeed, one of the responsibilities of the WP3 team is to decide how to represent the information

managed by the system, then analyze the FHIR specification to identify and select the most suitable resources that best represent the CAPABLE data.

To initially identify all the possible data needed, project partners took care of creating a first scenario that was relatively simple to use as a simulation of the system during the M12 Demo (1st Proof of Concept) for the first annual review of the project. This scenario is based on the experience of the CAPABLE medical partners and is an extract of a CIG chosen among those that the CAPABLE system will model in its final phase. This guideline is "Diarrhea in adult cancer patients: ESMO Clinical Practice Guidelines" [20] and is used in the clinical setting for the treatment of diarrhea in cancer patients. In fact, diarrhea is one of the possible side effects of most cancer treatments. Detailed information on this scenario can be found in [21].

Starting from the aforementioned guideline, the activity carried out first involved building a data model, whose concepts were then mapped with the corresponding attributes of the most relevant FHIR resources. The resources that have been selected and implemented so far in CAPABLE are the following.

- *Patient* includes demographic data and other information of the assisted individual necessary to support the procedures of health care services. Patient is used as the main resource since all clinical data acquired must refer to the enrolled patients.
- *Observation* represents measurements and simple assertions consisting of name/value pairs relating to a patient. These resources are central to healthcare because they are used to support diagnosis and monitor progress. Uses of the Observation resource include vital signs (blood pressure, body temperature, etc.), laboratory data (e.g., LDH level), symptoms (e.g., diarrhea, fever, headache), personal characteristics (e.g., weight, height), personal history (e.g., smoking and alcohol consumption) and device measurements (e.g., ECG data).
- *MedicationRequest* is used to indicate a prescription for a drug and instructions for administering it to a patient. This resource covers both hospital medications and prescriptions made by a physician, while it is not intended for diet prescriptions. *MedicationRequest* allows to indicate only one drug at a time.
- *Goal* describes the intended objectives for a patient care (Sleep improvement, mental and physical well-being, etc.). In general, Goal is a desired health state that must be achieved by a patient over a period or at a specific point of time as a result of healthcare interventions. This resource is used in CAPABLE to represent the well-being goals set by the clinician (such as improve sleeping or increasing physical activity) as well as the appropriate goals for a patient suggested by the system; for example, when drug-drug interactions are found, the Goal resources contain solutions to mitigate them.
- *Communication* represents the transmission of information from a sender to a recipient. These can be any entity such as patients, professionals, or devices. Examples of Communications sent may include alert, recommendations, reminders, educational material, etc. Within CAPABLE, the Communication resource is also used to store into the Data Platform the messages exchanged among the system components. In particular, by exploiting the Case Manager's event generation mechanism, it is possible to notify the components about each message received, providing a sort of private communication channel.
- *List* is a curated collection of resources. In CAPABLE, it is mainly used to represent option-sets of recommendations or alternative solutions for the patient or physician.

In parallel, it was also necessary to define the terminological constraints, adopting coding systems including mainly SNOMED CT, LOINC, RxNorm and UCUM to encode the semantics of the clinical concepts used in resources. Many elements of the FHIR resources, in fact, must have an encoded value, which cannot be indicated as a string but must be a code, chosen from those defined in the standard terminologies.

Table 6 shows a partial list of patient information (including demographic and personal data, symptoms, measurement and drug prescriptions) to be collected in CAPABLE. For each data item, the FHIR resource chosen and the terminology code (only if required) to represent it are reported. To store the demographic data (name, surname, date of birth and gender) of a patient, the FHIR Patient resource is used, whose corresponding instance in JSON is shown in Figure 7. The height of a patient, for example, as well as other personal characteristics, are instead represented in FHIR through the Observation resource, as seen in Figure 8. The relevant fields are the "code" attribute, which describes what was observed, the "subject" attribute, which indicates the referenced patient id, and the "valueQuantity" attribute, which is the simple value determined as a result of what was observed. The prescription of antitumoral treatments, such as Nivolumab, is instead modeled using the MedicationRequest resource shown in Figure 9. The significant attributes of this resource are "medicationCodeableConcept", which indicates the name of the drug that has been prescribed, and "dosageInstruction", which is a composite attribute indicating how the drug should be taken by the patient (i.e. the time frame for the prescription, the frequency and the dosage of the drug to be taken).

Table 6. List of FHIR resources, OMOP tables and terminology codes used to represent some significant data extracted from the guideline to treat diarrhea. The full list can be found in the Annex in Section 12.

Data item	Data type	FHIR resource	Terminology code	OMOP table
Name	Text	Patient	-	Person (extended)
Surname	Text		-	Person (extended)
Date of birth	Date		-	Person
Gender	Text (M/F)		In OMOP: coded using OHDSI-specific codes 8507 and 8532	Person
Height	Double [cm]	Observation	50373000 | Body height measure (SNOMED) cm | centimeter (UCUM)	Measurement
Weight	Double [kg]	Observation	27113001 | Body weight (SNOMED) kg | kilogram (UCUM)	Measurement
Body mass index	Double [kg/m^2]	Observation	60621009 | Body Mass Index (SNOMED) kg/m2 | kilogram per square meter (UCUM)	Measurement
Location primary tumor: trunk	Text (eye/head and neck/trunk/extremities/acral/mucosal/unknown)	Observation	399687005 | Primary tumor site (SNOMED) 22943007 | Trunk structure (SNOMED)	Observation
Hypertension	Boolean (absent/ present)	Observation	38341003 | Hypertensive disorder (SNOMED)	Condition
Years as smoker	Int [years]	Observation	77176002 | Smoker (SNOMED) a | year (UCUM)	Observation

Data item	Data type	FHIR resource	Terminology code	OMOP table
LDH level	Double [U/L]	Observation	250644007 | LDH blood measurement (SNOMED) [U]/L | unit per liter (UCUM)	Measurement
Diarrhea: CTCAE grade 1	The severity level is measured in CTCAE grades (1=mild/ 2=moderate/3=severe/4=life threatening)	Observation	62315008 | Diarrhea (SNOMED) 446411000124101 | Common terminology criteria for adverse events grade 1 (SNOMED)	Observation
Blood in faeces: CTCAE grade 3	The severity level is measured in CTCAE grades (1=mild/ 2=moderate/3=severe/4=life threatening)	Observation	405729008 | Hematochezia (SNOMED) 446431000124107 | Common terminology criteria for adverse events grade 3 (SNOMED)	Observation
Nivolumab injection prescription	Consisting of validity period, dosage [mg], method and frequency of administration	Medication Request	1597876 | nivolumab (RxNorm) mg | milligram (UCUM) 43060002 | Intravenous injection (SNOMED)	Drug_Exposure (extended)
Loperamide prescription	Consisting of validity period, dosage [mg], maximum daily dose [mg] and frequency of administration	Medication Request	6468 | loperamide (RxNorm)	Drug_Exposure (extended)
Sleep improvement	Text	Goal	404950004 | Sleep behavior (SNOMED)	Observation
Physical well-being improvement	Text	Goal	68130003 | Physical activity (SNOMED)	Observation
Hobbies: Tai Chi	Text (photography/ gardening/listening to music/walking/tai chi/writing)	Observation	300758009 | Does engage in a hobby (SNOMED) 30900002 | Tai-chi (SNOMED)	Observation
Daily sleep duration	Int [hours]	Observation	93832-4 | Sleep duration (LOINC) h | hour (UCUM)	Observation
Hospital visit reminder	Text	Communication	-	Custom table
Symptom reporting intervention	Text	Communication	-	t.b.d.

```
{
    "resourceType": "Patient",
    "meta": {
        "lastUpdated": "2021-05-25T10:46:00+02:00"
    },
    "identifier": [
        {
            "value": "CAPABLE Internal ID"
        }
    ],
    "active": true,
    "name": [
        {
            "family": "Rossi",
            "given": [
                "Maria"
            ]
        }
    ],
    "gender": "female",
    "birthDate": "1968-03-23"
}
```


Figure 7. Example of a Patient resource instance in JSON format to capture a patient's personal details

```
{
    "resourceType": "Observation",
    "meta": {
        "lastUpdated": "2021-05-25T13:06:00+02:00"
    },
    "identifier": [
        {
            "value": "Internal EHR SerialCode"
        }
    ],
    "status": "final",
    "code": {
        "coding": [
            {
                "system": "http://snomed.info/sct",
                "code": "50373000",
                "display": "Body height measure"
            }
        ]
    },
    "subject": {
        "reference": "Patient/4"
    },
    "effectiveDateTime": "2021-05-25T13:06:00+02:00",
    "valueQuantity": {
        "value": 162,
        "unit": "centimeter",
        "system": "http://unitsofmeasure.org",
        "code": "cm"
    }
}
```


Figure 8. Example of an Observation resource instance in JSON format to capture a patient's height

We plan to extend the number of supported FHIR resources based on the new data that the system will need to manage as additional guidelines are implemented in CAPABLE. We are currently working on the creation of a new clinical scenario, with a more realistic and complex user story than the one used for the M12 Demo. This new scenario will make use of additional CIGs and will be used for the M18 Demo (which will be described in *Deliverable No. 4.2 – "2nd iteration of the Platform POC"*).

Finally, FHIR provides a set of resources, known as conformance resources, which can be used to represent and share the implementation choices that have been made within the project. These conformance resources allow to define, for each supported resource, the attributes used, and which terminologies are allowed in the encoded elements. This is done through the resources called *StructureDefinition* for profiling. All conformance resources are then combined to create an *ImplementationGuide*, which is a structured document describing how FHIR is adapted to support a specific use case, and a *CapabilityStatement*, that documents which aspects of the FHIR specification and API are implemented and how.

This set of elements describing how FHIR has been used within CAPABLE will be created and published once all the data items we plan to manage in the project have been defined and the corresponding FHIR resources have been implemented.

```
{
    "resourceType": "MedicationRequest",
    "meta": {
        "lastUpdated": "2021-05-26T14:38:00+02:00"
    },
    "status": "active",
    "intent": "order",
    "medicationCodeableConcept": {
        "coding": [
            {
                "system": "http://www.nlm.nih.gov/research/umls/rxnorm",
                "code": "1597876",
                "display": "Nivolumab"
            }
        ]
    },
    "subject": {
        "reference": "Patient/4"
    },
    "authoredOn": "2021-05-26T14:38:00+02:00",
    "dosageInstruction": [
        {
            "text": "240 mg for an hour once every two weeks",
            "timing": {
                "repeat": {
                    "boundsPeriod": {
                        "start": "2021-05-26",
                        "end": "2021-11-26"
                    },
                    "duration": "1",
                    "durationUnit": "h",
                    "frequency": "1",
                    "period": 2,
                    "periodUnit": "wk"
                }
            },
            "asNeededBoolean": false,
            "route": {
                "coding": [
                    {
                        "system": "http://snomed.info/sct",
                        "code": "43060002",
                        "display": "Intravenous injection"
                    }
                ]
            },
            "doseAndRate": [
                {
                    "doseQuantity": {
                        "value": 240,
                        "unit": "milligram",
                        "system": "http://unitsofmeasure.org",
                        "code": "mg"
                    }
                }
            ]
        }
    ]
}
```


Figure 9. Example of a MedicationRequest resource instance in JSON format to represent the prescription of Nivolumab (an immunotherapy drug)

6.3. Vocabularies

As described in the previous sections, the choice for specific vocabularies is driven by various criteria. First, the vocabularies should provide proper codes for information being represented, second, use of the vocabularies should be supported in OMOP CDM and HL7 FHIR. For OMOP CDM, this means that they are included in OHDSI's Athena vocabulary [17] as valid standard concepts. For example, the drug Nivolumab is represented using concepts from many vocabularies, including RxNorm, SNOMED CT, the Anatomical Therapeutic Chemical (ATC) Classification, and Medical Subject Headings (MeSH). Of these, the RxNorm concept is the standard concept, while all others are non-standard. Applying this approach resulted in the following choice of vocabularies.

SNOMED CT – for coding of findings, procedures, and events. For example:

38341003 | Hypertensive disorder

77176002 | Smoker

LOINC or SNOMED CT – for coding of types of observations and measurements. For example:

27113001 | Body weight

404950004 | Sleep behavior

93832-4 | Sleep duration

RxNorm – for coding of drugs. For example:

6468 | loperamide

UCUM – for coding of units of measure. For example:

cm | centimeter

[U]/L | unit per liter

Data, especially regarding findings and procedures, can be captured at varying levels of granularity. For example, using SNOMED CT one can capture *38341003 | Hypertensive disorder* as indicated above, but also a specific hypertensive disorder, such as *194783001 | Secondary malignant renovascular hypertension*. This also enables capturing data at the various levels of the ontological model for clinical problem solving that was introduced in Section 4.2. Such data can either be entered by users or be generated by the AI modules and then stored on the Data Platform. An example of the latter would be the following.

- A clinician has prescribed:
  (RxNorm) 1597876 | Nivolumab
- The patient then reports a finding:
  (SNOMED CT) 418363000 | Pruritus
- Decision support may suggest a diagnosis to the clinician:
  (SNOMED CT) 724843002 | Drug-induced pruritus
- This diagnosis may then be accepted by the clinician and stored on the Data Platform

7.Metadata Models, Vocabularies and Deposition

CAPABLE opted for four metadata standards which are Dublin core [22], Data Catalog (DCAT) [23], DataCite [24] and Generic Dataset Metadata Template (GDMT) [25]. Choosing these four metadata standards should cover the CAPABLE metadata needs and increase the findability of the CAPABLE dataset. These are addressed in more detail below, after which we reflect on possibilities for deposition of the metadata.

7.1. Metadata Models and Vocabularies

Dublin core is one of the most generic metadata standards available and contains fifteen data elements. Dublin core was published as ISO standard in 2009 [26] and is still deemed relevant [27]. Dublin Core is also one of the most best known and widely used standards [26]. Dublin core covers the basic metadata needs of CAPABLE for these reasons.

CAPABLE also opted for DataCatalog (DCAT) besides Dublin Core. DCAT is a RDF vocabulary designed to facilitate interoperability of datasets that are published on the web. A publisher is enabled to describe datasets and data services in a catalog with DCAT using a standard model and vocabulary that can facilitate metadata from multiple catalogs [23]. This will increase the findability of datasets that make use of DCAT. This approach supports a decentralized manner to publishing data catalogs and also makes federated searches for datasets across catalogs in multiple sites possible using the same query mechanism and structure [23]. As CAPABLE has multiple data sets, DCAT will be applied to provide metadata for the catalog of datasets, as well as each of these data sets.

DataCite has been developed with the aim of accurately and consistently identifying resources for citation and retrieval. A resource is typically a dataset, but can be of any kind. DataCite is intended to be a generic metadata schema to support a broad range of research. DataCite primarily supports citation and findability of data and is not meant to replace domain specific meta data standards [28]. DataCite is considered state of the art regarding metadata schemas and is therefore interesting to CAPABLE.

GDMT is the last chosen metadata schema for the CAPABLE dataset. GDMT is compatible with the Center for Expanded Data Annotation and Retrieval (CEDAR) Workbench [29], is capable of evaluating and supporting FAIR Data Points, supports RDF production that can be presented as nanopublications, satisfies FAIR principles and metadata evaluators, and securely generates Digital Object Identifiers (DOI) [25]. GDMT is one of the newest metadata schemas and might prove to be a good choice for CAPABLE in the future.

7.2. Metadata Deposition

Zenodo is a universal open-access repository owned by the European Organization for Nuclear Research CERN, and developed within the OpenAIRE project that is commissioned by the European Commission [30]. Zenodo also provides a DOI to datasets to make the dataset more findable and easier to cite. DataCite Commons is a web search interface for a graph formed by the collection of scholarly resources such as publications, datasets, people and research organizations and their connections. This graph is called the Persistent Identifiers (PID) graph. The PID graph makes use of persistent identifiers and GraphQL (a query language for APIs). The PIDs and metadata are provided by DataCite, Crossref, ORCID, Zenodo and others [31]. Consequently, using various metadata schemas and various data repositories greatly increases the findability and exposure of the CAPABLE (meta)data.

8. Architectural decisions and current status

The data and metadata specified in this deliverable and the selected underlying models will be implemented to establish the CAPABLE FAIR infrastructure. This section provides a brief overview of this infrastructure, the functionality of which is described in Deliverable 3.2 [19].

8.1. Data storage and access

Data is stored in an OMOP CDM-compliant PostgreSQL database. This provides an increasingly accepted community standard information model, together with related vocabularies. At Month 18 of the CAPABLE project, a demo system has been delivered, that forms the basis for the system that will be used once data collection starts. This system provides access to the data through a HAPI FHIR [32] server, an HL7 FHIR-compliant RESTful server. Responses to data requests conform to the FHIR-specified resources that are mentioned in Table 6: Patient, Observation, MedicationRequest, Communication, and Goal.

As the CAPABLE infrastructure will contain privacy-sensitive data, access is restricted by an authentication and authorization mechanism, provided by HL7 FHIR.

The data infrastructure is provided as a dedicated service that will be available throughout the duration of the project. If upon completion of the CAPABLE project this service is discontinued, data will be archived in compliance with regulations as described in the data management plan [33]. It may then no longer be accessible, which will then be made explicit in the metadata.

8.2. Metadata storage and access

Currently, a variety of highly generic metadata models exist. None of these provides a model to completely cater for the metadata of the CAPABLE infrastructure. This means that it is required to combine various metadata models: DCAT, Dublin Core, DataCite, and GDMT, and may expand these.

Current (meta)data deposition services, such as Zenodo or figshare [34], support restricted metadata models, largely focusing on the DataCite metadata schema. Consequently, we will establish a proprietary metadata repository for CAPABLE, by implementing a FAIR Data Point. This enables description of rich metadata, based on community accepted standards.

The metadata is provided by a dedicated FAIR Data Point that will be available throughout the duration of the project. If upon completion of the CAPABLE project this service is discontinued, metadata will be made available in a third-party deposition service to ensure continued availability of the metadata. If such services do not sufficiently support the metadata models used, the metadata can be deposited as a dataset in one of the deposition services, indicating that the dataset consists of metadata. While such indirection is suboptimal, it does provide persistence of the metadata.

9. Assessment of adherence of the Information Architecture to the FAIR Principles

9.1. Findability

F1. (Meta)data are assigned a globally unique and persistent identifier

- ✔ Data identifiers are FHIR resource identifiers, which are globally unique. These identifiers can be made persistent and resolvable using persistent redirection services, such as w3id.org [35], to provide a persistent URL for the service.
- ✔ Metadata identifiers will be made persistent, globally unique, and resolvable, by using persistent redirection services, such as w3id.org.

F2. Data are described with rich metadata (defined by R1 below)

- ✔ The metadata defined in this Deliverable are aimed at fulfilling this Principle.

F3. Metadata clearly and explicitly include the identifier of the data they describe

- ✔ Metadata will address one or more catalogs, datasets, and distributions, each of which will be explicitly identified.

F4. (Meta)data are registered or indexed in a searchable resource

- ✔ Data are provided via a FHIR server (if access is granted), which is described as a distribution of one or more datasets in the metadata.
- ✔ Metadata is provided via a FAIR Data Point, that is registered with the FAIR Data Point index.

9.2. Accessibility

A1. (Meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 The protocol is open, free, and universally implementable

- ✔ Data are retrievable using HL7 FHIR, that follows this Principle.
- ✔ Metadata are retrievable by the FAIR Data Point, that follows this Principle.

A1.2 The protocol allows for an authentication and authorization procedure, where necessary

- ✔ HL7 FHIR allows for authentication and authorization, which is included in the infrastructure.
- ✔ For metadata no authentication and authorization is required.

A2. Metadata are accessible, even when the data are no longer available

- ✔ A deposition service will be used for the metadata if the FAIR Data Point is discontinued.

9.3. Interoperability

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

- ✔ Data are represented as HL7 FHIR resources, which, especially if expressed in RDF, e.g., using the turtle syntax, follow this Principle.
- ✔ Metadata are represented in RDF, which follows this Principle.

I2. (Meta)data use vocabularies that follow FAIR principles

- ✔ Data are represented using well-established vocabularies, that can be considered as FAIR to some extent, such as SNOMED, LOINC, RxNorm, and UCUM.
- ✔ Metadata are represented using community accepted schemas and accompanying vocabularies, which can be considered as FAIR.

I3. (Meta)data include qualified references to other (meta)data

- ✔ Data are represented as HL7 FHIR Resources, which provide qualified references to other Resources.
- ✔ Metadata are represented using schemas that specify qualified references to other metadata.

9.4. Reusability

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

- ✔ As described in the Data Management Plan, sharable data will be licensed with Creative Commons CC-BY.
- ✔ Metadata will be licensed with CC-0.

R1.2. (Meta)data are associated with detailed provenance

- ✔ As described in Deliverable 3.2 [19], provenance of data will be described using the provenance ontology, PROV-O.
- ✔ Where relevant, also provenance of metadata will be described using PROV-O.

R1.3. (Meta)data meet domain-relevant community standards

- ✔ Data are represented as HL7 FHIR Resources, and hence meet domain-relevant community standards.
- ✔ Metadata are represented with a selection of established standards, to follow this Principle.

10. Glossary

AI	Artificial Intelligence
AIMAC	Associazione Italiana Malati di Cancro
API	Application Programming Interface
ATC	Anatomical Therapeutic Chemical
CAPABLE	Cancer Patients Better Life Experience
CDM	Common Data Model
CEDAR	Center for Expanded Data Annotation and Retrieval
CIG	Computer-Interpretable Guideline
CR	Create and Read
CRUD	Create, Read, Updated and Delete
CTCAE	Common Terminology Criteria for Adverse Events
DSS	Decision Support System
ECG	Electrocardiogram
EHR	Electronic Health Record
ETL	Extract, Transform and Load
FAIR	Findability, Accessibility, Interoperability, Reusability
FHIR	Fast Healthcare Interoperability Resources
GUI	Graphical User Interface
HL7	Health Level Seven
HTTP	HyperText Transfer Protocol
ICSM	Istituti Clinici Scientifici Maugeri
JSON	JavaScript Object Notation
KDOM	Knowledge-Data Ontology Mapper
LDH	Lactate Dehydrogenase
LOINC	Logical Observation Identifiers Names and Codes
MeSH	Medical Subject Headings
NISO	National Information Standards Organization
NKI	Netherlands Cancer Institute

OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
PPG	Photo-plethysmography
PRE	Patient Reported Experience
PRO	Patient Reported Outcome
RDF	Resource Description Framework
REDCap	Research Electronic Data Capture
REST	Representational State Transfer
UCUM	Unified Code for Units of Measure
XML	eXtensible Markup Language

11. References

[1] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 160018, 2016, doi: 10.1038/sdata.2016.18.
[2] Peleg, Mor *et al.*, "CAPABLE D2.1: Requirements Table and Use Case Description," Dec. 2020, doi: 10.5281/ZENODO.4540452.
[3] E. Barkan, F. Gilboa-Solomon, and S. Quaglini, "CAPABLE D5.2: Framework Defined (Including Patients' Needs) Based on Available Data and Modelling Approaches," Jun. 2020.
[4] Sacchi, Lucia *et al.*, "CAPABLE D7.1: Study Plan, Protocols Definition, and Informed Consent/Assent Drafts," Dec. 2020, doi: 10.5281/ZENODO.4540503.
[5] *Asus VivoWatch BP (HC-A04)*. https://www.asus.com/Mobile/Wearable-Healthcare/VivoWatch/ASUS-VivoWatch-BP-HC-A04/ (accessed Jun. 18, 2021).
[6] ASUS Life Corporation, "OmniCare Medical IoT Platform." https://omnicare.asus.life/ (accessed Jul. 14, 2021).
[7] Barkan, Ella, Gilboa-Solomon, Flora, and Quaglini, Silvana, "CAPABLE D5.1: Data Ready for Modelling and Reasoning Development," Jun. 2020, doi: 10.5281/ZENODO.4540571.
[8] J. F. Arocha, D. Wang, and V. L. Patel, "Identifying reasoning strategies in medical decision making: A methodological guide," *J. Biomed. Inform.*, vol. 38, no. 2, pp. 154–171, Apr. 2005, doi: 10.1016/j.jbi.2005.02.001.
[9] J. Riley and National Information Standards Organization (Estats Units d'Amèrica), *Understanding metadata: what is metadata, and what is it for?* 2017. Accessed: Jun. 25, 2021. [Online]. Available: https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Met%E2%80%A6
[10] S. Canham and C. Ohmann, "A metadata schema for data objects in clinical research," *Trials*, vol. 17, no. 1, p. 557, Dec. 2016, doi: 10.1186/s13063-016-1686-5.
[11] *OMOP Common Data Model*. https://www.ohdsi.org/data-standardization/the-common-data-model/ (accessed Jun. 30, 2021).
[12] *HL7 FHIR Release 4*. https://www.hl7.org/fhir/ (accessed Jun. 30, 2021).
[13] *HL7 FHIR RESTful API*. https://www.hl7.org/fhir/http.html (accessed Jun. 30, 2021).
[14] OHDSI Consortium, *OMOP CDM v5.3.1*. https://ohdsi.github.io/CommonDataModel/cdm531.html (accessed Jul. 15, 2021).
[15] G. Jiang, R. C. Kiefer, D. K. Sharma, E. Prud'hommeaux, and H. R. Solbrig, "A Consensus-Based Approach for Harmonizing the OHDSI Common Data Model with HL7 FHIR," *Stud. Health Technol. Inform.*, vol. 245, pp. 887–891, 2017.
[16] OHDSI Consortium, *SQL scripts*. http://ohdsi.github.io/CommonDataModel/sqlScripts.html (accessed Jun. 30, 2021).
[17] OHDSI Consortium, *Athena*. https://athena.ohdsi.org/ (accessed Jun. 30, 2021).
[18] K. Kostka, G. Klebanov, and S. Dempster, *OHDSI Network Research*. https://ohdsi.github.io/TheBookOfOhdsi/NetworkResearch.html (accessed Jun. 30, 2021).
[19] Cornet, Ronald *et al.*, "CAPABLE D3.2: Data-related Functionality to Realize a FAIR Infrastructure," Jun. 2021, doi: 10.5281/ZENODO.5005509.
[20] P. Bossi *et al.*, "Diarrhoea in adult cancer patients: ESMO Clinical Practice Guidelines," *Ann. Oncol.*, vol. 29, pp. iv126–iv142, Oct. 2018, doi: 10.1093/annonc/mdy145.
[21] Gabetta, Matteo and Parimbelli, Enea, "CAPABLE D4.1: 1st Iteration of the Platform Proof Of Concept," Dec. 2020, doi: 10.5281/ZENODO.4540457.
[22] *Dublin Core metadata initiative*. https://www.dublincore.org/ (accessed Jun. 30, 2021).
[23] R. Albertoni, D. Browning, S. Cox, A. Beltran, A. Perego, and P. Winstanley, *Data Catalog Vocabulary (DCAT) - Version 3*. https://www.w3.org/TR/vocab-dcat-3/ (accessed Jun. 30, 2021).

[24] *DataCite Metadata Schema*. https://schema.datacite.org/ (accessed Jun. 30, 2021).
[25] N. Vasiljevic and J. Graybeal, *Generic Dataset Metadata Template*. https://github.com/fair-data-collective/generic-dataset-metadata-template (accessed Jun. 30, 2021).
[26] DCC, *Dublin Core*. https://www.dcc.ac.uk/resources/metadata-standards/dublin-core (accessed Jun. 30, 2021).
[27] SURF, *Dublin Core*. https://standaarden.surf.nl/index.php/Dublin_Core (accessed Jun. 30, 2021).
[28] DataCite Metadata Working Group, "DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4," p. 82 pages, 2021, doi: 10.14454/3W3Z-SA82.
[29] R. S. Gonçalves *et al.*, "The CEDAR Workbench: An Ontology-Assisted Environment for Authoring Metadata that Describe Scientific Experiments," in *The Semantic Web – ISWC 2017*, vol. 10588, C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, and J. Heflin, Eds. Cham: Springer International Publishing, 2017, pp. 103–110. doi: 10.1007/978-3-319-68204-4_10.
[30] European Organization For Nuclear Research and OpenAIRE, "Zenodo: Research. Shared.," 2013, doi: 10.25495/7GXK-RD71.
[31] DataCite, *About DataCite Commons*. https://commons.datacite.org/about (accessed Jun. 30, 2021).
[32] *HAPI FHIR*. https://hapifhir.io/ (accessed Jun. 13, 2021).
[33] Parimbelli, Enea *et al.*, "CAPABLE D1.2: Data Management Plan," Jun. 2020, doi: 10.5281/ZENODO.3970579.
[34] *FigShare*. https://figshare.com/ (accessed Jun. 13, 2021).
[35] W3C Permanent Identifier Community Group, *Permanent Identifiers for the Web*. https://w3id.org/ (accessed Jul. 14, 2021).

12. Annex

Data item (orange -> pred. models, blue -> KDOM abstraction)	**Data type**	**Value in the scenario**	**Terminology codes** (SNOMED CT, unless otherwise stated)	**FHIR resources**
Cancer diagnosis	text (melanoma stage III / melanoma stage IV / kidney)	Melanoma stage IV	93655004 Malignant melanoma of skin 363518003 Malignant tumor of kidney Values: 50283003 | Clinical stage III (finding) 2640006 | Clinical stage IV (finding) 52101004 | Present (qualifier value)	Observation (Condition if it will be implemented) https://drive.google.com/file/d/109MwFRv-tEvr5n-5aA1Z6u5CPfedVswo/view?usp=sharing
BRAF mutation	boolean	FALSE	93690-6 | BRAF gene V600 mutations [Identifier] in Plasma cell-free DNA by Molecular genetics method Nominal (LOINC) *-> assuming that V600E and V600K mutations means and/or, i.e., no further detail*	Observation (Condition if it will be implemented) https://drive.google.com/file/d/1-m7IX-HxuMAklbqJTYUhayk9v7dg68hj/view?usp=sharing
LDH level	float	265 U/L	250644007 | LDH blood measurement [U]/L | unit per liter (UCUM)	Observation https://drive.google.com/file/d/1orWGaRP9tf58Dq7w0oT49wKCmdpwyCXq/view?usp=sharing
Elevated LDH	binary (increased / normal)	Increased	250644007 | LDH blood measurement Values: 35105006 | Increased 260395002 | Normal range	Observation https://drive.google.com/file/d/1JJUiHctEP5iOuST0xck3YpXuXzkxfs-X/view?usp=sharing
Metastasis site	text (brain/ liver / lung/ other)	Lung	399462009 | Secondary tumor site Values: 12738006 | Brain structure (body structure) 10200004 | Liver structure (body structure) 39607008 | Lung structure (body structure) 74964007 | Other (qualifier value) 2667000 | Absent	Observation https://drive.google.com/file/d/1zL55JLSBkhuzvvO9QUpWSpu5fMXlAo5O/view?usp=sharing
Option 1 - Nivolumab monotherapy			1597876 Nivolumab (RxNorm) 708255002 First line treatment (Snomed)	Goal https://drive.google.com/file/d/1wZShTFo2ZAYZMfAIoVkvePyf9s9ts0ZN/view?usp=sharing
Option 2 - Nivolumab-Ipilimumab combination therapy			1094833 Ipilimumab (RxNorm) 708255002 First line treatment (Snomed)	Goal https://drive.google.com/file/d/1XenSMxl0cT

				MllaWvMM_CrPYyZDAmMnNc/view?usp=sharing
Option-set that contains Option 1 and Option 2			385642001 Under consideration	List https://drive.google.com/file/d/1Q3rLginvauDizC9jB-nX7M0Vc3JySBeD/view?usp=sharing
Cancer treatment decision - Medication request	set of text	Nivolumab monotherapy - Reference to Option 1	420072004 | Decision status (Snomed) 1597876 | Nivolumab (RxNorm) mg | milligram (UCUM) 43060002 | Intravenous injection (Snomed)	Observation + MedicationRequest https://drive.google.com/file/d/1x0dGb2hXY_4q1DhUTlmFZksrXbnowuIL/view?usp=sharing and https://drive.google.com/file/d/1SFBqTNCLWcNgAlm6WO1pgc04Nb5pzwGR/view?usp=sharing
Hobbies: walking	text (photography / gardening / listening to music / walking / tai chi / writing)	Walking	300758009 | Does engage in a hobby (finding) Values: LA15948-5 | Hobbies developing photographs (LOINC) 300752005 | Does perform gardening activities LA25353-6 | Music (LOINC) 129006008 | Walking 30900002 | Tai-chi 12261009 | Writing	Observation https://drive.google.com/file/d/1_0ygjsjP34zK3ZA4GuvB5GNLCxWjkZc-/view?usp=sharing
Daily sleep duration	float	4.5 hours	93832-4 Sleep duration (LOINC) h | hour (UCUM)	Observation https://drive.google.com/file/d/1vTaSlwWyIytOOcegGTx63bmF3l9pVNLj/view?usp=sharing
			93829-0 REM sleep duration (LOINC)	Observation https://drive.google.com/file/d/1EFSPjtP_LMA-eZynGGMY11rl-GKsaSGe/view?usp=sharing
			93831-6 Deep sleep duration (LOINC)	Observation https://drive.google.com/file/d/1-TSZA9GrePOIloGxIi7TRLYsKFgQ2ti5/view?usp=sharing
			93830-8 Light sleep duration (LOINC)	Observation https://drive.google.com/file/d/1QDRiHjSuR4QgjHk-CBuTxtLU_qTW-5B8/view?usp=sharing
Daily number of steps	int	1000	41950-7 Number of steps in 24 hour Measured (LOINC)	Observation https://drive.google.com/file/d/1cxONXOCW24lKXdbx6TpcH5jiOlyfFyZ_/view?usp=sharing
Weekly average steps	float	2000	41952-3 Number of steps in 1 week Measured (LOINC)	Observation https://drive.google.com/file/d/14L4Eo1t3u0SspFje0Shhb75j9Ly_TAKJ/view?usp=sharing
Goal 1 - Sleep Improvement	text	Sleep improvement= "prescribed (by physician)"	404950004 | Sleep behavior	Goal https://drive.google.com/file/d/19fkEKqNI1SRD39N4eK4i9KzdcTpSCFMC/view?usp=sharing

Goal 2 - Physical activity improvement	text	Physical activity improvement= "prescribed (by physician)"	68130003 | Physical activity	Goal https://drive.google.com/file/d/1FPMiCyr37vr1YqAHMizeheERr7uGtEto/view?usp=sharing
New Goals for patient				Communication https://drive.google.com/file/d/15piJqStYR3H6mrZxmLjKZmEXEAL7Zb_b/view?usp=sharing
Intervention sent - Capsule	text	"Consistent with the goals you have decided with your doctor, improve the quality of sleep and increase physical well-being, we propose you to start a journey with the 30x30 Nature Challenge capsule. Are you there?"	310882002 | Exercise on prescription (regime/therapy)	Communication + MedicationRequest (ServiceRequest if it will be implemented) https://drive.google.com/file/d/1OmV1XiUcSgDTSlOiCaf-IB0HrYhrN4rf/view?usp=sharing and https://drive.google.com/file/d/1I6Y1clYU7H9ptDvqcJNEeNOj7-A8OTCr/view?usp=sharing
Answer to intervention sent - Capsule	boolean	TRUE	385645004 Accepted 443390004 Refused	Observation https://drive.google.com/file/d/1OMo36_B9YowLJ8xuijWvebn-364yeXyx/view?usp=sharing
Capsule start				MedicationRequest https://drive.google.com/file/d/1BfB-Mm6fyLuUGwWddiMZ4YZnh5bmcBoU/view?usp=sharing
Capsule activity reporting	int (duration in minutes)	30 minutes	55411-3 | Exercise duration (LOINC) min | minute (UCUM)	Observation https://drive.google.com/file/d/1piWzX6fnJYsogKwXYOJcQtkXEsGFul_-/view?usp=sharing
Capsule satisfaction score	int (0/1/2/3/4/5)	2	273689006 | Patient satisfaction score (assessment scale)	Observation https://drive.google.com/file/d/1ofhPiu5FSyhX8kPg7j5EEs034K3fMj7t/view?usp=sharing
Pruritus			279333002 Pruritus of skin	Observation https://drive.google.com/file/d/1oFqPf56KV3F66Vma45EtoHqbbX0vKYTI/view?usp=sharing
Pruritus grade	set of grades (0/1/2/3/4)	CTCAE grade 1	2667000 | Absent 446411000124101 | Common terminology criteria for adverse events grade 1 446421000124109 | Common terminology criteria for adverse events grade 2 446431000124107 | Common terminology criteria for adverse events grade 3 446441000124102 | Common terminology criteria for adverse events grade 4	

Pruritus body part	text	Left arm	368208006 Left upper arm structure	
Answer to intervention sent - Symptom Rash	boolean	FALSE	271807003 Eruption of skin 2667000 Absent	Observation https://drive.google.com/file/d/1JGm3_GU2mP9pRt3zLUsxZIQJvlSH6DLE/view?usp=sharing
Intervention sent to patient - Pharmacological		"You reported pruritus.If you haven't since the symptom has started, you can use OTC topical emollient (moisturizer) without perfume. Did you use it or are you going to use it?"	225129006 | Application of emollient to skin (procedure)	Communication + MedicationRequest https://drive.google.com/file/d/1XPjNudoA8BzSlmGbr_gGiwjcdS2ZVmtE/view?usp=sharing and https://drive.google.com/file/d/1fgpx9Jawxn4do0pmQfhzdxFKImj0_n_e/view?usp=sharing
Answer to intervention sent - Pharmacological	boolean	TRUE	385645004 Accepted 443390004 Refused	Observation https://drive.google.com/file/d/1OMo36_B9YowLJ8xuijWvebn-364yeXyx/view?usp=sharing
Predictive model result	vector		Observation, similar to the approach used for ECG, described at https://www.hl7.org/fhir/observation-example-sample-data.html 225953001 Evaluating response to treatment	Observation with attribute "valueString" containing the vector of graph points, i.e. (x,y) pairs. Nivolumab: https://drive.google.com/file/d/1156l1FyoFPWZBJU00rj0C8HOyCu_ucY4/view?usp=sharing Nivolumab+Ipilimumab: https://drive.google.com/file/d/1mwLaeTzNfPiLuEJnTGHmdH3Ep5OBMCff/view?usp=sharing
Features score	vector		LP148736-4 | Explanations (LOINC)	2 Observations: https://drive.google.com/file/d/1vPDMdCLxuPwVXkRtnGJIVYNT1VLyMRSe/view?usp=sharing and https://drive.google.com/file/d/17TRwNgeLk298R2XUK_OgloB7kOwXZAcC/view?usp=sharing
Age	int	74 years	397669002 | Age (qualifier value) a | year (UCUM)	Observation https://drive.google.com/file/d/1MoHpsFDxFYjuKBoR1BvzpdP03XxL3V1f/view?usp=sharing
Sex	text	Female		Patient.gender https://drive.google.com/file/d/1YI3TDrcomwX7t8skojzJmCyK8XDt4iWg/view?usp=sharing

Location primary tumor	text	Head and neck	399687005 | Primary tumor site (observable entity) Values: 81745001 | Eye structure 774007 | Structure of head and/or neck 22943007 | Trunk structure 2592007 | All extremities 255414000 | Acral 313268005 | Mucosal 54690008 | Unknown (origin)	Observation https://drive.google.com/file/d/1U7_Cd9wriPzg_IKNgLC5tjJGAz9YDmZ3/view?usp=sharing
Type of melanoma	text	Superficial spreading	372156000 | Malignant melanoma - category (morphologic abnormality) Values: 1 = 55320002 | Superficial spreading melanoma (morphologic abnormality) 2 = 2142002 | Nodular melanoma (morphologic abnormality 3 = 16974005 | Acral lentiginous melanoma, malignant (morphologic abnormality 4 = 44474009 | Malignant melanoma in Hutchinson's melanotic freckle (morphologic abnormality) 5 = 51757004 | Desmoplastic melanoma, malignant (morphologic abnormality) 7 = 2092003 | Malignant melanoma, no International Classification of Diseases for Oncology subtype (morphologic abnormality) 9 = 424190005 | Malignant melanoma of unknown origin (disorder)	Observation https://drive.google.com/file/d/1qTes6s4P28CJJrI5bFhJFAm_MDlVt-pP/view?usp=sharing
Breslow thickness in millimeters	float	1.02 mm	385348009 | Breslow depth finding for melanoma (finding) mm | millimeter (UCUM)	Observation https://drive.google.com/file/d/1lRQU-oPXx9ctVH8z-V2vRfvGU6eBuEB4/view?usp=sharing
Ulceration	boolean	FALSE	263913002 | Ulceration (qualifier value)	Observation https://drive.google.com/file/d/1GYPljr7c8zifzACoTQdtrNlZkYqP5Jdp/view?usp=sharing
Dermal mitoses	text (none / 1 or more / unknown)	1 or more mitotic figure per mm2	~~LP247956-8 | Mitotic rate (LOINC)~~ 371472000 | Mitotic count score (observable entity) Values: 260413007 | None (qualifier value) 396446002 | Less than 1 mitotic figure per mm2 (finding) 396447006 | 1 or more mitotic figure per mm2	Observation https://drive.google.com/file/d/1b6_Y2rRXnKuscEzY2vY3Ukb5W6IFaIzG/view?usp=sharing

			(finding) 261665006 | Unknown (qualifier value)	
Satellitosis / in-transits metastases	text (satellitosis / in transit / none / both)	Satellitosis	396388007 | pN2c: Satellite or in-transit metastasis without nodal metastasis (melanoma of the skin) (finding) Values: 52214007 | Satellitosis (morphologic abnormality) 398101002 | In transit (attribute) 2667000 | Absent 52101004 | Present *-> both of them*	Observation (Condition if it will be implemented) https://drive.google.com/file/d/18FZs1jkfEsSuFsXK8Z9AFjAyLvSFx0-0/view?usp=sharing
Lymph node metastases (macroscopic)	boolean	FALSE	94392001 | Secondary malignant neoplasm of lymph node (disorder) Values: 2667000 | Absent 52101004 | Present	Observation (Condition if it will be implemented) https://drive.google.com/file/d/1IBhF1KKQYsm4IPlujTxzxcamignJl3cW/view?usp=sharing
Distance metastases	booelan	TRUE	399409002 | Distant metastasis present (finding)	Observation (Condition if it will be implemented) https://drive.google.com/file/d/1ZQEPNoU17CpqsDFDERA8KlRwn9BN4OcP/view?usp=sharing
WHO score	set of grades (0/1/2/3/4)	ECOG grade 0	103319000 | Patient performance (contextual qualifier) (qualifier value) Values: 425389002 | ECOG performance status - grade 0 (finding) 422512005 | ECOG performance status - grade 1 (finding) 422894000 | ECOG performance status - grade 2 (finding) 423053003 | ECOG performance status - grade 3 (finding) 423237006 | ECOG performance status - grade 4 (finding) 423409001 | ECOG performance status - grade 5 (finding)	Observation https://drive.google.com/file/d/1Wu2IhiZopxkXHGQvysEuwP80zj-xtxus/view?usp=sharing
Comorbidity?	boolean	FALSE	398192003 | Co-morbid conditions (finding)	Observation https://drive.google.com/file/d/12_Ckz3kUMNy6iNX_7jrZkhKtRLKgkNlH/view?usp=sharing
Medication use	boolean	FALSE	8677-7 | History of Medication use (LOINC)	Observation https://drive.google.com/file/d/112p8eyZch4-aQVyAPBXFEbHqBCfKeG99/view?usp=sharing

Lab determination: S100	float	0.06 µg/L	415357005 | S-100 neuronal cell marker measurement ug/L microgram per liter (UCUM)	Observation https://drive.google.com/file/d/1F8dMuV3QWqZlChu9Ye3pJBfF-wD0VXlN/view?usp=sharing
Increased S100	binary (increased / normal)	Normal	415357005 | S-100 neuronal cell marker measurement (procedure) Values: 35105006 | Increased (qualifier value) 260395002 | Normal range (qualifier value)	Observation https://drive.google.com/file/d/1NOtMpVvPT62qPDgbso4dBGXiFWvvYCVT/view?usp=sharing
Total number of metastases	binary (none / multiple)	Multiple	252105006 | Number of metastases (observable entity) Values: 255204007 | Multiple (qualifier value) 260413007 | None (qualifier value) 261665006 | Unknown (qualifier value)	Observation https://drive.google.com/file/d/1-JHlZJ3cFEmvGC3Vy05x_kw_hNAUaksV/view?usp=sharing
NRAS mutation?	boolean	FALSE	21719-0 | NRAS gene mutations found [Identifier] in Blood or Tissue by Molecular genetics method Nominal (LOINC)	Observation https://drive.google.com/file/d/1q_r2eYomaj8Ep5DhPXFz9QiG6tViTYKg/view?usp=sharing
Sequence (Treatment Line)	text (active surveillance / 1st line treatment / 2nd line treatment / 3rd line treatment / no treatment with curative intentions)	First line treatment	182991002 | Treatment given Values: 424313000 | Active surveillance (regime/therapy) 708255002 | First line treatment (procedure) 708256001 | Second line treatment (procedure) 708257005 | Third line treatment (procedure) 260413007 | None (qualifier value)	Observation https://drive.google.com/file/d/1IpoyExP_Fa47IMkwsEWQ61dDyGAThRP2/view?usp=sharing
Intervention sent - Pruritus	text	"You reported mild or localized symptom pruritus. Do you still have the symptom today?"	279333002 Pruritus of skin	Communication https://drive.google.com/file/d/1roE7_oZ7pEokSI_CnHauIXjqKB3oK06H/view?usp=sharing
Invervention response		Yes, grade 1	279333002 Pruritus of skin 446411000124101 Common terminology criteria for adverse events grade 1	Observation https://drive.google.com/file/d/12XM4T62UmBtjV40oJovjOT3ItZwqKlL1/view?usp=sharing
Rash		TRUE	271807003 | Eruption of skin (disorder)	Observation https://drive.google.com/file/d/1EtCHOQyMP

			Values: 2667000 | Absent 52101004 | Present	57jFrSsO0uCwhxzXEwW6sz_/view?usp=sharing
Rash grading	text	"Presence of symptoms (pruritus, tightness or burning skin), limitation of instrumental activities of daily living"		
Symptom note	text	"I have macule in 9% of my body"		
Interaction found	text	"An interaction was found between recommendations [30x30 nature walk and avoid sun exposure]"	410342001 | Interaction case management 272385004 | Outdoor pursuit 418521000 | Avoid exposure of skin to direct sunlight or sun lamps	List https://drive.google.com/file/d/1maiFBiMjAkAJVy1JPnG1m51rFs-9_GKG/view?usp=sharing + 2 Goals : https://drive.google.com/file/d/1O0x9cM-UICxRERWlDFnjpOIDaYRe3Cib/view?usp=sharing and https://drive.google.com/file/d/1jxl6mMw67TUXhIrOQWnIjOVRxqzf8A4s/view?usp=sharing
Intervention sent to patient from VC - GL + Capsule	text	"Rash symptom requires a medical evaluation by the clinician, so please contact your doctor, if you haven't already.You can send an email to your clinician with a picture of your symptom.In the meanwhile avoid skin irritants and sun exposure. It is recommended to use OTC topical emollient (moisturizer) without perfume. You are performing 30x30 Nature	225129006 | Application of emollient to skin (procedure)	Communication + MedicationRequest https://drive.google.com/file/d/1mkiSZYcPuBeQ21lMMLZpx8wW7X2jTGBH/view?usp=sharing and https://drive.google.com/file/d/1RrAU5PDxYDQZYnZq1yhCnlyTvi48NPkj/view?usp=sharing

		Challenge capsule, that include outdoor activities. Consider suspending it, if it's a sunny day, because you should avoid exposing your rash to the sun."		
Ongoing symptom	boolean/ number of days (as valueQuantity with UCUM days) + link to the actual symptom through the attribute Observation.focus	TRUE for pruritus (since 3 days)	303350001 | Ongoing episode (qualifier value)	Observation https://drive.google.com/file/d/1Hdf4mBkTflRs1B3qUyaelSroSo301P9i/view?usp=sharing
Clinician recommendation	text	Please contact the patient		Communciation https://drive.google.com/file/d/1MjG6UObqxBwGOqBhD6n0nLJ9KzY9Wult/view?usp=sharing
New visit		Set new visit on 2021-07-21 12:00 p.m.	308335008 | Patient encounter procedure 5880005 | Physical examination 416151008 | Scheduled - procedure status	Appointment https://drive.google.com/file/d/1JAH5Z1jjCvOT2lHiN8j4VOdgftvUNZ6b/view?usp=sharing
Reminder - New visit	text	Tomorrow (July 21st) you have a visit to doctor at 12:00		Communication https://drive.google.com/file/d/11Tsl4RVbwfoiJQBhKBCAPLsoDoG12L6w/view?usp=sharing
Symptom grade by clinician - Rash	set of grades (0/1/2/3/4)	CTCAE grade 2	271807003 | Eruption of skin 446421000124109 | Common terminology criteria for adverse events grade 2	Observation https://drive.google.com/file/d/1cKhM0BRSB8i04XBu2HAE4KXdgMKlZIcF/view?usp=sharing
Clinical recommendation for physician - GL	text	Treatment includes topical emollients, oral antihistamines and high strength		Communication https://drive.google.com/file/d/1i7G2jFatAqG40mqwjOR4R2s0_05-aTXz/view?usp=sharing

		topical steroids [II, B]. Systemic corticosteroids 0.5–1 mg/kg can be considered, depending on the severity of the symptoms. In the case of grade 2 skin AEs, treatment with checkpoint inhibitors can be continued but should be checked weekly for improvement. If not resolved, treatment should be interrupted until the skin AE has reverted to grade 1.		
New treatment	text	Oral antihistamines: loratadine 10 mg tablet, once daily	311372 | loratadine 10 MG Oral Tablet (RxNorm) 738956005 | Oral	MedicationRequest https://drive.google.com/file/d/1jp3enzwU9ivb8ZxsnTXegVu192kLkSLT/view?usp=sharing
Intervention sent - Pruritus	text	"You reported mild or localized symptom pruritus. Do you still have the symptom today?"	279333002 | Pruritus of skin	Communication https://drive.google.com/file/d/1roE7_oZ7pEokSI_CnHauIXjqKB3oK06H/view?usp=sharing
Invervention response		FALSE	279333002 | Pruritus of skin 2667000 | Absent	Observation https://drive.google.com/file/d/1jjVV2BTHFSHcI-bKMIV446omX6dkO9N0/view?usp=sharing
Intervention sent - Rash	text	"You reported symptom rash. Do you still have the symptom today?"	271807003 | Eruption of skin	Communication https://drive.google.com/file/d/1CX3znQdScL5xAcatarCU4EfUY_NdmRRi/view?usp=sharing
Invervention response		FALSE	271807003 | Eruption of skin 2667000 | Absent	Observation https://drive.google.com/file/d/1JGm3_GU2mP9pRt3zLUsxZIQJvlSH6DLE/view?usp=sharing

Intervention sent - Forum	text	It's good that your symptom is solved. If you have any suggestion for other patients to solve the symptom, write it on the AIMAC forum (https://forumtumore.aimac.it).		Communication https://drive.google.com/file/d/1s3zotGn6m7_gx5WDWJkzUbrPJF0zOYwR/view?usp=sharing