Co-Funded by the Horizon 2020 programme of the European Union

Grant Agreement: 875358

Deliverable D3.2

FAITH Common Data Model

Work package:	WP3 - Hospital Infrastructures, Visualisation & Distributed Ledger Technology (DLT)
Prepared By/Enquiries To:	Diego Carvajal – UPM Maria Eugenia (Xenia) Beltrán – UPM
Reviewers:	Fernando Ferreira – UNINOVA Tom Flynn - TFC
Status:	Final
Date:	18/06/21
Version:	1.0
Classification:	Public

Authorised by:

Philip O'Brien
WIT

Authorised date: 18/06/21

Disclaimer:

This document reflects only authors' views. Every effort is made to ensure that all statements and information contained herein are accurate. However, the Partners accept no liability for any error or omission in the same. EC is not liable for any use that may be done of the information contained therein.

FAITH Project Profile

Contract No H2020-ICT- 875358

Acronym	FAITH
Title	a Federated Artificial Intelligence solution for moniToring mental Health status after cancer treatment
URL	https://h2020-faith.eu/
Twitter	https://twitter.com/H2020_Faith
LinkedIn	linkedin.com/company/faith-project
Facebook	https://fb.me/H2020.FAITH
Start Date	01/01/2020
Duration	36 months

FAITH Partners

List of participants

Participant No	Participant organisation name	Short Name	Country
1 (Coordinator)	WATERFORD INSTITUTE OF TECHNOLOGY.	WIT	Ireland
2	UPMC Whitfield, Euro Care Healthcare Limited.	UPMC	Ireland
3	Universidad Politécnica de Madrid.	UPM	Spain
4	Servicio Madrileño de Salud.	SERMAS	Spain
5	UNINOVA, Instituto de Desenvolvimento de Novas Tecnologias.	UNINOVA	Portugal
6	Fundação D. Anna de Sommer Champalimaud e Dr. Carlos Montez Champalimaud.	CF	Portugal
7	Deep Blue.	DBL	Italy
8	Suite5 Data Intelligence Solutions Limited.	SUITE5	Cyprus
9	TFC Research and Innovation Limited.	TFC	Ireland

SC1-DTH-01-2019: Big data and Artificial Intelligence for monitoring health status and quality of life after the cancer treatment

H2020-SC1-DTH-2019

FAITH is co-funded by the European Commission - Agreement Number 875358 (H2020 Programme)

Document Control

This deliverable is the responsibility of the Work Package Leader. It is subject to internal review and formal authorisation procedures in line with ISO 9001 international quality management system procedures.

Version	Date	Author(s)	Change Details
0.1	21/01/21	Diego Carvajal (UPM)	Table of Contents defined
0.2	25/03/21	M.E. Beltrán, Samanta Villanueva, Diego Carvajal (UPM)	Table of Contents updated
0.3	10/04/21	Diego Carvajal (UPM)	First version added section 3 and 5
0.4	16/04/21	Fernando Ferreira (UNINOVA)	Updated section 3
0.5	26/04/21	Samanta Villanueva (UPM)	Added section 4
0.6	27/04/21	M.E. Beltrán, Samanta Villanueva, Diego Carvajal (UPM)	Updated section 4 and added section 1
0.7	28/04/21 – 28-05/2021	Diego Carvajal (UPM)	Added annex with data variables discussed as preliminary selected ones as for the FAITH Protocol
0.8	30/05/21	Diego Carvajal, M.E. Beltrán (UPM)	Final revisions and integrations
0.9	14/06/21	Fernando Ferreira (Uninova)	Internally discussed and reviewed.
0.9.1	15/06/21	Tom Flynn (TFC)	QA'd version.
1.0	18/06/21	Philip O`Brien (WIT)	Final release for submission to European Commission portal.

Executive Summary

Objectives:

The main objective of this report is to describe the data model that structures the data collected from the observational trial Phase (Alpha Version Data Model), which collects data from the trial (as defined in the project protocol) to support the creation of the FAITH data lake (aggregated data set) to create the federated model. The data lake component in the project stores and models data structures used for annotating raw and aggregated data of diverse types including sensor data, app data and hospital data and how they are used by services in the FAITH framework supporting the observational trial and the prospective data processes for depression, of the initial model to be federated. This deliverable links with D3.1 Network Infrastructures & Integration Methodology which describe the overall infrastructure of the hospitals providing the clinical Trials; D2.4 (2.3a) Conceptual architecture and supported by D3.3 Data Privacy & Protection.

This is the first report from a two series report which aims at providing a first version of conceptual data services model, which will establish a data harmonization methodology to harmonize the diverse data types (hospital data, sensor data, app data) solving the interoperability aspects between the different sources so that analysis can be performed..

The second report will comprise the data model for the Beta Version of the FAITH infrastructure which comprise the Federated Machine learning modelling, under a decentralized data approach which collaboratively learns on the local repositories to provide an optimized shared prediction of depression.

Results:

The primary results of this deliverable, and particularly this iteration (M16) is the preliminary definition of a Common Data Model for the scope of the Alpha version of the project infrastructure and define an implementation strategy based on an Industrial Standard Data Model (ISDM) approach. This deliverable comprises the work and activities undertaken as part of T3.2 to M16.

TABLE OF CONTENT

TABLE OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Data modelling is the way of structuring and organizing data so that it can be easily used by the different entities that need to exploit it. The purpose of the data model is to allow a long-term effort to be coordinated between multiple groups of developers. The Data Model (DM) consists of definitions, relationships, keys, data groupings, attributes, "type of" relationships, multiple-occurrence relationships, and foreign keys. Due to the similarity of businesses between different organizations, there are generic data models [1].

1.1 FAITH project phases, data infrastructure and data modelling

The FAITH Project is divided in three phases with different levels of infrastructure development supporting the project progress:

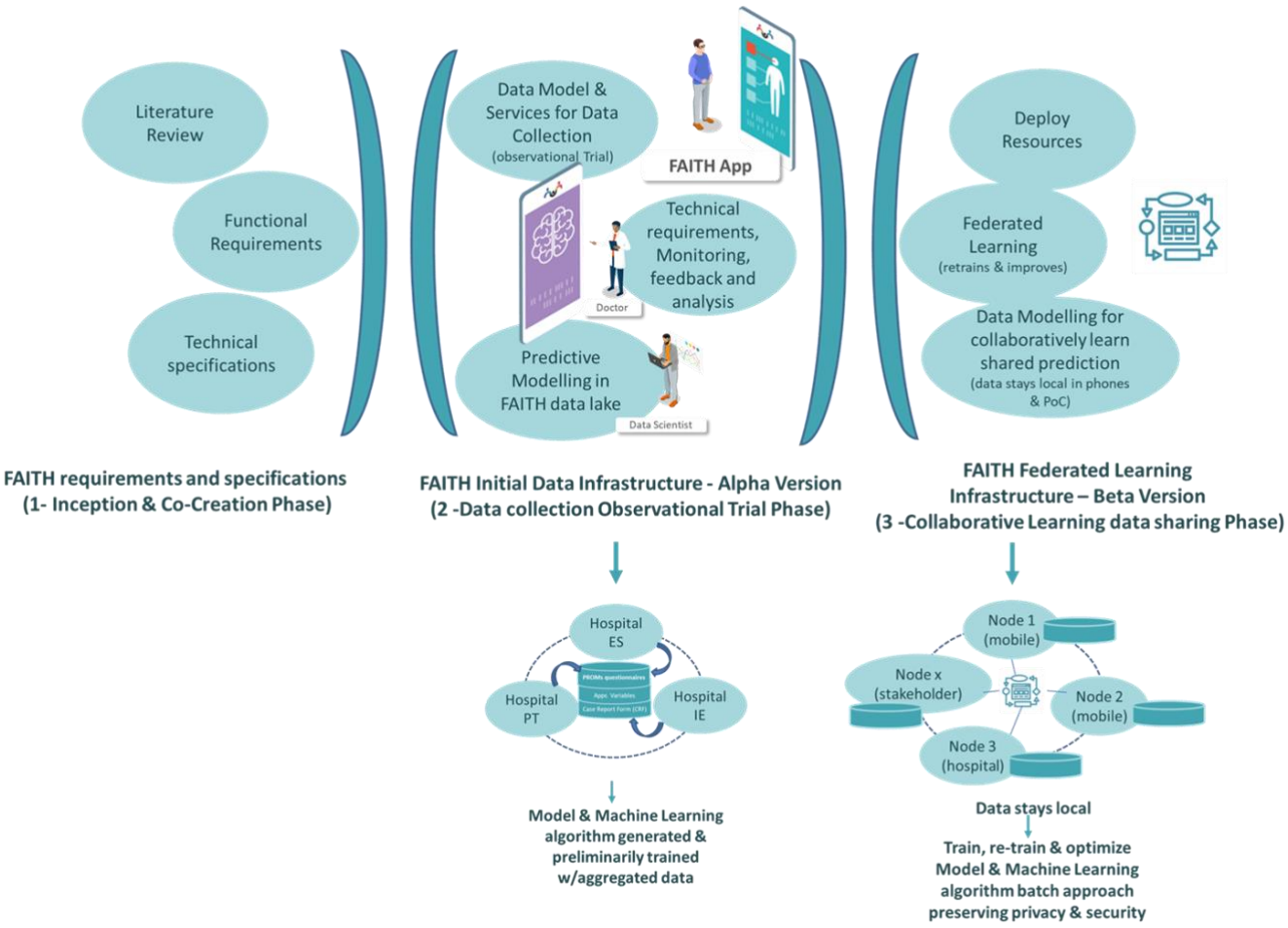

Figure 1 FAITH Project phases and data approaches

1) **Inception & Co-creation phase**, which comprised literature review and elicitation of functional and technical specifications;

2) **The data collection for the observational trial Phase**, which enables the needed initial infrastructure **(Alpha Version)** to gather and collect data from the scientific activities (observational prospective trial aiming at predicting depression in post-cancer survivors; as defined in the project protocol and D6.1). This phase`s output is the creation of a common algorithm or model trained or generated on the aggregated data set generated by the observational trial.

3) **The Collaborative Learning data sharing Phase**, which develops the Federated Learning Infrastructure (**Beta Version**) in order to retrain and improve a private and secure machine learning framework (model) that learns over distributed data, in each local node repository (e.g. participants mobiles). This phase`s output is the federated trained algorithm which is re-trained and improved (batch-wise) over the network of nodes (data stays local and it is not aggregated).

1.2 Data model in D3.2a (Faith Common Data Model)

This deliverable describes the data model that structures the **data collected from the observational trial Phase (Alpha Version Data Model, middle phase in figure 1)**, which collects data from the trial (as defined in the project protocol and D6.1) to support the creation of the FAITH data lake (aggregated data set). The data lake component in the project stores and models the information managed in the observational trial and supports the prospective data processes for depression. It gathers data from to get data from user actions (forms, detected activities, etc.). Within FAITH, the Alpha Version Common Data Model goes hand in hand with the observational trial protocol and the needs that arise.

This document presents the path followed to define the data model, which comprised:

a) review and collection of data sources defined so far, with emphasis on clinical data, Patient Reported Outcomes (PROMs) from questionnaires and from sensors and the FAITH application.

b) Review of Common Data Models (CDM) frameworks and standards developed for clinical research data options to define the data model; while complying with the necessary data privacy requirements (e.g. standards, ethical, data privacy –pseudonymization - and GDPR, as described in D3.3).

c) Implementation of the FAITH data model and services that shall be available to clinicians and trial administrators, where they can interact, provide requests and issue recommendations. In this way, the use of confidential data is avoided and the problems of possible non-compliance with the protection of personal data are reduced since such information is not collected from hospitals.

To extend interoperability among the FAITH`s inter-connected components, F.I.H.R has been selected as the interoperability standard to integrate different resources that have already been developed and to guarantee that the developments that are carried out are not isolated products that become obsolete because they cannot interact or integrate with other systems that are developed and can contribute to create something with greater potential.

2 ABBREVIATIONS AND ACRONYMS

Abbreviation	Description
API	Application Programming Interface
CDM	Common Data Model
DLM	Data Lifecycle Model
DMP	Data Management Plan
EORTC QLQ-C30	European Organization for Research and Treatment of Cancer Core Quality of Life Questionnaire for cancer patients
EORTC QLQ-BR45	European Organization for Research and Treatment of Cancer Quality of Life Questionnaire for breast cancer patients
EORTC QLC-LC29	European Organization for Research and Treatment of Cancer Quality of Life Questionnaire for lung cancer patients
FAIR	Findable, Accessible, Interoperable, Reusable
FFQ	Food Frequency Questionnaire
FHIR	Fast Interoperable Healthcare Resources
HADS	Hospital Anxiety and Depression Scale
HL7	Health Level 7
HMO	Health Maintenance Organization
HMORN	Health Maintenance Organization Research Network
HTML	Hyper Text Markup Language
IG	Implementation Guideline
ISO9001-2015	International Quality Management Systems.
I2b2	Informatics for Integrating Biology and the Bedside
NLP	Natural Language Processing
OMOP	Observational Medical Outcomes Partnership
PG-SGA	Scored Patient-Generated Subjective Global Assessment
SEDIA	Single Electronic Data Interchange Area
UML	Unified Modelling Language
URL	Uniform Resource Locator
WP	Work Package
WPL	Work Package Leader

3 FAITH DATA SOURCES AND DATA MANAGEMENT INFRASTRUCTURE

The FAITH observational trial prospective study uses pseudo-anonymized data. Pseudonymization techniques are expected to take place when users' data are collected through the Hospital or Point of Care (PoC) tool. The FAITH prospective trial study, deals with several types of data, transmitted and/or collected from mainly three data sources, where all data sources or containers generate and process the Prospective Datasets for the patients enrolled in the observational trial.

The data sources can be characterized as follows:

Source of origin	Data Categories managed
Sensors, Mobile App & Smartphone	Depending on the type of information, data gathering can be done in two ways: a) Passive. Information is collected in the background, without direct user interaction e.g. through sensors or the app; b) Active. The information is collected in an interactive manner, i.e., user interaction is necessary (e.g. PROMs) The sensors provide data categories related with sleep (variables related with sleep cycle and monitoring. The smartphone provides data categories related with the Physical activity (Smartphone sensors track user activity), daily routines, smartphone usage, or connection (e.g., Wi-Fi connection). Also, the App provides reminders, alerts and supports the prospective study activities flow (scheduled activities, e.g., visits or questionaries), as well as variables for outlook (e.g. voice)
Hospitals (within Data collection functionality)	This relates with data collected at hospitals as Point of Care (PoC) which during visits gather basal information and the Case Report Form (CRF) data, which comprise Clinical-pathological data, treatment data, follow-up data, activity, behavioral data, and information useful for patient follow-up management such as phone usage. Also, at the PoC data such as alerts or visits for the trial pathway is originated. At the hospitals the web tool will generate Inferred risk scores, other inferred information to follow-up patients, extract behavioural markers, process and generate scoring of PROMs questionaries or depression predictors.

Patients Reported Outcomes (PROMs)	Data from questionnaires (such as the ones filled each three months: Hospital Anxiety and Depression Scale (HADS) questionnaire, Scored Patient-Generated Subjective Global Assessment (PG-SGA), Quality of Life (QoL- QoL data is gathered through two questionnaires every 3 months EORTC QLQ-C30, EORTC QLQ-BR45 (Breast Cancer) and EORTC QLQ-C30, EORTC QLQ-LC29 (Lung cancer) or the yearly nutrition/appetite (Food Frequency Questionnaire)

All data sources address privacy, ethical and security, supported by an architecture provided in D3.3 (Data Privacy Protection), which also describes the data access policies and measures to be implemented in the project in order to deliver a best-of-class, secure solution, in full alignment regulations.

3.1 DATA MANAGEMETN INFRASTRUCTURE

The following graph (figure 2) illustrates the data management infrastructure being established in FAITH project, which supports the collection of data at three data sources defined above.

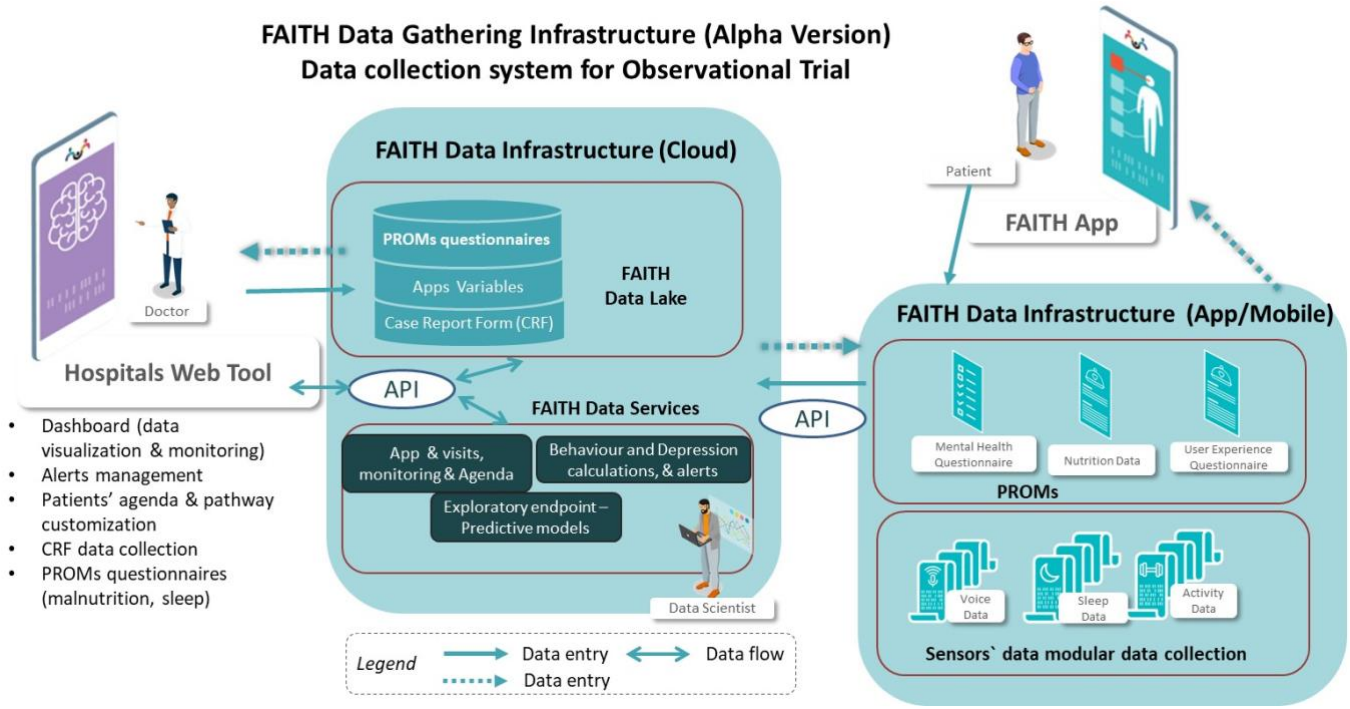

Figure 2 FAITH Data collection and Infrastructure

On the left side of figure, represents the data collected during the non-interventional/observational trial at the hospital or PoC (basal data and/or visits' data). Data is visualized or represented through a dashboard, so the nurses and doctors can see and monitor data from patients, see the completion of PROMs or set the alarms, notifications, visits, etc.

On the right side of the picture (the users' side) shows the data collection through the sensors, app or smartphone; including the data coming from PROMS (such as questionnaire gathered data) and app.

The FAITH data cloud infrastructure (platform) enables to implement the data flow through different capabilities, such as relational databases, NoSQL databases, files stored in the filesystem, etc. This storage is defined using a set of common databases for the prospective data and/or information from all the applications (patients mobile app or PoC web tool), and based in all variables defined in the protocol (e.g. see Annex 2 -preliminary list of variables-. This model can be accessed through different APIs defined by each application (PoC and mobile), depending on its needs.

Each of the applications/web tool define its own processes sending information to the cloud to generate the initial predictive model. The non-relational storage is implemented through file system' storage and the APIs could be used by the rest of the applications to consume the generated information by them.

The relevant data containers that link from data infrastructure (previous figure) and support the high-level description of information flow for the FAITH prospective study are represented in the following figure (figure 3).

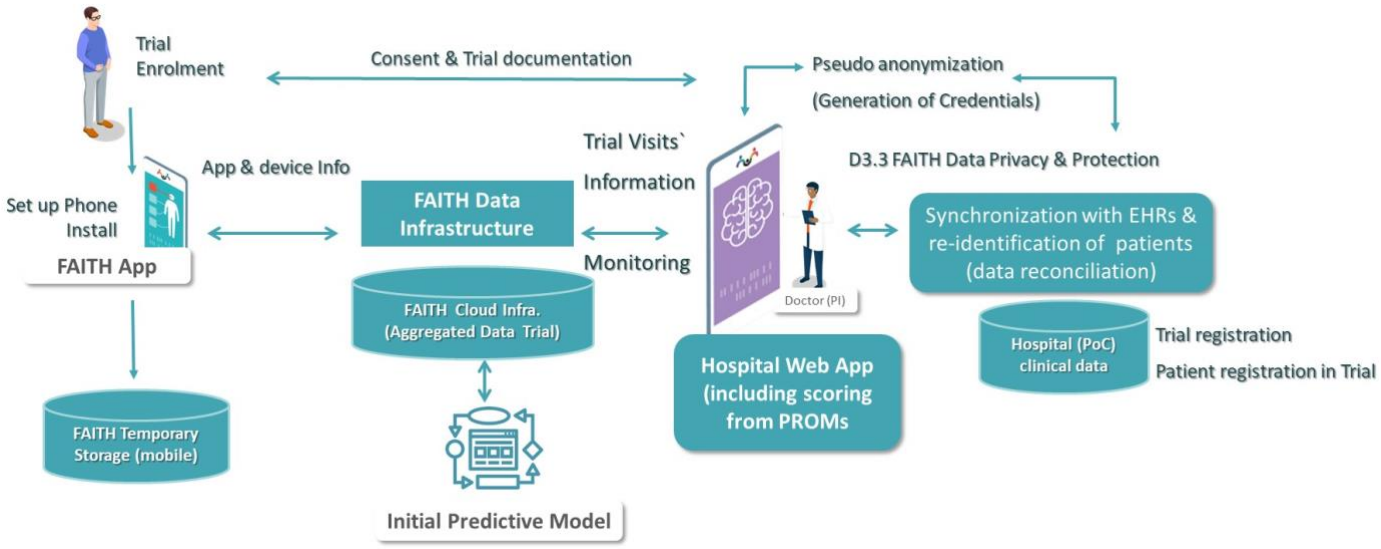

Figure 3 FAITH High Level of information flow

The high-level description of information flow for the FAITH prospective study, is as follows: the patient enrols and reads the study informative sheet, discusses with the doctor (Principal Investigator -PI) the

study protocol and its objective, methods, tools, risks and benefits of the participation to the study. The patient signs the consent form accepting to take part in the observational trial study, he/she is registered in the trial observational study. Then patient data is pseudorandomized and credentials are generated. The patient comes for the first visit and receives the project documentation, provides needed personal, clinical and psyche-social data required as part of the first visit and basal data for the study. The patient is helped by the PI or the nurse /study IT support to configure and set up the smartphone, downloading the FAITH Application and walked through the app and questionnaires. Three main containers are underpin the data flow:

- The Relevant Data Containers are managed inside the information domain of each partner hospital where the connection of the data from patient and trial participant are connected.
- A temporary set of personal data is stored in the mobile application on the smartphone of the study participant; this data, through the cloud API will be accessible by the hospital application.
- The FAITH cloud infrastructure host a set of common databases collecting in a data lake the aggregated pseudorandomized data the patients` visits (including basal data), the information coming from the sensors, applications/smartphone or PROMs. The FAITH cloud infrastructure host and integrates data/information collected from all the three FAITH hospitals or PoCs. The API can be used by the rest of the applications to consume the generated information by them.

The corresponding main data processes are the following:

- Day-by-day and frequent data gathered by sensors and FAITH app regarding physical activities, sleep, nutrition, outlook and related information are stored in the temporary patient smartphone and sent to the FAITH Cloud.
- The Hospital tool integrates all available information in the cloud through the web tool, which provides feedback to the clinicians for follow-up and alerts management through an interactive dashboard.

4 COMMON DATA MODELS BACKGROUND

The Common Data Model (CDM) is an information model that provides uniform definitions for managed resources, business systems and processes, and other data, and the relationships between those elements. CDM is based on the Unified Modelling Language (UML) [2]. In research, CMDs are often used when there is a need to exchange or share a set of data for some particular use. These data models are responsible for defining the structure, format, and content of data to be grouped or shared, and have been used in clinical research since the early days of multi-center registries and data reporting, when minimal data sets were defined and exchanged [3].

In recent years, various CDM frameworks and standards have been developed for clinical research data. These include HL7-FHIR [4], the Observational Medical Outcomes Partnership (OMOP) CDM [5], the Health Maintenance Organization Research Network (HMORN) [6], and I2b2 [7].

4.1 HL7 - FHIR

Fast Healthcare Interoperability Resources (FHIR) is a next generation standards framework created by HL7 [8]. FHIR combines the best features of HL7's v2, HL7's v3 and Clinical Document Architecture product lines while leveraging the latest web standards and applying a tight focus on implementation [4].

FHIR is based on the concept of "resources", where a "resource" is the basic unit of interoperability, the smallest unit that makes sense to exchange. “Resources” are representations of concepts from the healthcare world: patient, doctor, health problem, etc [9].

The “Resources” have the following common characteristics:

- A set of core properties that most current systems support.
- An extension mechanism that allows implementers to easily add new properties.
- An identification through which it can be registered, located and recovered.
- A component that allows a readable view of the data stored in the resource.

Resources can be used in their simplest form or grouped together in the form of messages, documents or even services that use one or more resources.

Figure 4, shows an example where the important parts of a "resource" can be identified: a local extension, the human-readable HTML presentation, and the standard-defined data content.


```
<Patient xmlns="http://hl7.org/fhir">
  <id value="glossy"/>
  <meta>
    <lastUpdated value="2014-11-13T11:41:00+11:00"/>
  </meta>
  <text>
    <status value="generated"/>
    <div xmlns="http://www.w3.org/1999/xhtml">
      <p>Henry Levin the 7th</p>
      <p>MRN: 123456. Male, 24-Sept 1932</p>
    </div>
  </text>
  <extension url="http://example.org/StructureDefinition/trials">
    <valueCode value="renal"/>
  </extension>
  <identifier>
    <use value="usual"/>
    <type>
      <coding>
        <system value="http://hl7.org/fhir/v2/0203"/>
        <code value="MR"/>
      </coding>
    </type>
    <system value="http://www.goodhealth.org/identifiers/mrn"/>
    <value value="123456"/>
  </identifier>
  <active value="true"/>
  <name>
    <family value="Levin"/>
    <given value="Henry"/>
    <suffix value="The 7th"/>
  </name>
  <gender value="male"/>
  <birthDate value="1932-09-24"/>
  <careProvider>
    <reference value="Organization/2"/>
    <display value="Good Health Clinic"/>
  </careProvider>
</Patient>
```

Figure 4 Example Resource: Patient [29].

4.2 Observational Medical Outcomes Partnership Common Data Model

Designing medical studies is based in requirements that, beyond the medical criteria, require decisions about population from the geographical diversity for engagement to the different centres and local specificities. Common data models (CDM) are enablers for the simultaneous analysis of disparate and large data sources. CMDs enable data to be standardized, and facilitate data exchange, sharing, and storage, particularly when the data have been collected via distinct, heterogeneous systems. Moreover, CDMs provide tools for data quality assessment, integration into models, visualization, and analysis.

The Observational Medical Outcomes Partnership (OMOP) Common Data Model enables the capture of information (e.g., encounters, patients, providers, diagnoses, drugs, measurements and procedures) in the same way across different institutions. This is a foundational organization when it is aimed to converge to a federated model of studies where the study aims to benefit from different geometries in the studies to be carried as to cover different populations with its own specificities.

4.3 Health Maintenance Organization Research Network

Health Maintenance Organization Research Network (HMORN) are integrated care health maintenance organizations (HMOs) that provide electronic medical and financial databases, and longitudinal observation for health research. The aim for these networks is the support for research mostly in addressing issues such as the costs and effectiveness of prevention and treatment practices, organization of care, secular trends in diseases, and relative priorities on how to apportion scarce resources. The main benefits for such networks rely in several pillars, among those, the engagement of a stable population base. In integrated care systems, many cohorts exist naturally, and data on those cohorts are present in electronic form. This permits long-term cohort studies to be conducted retrospectively and at reasonable cost. The experiences of health plan members can identify secular trends, outcomes of system interventions, and risk factors and their interactions networks. Another supportive consequence of the design and configuration of such networks is the trust they motivate, leading to high voluntary rates. Health plan members are more likely to respond to an appeal for research volunteers when that appeal is from their own health plan. Recruitment and retention rates from integrated care system research cohorts are substantially higher than those from community recruitment and, therefore, afford high external validity. These cohorts become highly valuable for their representativeness. Large, often non-profit, integrated care systems are usually demographically representative of their geographic populations, enabling findings that are generalizable to other defined populations. Also, to consider is the diversity of these research cohorts' population. The members of the HMORN are diverse ethnically, culturally, and geographically, facilitating studies that address racial and other disparities and are generalizable to other populations.

4.4 Informatics for Integrating Biology and the Bedside (I2b2)

I2B2 is a scalable informatics framework that is designed to bridge clinical research data and the vast data banks arising from basic science research to better understand the genetic bases of complex diseases. i2b2 was designed primarily for cohort identification, allowing users to perform an enterprise-wide search on a de-identified repository to determine the existence of a set of patients meeting certain inclusion or exclusion criteria. A major aim of the i2b2 data informatics framework aims to create an efficient structure within which patients can be identified for clinical and translational research projects. The first version of code for i2b2 was released publicly in 2007 with support from an NIH-funded National Center for Biomedical Computing. i2b2 is aimed to create a cost-effective and efficient way to identify patients for many types of clinical and translational research but providing that while running the trial healthcare data will be handled in a de-identified model. As a good practice for i2b2 would be a cohort discovery tool, meaning only patient counts can be retrieved, as such, individual patient-level data should not be accessed through this tool.

5 MODELLING TOOLS & TECHNIQUES

For the definition of FAITH's CDM it is necessary to achieve a harmony between the data sources and the different types of data identified. For this reason, it is important to use adequate techniques and tools for data modelling.

To achieve a better contribution and compilation of information, it is necessary that there be a coordinated work and that it can be enriched in a collaborative way by the different specialists in charge of this definition.

5.1 Collaborative Data Modelling

The demand for collaborative tools and the power of today's web interfaces have made it possible to create collaborative modelling tools. In fact, most of these tools have much more drawing-oriented functionalities, but at least they offer the possibility of drawing and sharing online models, especially UML models (mainly class diagrams, sequence, use cases and state machines) and database schemas.

Below are some of the most relevant collaborative tools at this time, one of which will be used in FAITH's data modelling process.

5.1.1 Lucidchart

Lucidchart [10] is an HTML5-based modelling tool. It has support for UML and allows online collaboration in real time. This tool allows the import of Visio files so the work done in this software would not be lost. In addition to UML, it also includes templates to create Entity Relationship Diagrams, business processes, and network diagrams, among others.

In the Lucidchart blog this tool is defined as an essential visual productivity platform that helps anyone understand and share ideas, information, and processes with clarity. With this intuitive, cloud-based solution, anyone can learn to work visually and collaborate in real time while building flowcharts, mock-ups, UML diagrams, and more [11].

You can use this tool for free but with certain limited features, such as it is not possible to create more than three editable documents and only 100 professional templates are available in this version.

5.1.2 Diagrams.net

Diagrams.net [12] is an open-source web application that allows the creation of a wide variety of diagrams. This tool also has a desktop version available for Windows, Linux, and macOS. From the web page or the application, it is possible to create and edit a wide variety of diagrams such as: flow diagrams, entity-relationship diagrams, UML diagrams, organization charts, process diagrams, mind maps, business process models, among others. It also incorporates mathematical notations and layers that facilitate editing.

This is a security-first diagramming app for teams that provides diagramming functionality and the ability to save diagram data to a service of choice. There are many different integrations with other platforms and applications, including Atlassian Confluence Cloud, Google Documents, GitHub, Microsoft Word [13].

Some of the main features offered by this application are:

- Unlimited collaboration that allows work to be shared with all team members.
- Advanced features to make it possible to create professional diagrams using the libraries that the application has.
- The user chooses where he wants to store the data, in this way the application does not have direct access to it and guarantees privacy.

Being an open-source tool, it is possible to use all its functionalities for free and always available online or through its desktop applications.

5.1.3 Creately

Creately [14] is a visual workspace for team collaboration. With this tool it is possible to create diagrams, drawings, images, and text on an infinite canvas to collaborate online in real time. Supports over 50 types of flowchart diagrams, mind maps, org charts, UML and database layouts, network diagrams, Gantt charts, business process diagrams, and more. This application can also be used as a collaborative whiteboard for teams working remotely or at the same location [15].

This tool offers multiple features to simplify the online drawing and collaboration experience. It is a cloud-based application; therefore, all your files can be accessed from anywhere, from any device. Focused on engineering, Creately offers a variety of tools from flowcharts to UML diagrams with libraries of standard shapes to help development teams collaborate on simplifying these processes.

Users can access a basic plan for free in which it is possible to create 5 free public documents and work with up to 3 collaborators. If these features are not enough for the work to be done, it is necessary to pay for a plan that includes more features, thus making it a less accessible tool.

5.2 UML

The Unified Modelling Language (UML) refers to a set of standard symbols and diagram types used in data modelling, workflow visualization, and systems modelling. UML notation is the standard used in the fields of software development, IT infrastructure, business systems, and many others. Many languages, such as SysML, SoaML, and several architecture frameworks use and extend UML [16].

Figure 2 shows an example of a UML diagram representing Alice's interaction with her parents. There it is possible to observe the use of various standard elements that represent this event.

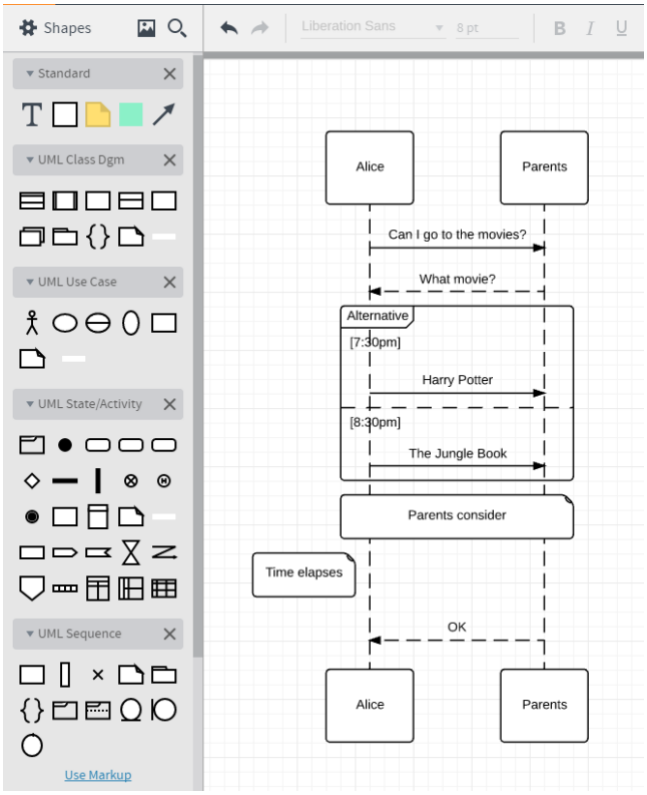

Figure 5 UML Diagram Example [30].

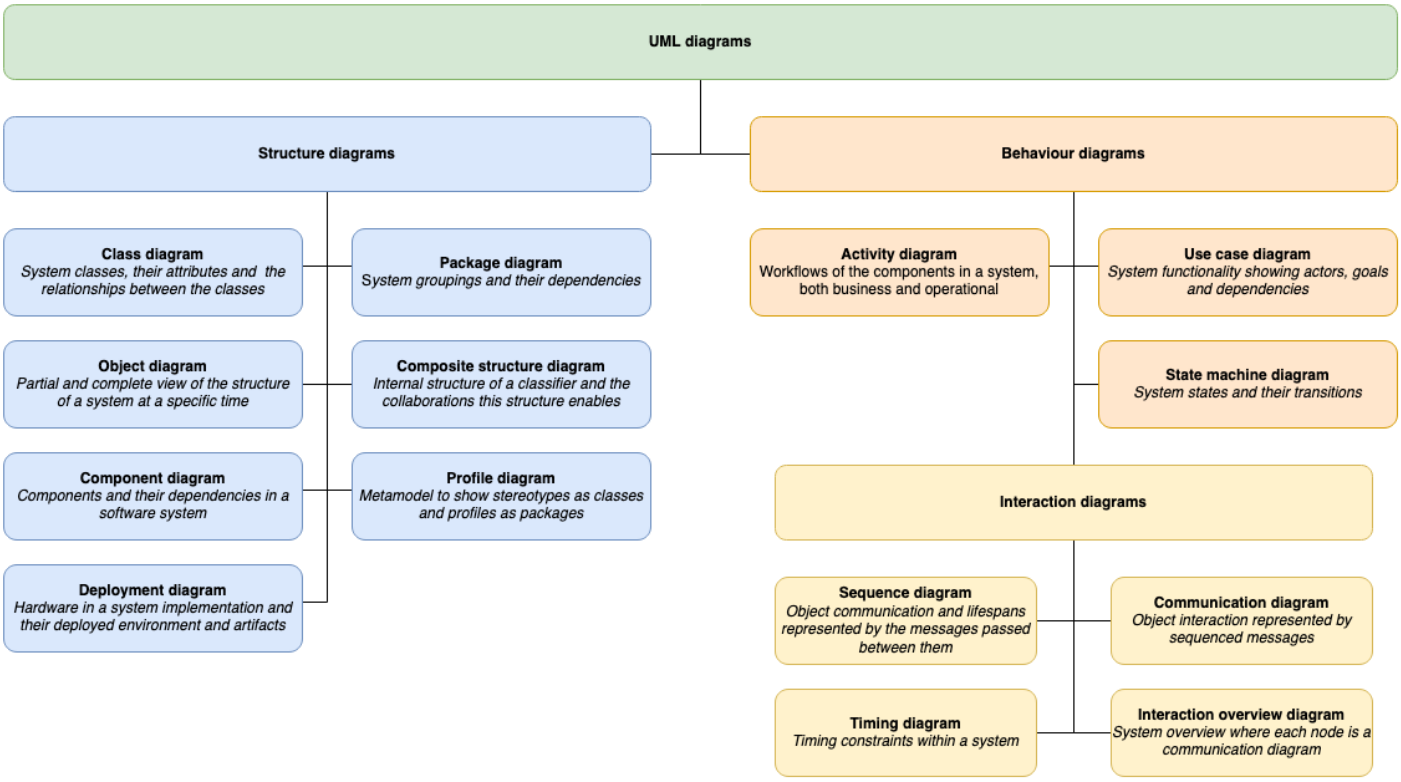

Figure 6 UML diagram types [16].

As can be seen in figure 5 and figure 6, there are two types of diagrams that provide static or dynamic views of a system.

The structure diagrams are:

- Class diagrams
- Component diagrams
- Composite structure diagrams
- Deployment diagrams
- Object diagrams
- Package diagrams
- Profile diagrams

The behaviour diagrams are:

- Activity diagrams
- Use case diagrams
- State machine diagrams

Within the behaviour diagrams are included the interaction diagrams which are:

- Sequence diagrams
- Communication (collaboration) diagrams
- Interaction overview diagrams
- Timing diagrams

In Annex 1 it is possible to observe some UML diagrams made from the variables identified in the FAITH protocol.

6 IMPLEMENTATION STRATEGY

From the definition of the FAITH protocol, at this stage of the project it has been possible to identify the necessary variables and the different types of data present in the Data Model. Among the variables identified are the demographic, clinical or compliance variables and the variables that will be collected by the clinical staff or from the mobile application. In Annex 2 it is possible to view the list of variables that have been mentioned above.

For it to be possible to implement the Common Data Model, it is necessary to carry out a definition and integration work that relates to the work carried out in other project tasks related to this task. With everything mentioned before and with the information described in the previous sections of this document, now it is necessary to talk about the definitions that have been made up to this point.

6.1 FHIR Resources in FAITH

As mentioned in section 3.1 of this document, the FHIR standard framework is based on the concept of "resources", this being the smallest unit that makes sense to exchange to generate interoperability. These resources are representations of concepts from the world of health care and have common characteristics that allow them to be used separately or grouped as appropriate.

Specifically, a FHIR “resource” is an entity that [17]:

- Has a known identity (a URL) by which it can be addressed.
- Identifies itself as one of the types of resource defined in this specification.
- Contains a set of structured data items as described by the definition of the resource type.
- Has an identified version that changes if the contents of the resource change.

All "resources" can have an identity, meta data, a base language and a reference to "Implicit Rules". Most "resources" can also contain text, contained resources, extensions, and data elements specific to the particular domain of the resource. There is a special type of resource called a Bundle for resource collections.

6.1.1 Resource Identity

A "resource" can be identified using a "Location" URL or through an inherent identifier. Using the "Location" URL will identify where it can be accessed (based on the "logical ID"). This location will change as it is copied / moved. Instead, the inherent identifier ("Business Identifier" or "Canonical URL") will be an element that is part of the resource and remains fixed as it is copied / moved.

6.1.2 Logical ID

Each "resource" has an id element that contains the "logical ID" of the resource assigned by the server responsible for storing it. Resources always have a known logical identification, except in some special

cases. The logical identification is unique within the space of all resources of the same type on the same server. Once assigned by the server, the ID is never changed.

6.1.3 Consistent Resource Identification

Business identifiers are commonly used to recognize the same content in different systems. Figure 7 shows an example that shows the same patient described in two different registration systems. In this case, if the identifier is the same, it is understood that the patient's resources mean that they describe the same patient.

For the reasons mentioned above, systems should ensure that identifiers are only assigned to resources when they uniquely identify the real-world entity that the resources match (for example, do not use account numbers as patient identifiers when multiple different patients share the same account number). It is also important that systems keep the identifiers as much as possible and do not throw them away. Even if identifiers are not useful to the system itself, they are likely useful to downstream consumers. And when a system serves resources, it should use consistent identifiers from the master persistent store whenever possible (for example, when creating a resource from a parent record, fill in the identifier and store the identifier somewhere for the same identifier next time).

```
GET http://a.company.example.com/Patient/23

<Patient xmlns="http://hl7.org/fhir">
  <id value="23"/>
  <identifier>
    <system value="http://a.particular.system/identifier"/>
    <value value="123456"/>
  </identifier>
</Patient>

GET http://other.company.example.com/fhir/Patient/5860200e-0ee3-42f5-8095-506e18dc9ca2

<Patient xmlns="http://hl7.org/fhir">
  <id value="5860200e-0ee3-42f5-8095-506e18dc9ca2"/>
  <identifier>
    <system value="http://a.particular.system/identifier"/>
    <value value="123456"/>
  </identifier>
</Patient>
```

Figure 7 Example of the same patient described in two different systems [17].

6.1.4 Implicit Rules

A custom agreement is referenced here to describe how the "resource" is being used, for example, an implementation guide that was followed when the "resource" was built, where the implementation guide should be known and understood to process the content safely.

The creation of these rules restricts the content to be understood by only a limited set of partners. This inherently limits the usefulness of data in the long term and should be avoided whenever possible.

However, the existing healthcare ecosystem is severely fractured and not yet ready to define, collect and exchange data in a generally interchangeable sense.

6.1.5 Language

Each resource can have a language element that specifies the base language of the content using a code defined in BCP 47 [18]. The language element is provided to support indexing and accessibility.

Although it can be inferred from the context of use, a default language has not been defined. Not all content in the "resource" must be in the specified language as it is possible to support multiple languages.

6.1.6 Resource Metadata

Within each "resource" is contained a "meta" element, of type "Meta", which defines a set of metadata that provides technical and workflow context to the "resource". In figure 8 it is possible to observe the metadata elements, they are all optional, although some or all may be necessary in particular implementations or contexts of use.

Metadata Item	Type	Usage
versionId (0..1)	id	Changes each time the content of the resource changes. Can be referenced in a resource reference. Can be used to ensure that updates are based on the latest version of the resource. The version can be globally unique, or scoped by the Logical Id of the resource. Version identifiers are generally either a serially incrementing id scoped by the logical id, or a uuid, though neither of these approaches is required. There is no fixed order for version ids - clients cannot assume that a versionId that comes after another one either numerically or alphabetically represents a later version. The same versionId can never be used for more than one version of the same resource. On the RESTful API: On receiving a write operation, the server SHALL update this item to the current value, or remove it. Note that servers SHOULD support versions, but some are unable to
lastUpdated (0..1)	instant	If populated, this value changes each time the content of the resource changes. It can be used by a system or a human to judge the currency of the resource content. Note that version aware updates do not depend on this element. Note that a timezone code extension may be present on Meta.lastUpdated. If present, the timezone code applies to the server copy of the resource, and not necessarily to other time related elements within the resource even if the timezone offsets are the same. On the RESTful API: On receiving a write operation, the server SHALL update this item to the current time on the server
source (0..1)	uri	A uri that identifies the source system of the resource. This provides a minimal amount of Provenance information that can be used to track or differentiate the source of information in the resource. The source may identify another FHIR server, document, message, database, etc. In the provenance resource, this corresponds to Provenance.entity.what[x]. The exact use of the source (and the implied Provenance.entity.role) is left to implementer discretion. Only one nominated source is allowed; for additional provenance details, a full Provenance resource should be used. On the RESTful API: On receiving a write operation, the server SHOULD generally leave this unchanged, unless applicable business rules, along with available provenance, dictate otherwise
profile (0..*)	canonical	An assertion that the content conforms to a resource profile (a StructureDefinition). See Extending and Restricting Resources for further discussion. Can be changed as profiles and value sets change or the system rechecks conformance. The profile can be used to indicate which version(s) of FHIR a resource conforms to. On the RESTful API: On receiving a write operation, the server MAY elect to remove invalid claims, SHOULD retain claims that are correct or untested, and MAY add additional claims it believes are valid
security (0..*)	Coding	Security labels applied to this resource. These tags connect resources in specific ways to the overall security policy and infrastructure. Security tags can be updated when the resource changes, or whenever the security sub-system chooses to. On the RESTful API: On receiving a write operation, the server SHOULD preserve the labels unless applicable business rules dictate otherwise
tag (0..*)	Coding	Tags applied to this resource. Tags are used to relate resources to process and workflow. Applications are not required to consider the tags when interpreting the meaning of a resource. On the RESTful API: On receiving a write operation, the server SHOULD preserve the labels unless applicable business rules dictate otherwise

Figure 8 Resource Metadata [17].

At this point in the development of the project, all the “resources” that will be used have not yet been defined. Annex 1 presents some examples that help to understand the relationship between the FAITH CDM and the FHIR "resources".

6.2 Implementation Guide Registry

An FHIR Implementation Guide (IG) is a set of rules for how FHIR resources are used to solve a particular problem, with associated documentation to support and clarify its use. An FHIR IG can include many different types of elements, such as FHIR logic models, FHIR API conformance facility, FHIR profiles, and many other FHIR and non-FHIR elements.

The implementation guides contain two different types of resource references [19]:

- Content: A set of logical statements that implementations must conform to. These are almost always compliance resources.
- Examples: Examples that illustrate the intent of the profiles defined in the implementation guide. These can be any type of resource.

Implementation Guides are published through the FHIR Package distribution system. Most implementation guides point to a single version, that is, they describe how to use a particular version, and all the profiles, value sets, and examples they contain are valid for that version. However, in other cases an implementation guide is not limited to a single version. It is common for the requirement to support multiple versions to emerge as the deployment matures and different deployment communities get stuck on different versions due to regulation or market dynamics.

In order to define a FHIR IG, it is necessary to have completely defined the Data Model and with it the FHIR "resources" that will be used in FAITH. The initial set of variables being defined in the protocol are defined in Annex 2 of this deliverable. As the definition of the protocol has not yet been closed, it is not possible to present a FHIR IG in this first document, having this presented in D3.3.

6.3 Data Lifecycle Model

In all Research projects a technical planning of activities and resources must be carried out. These processes require the incorporation of data management principles and best practices for the correct execution of the investigation. A Data Lifecycle Model (DLM) represents the stages of data management and describes how data is exchanged within a research project from start to finish. The DLM provides a high-level framework of individual actions, operations, or processes to be undertaken at different stages. The goal is to optimize data management, from efficient organization to elimination of any kind of waste, to provide meaningful and high-quality data according to user expectations and requirements [20].

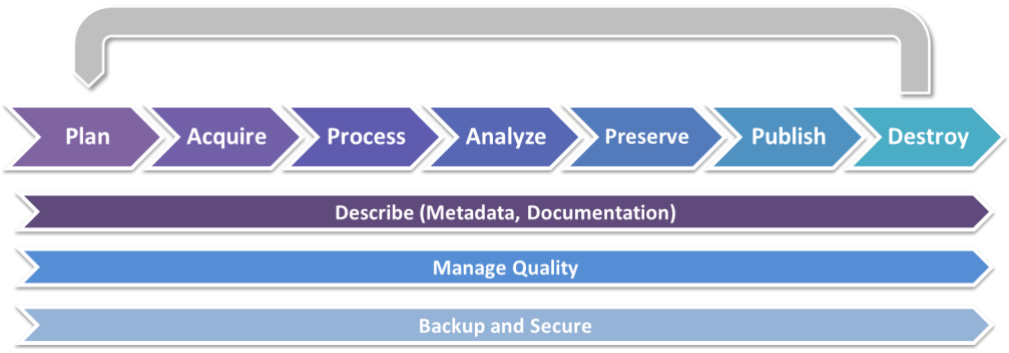

Figure 9 Data Lifecycle Model [20].

Figure 9 shows the elements that are part of the Data Lifecycle Model. The DLM elements are described below, and which will be implemented in FAITH supporting the Data Management Plan (D7.6).

6.3.1 Plan

The first stage is to define planning that allows the creation of a Data Management Plan (DMP) to control how the data will be managed throughout the life cycle. According to the European Single Electronic Data Interchange Area (SEDIA), a DMP describes the life cycle of data management for data to be collected, processed and / or generated by a Horizon 2020 project. As part of doing that research data is easy to find, accessible, interoperable and reusable (FAIR).

6.3.2 Acquire

It is important to define how the data is acquired within the project. This is something that is still in definition for the beta version. In all cases the Data Privacy Protection is framed to support data acquisition (D3.3). For the moment you can get an idea of the FAITH Data Sources that have been described in section 4 of this document.

6.3.3 Process

Data processing involves various activities associated with preparing new or previously collected data entries, including: Validation, summarization, transformation, integration, sub-configuration, and derivation of data. All these activities must be carried out constantly to guarantee the quality of the data used or created within FAITH.

6.3.4 Analyse

Data analysis involves various activities associated with the exploration and interpretation of processed data, which are: statistical analysis, visualization, spatial analysis, image analysis, modelling, and interpretation. In this sense, FAITH will be able to incorporate all the tools mentioned above to take better advantage of its functions.

6.3.5 Preserve

Data preservation involves a series of actions and procedures that are used to ensure the long-term sustainability and accessibility of the data archive and repositories.

6.3.6 Publish/Share

The set of data collected in the studies will be confidential and will be shared in compliance with confidentiality agreements and all applicable laws governing the protection of personal data.

6.3.7 Destroy

The data destruction policy within FAITH will be determined in the next version of this document based on the discussion and agreement of the consortium.

7 CONCLUSIONS

This deliverable provides the initial version of the FAITH Common Data Model (Alpha Version of the FAITH infrastructure) as well as the description of the process, resources and tools that are being used for this definition.

Using FHIR as a standard for extending interoperability and modelling the data is an advantage of this project. In this way, a common language is found for the exchange of information and a model to follow to achieve a correct definition of the Data Model and its extension.

The implementation strategy allows the work to be done in an orderly manner. In this way, it must first be concluded with a definition of the FHIR "resources" applied to this specific case. Once this definition is made, it is important to create the FHIR Implementation Guide when necessary.

Having defined a Data Lifecycle Model is important to ensure that the data obtained will go through an orderly and rigorous process. In this way it can be ensured that the data treatment will be according to the needs of the project.

This deliverable will be updated in M30, providing the final version of the Data Model, related updates and needed adaptations to move to the collaborative federated machine learning approach (Beta Version of the infrastructure).

8 BIBLIOGRAPHY

[1] W. Inmon and D. Linstedt, "5.3 - Data Modeling for the Structured Environment," in *Data Architecture: a Primer for the Data Scientist*, Morgan Kaufmann, 2015, pp. 181-188.

[2] IBM Corporation, "Vistas de modelo de datos común," 2021. [Online]. Available: https://www.ibm.com/docs/es/networkmanager/4.2.0?topic=dictionary-common-data-model-views. [Accessed April 2021].

[3] M. Abdelhak, S. Grostick and M. Hanken, Health Information: Management of a Strategic Resource, St. Louis, MO: Elsevier Health Sciences, 2007.

[4] HL7.org, "HL7 FHIR," 1 November 2019. [Online]. Available: https://www.hl7.org/fhir/. [Accessed April 2021].

[5] Observational Health Data Sciences and Informatics, [Online]. Available: https://www.ohdsi.org/data-standardization/the-common-data-model/. [Accessed April 2021].

[6] The Health Care Systems Research Network Web site is hosted by Kaiser Permanente Division of Research (Northern California)., "VDW Data Model," [Online]. Available: http://www.hcsrn.org/en/Tools%20&%20Materials/VDW/. [Accessed April 2021].

[7] Partners Healthcare, "i2b2 - Informatics for Integrating Biology & the Bedside," [Online]. Available: https://www.i2b2.org/. [Accessed April 2021].

[8] hl7.org, "HL7," Health Level Seven International, [Online]. Available: https://www.hl7.org/. [Accessed April 2021].

[9] Deloitte, "FHIR, ese gran desconocido," [Online]. Available: https://www2.deloitte.com/es/es/pages/technology/articles/fhir-ese-gran-desconocido.html. [Accessed April 2021].

[10] Lucid Software Inc., "Lucidchart," [Online]. Available: https://www.lucidchart.com/. [Accessed April 2021].

[11] Lucid Software Inc, "Lucidchart," [Online]. Available: https://www.lucidchart.com/blog/. [Accessed April 2021].

[12] diagrams.net, "diagrams.net," [Online]. Available: https://www.diagrams.net/. [Accessed April 2021].

[13] diagrams.net, "About diagrams.net," [Online]. Available: https://www.diagrams.net/about. [Accessed April 2021].

[14] Cinergix Pty Ltd (Australia), "Creately," [Online]. Available: https://creately.com/. [Accessed April 2021].

[15] Shiraz, "¿Qué es creately?," March 2021. [Online]. Available: https://support.creately.com/hc/es/articles/360001418815--Qu%C3%A9-es-creately-#h_01EEFPV5X4ZG28VX88ZDGB3899. [Accessed April 2021].

[16] diagrams.net, "UML 2.5 shape library with updated shapes," 21 January 2021. [Online]. Available: https://www.diagrams.net/blog/uml-2-5?fbclid=IwAR35bU5bmVNcuEmSM5vC-ZpLescLIbBDeO9PfYSBwkS9k5g0DV27rIwLQRg. [Accessed April 2021].

[17] hl7.org, “Base Resource Definitions,” [Online]. Available: https://www.hl7.org/fhir/resource.html. [Accessed April 2021].

[18] A. Phillips and M. Davis, “Tags for Identifying Languages,” 2009.

[19] hl7.org, “Resource ImplementationGuide - Content,” [Online]. Available: https://www.hl7.org/fhir/implementationguide.html. [Accessed April 2021].

[20] USGS, “Data LifeCycle,” [Online]. Available: https://www.usgs.gov/products/data-and-tools/data-management/data-lifecycle. [Accessed April 2021].

[21] Beck, Kent et al, “Manifesto for Agile Software Development,” 2001. [Online]. Available: http://agilemanifesto.org/. [Accessed 18th March 2015].

[22] Aquasmart Project, “Redmine,” 2015. [Online]. Available: https://www.aquasmartdata.eu/redmine. [Accessed 19th March 2015].

[23] J. McLaughlin, “Creating User Stories,” TSSG, 19th March 2015. [Online]. Available: https://www.aquasmartdata.eu/redmine/projects/aquasmart/wiki/UserStory. [Accessed 19th March 2015].

[24] Git contributors, “git --distributed-is-the-new-centralized,” 2015. [Online]. Available: http://git-scm.com/. [Accessed 19th March 2015].

[25] J. McLaughlin, “How to Use Git,” Aquasmart Project, 19th March 2015. [Online]. Available: https://www.aquasmartdata.eu/redmine/projects/aquasmart/wiki/UsingGit. [Accessed 19th March 2015].

[26] A. Weller, “Challenges for Transparency,” *arXiv,* 2017.

[27] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning,” *arXiv,* 2017.

[28] D. Doran, S. Schulz and T. R. Besold, “What Does Explainable AI Really Mean? A New Conceptualization of Perspectives,” *arXiv,* 2017.

[29] hl7.org, “Introducing HL7 FHIR,” [Online]. Available: https://www.hl7.org/fhir/summary.html. [Accessed April 2021].

[30] Lucid Software Inc., “UML Sequence Diagrams Made Easy,” [Online]. Available: https://www.lucidchart.com/blog/lucidchart-uml-sequence-diagram-markup. [Accessed April 2021].

Annexes

A.1. Annex 1: UML FAITH DIAGRAMS

This annex shows some UML diagrams that contain the classes that have been defined early in the project. It should be noted that this is not the final definition but rather a first approach to what will be the full version that will be presented in the update of this document in the M30.

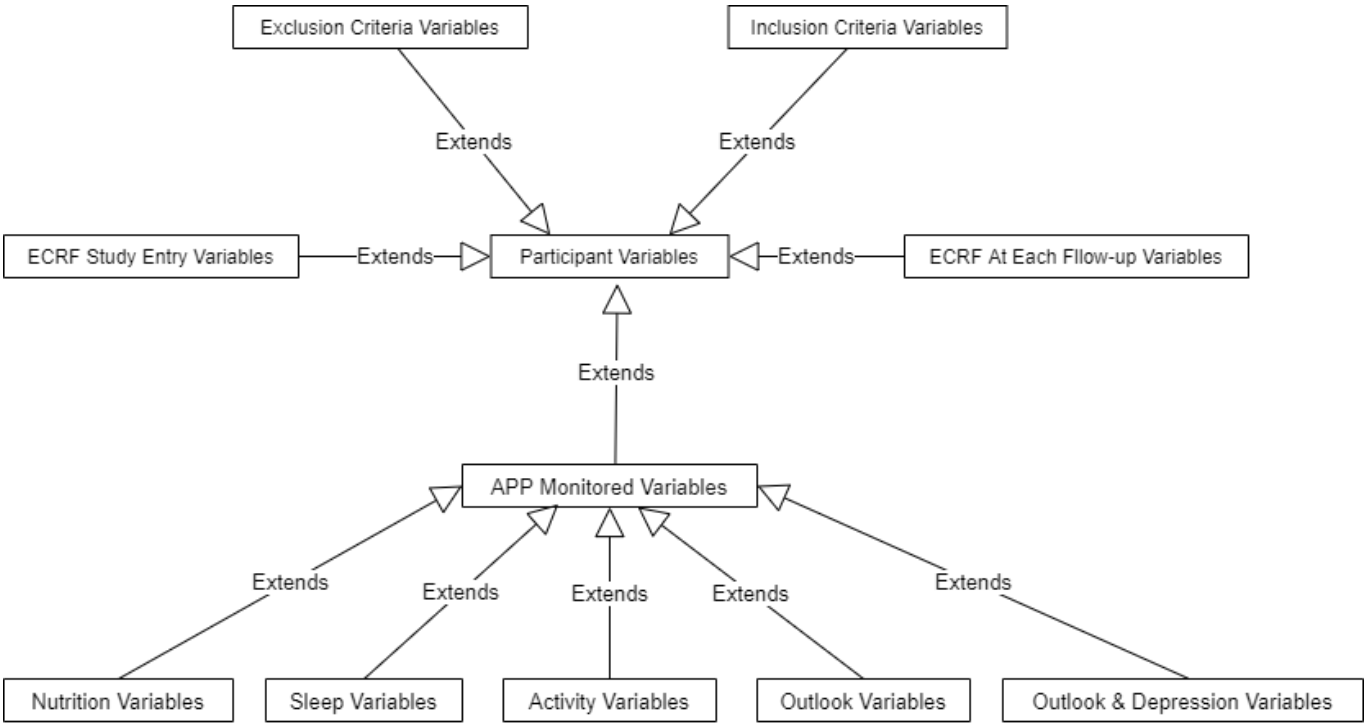

Figure 10 FAITH Data Model

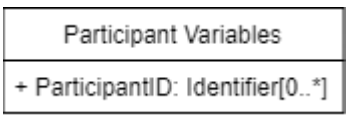

Figure 11 Participant Variables Class

ECRF At Each FIlow-up Variables
+ Is it a planned follow-up?: String/ValueSet [outpatient/visit/teleconsultation/unplanned]
+ Date of follow-up consultation: Date - Timestap
+ Timing in months after : Integer [0,...]
+ Weight: Double [0,...]
+ Body Mass Index (BMI): Double [0,...]
+ Recurrence: Boolean [Yes/No]
+ Date of recorrence: Date - Timestap
+ Type of recurrence: String [...]
+ Non-cancer-related medical event: String [...]
+ Type of new non-cancer-related medical event: String [...]
+ Grade adverse event: String [...]
+ Date of onset of new non-cancer-related medical event: Date - Timestap
+ Date of resolution of New non-cancer-related medical event: Date - Timestap
+ Anorexia: Boolean [Yes/No]
+ Obesity: Boolean [Yes/No]
+ Food intolerances: String/ValueSet [...]
+ Diet: String/ValueSet [Regular Diet, Low Fiber Diet, Salt-restricted (Low Sodium) Diet, Fat-Restricted Diet, Cholesterol-Restricted Diet, Vegetarian Diet, Consistent Carbohydrate (Diabetic Diet), other]
+ Endocrine basal diseases: Boolean [Yes/No]
+ What basal endocrine diseases?: String/ValueSet [...]
+ Osteoporosis: Boolean [Yes/No]
+ Physical activity regularly practice: Boolean [Yes/No]
+ Insomnia: Boolean [Never/Hardly]
+ Pain (ex. Artralgia, neuropathic pain, etc.): Boolean [Never/Hardly]
+ Hot flushes: Boolean [Never/Hardly]
+ Asthenia (lack of strength): Boolean [Yes/No]
+ Sweating (lack of strength): Boolean [Yes/No]
+ Stigmas of first treatment, chemo mainly: String/ValueSet [facial edema, alopecia, loss of eyebrows and eyelashes, nails, lymphedema, weight, hirsutism, etc]
+ Emotional: String/ValueSet [Anger, Fear, Apathy, Sadness, Confused, happy, strong, energized (hyperactivity)]

Figure 12 ECRF at each Follow-up Variables Class

Inclusion Criteria Variables
+ FAITH consent: Boolean [Yes/No]
+ Data consent: Date - Timestamp
+ Age in range: Boolean [Yes/No]
+ Language: Boolean [Yes/No]
+ Cancer survivor: Boolean [Yes/No]
+ Type of cancer / Tumoral Stage: Boolean [Breast/Lung]
+ Specific treatment: Boolean [Yes/No]
+ Completed treatment: Boolean [Yes/No]
+ Performance status (ECOG): Boolean [1/2]
+ GDPR consent: Boolean [Yes/No]
+ Acceptance participation: Boolean [Yes/No]
+ Acceptance smartphone: Boolean [Yes/No]
+ Acceptance share data: Boolean [Yes/No]

Figure 13 Inclusion Criteria Variables Class

Exclusion Criteria Variables
+ Unable protocol: Boolean [Yes/No]
+ Illiteracy: Boolean [Yes/No]
+ Digital illiteracy: Boolean [Yes/No]
+ Phone OS: Boolean [Yes/No]
+ Presence of distant metastases: Boolean [Yes/No]
+ Previous invasive malignanciesl: Boolean [Yes/No]
+ Accute illnessl: Boolean [Yes/No]
+ Concomitant diseases: Boolean [Yes/No]
+ Surgery: Boolean [Yes/No]
+ Major illness: Boolean [Yes/No]
+ Pregnancy/Breastfeeding: Boolean [Yes/No]
+ Psychiatric comorbidities: Boolean [Yes/No]
+ Previous hypomanic/manic: Boolean [Yes/No]
+ Psichiatric disorders: Boolean [Yes/No]
+ Substance dependence: Boolean [Yes/No]
+ Dementia: Boolean [Yes/No]
+ CNS structural lesion: Boolean [Yes/No]
+ Developmental disorders: Boolean [Yes/No]

Figure 14 Exclusion Criteria Variables Class

ECRF Study Entry Variables
+ Randomization date: Date - Timestamp
+ Birth date: Date - Timestamp
+ Sex: Boolean [Male/Female]
+ Education: String/ValueSet [Elemental/Undergraduate/Postgraduate/Vocational]
+ Marital status: String/ValueSet [Single, Married, Widowed, Divorced, Separated, Registered partnership, Cohabiting]
+ Number of children: Integer [0,...]
+ Employment status: Boolean [Employed/Unemployed]
+ Caregivers: Boolean [Yes/No]
+ Familial support: Boolean [Yes/No]
+ Baseline height: Double [0,...]
+ Baseline weight: Double [0,...]
+ Baseline Body Mass Index (BMI): Double [0,...]
+ Baseline smoking status: String/ValueSet [Current, Former, Never, Unknown]
+ Baseline smoking habits: String/ValueSet [Cigarettes, Cigar/pipes, Betel quid, Smokeless (spit) Tobacco]
+ Cigarrets a day : Integer [0,...]
+ Number of years as a smoker: Integer [0,...]
+ Alcohol consumption: String/ValueSet [Current, Former, Never, Unknown]
+ Average number of alcohol units per week : Integer [0,...]
+ Date of pathological procedure for primary tumor diagnosis: Date - Timestamp
+ Age at primary tumor diagnosis: Integer [0,...]
+ Date of first treatment: Date - Timestap
+ Node afectation: StringValueSet [Done/pos=1; Done/neg=2; Done/No value=3; Not Done =4; Not Done/PAAF-BAG negative= 5; Not Done/PAAF-BAG positive= 6;No available=N/A)]
+ Neoadjuvant (NAD) treatment: StringValueSet [Chemotherapy, Hormone therapy]
+ Date of last NAD treatment: Date - Timestamp
+ Surgery: String/ValueSet [Segmentectomy, Mastectomy, None]
+ Type of axillary surgery: String/ValueSet [1=SLNB only, 2=SLNB+ALND, 3=ALND only, 4=no axillary surgery]
+ Date of surgery: Date - Timestamp
+ Grading of tumor regression at surgery/last treatment previous Faith recruitment: String/ValueSet [RCB for breast cancer/Grade for lung cancer]
+ Last treatment outcome: Date - Timestamp
+ Concomitant medications: Boolean [Yes/No]
+ Developmental disorders: Boolean [Yes/No]
+ Date of last treatment: Date - Timestap

Figure 15 ECRF Study Entry Variables Class

APP Monitored Variables
+ QoL questionnaires: EORTC QLQ-30/ EORTC QLQ BR-45/EORTC QLQ-LC29
+ Depression questionaires: Hamilton Rating Scale for Depression

Figure 16 Monitored Variables Class

Nutrition Variables
+ Meals per day: Integer [0,...]
+ Liquid intake / Litres per day: Integer [0,...]
+ Loss of appetite: Boolean [Yes/No]
+ Portions per food group per day: Integer [0,...]
+ Calories burned: Integer [0,...]

Figure 17 Nutrition Variables Class

Sleep Variables
+ Hours of sleep or sleep range: Integer [0,...]
+ Sleep interruptions/disturbances: String/ValueSet [0/1-2/more than 3]
+ Do you awake tired?: Boolean [Yes/No]
+ Sleep duration (night): Integer [0,...]
+ Nightmares: Boolean [Yes/No]
+ Sleep drugs intake: Boolean [Yes/No]
+ Nap duration: Integer [0,...]

Figure 19 Sleep Variables Class

Activity Variables
+ Frequency of diary physical activity: String/ValueSet [none/once/twice/ more than twice]
+ Duration of the exercise: Integer [0,...]
+ Distancia: Integer [0,...]
+ Average Distance: Integer [0,...]
+ The geographic space (area) in which a person lives and conducts their roles and activities: Location Coordinates

Figure 18 Activity Variables Class

Outlook Variables
+ Hot Flushes: Boolean [Yes/No]
+ Emotional: String/ValueSet [Anger, Fear, Apathy, Sadness, Confused, happy, strong, energized (hyperactivity)]

Figure 21 Outlook Variables Class

Outlook & Depression Variables
+ Speech rate (SR): Double [0,...]
+ Mean pause duration (MPD): Double [0,...]
+ Number of pauses (NOP): Double [0,...]
+ Fundamental frequency (F0): Double [0,...]
+ Mean F0: Double [0,...]
+ Range of F0 Double [0,...]
+ Minimum F0: Double [0,...]
+ Speech signal: Signal

Figure 20 Outlook & Depression Variables Class

A.2. Annex 2: FAITH VARIABLE MODELLING

This annex shows the variables that have been defined so far, considering the meetings and definitions of the protocol and the needs of the studies to be carried out. As with the definition of classes and "resources", this variable modelling will be updated, and its final version will be presented in the update of this document in month 30.

Table 1 Inclusion Criteria Variables

Data (Variable)	Data Description	Values	Type of Variable	Frequency of Acquisition	Format	Rules
FAITH Consent	Has the patient signed the informed consent?	Yes/No	Compliance	once	Radio-Button	One option chosen. If Yes---> the Date of Informed consent [xxx] must be filled, if = No --> Exclude the patient from enrolment
Date Consent	Date of signed informed consent	Date - Timestamp	Compliance	once	DD/MM/YYYY	Date - Timestamp Mandatory
Age in range	Is the patient aged ≥ 18 years and <70 years?	Yes/No	Demographic	once	Radio-Button	One option chosen. if = No --> Exclude the patient from enrolment
Language	Native or fluent English/Spanish/Portuguese speakers	Yes/No	Demographic	once	Radio-Button	One option chosen.
Carcer survivor	Is patient cancer-free? (no clinical evidence of currently suffering cancer)	Yes/No	Clinical	once	Radio-Button	One option chosen. if = No--> Exclude the patient from enrolment

Type of cancer/Tumoral Stage	Diagnosis that applies to the patient. Disease stage refers to AJCC/UICC eighth edition.	Early breast cancer of the subtypes: - Luminal A-like ▪ Stage I [ ] ▪ Stage II [ ] ▪ Stage III [ ] - Luminal B-like ▪ Stage I [ ] ▪ Stage II [ ] ▪ Stage III [ ] - Luminal B HER2+ (HR+/-) ▪ Stage I [ ] ▪ Stage II [ ] ▪ Stage III [ ] Lung cancer patients, with non-small cells lung carcinoma, of subtypes: - Stages I-IIIA o Stage I [ ] o Stage II [ ] o Stage III [ ] - Stages I-IIIB o Stage I [ ] o Stage II [ ] o Stage III [ ]	Clinical	once	Radio-Button	2 nested buttons; (only one option chosen)
Specific Treatment	Disease-specific treatments, to reduce the minimal risk of recurrence or relapse, are accepted according to each cancer subtype	Yes/No	Clinical	once	Radio-Button	One option chosen. No rule

Completed treatment	Has the patient completed treatment with curative intent (chemotherapy, radiotherapy, or surgery with curative purposes treatment) and in follow up after first year of treatments and less than 5 years (i. e., between the second and fifth year of follow-up)?	Yes/No	Clinical	once	Radio-Button	One option chosen. if = No--> Exclude the patient from enrolment
PS	Performance status (ECOG)	1,2	Clinical	once	Radio-Button	One option chosen. if = xxxx--> Exclude the patient from enrolment
GDPR Consent	Consented to use their data	Yes/No	Compliance	once	Radio-Button	One option chosen. if = No--> Exclude the patient from enrolment
Acceptance Participation	Acceptance to be followed-up for up to 1?? year answering questionnaires and data from devices?	Yes/No	Compliance	once	Radio-Button	One option chosen. if = No--> Exclude the patient from enrolment
Acceptance Smart Phone	Acceptance to use and install apps in the smartphone and internet	Yes/No	Compliance	once	Radio-Button	One option chosen. if = No--> Exclude the patient from enrolment
Acceptance share data	Acceptance to provide data after project to be re-used for research purposes	Yes/No	Compliance	once	Radio-Button	One option chosen. if = No--> Exclude the patient from enrolment

Table 2 Exclusion Criteria Variables.

Data (Variable)	Data Description	Values	Type of Variable	Frequency of Acquisition	Format	Rules
Unable Protocol	Patient unable to comply with Protocol	Yes/No	Compliance	once	Radio-Button	One option chosen. if = No--> Exclude the patient from enrolment
Illiteracy	illiteracy or otherwise not understanding the study's instructions	Yes/No	Compliance	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment
Not able to find, evaluate, gather, or compose clear information in digital devices (medical devices, smartphones, etc)	digital illiteracy	Yes/No	Compliance	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment
Phone software	different from android and iOS	Yes/No	Compliance	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment
Presence of distant metastases	Has the patient ever been diagnosed with distant metastases?	Yes/No	Clinical	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment
Previous invasive malignancies	prior invasive malignance whose treatment was completed within 5 years before study entry	Yes/No	Clinical	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment (exceptions: patients with adequately treated, basal or squamous cell skin carcinoma or curatively resected cervical cancer in situ, are eligible)
Acute illness	Any acute medical illness	Yes/No	Clinical	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment

Concomitant diseases	Patient affected by any known or underlying medical conditions that, could adversely affect the ability of the participating subject to comply with the study? serious other diagnosed concomitant diseases such as clinically significant (i.e. active) cardiac disease (e.g. congestive heart failure, symptomatic coronary artery disease or cardiac arrhythmia not well controlled with medication) or myocardial infarction within the last 12 months	Yes/No	Clinical	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment
Surgery	major surgery for a severe disease or trauma which could affect patient's psychosocial wellbeing (for example, major heart or abdominal surgery) within 4 weeks prior to study entry or lack of complete recovery from the effects of surgery	Yes/No	Clinical	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment
Major illness	Treatment for any major illness in the last half year	Yes/No	Clinical	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment
Pregnancy breastfeeding	pregnancy or breastfeeding at time of recruitment	Yes/No	Clinical	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment
Psychiatric comorbidities	Psychiatric comorbidities such as diagnosis of a moderate to severe major	Yes/No	Clinical	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment

	depressive episode at baseline according to the M.I.N.I. (diagnosis) and HAM-D (severity)					
Previous hypomanic/manic	current or previous hypomanic or manic episode, current or previous psychotic disorder or current mood disorder with psychotic symptoms as screened by the M.I.N.I.	Yes/No	Clinical	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment
Psychiatric disorders	individuals presenting any psychiatric disorder requiring urgent care or hospitalization at that moment	Yes/No	Clinical	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment
Substance dependence	substance abuse or dependence in the last 12 months as screened by the M.I.N.I.	Yes/No	Clinical	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment
Dementia	dementia or other active neurodegenerative disease	Yes/No	Clinical	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment
CNS structural lesion	previously known structural lesion of the central nervous system (e.g., stroke)	Yes/No	Clinical	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment
Developmental Disorders	developmental disorders with low intelligence quotient or any other form of cognitive impairment	Yes/No	Clinical	once	Radio-Button	One option chosen. if = Yes--> Exclude the patient from enrolment

Table 3 ECRF Study Entry Variables

Data (Variable)	Data Description	Values	Type of Variable	Frequency of Acquisition	Format	Rules

Randomization Date	Date of randomization	Date - Timestamp	Compliance	once	DD/MM/YY	NA
Birth Date	Date of patient`s birth	Date - Timestamp	Demographic	once	DD/MM/YY	Check whether patient age is between 18 and 70 years (≥ 18 years and ≤ 70 years).
Sex	Sex - biological variable	Female /Male	Demographic	once	Radio-Button	One option chosen.
Education	level of education as a research variable including vocational	Elemental/Undergraduate /Postgraduate/Vocational	Demographic	once	Radio-Button	One option chosen.
Marital status	variable that describes a person's relationship with a significant other.	● Single ● Married ● Widowed ● Divorced ● Separated ● Registered partnership ● Co-habiting	Demographic	once	Radio-Button	One option chosen.
Number of Children	Number of children born	Integer	Demographic	once	00 [children]	from 0 to 99
Employment status	Status of employment variable	Not employed/ Employed	Demographic	once	Radio-Button	One option chosen.
Caregivers	Caregiver residing with patient, paid caregiver or another caregiver	Yes/no	Demographic	once	Radio-Button	One option chosen.

Familial support	Receiving family support as caregiver	Yes/no	Demographic	once	Radio-Button	One option chosen.
Baseline height	Between 0,5 m and 2,3 m (error if less than 0,5 m or more than 2,3 m).	decimal	Clinical	once	0,00 [m]	0,5 m ≤ baseline height ≤ 2,3 m
Baseline weight	Between 25 Kg and 200 Kg (error if less than 25 kg or more than 200 kg).	decimal	Clinical	once	00,0 [Kg]	25 kg ≤ baseline weight ≤ 200 kg
Baseline BMI	The body mass index (BMI) is a measure that uses the height and weight to work out if a person's weight is healthy. The BMI calculation divides an adult's weight in kilograms by their height in metres squared. The BMI ranges are: ● BMI < 18.5 – Underweight ● 18.5 ≤ BMI < 25 – Normal weight ● 25 ≤ BMI < 30 – Overweight	decimal	Clinical	once	00 [???]	Value is automatically calculated based on fields Baseline height [xxx] and Baseline weight [xxx], using the formula: BMI = weight (Kg) / [height(m)]2
Baseline Smoking status	Status for smoking as categorical variable	● Current ● Former ● Never ● Unknown	Clinical	once	Radio-Button	One option chosen.
Baseline Smoking habits	type of smoking behaviour	● Cigarettes ● Cigar/pipes ● Betel quid ● Smokeless (spit) Tobacco	Clinical	once	Radio-Button	One option chosen. Dependent to "Smoking status" [xxx] if is selected "Current" or "Former".

Cigarrets a day	number of cigarettes /packages of cigarettes/related consumed on one day 1 pack = numbers per Day/20 for cigarettes or /4 for cigars	Integer	Clinical	once	00 [cigarettes/ Packages]	Dependent to “Smoking status” [xxx] if is selected “Current” or “Former”.
Number of years as a smoker	Total number of years since starting smoking	Integer	Clinical	once	00 [Years]	Dependent to “Smoking status” [xxx] if is selected “Current” or “Former”.
Alcohol consumption	An indication of a person's current alcohol consumption as well as some indication of alcohol consumption history.	Current ● Former ● Never ● Unknown	Clinical	once	Radio-Button	One option chosen.
Average number of alcohol units per week	Refers to the standard units of alcohol consumed in a week	Integer	Clinical	once	00 [unit selected]	Dependent to “Alcohol consumption” [xxx] if is selected “Current” or “Former”.
Date of pathological procedure for primary tumour diagnosis	date of diagnosis. It refers to the very first point in patient’s history	Date - Timestamp	Clinical	once	DD/MM/YYYY	Dependent to "Birth Date". Check whether patient date for primary tumour diagnosis<birth date
Age at primary tumour diagnosis	Age of the patient when he/she was diagnosed with lung/breast cancer.	Integer	Clinical	once	00 [Years]	Dependent to "Birth Date". Check whether patient age at primary tumour diagnosis< current age
date of first treatment	date of diagnosis in which first treatment for first	Date - Timestamp	Clinical	once	DD/MM/YYYY	Dependent to "date of diagnosis". Check whether patient date for

	pathological procedure started. It refers to the very first point in patient's history					primary tumour diagnosis<date of 1st treatment
Node affectation	Node affectation: SLNB PRE CHEMO Details	Done/pos=1; Done/neg=2; Done/No value=3; Not Done =4; Not Done/PAAF-BAG negative= 5; Not Done/PAAF-BAG positive= 6; No available=N/A)	Clinical	once	Radio-Button	One option chosen.
Neoadjuvant (NAD) treatment	Neoadjuvant (NAD) Cancer treatment: Chemotherapy, Hormone therapy	▪Chemotherapy [] ▪Hormone therapy []	Clinical	once	Radio-Button	2 nested buttons; (only one option chosen)
Date of last NAD treatment	Date of last NAD treatment	Date - Timestamp	Clinical	once	DD/MM/YY	Dependent to "date of diagnosis". Check whether patient date for primary tumour diagnosis<date NAD
Surgery	Type of surgery if any	segmentectomy/mastectomy/NON	Clinical	once	Under definition	Under definition
Type of axillary surgery	Under definition	1=SLNB only, 2=SLNB+ALND, 3=ALND only, 4=no axillary surgery	Clinical	once	Under definition	Under definition
Date of surgery	Under definition	Date - Timestamp	Clinical	once	DD/MM/YY	Under definition
Grading of tumour regression at surgery/last treatment previous Faith recruitment	Under definition	RCB for breast cancer/Grade for lung cancer	Clinical	once	Under definition	Under definition

Adjuvant Cancer treatment: Chemotherapy, Hormone therapy	▪Chemotherapy [] ▪Hormone therapy []	M	Clinical	once	2 nested buttons; (only one option chosen)	New included to categorize the patients
Date of last NAD treatment	Under definition	Date - Timestamp	Clinical	once	DD/MM/YYYY	NA
Last treatment outcome	Under definition	Date - Timestamp	Clinical	once	DD/MM/YYYY	NA
Concomitant medications	Use drugs --> Concomitant medications can affect to patient's status (depression levels, ..)	Under definition	Clinical	once	Radio-Button	One option chosen.
Date last treatment	Last date treatment completion	Date - Timestamp	Clinical	once	DD/MM/YYYY	NA

Table 4 ECRF AT EACH FOLLOW-UP VARIABLES

Data (Variable)	Data Description	Values	Type of Variable	Frequency of Acquisition	Marker Category	Format
Is it a planned follow-up?	It refers to a planned/unplanned follow-up	outpatient/visit/teleconsultation/unplanned	Clinical	every three months	NA	Radio-Button
Date of follow-up consultation	Date of follow-up visit	Date - Timestamp	Clinical	every three months	NA	Time elapsed from the date of treatment end to the date of follow-up consultation. This time is measured in months.

Timing in months after treatment completion	Time elapsed from the date of treatment end to the date of follow-up consultation. This time is measured in months.	Under definition	Clinical	every three months	NA	Under definition
Weight	Weight at follow-up visit.	Under definition	Clinical	every three months	NA	Under definition
Body mass index (BMI)	BIM at follow-up visit.	Under definition	Clinical	every three months	NA	Under definition
Recurrence	Confirmed recurrence by unequivocal radiological confirmation or histological confirmation; no clinical suspicion	Under definition	Clinical	every three months	NA	Under definition
Date of recurrence	Date of radiological or histological confirmation	Under definition	Clinical	every three months	NA	Under definition
Type of recurrence	Under definition	Under definition	Clinical	every three months	NA	Under definition

Non-cancer-related medical event	New non-cancer event?	Under definition	Clinical	every three months	NA	Under definition
Type of new non-cancer-related medical event	Checklist	Under definition	Clinical	every three months	NA	Under definition
Grade adverse event	Grade according to CTCAE	Under definition	Clinical	every three months	NA	Under definition
Date of onset of new non-cancer-related medical event	Under definition	Date - Timestamp	Clinical	every three months	NA	Under definition
Date of resolution of New non-cancer-related medical event	Under definition	Date - Timestamp	Clinical	every three months	NA	Under definition
Anorexia	Under definition	Yes/No	Clinical	every three months	All	Under definition
Obesity	Under definition		Clinical	every three months	All	Under definition
Food intolerances	Under definition	**Checklist**	Clinical	every three months	Nutrition	Under definition
Diet	Under definition	Regular Diet, Low Fibber Diet, Salt-restricted (Low Sodium) Diet, Fat-Restricted Diet, Cholesterol-Restricted Diet, Vegetarian Diet, Consistent Carbohydrate (Diabetic Diet), other	Clinical	every three months	Nutrition	Under definition

Endocrine basal diseases (ex. diabetes)	Under definition	yes/no. If yes, specify Checklist	Clinical	every three months	Nutrition	Under definition
Osteoporosis	Under definition	Yes/no	Clinical	every three months	Nutrition, Activity	Under definition
Physical activity regularly practices	Under definition	Yes/no	Clinical	every three months	Nutrition, Activity	Under definition
Insomnia	Under definition	never/hardly ever/ sometimes/frequently/always	Clinical	every three months	Sleep	Under definition
Pain (ex. Arthralgia, neuropathic pain, etc.)	Under definition	never/hardly ever/ sometimes/frequently/always	Clinical	every three months	Sleep, Activity, Outlook	Under definition
Hot flushes	Under definition	never/hardly ever/ sometimes/frequently/always	Clinical	every three months	Sleep, Activity, Outlook	Under definition
Asthenia (lack of strength)	Under definition	Yes/no	Clinical	every three months	Activity, Outlook	Under definition
Sweating	Under definition	Yes/no	Clinical	every three months	Outlook	Under definition
Stigmas of first treatment, chemo mainly	Under definition	Facial swelling, alopecia, loss of eyebrows and eyelashes, nails, lymphedema, weight, hirsutism, etc	Clinical	every three months	Outlook	Under definition
Emotional	Under definition	Anger, Fear, Apathy, Sadness, Confused, happy,	Clinical	every three months	Outlook	Under definition

		strong, energized (hyperactivity).				
Asthenia (lack of strength)	Yes/no	Yes/no	Clinical	every three months	Activity, Outlook	Under definition
Weight	Under definition	Kg	Clinical	every 3 MONTHS	Nutrition, Outlook, Activity	Under definition

Table 5 Monitored Variables.

Data (Variable)	Data Description	Type of Variable	Frequency of Acquisition	Marker Category
Food Intake	meals per day (number)	APP	WEEKLY	Nutrition
Intake frequency		APP	daily	Nutrition
Intake time		APP	daily	Nutrition
Do you skip any food?	Yes/no	APP	daily	Nutrition
Liquid intake (quantity)	Litres per day (media)	APP	WEEKLY	Nutrition
Concomitant medications	specify	APP	monthly	All

Anorexia (basal and intermittent)	Yes/no	Oncology Nurse consulting	every three months	Nutrition, Outlook
Endocrine basal diseases (ex. diabetes)	Yes/no	Oncology Nurse consulting	every three months	Nutrition
Hours of sleep or sleep range?	number of hours	APP	daily	Sleep
Sleep interruptions/disturbances	0/1-2/more than 3	APP	daily	Sleep
Loss of appetite	Yes/no	APP	Monthly	Nutrition
Portions eaten per food group	Portions per food group per day (number) by food group type per day	APP	WEEKLY	Nutrition
Kind of sleep	More or less than 5 hours	APP	daily	Sleep
Sleep duration (night)	minutes/hours	APP	daily	Sleep

Nightmares	yes/no	APP	MONTHLY	Sleep
sleep drugs intake	yes/no	APP	MONTHLY	Sleep
Nap duration	minutes/hours	APP	daily	Sleep
Frequency of diary physical activity	once/twice/ more than twice	APP	daily	Activity
Type of Exercise/Workout	aerobic, cardio, strength	APP	daily	Activity
Duration of the exercise	minutes/hours	APP	daily	Activity
Exercise metrics	Steps, distance and avg distance	APP	daily	Activity
Calories burnt	Resting Calories / Active Calories	APP	daily	Nutrition/Activity
The geographic space (area) in which a person lives and conducts their roles and activities (without storing the coordinates/locations)	GPS location/Important Locations	APP	daily	Activity
Hot flushes and/or sweating	Yes/no	APP	daily	Outlook

Stigmas of first treatment (chemo mainly)	Facial swelling, alopecia, loss of eyebrows and eyelashes, nails, lymphedema, weight, hirsutism, etc	Oncology Nurse consulting	every three months	All
Emotional	Anger, Fear, Apathy, Sadness, Confused, happy, strong, energized (hyperactivity).	APP	MONTHLY	Outlook