Personalised Health Monitoring and Decision Support Based on Artificial Intelligence and Holistic Health Records

D2.2 – State of the art and requirements analysis II

WP2 Requirements, State of the Art Analysis and User Scenarios in iHelp

Dissemination Level:	Public
Document type:	Report
Version:	1.0.0
Date:	December 20, 2021

The project iHelp has received funding from the European Union’s Horizon 2020 Programme for research, technological development, and demonstration under grant agreement no 101017441.

Document Details

Project Number	101017441
Project Title	iHelp - Personalised Health Monitoring and Decision Support Based on Artificial Intelligence and Holistic Health Records
Title of deliverable	State of the art and requirements analysis II
Work package	WP2
Due Date	30/12/2021
Submission Date	20/12/2021
Start Date of Project	January 1, 2021
Duration of project	36 months
Main Responsible Partner	UPRC
Deliverable nature	Report
Author name(s)	Ainhoa Azqueta (UPM), Miriam Cabrita (iSprint), Calogero Casa (FPG), Andrea Damiani (FPG), Sakis Dalianis (ATC), Krasimir Filipov (KOD), Giorgos Giotis (ATC), Lars Gustaffson (KI), Pavlos Kranas (LXS), Dimosthenis Kyriazis (UPRC), Rostislav Kostadinov (MUP), Chris Orton (ICE), Artitaya Lophatananon (UNIMAN), Borja Liobell Crespo (HDM), George Marinos (UPRC), George Manias (UPRC), Shwetambara Malwade (TMU), Nikolay Mehandjiev (DS4), Krisztina Mekli (UNIMAN), Kenneth Muir (UNIMAN), Anna Nanou (UPRC), Petia Nikolova (DS4), Marzena Nieroda (UNIMAN), Harm Op den Akker (iSprint), Anastasios Pantazidis (iSprint), Aristodemos Pnevmatikakis (iSprint), Usman Rashid (ICE), Shabbir Syed-Abdul (TMU), Tanja Tomson (KI), Usman Wajid (ICE), Nerea Aguado Lopez (HDM), Eshita Dhar (TMU), Te-Min Ke (UNIMAN)
1st Reviewer name	Usman Wajid (ICE)
2nd Reviewer name(s)	Kenneth Muir (UNIMAN), Te-Min Ke (UNIMAN)

Document Revision History

Version History			
Version	**Date**	**Author(s)**	**Changes made**
0.1	2021-09-06	George Marinos (UPRC), George Manias (UPRC)	Initial version of deliverable
0.2	2021-10-22	Krasimir Filipov (KOD)	Update the 4.13.4 & 4.13.5 sections
0.3	2021-10-26	Eshita Dhar (TMU)	Update the secondary data requirements for TMU pilot
0.4	2021-10-26	Nerea Aguado Lopez (HDM)	Update the secondary data requirements for HDM pilot
0.5	2021-11-03	Artitaya Lophatananon (UNIMAN)	Update the secondary data requirements for UNIMAN pilot
0.6	2021-11-05	Chris Orton (ICE)	Update Section 4.12.4
0.7	2021-11-12	George Marinos (UPRC), Anna Nanou (UPRC)	Update the Executive Summary, Introduction and Conclusion sections. Added a new section summarizing the changes compared to the previous version of this report
0.8	2021-11-23	George Marinos (UPRC)	Updates based on internal review comments
0.9	2021-12-19	Pavlos Kranas (LXS)	Quality check
1.0	2021-12-20	Dimosthenis Kyriazis (UPRC)	Final version

Table of Contents

Executive Summary 7

1 Introduction 8

1.1 Summary of Changes 8

2 Pilot & Scenarios - Scenarios to Requirements 11

2.1 Pilot #1 - Study of Genomics and Epigenomics Markers for Early Risk Assessment of Pancreatic Cancer 12

2.1.1 Goals and objectives 12

2.1.2 “As-Is vs To-Be” Scenario 12

2.1.3 Description of scenarios 13

2.1.4 Scenarios to requirements 16

2.2 Pilot #2 - Interventional Monocentric Study based on Patient Reported Outcomes 19

2.2.1 Goals and objectives 19

2.2.2 “As-Is vs To-Be” Scenario 19

2.2.3 Description of scenarios 20

2.2.4 Scenarios to requirements 21

2.3 Pilot #3 - Study of Lifestyle Choices on Elevating the Risk Factors for Pancreatic Cancer 25

2.3.1 Goals and objectives 25

2.3.2 “As-Is vs To-Be” Scenario 28

2.3.3 Description of scenarios 28

2.3.4 Scenarios to requirements 30

2.4 Pilot #4 - Study of Risk, Personalised Recommendations and Measures to Raise Awareness of Relevant Factors 39

2.4.1 Goals and objectives 39

2.4.2 “As-Is vs To-Be” Scenario 39

2.4.3 Description of scenarios 40

2.4.4 Scenarios to requirements 41

2.5 Pilot #5 - Study of Improved Risk Prediction Models and Targeted Interventions that can Delay the Onset of Cancer 45

2.5.1 Goals and objectives 45

2.5.2 “As-Is vs To-Be” Scenario 46

2.5.3 Description of scenarios 48

2.5.4 Scenarios to requirements 50

3 Pilot datasets and data regulatory constraints 55

3.1 Pilot #1 Study of Genomics and Epigenomics Makers for Early Risk Assessment of Pancreatic Cancer 55

3.1.1 Data Collection 55

3.1.2 Dataset Description & Requirements 55

3.2 Pilot #2 Interventional Monocentric Study based on Patient Reported Outcomes 61

3.2.1 Data Collection 61

3.2.2 Data Description 61

3.3 Pilot #3 Study of the Lifestyle choices on Elevating the Risk Factors for Pancreatic Cancer 64

3.3.1 Data Collection 64

3.3.2 Data Description 64

3.4 Pilot #4 Study of Risk, Personalised Recommendations and Measures to Raise Awareness of Relevant Factors 67

3.4.1 Data Collection 67

3.4.2 Data Description 67

3.5 Pilot #5 Study of Improved Risk Prediction Models and Targeted Interventions that can Delay the Onset of Cancer 74

3.5.1 Data Collection 74

3.5.2 Data Description 74

4 Modules 77

4.1 Data Modelling and Integrated Health Records 78

4.1.1 Goals and Objectives 78

4.1.2 State of the Art 78

4.1.3 Background technology 79

4.1.4 Module to User Requirements 79

4.1.5 Module to Technical Requirements 80

4.2 Primary Data Capture and Ingestion 82

4.2.1 Goals and Objectives 82

4.2.2 State of the Art 82

4.2.3 Background technology 83

4.2.4 Module to User Requirements 84

4.2.5 Module to Technical Requirements 84

4.3 Secondary Data Extraction and Interoperability 88

4.3.1 Goals and Objectives 88

4.3.2 State of the Art 88

4.3.3 Background technology....88
4.3.4 Module to User Requirements....89
4.3.5 Module to Technical Requirements....97
4.4 Standardisation and Quality Assurance of Heterogeneous Data....104
4.4.1 Goals and Objectives....104
4.4.2 State of the Art....104
4.4.3 Background technology....106
4.4.4 Module to User Requirements....107
4.4.5 Module to Technical Requirements....109
4.5 Personalised Health Modelling and Predictions....119
4.5.1 Goals and Objectives....119
4.5.2 State of the Art....119
4.5.3 Background technology....119
4.5.4 Module to User Requirements....120
4.5.5 Module to Technical Requirements....121
4.6 Model Library: Implementation and Recalibration of Adaptive Models....124
4.6.1 Goals and Objectives....124
4.6.2 State of the Art....124
4.6.3 Background technology....124
4.6.4 Module to User Requirements....125
4.6.5 Module to Technical Requirements....126
4.7 Clinical DSS Suite with Visual Analytic Tools....128
4.7.1 Goals and Objectives....128
4.7.2 State of the Art....128
4.7.3 Background technology....136
4.7.4 Module to User Requirements....137
4.7.5 Module to Technical Requirements....138
4.8 Big Data Platform and Knowledge Management System....141
4.8.1 Goals and Objectives....141
4.8.2 State of the Art....141
4.8.3 Background technology....142
4.8.4 Module to User Requirements....143
4.8.5 Module to Technical Requirements....144

4.9 Techniques for Early Risk Identification, Predictions and Assessment 147
4.9.1 Goals and Objectives 147
4.9.2 State of the Art 147
4.9.3 Background technology 147
4.9.4 Module to User Requirements 148
4.9.5 Module to Technical Requirements 150
4.10 Design of Personalised Prevention and Intervention Measures 153
4.10.1 Goals and Objectives 153
4.10.2 State of the Art 153
4.10.3 Background technology 153
4.10.4 Module to User Requirements 153
4.10.5 Module to Technical Requirements 156
4.11 Delivery Mechanisms for Personalised Healthcare and Real-time Feedback 157
4.11.1 Goals and Objectives 157
4.11.2 State of the Art 158
4.11.3 Background technology 161
4.11.4 Module to User Requirements 161
4.11.5 Module to Technical Requirements 164
4.12 Social Analytics for the Study of Societal Factors and Policy Making 168
4.12.1 Goals and Objectives 168
4.12.2 State of the Art 168
4.12.3 Background technology 168
4.12.4 Module to User Requirements 170
4.12.5 Module to Technical Requirements 170
4.13 Monitoring, Alerting, Feedback and Evaluation Mechanisms 176
4.13.1 Goals and Objectives 176
4.13.2 State of the Art 176
4.13.3 Background technology 176
4.13.4 Module to User Requirements 177
4.13.5 Module to Technical Requirements 179
5 Conclusions 183
Bibliography 184
List of Acronyms 186

Executive Summary

This deliverable summarizes the work that has been done in the score of Task 2.1 ("State of the art and requirement analysis") until M12. This is the second version of a series of deliverables of this task, whose main objective is to specify the pilot scenarios, their involved datasets, and their corresponding user requirements, as well as the system and technical requirements that are being imposed by the platform. As mentioned in the previous version of this deliverable (i.e. D2.1 – State of the art & Requirement Analysis I), the purpose of these series is to track those requirements throughout the project and update them during the progress of the project. As a result, this second version depicts the enhanced (primary & secondary) data requirements that have been identified after the delivery of the first versions of the scientific reports and provides an updated version of some of the technical requirements of the iHelp platform. As mentioned in the first version of this deliverable the approach that has been followed is twofold: A top-down approach that is followed with respect to the user requirements that were collected by the use case providers themselves, after specifying the business goals and objectives of the use case, along with a concrete definition of the scenario. Moreover, a bottom-up approach is additionally complemented that aims to identify and analyse the technical requirements with respect to the technical work packages that are focusing on the platform technological needs.

The result of this analysis is an updated list of measurable an unambiguous requirement that will drive the design of the overall architecture of the iHelp platform, focusing on serving all different needs of the various use cases of the project. As the project is still progressing, a final version of this deliverable will help the architecture designers of the platform and its software developers to adjust the overall architecture and its implementation accordingly, with respect to the principles of the agile methodology. Moreover, as our scope is to continuously keep track with the latest technological advances and how we can take advantage of them while developing our -a state-of-the-art analysis has been performed regarding the major technologies that are envisioned to be exploited.

This deliverable is being released on M12 of the project and its main aim is to specify the scenarios of the use cases, which drove the user requirements based on their perspective, the updated version of the secondary data requirements and the technical requirements as foreseen by the technical partners that are being involved in the design of the overall architecture of the platform. A final version of this document will be released on M18.

1 Introduction

This document describes measurable and specific user, system and data requirements that drives the design of the architecture of the iHelp platform. These requirements are used as the basis for the implementation of the various software components of the iHelp platform.

This report is the second deliverable of those that need to be produced in the context of the work that will be carried out in iHelp's task T2.1 "*State of the Art and Requirement Analysis*". The main objective of this task is to collect the user and system requirements that are identified during the course of the project. The analysis of the requirements is guiding the technical developments during the project lifecycle in order to ensure that the iHelp objectives are fully addressed and properly considered. Another important objective of task T2.1 is to investigate and analyse the State-of-the-Art (SotA) for iHelp technologies. Both these two objectives provide valuable input for the design of platform's overall architecture and all research and innovation activities of the project.

- As this task duration ends in M18, there will be an additional/final version of this document that will refine the previous two deliverables, named: D2.3: State of the art and requirements analysis III (M18)

The analysis and elicitation of the requirements for the technology components have been carried out considering the exact needs and concerns identified by the project partners, particularly the iHelp pilots and technology providers. As a result, the analysis described in this deliverable not only covers specifies **user requirements**, that can be also considered as *stakeholder requirements*, but also **technical requirements** that can be considered as *system* and *software requirements*.

This analysis, even if it is not addressing strictly technical perspectives of the project, will be a valuable input for the other tasks of WP2, mainly on what concerns the project's reference architecture. Moreover, in order to better understand the software technology requirements, this deliverable includes in Section 4 an analysis of the state-of-the-art related technologies. Also, lists of relevant research initiatives and projects has been provided in this deliverable, along with the description of the baseline technologies that the technical partners will bring to the iHelp project.

This document is organized as follows: Section 1.2 states that differences and updates with the previous version of this deliverable; Section 2 provides an analytical description of the pilots, along with the initial list of the user requirements, while Section 3 describes the various primary datasets that each pilot intends to use, along with the updated secondary data requirements which will be used for the necessary secondary data collection during the project. Section 4 provides the state-of-the-art analysis and specify a list of the baseline technologies that are intended to be used in the development and implementation of the iHelp platform. Moreover, Section 4 delivers the technical requirements of the iHelp platform and finally Section 5 concludes this deliverable.

1.1 Summary of Changes

This section highlights the updates in the current document compared to the previous version of this deliverable, D2.1 – State of the art & Requirement Analysis I.

- Section 1:
  - Updates on section 2.3.1 in table 17 which lists various categories of risk factors:

        - Redistribution of some of the risk factors listed in this table. Alcohol has been moved in the first column of the table (factors that can be changed) and infections factor has been moved in the 4th line of the table.
- Section 2:
    - Updates on the pilot scenarios and requirements:
        - More information was added in the UNIMAN's 1st requirement at section 2.1.4
        - More information was added in the FPG's 1st & 4th requirements at section 2.2.4. Some modifications on the tables' names at the same section were performed in order to be more descriptive.
        - More information was added in the MUP's 1st requirement, and some modifications were performed on the tables' names at the same section (2.4.4).
        - Three additional requirements were added for the TMU pilot at section 2.5.4. The new requirements are described in tables with numbers 41, 42, 43.
- Section 3:
    - Updates on the primary and secondary data requirements of the pilots. More specifically, Section 3.1.2.1:
        - Biological data table has been renamed to Biomarker data, as this name describes more adequately the content and the scope of the table. The Licence/Privacy field of the aforementioned table has also changed.
        - Risk mitigation data table which was included in D2.1, has been removed in the current version of the deliverable since this kind of data are included in the secondary data section.
    - Section 3.1.2.2:
        - The 2nd measurement requirement of secondary data for the UNIMAN pilot has been described in detail.
        - The 3rd and 4th measurement requirements of secondary data for the UNIMAN pilot have been removed since are now summarized in the previous measurement requirement.
    - Section 3.4.2.1:
        - More information regarding data storage, data security and regulatory constraint requirements were added for the MUP pilot in this section in the primary data tables.
    - Section 3.3.2.2:
        - Changes in the 2nd measurement requirement for the HDM pilot.
        - Changes in the 1st application requirement of secondary data for the HDM pilot.
    - Section 3.5.2.2:
        - Additional input in the 3rd measurement requirement of secondary data for the TMU pilot.
        - Changes in the 3rd application requirement of secondary data for the TMU pilot.
- Section 4:
    - Section 4.12.4.:
        - So far, there are no specific user requirements collected concerning the social analysis and support for policy making, simply because the disruptive nature of

this solution requires time for familiarisation and the project is currently focusing on the core objectives of assessing risks through healthcare data (primary and secondary) that is directly gathered from users. However, the social media analysis solution is seen as an additional form of risk assessment as it allows the analysis of users' view on the existing risk assessment techniques, the usefulness of iHelp solutions and in general the healthcare services. A dedicated table which summarizes the mapping of general user requirement with the Social Analytics component has been added in that section.

- Section 4.13.4:
    - Specific tables which are used for the mapping of the user requirements collected in Section 2 and the technical requirements that are described in Section 4.

2 Pilot & Scenarios - Scenarios to Requirements

The purpose of this section is to present the pilot specific scenarios along with the list of the initial requirements that have been defined by each of the five pilots of the iHelp project.

Each of the five pilots describes the exact usage from a use case perspective at a high-level description. The scope of the work that is being reported in this section is the definition of the detailed scenarios as well as the definition of the behaviour and identification of the important necessities that the architecture should comply with, so that they can be taken into account from the very beginning of the project.

The following subsections firstly give an introductory overview of the purpose of each pilot, followed by a detailed description of the most common approaches and the envisioned novelties that each pilot plans to deliver throughout the project, along with the corresponding UML diagram. Then, the next subsection presents a list of the different scenarios for each pilot and finally the initial list of the user requirements extracted from these scenarios is reported.

The table below describes the different types of the requirements and each user requirement identified in 2.X.4 section is characterized using one of the available types.

Table 1: Types of the requirements

FUNCTIONAL	*Functional*	**FUNC**
	Data	**DATA**
NON-FUNCTIONAL	*Look and Feel Requirements*	**L&F**
	Usability Requirements	**USE**
	Performance Requirements	**PERF**
	Operational / Environment Requirements	**ENV**
	Maintainability and Support Requirements	**SUP**

2.1 Pilot #1 - Study of Genomics and Epigenomics Markers for Early Risk Assessment of Pancreatic Cancer

2.1.1 Goals and objectives

The goal of the University of Manchester (UNIMAN) pilot is to provide an efficient platform to identify people at high risk of Pancreatic Cancer (PC) and facilitate cancer risk mitigation. The UNIMAN pilot aims:

- To utilise their on-line platform for cancer risk prediction to identify people at risk.
- To explore the added value of omics-based markers in enhancing the cancer prevention approach.
- To explore to implementation of an interactive iHelp platform for delivery and monitoring of targeted recommendations/interventions.

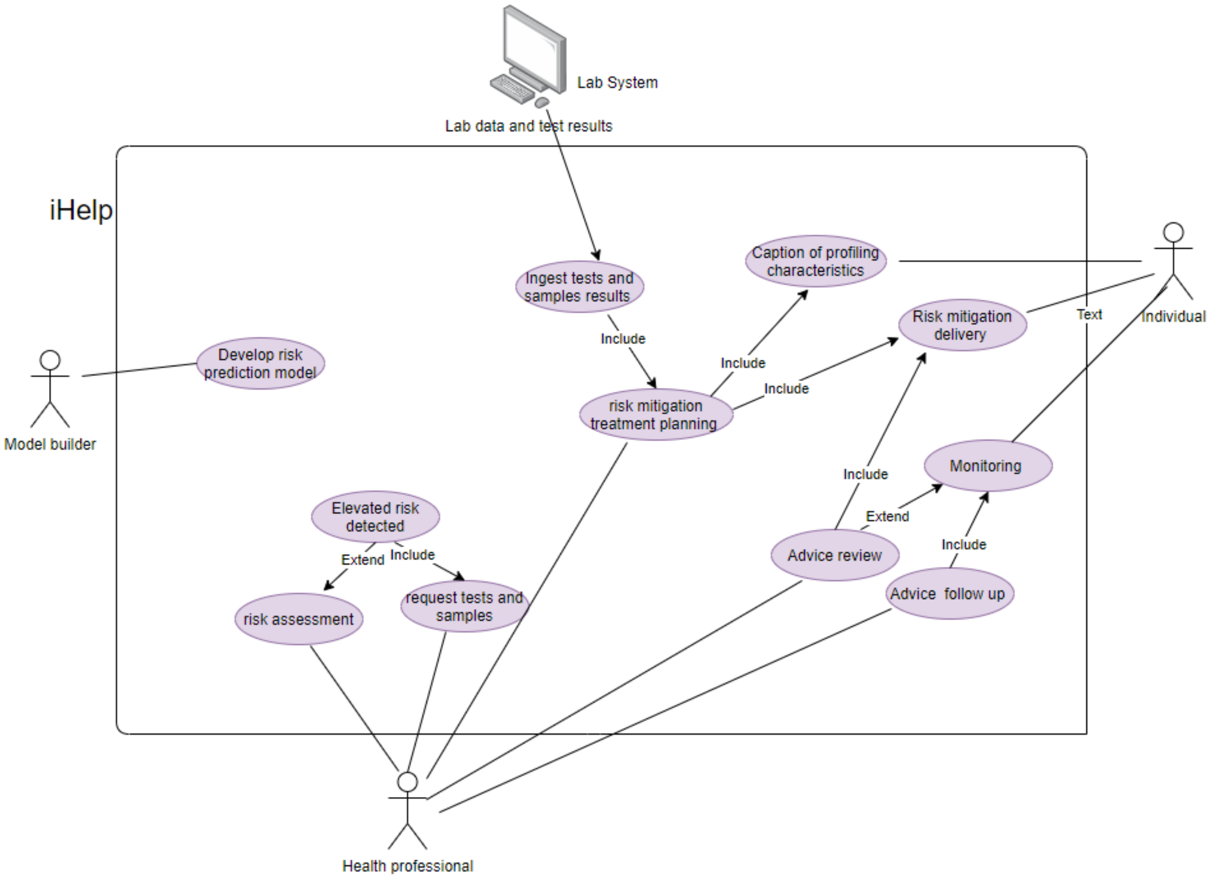


Figure 1: UNIMAN Pilot use case diagram

2.1.2 "As-Is vs To-Be" Scenario

The risk prediction functionality in the UNIMAN pilot will be developed using available larger datasets such as the UKBiobank[1]. The UNIMAN pilot will identify risk factors related to Pancreatic cancer and will also test and integrate a genetic/epigenetic predisposition score. Both components are essential to establish an accurate risk prediction model. The pilot will then implement the UNIMAN prediction model with eligible

[1] https://www.ukbiobank.ac.uk/

consenting participants within in a range of community health check settings including the National Health Service health check.

The developments sought in the iHelp project include the ability to notify the level of risk in different individuals and only high-risk subjects will further be guided to provide their biological sample for Omic assessment. All subjects will be directed to risk mitigation advice (iHelp). The interactive iHelp (the state-of-the-art approach) will have a built-in automated alert for feedback at regular appropriate intervals. The UNIMAN pilot will continue risk monitoring and mitigation for 6 months, while all data will be captured and will be used for analysis in the iHelp platform. The pilot will look for compliance rate and level of achievement compared to goals set for each participant. Specifically, for the high-risk group, the pilot will explore if by informing people on their relevant biology, will their risk mitigation activity be enhanced. At two time points in the follow-up, participants will be invited to use the risk prediction platform again to allow participants to monitor changes in their personalised risk before and after and their progress with their risk mitigation activities.

2.1.3 Description of scenarios

Table 2: UNIMAN Pilot – 1st Scenario

Section	Description
ID	SCE-P1-01
Title	Development of risk prediction (epidemiological and genetic factors)
Description	To develop a comprehensive cancer risk prediction model
Actors	Risk model-builders (Epidemiologist, statistician)
Objectives	To develop the risk prediction model and to validate the model.
Pre-conditions	Access to large scale dataset- UKBiobank, Office of National Statistic, CRUK incidence and mortality data, Albertas Tomorrow Data.
Process Dialog	Accessing the data, process the data using the statistical modelling package. Conversion of probability risk to absolute risk using the population registry data.
Variations	1st model epidemiological factors only- to be used as first line of risk assessment and only high-risk group will be further risk assessed using the extended model. The extended model with include epidemiological and genetic factors (to be used in high-risk group).
Post-conditions	Pancreatic Cancer
Diagrams	This scenario is covered in **Figure 1**
Issues and Notes	-

Table 3: UNIMAN Pilot – 2nd Scenario

Section	Description
ID	SCE-P1-02
Title	Capture of profiling (Personalisation) characteristics
Description	To develop questionnaire enabling to profile individuals according to the identified personal characteristics and behaviour journey stage
Actors	- Healthy individuals who participate in community health checks including the NHS-Health Check and with no history of any type of cancer - Model user- Health check staff
Objectives	To profile individuals according to the identified personal characteristics and behaviour journey stage
Pre-conditions	Access to the profiling rules to be developed in WP 5.2. and WP 5.3.
Process Dialog	Enabling individuals to fill in questionnaire
Variations	Individuals will be classified into different user groups, as dictated by the findings from WP 5.2 and 5.3. Different groups will be communicated risk in form of (1) increasing lifespan, or (2) reducing risk of future diseases/health decline
Post-conditions	High risk will be asked to provide biological sample for further omics-based assessments.
Diagrams	This scenario is covered in **Figure 1**
Issues and Notes	-

Table 4: UNIMAN Pilot – 3rd Scenario

Section	Description
ID	SCE-P1-03
Title	Risk assessment and risk communication
Description	To utilise a comprehensive cancer risk prediction model to motivate participants to maintain healthy lifestyle, engage more in healthy lifestyle
Actors	- Healthy individuals who participate in community health checks including the NHS-Health Check and with no history of any type of cancer - Model user- Health check staff
Objectives	To be able to communicate the assessed cancer risk in the most motivational way to the target participant
Pre-conditions	Risk prediction will be an online platform and registration for each participant is required so that identification number will be provided for future tracing, monitoring and communication.
Process Dialog	Data on individual risk factors to be presented/communicated as low, medium, and high risk. The algorithm will provide individualized risk result.
Variations	Recommended behaviours to be presented as a way to (1) increase one's lifespan or (2) reduce risk of future diseases/health decline.

Post-conditions	Low/medium risk individuals to be monitored on their behaviour change following the risk assessment.
Diagrams	This scenario is covered in **Figure 1**
Issues and Notes	-

Table 5: UNIMAN Pilot – 4th Scenario

Section	Description
ID	SCE-P1-04
Title	Communication of genetic and epigenetic measures in the high-risk group
Description	To communicate individual measurement on relevant omics-based assessments.
Actors	- Individuals profiled as having high risk - iHelp or healthcare practitioner
Objectives	To communicate genomics and epigenomic markers in order to motivate individuals most to change their behaviour
Pre-conditions	Access to the profiling rules to be developed in WP 5.2. and WP 5.3.
Process Dialog	–
Variations	Recommended behaviours to be presented as a way to (1) increase one's lifespan or (2) reduce risk of future diseases/health decline.
Post-conditions	High risk individuals to be monitored on their behaviour change following the risk assessment.
Diagrams	This scenario is covered in **Figure 1**
Issues and Notes	-

Table 6: UNIMAN Pilot – 5th Scenario

Section	Description
ID	SCE-P1-05
Title	Risk monitoring and behavioural feedback
Description	To facilitate risk mitigation using iHelp platform
Actors	- Individuals profiled as having high risk, intermediate and low iHelp work packages for AI communication and recommendation
Objectives	To provide feedback in a way that it is motivational and leads to habit building
Pre-conditions	iHelp platform for risk mitigation will be constructed and co-designed between WP6 pilto#1 and WP5. Profiling questionnaires need to be utilised (WP5) and feedback provided to iHelp users in the most motivational way.

Process Dialog	Participants will be using the platform to monitor/enter their activities, set targets and will receive constant feedback on those activities, as advised by the gamification literature.
Variations	Each mitigation package will be tailored to fit the suite of target sets for each participant based on their starting profile (pre-defined by pilot team).
Post-conditions	All data will be stored in iHelp central data storage for AI iterations.
Diagrams	This scenario is covered in **Figure 1**
Issues and Notes	-

2.1.4 Scenarios to requirements

The following tables contain the initial list of the requirements for the scenarios of this pilot that were described in the previous subsection.

Table 7: UNIMAN Pilot – 1st Requirement

Section	**Description**
ID	REQ-P1-01
Type	DATA
Short Name	Data access
Description & quantification	Access to large scale genomics and epidemiological data will be mandatory for the risk prediction of Pancreatic Cancer
Additional information	Large-scale genomics and epidemiological data can be accessed by data download from sources (we have permission to use the data) via designated portal, checksum method etc... These data are considered as raw data so the UNIMAN team will transform the data into workable format for example genotype data will be extracted only SNPs of interest for pancreatic cancer and then quality control will be in place. The data will then be ready for further processing. For epidemiological data, outliers will be diagnosed and meaningful code for each answer will be constructed, this procedure will make the dataset interpretable and analysable
Priority	MAN
Reference Scenarios	SCE-P1-01
Success Criteria	Successful access to the provided data

Table 8: UNIMAN Pilot – 2nd Requirement

Section	**Description**

ID	REQ-P1-02
Type	DATA
Short Name	Capture personality types
Description & quantification	Apply questionnaire to 700 participants to capture the data parameters that specifically enable the capture of personality types of participants
Additional information	-
Priority	MAN
Reference Scenarios	SCE-P1-02
Success Criteria	Successful profiling of participants

Table 9: UNIMAN Pilot – 3rd Requirement

Section	Description
ID	REQ-P1-03
Type	DATA
Short Name	Risk assessment
Description & quantification	Using innovative risk analysis algorithms to evaluate pancreatic cancer risk in 700 participants
Additional information	-
Priority	MAN
Reference Scenarios	SCE-P1-03
Success Criteria	Successful risk evaluation

Table 10: UNIMAN Pilot – 4th Requirement

Section	Description
ID	REQ-P1-04
Type	FUNC
Short Name	Risk communication

Description & quantification	Communication of omics profile back to participants and frame targeted/personalised messages to enhance the behavioural change. This will only apply to high-risk group (expect 10% of 700)
Additional information	-
Priority	MAN
Reference Scenarios	SCE-P1-04
Success Criteria	Behavioural change motivation enhancement by genetic and epigenetic risk feedback.

Table 11: UNIMAN Pilot – 5th Requirement

Section	Description
ID	REQ-P1-05
Type	FUNC
Short Name	Risk mitigation
Description & quantification	Usage of co-designed iHelp platform to provide feedback on targeted recommendations and level of activity
Additional information	-
Priority	MAN
Reference Scenarios	SCE-P1-05
Success Criteria	Successful implementation cycle and looping data for feedback at defined time interval.

2.2 Pilot #2 - Interventional Monocentric Study based on Patient Reported Outcomes

2.2.1 Goals and objectives

The Agostino Gemelli University Policlinic (FPG) pilot aims to perform a real-world data (RWD) analysis using a mobile application connected with Internet of Thing (IoT) devices to systematically acquire Patient Reported Experience Measures (PREMs) and Patient-Reported Outcome Measures (PROMs) for patients affected by pancreatic cancer with indication to radiotherapy, to predict outcomes and toxicity.

In this pilot, patients will be enrolled only with a histologically confirmed diagnosis of Pancreatic Cancer with the indication to a radiotherapy treatment. The risk evaluation of PC could not be assessed in this pilot. As a result of the pilot the risk of toxicity for patients affected by PC who underwent radiotherapy could be evaluated. Also, clinicians will invite users for an appointment if needed.

The system will share with clinician patient's lifestyle information and further elaboration and will integrate the current clinical information. The clinician will evaluate, according to the current clinical practice, any advice, or the need for a visit. Clinicians will provide oversight over every AI-sourced decision as the AI output will reach only the clinician and not the patient.

2.2.2 "As-Is vs To-Be" Scenario

Currently, clinical decisions regarding radiotherapy for patients affected by PC are based on information collected during the clinical examination (C., A., C., + 20) (mainly on laboratory (M., M., C., + 19) and instrumental examinations, objective assessments, and anamnestic interview). While important improvements in the technological point of view have been made to allows better outcomes (M., V., M., 07), (V., M., M., + 08), (M., M., C., + 12), (M., M., V., + 10), (M., B., N., + 20), (B., C., C., 19), (P., R., C., 20), just a few elements are considered to personalise the treatment and propose, for patients belonging to different risk category, a treatment suited on them. The omics technology is trying, on the one hand, to allows the building of solid predictive models (C., B., Y., + 21), but on the other hand the toxicity that can be obtained because of radiation treatment seems difficult to predict using a study approach based on information from usual research protocols. To improve this lack of information and predictive power and to allows a consequent personalization of the treatment, a systematic acquisition of RWD with continuous monitoring of parameters easily acquired through an application connected with IoT will be tested to evaluate patients with pancreatic neoplasia with radiotherapy indication to identify new toxicity risk factors or to early identify high-risk patients.

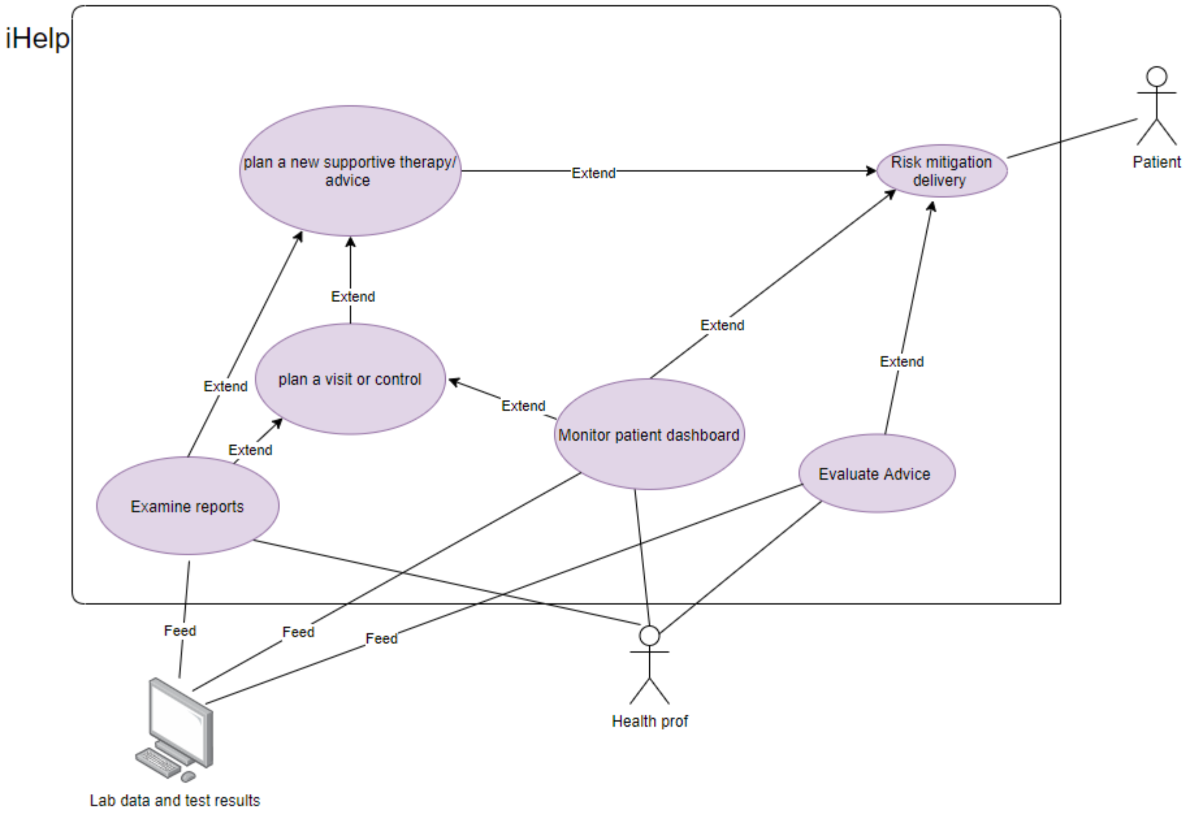


Figure 2: FPG Pilot use case diagram

2.2.3 Description of scenarios

Patients affected by PC with indication to undergo exclusive or adjuvant radiotherapy or chemoradiation, stereotactic body radiotherapy, radiotherapy for metastatic disease will be prospectively enrolled in this monocentric interventional study with device. Data regarding PROMs and PREMs will be acquired with the use of a mobile app and IoT devices, data regarding clinical outcomes, staging, blood, and radiological examinations will be used to identify different risk categories for outcomes and toxicity.

Table 12: FPG Pilot – 1st Scenario

Section	Description
ID	SCE-P2-01
Title	Interventional Monocentric Study based on Patient Reported Outcomes GENERATOR iHelp PROTECT – GENERATOR iHelp Patient Reported Outcomes to personalize radiotherapy for pancreatic cancer
Description	Patient-reported outcomes based real world data to perform a risk assessment of patients affected by Pancreatic Cancer
Actors	Radiation Oncologists, surgeons, data scientists and analysts, patients
Objectives	This pilot aims to perform a RWD analysis using a mobile application connected with IoT devices to systematically acquire PREMs and PROMs to predict outcomes and toxicity of patients affected by PC.

Pre-conditions	Technical specifications of App and IoT's, with CE certification, Ethical Committee approval, Patient informed consensus, GENERATOR infrastructure, App and IoT's
Process Dialog	Enroll patients with PC who will undergo radiotherapy treatments, acquire, and store real world data, finally evaluate results.
Variations	Patients with pancreatic cancer will be enrolled after the indication of: ▪ Exclusive radiotherapy or chemoradiation ▪ Stereotactic body radiotherapy ▪ Adjuvant radiotherapy or chemoradiation ▪ Palliative radiotherapy for primary tumour or nodal/distant metastasis
Post-conditions	-
Diagrams	This scenario is covered in 2
Issues and Notes	-

2.2.4 Scenarios to requirements

The following tables contain the initial list of the requirements for the scenarios of this pilot that were described in the previous subsection.

Table 13: FPG Pilot – 1st Requirement

Section	Description
ID	REQ-P2-01
Type	DATA
Short Name	Patient reported outcomes caption using IoT
Description & quantification	Data regarding correlation between outcomes (in terms of toxicity and survival outcomes) and PREMs/PROMs and RWD are not currently available. The knowledge of the correlation between those kinds of parameters and outcomes could play a role in the personalization of the treatment, therefore, we will use IoT devices to systematically collect RWD from patients during the treatment to identify the risk category of each patient. A wearable device to acquire: ▪ Steps [MAN] ▪ Climbs [MAN] ▪ energy expenditure [MAN] ▪ heart rate Min - Max - Resting heart rate - Average daily heart rate, automatically acquired from IoT [OPT/ENH]; or, if it is not possible, sporadic automatic heart rate acquisition and storage [MAN]

	▪ sleep metrics [MAN] ▪ blood oxygen saturation [DES] Automatic periodic (daily) measurement of blood oxygen saturation [DES]; or, if it is not possible, a manual acquisition and reporting by the patient [MAN] ▪ body temperature [OPT] ▪ blood pressure [OPT] ▪ exercise sessions autodetected by the device [OPT] NB: according to the need to allows patient compliance this feature is interesting only if it is possible to automatically acquire without any patient active action.
Additional information	For ethical committee also the following documents are needed: ▪ CE certificate: IoT ▪ Technical docs: IoT
Priority	Previously reported point by point ▪ MAN: Mandatory requirement ▪ DES: Desirable requirement ▪ OPT: Optional requirement ▪ ENH: Possible future enhancement
Reference Scenarios	SCE-P2-01
Success Criteria	Previously specified in the comments point by point where necessary

Table 14: FPG Pilot – 2nd Requirement

Section	**Description**
ID	REQ-P2-02
Type	DATA
Short Name	iHelp Mobile Application - Questionnaire
Description & quantification	▪ BMI [MAN] The patient periodically enters his/her weight and having entered the height at registration, BMI can be computed. ▪ EORTC QLQ-C30 questionnaire [MAN] Authorization already acquired from EORTC ▪ EORTC QLQ-PAN26 questionnaire [MAN] Authorization already acquired from EORTC ▪ COVID-19 prevention daily questionnaire [MAN]

	- eventual psychological condition extracted from social media [OPT] NB: according to the need to allows patient compliance this feature is interesting only if it is possible to automatically acquire without any patient active action. - possible smartphone utilization features automatically acquired that can predict psychological assessment (number of messages, number of phone calls, minutes of phone calls, number of whatsapp, number of SMS) [OPT] NB: according to the need to allows patient compliance this feature is interesting only if it is possible to automatically acquire without any patient active action.
Additional information	For ethical committee submission and consequent approval, the following documents are needed: - CE certificate: Healthentia app - Technical docs: Healthentia app
Priority	Previously specified point by point - MAN: Mandatory requirement - DES: Desirable requirement - OPT: Optional requirement - ENH: Possible future enhancement
Reference Scenarios	SCE-P2-01
Success Criteria	Previously specified in the comments point by point where necessary

Table 15: FPG Pilot – 3rd Requirement

Section	Description
ID	REQ-P2-03
Type	FUNC
Short Name	Targeted advice provided by iHelp platform
Description & quantification	Both personal dashboard and app-advice are sensitive issues for - Personal dashboard of activities [DES] NB: no data regarding other patients (also in aggregate form) must be proposed because no benchmark of ideal activities have been previously reported in scientific literature for patients who will undergo radiotherapy. Global inter-patient statistics or gamification features must be made available only for MDs. - Possibility to communicate to the patient and to see if the patient read the content of the message [DES]

	NB: good clinical practice advice will be proposed by clinicians to the patient. AI based advice will be proposed first to the clinician and, if coherent with a good clinical practice recommendation, will be proposed to the patient.
Additional information	For ethical committee submission and consequent approval, the following documents are needed: ▪ CE certificate: Healthentia app ▪ Technical docs: Healthentia app
Priority	Previously specified point by point ▪ MAN: Mandatory requirement ▪ DES: Desirable requirement ▪ OPT: Optional requirement ▪ ENH: Possible future enhancement
Reference Scenarios	SCE-P2-01
Success Criteria	Previously specified in the comments point by point where necessary

Table 16: FPG Pilot – 4th Requirement

Section	**Description**
ID	REQ-P2-04
Type	DATA
Short Name	Additional data obtained during current clinical practice
Description & quantification	Data collected using IoT and mobile applications will be integrated with data obtained during the current clinical practice (real world data), to allow a better and holistic evaluation of the patient. Clinical database containing data regarding clinical and pathological staging, blood examination before, during and after radiotherapy, surgery and possible complications, chemotherapy, radiotherapy, toxicity outcomes.
Additional information	-
Priority	[MAN]
Reference Scenarios	SCE-P2-01
Success Criteria	Clinical database maintained by clinicians (radiation oncologist and surgeons)

2.3 Pilot #3 - Study of Lifestyle Choices on Elevating the Risk Factors for Pancreatic Cancer

2.3.1 Goals and objectives

The main objective of the Hospital de Dénia-Marina Salud (HDM) pilot is to obtain relationships between the known risks that science has currently identified and to discover new ones if there are new factors that can affect negative or positive modulation. It is currently known that a risk factor is anything that increases the chance of getting a disease such as cancer. Different cancers have different risk factors. Some risk factors, like smoking, can be changed in contrast with others, like a person's age or family history that cannot be changed. In some cases, there might be a factor that may decrease the risk of developing cancer or has an unclear effect. That is not considered a risk factor, but they may be noted clearly on this page as well. Having a risk factor, or even many, does not mean that you will get cancer. And some people who get cancer may have few or not known risk factors.

In the HDM pilot, a group of patients made up of various populations will be selected, but they will mainly be divided between patients diagnosed with pancreatic cancer and patients who have not been diagnosed but have certain risk conditions. Therefore, these two populations will be invited by two individual teams of doctors, the first by oncology specialists and the second by family doctors.

When patients agree to participate in the study, they will be provided with the clinical and technological devices to monitor the parameters to be studied. It has been defined that a group of doctors will monitor the evolution of the patients, so that consultations will be carried out (virtual or face-to-face for personal evaluation). In the case of patients diagnosed and treated, it will be studied whether lifestyle habits improve the patient's situation both at the level of general or emotional health status. In the case of undiagnosed patients, the clinical variables will be studied together with the habits, which will be compared with the variables available in the primary data of diagnosed patients in order to obtain risk factors and modify these when possible. For this, it must be considered that today it is known that the risk factors can be classified as like in the table below:

Table 17: **Rik factors can be classified in the following three categories.**

Risk factors that can be changed	Risk factors that cannot be changed	Risk factors with unclear effect on risk
▪ Tobacco use Smoking is one of the most important risk factors for pancreatic cancer. The risk of getting pancreatic cancer is about twice as high among smokers compared to those who have never smoked. About 25% of pancreatic cancers are thought to be caused by cigarette smoking. Cigar smoking and the use of smokeless tobacco products also increase the risk. However, the risk of pancreatic cancer starts to drop once a person	▪ Age The risk of developing pancreatic cancer goes up as people age. Almost all patients are older than 45. About two-thirds are at least 65 years old. The average age at the time of diagnosis is 70.	▪ Diet Diets with red and processed meats (such as sausage and bacon) and saturated fats may increase the risk of pancreatic cancer. Sugary drinks may also increase this risk. More research is needed in this area.

stops smoking. See Can Pancreatic Cancer Be Prevented?		
▪ Being overweight Being very overweight (obese) is a risk factor for pancreatic cancer. Obese people (body mass index [BMI] of 30 or more) are about 20% more likely to develop pancreatic cancer. Gaining weight as an adult can also increase risk. Carrying extra weight around the waistline may be a risk factor even in people who are not very overweight.	▪ Gender Men are slightly more likely to develop pancreatic cancer than women. This may be due, at least in part, to higher tobacco use in men, which raises pancreatic cancer risk (see above).	▪ Physical inactivity Some research has suggested that lack of physical activity might increase pancreatic cancer risk. But not all studies have found this. Regular physical activity may help reduce the risk of pancreatic cancer.
▪ Diabetes Pancreatic cancer is more common in people with diabetes. The reason for this is not known. Most of the risk is found in people with type 2 diabetes. This type of diabetes is increasing in children and adolescents as obesity in these age groups also rises. Type 2 diabetes in adults is also often related to being overweight or obese. It's not clear if people with type 1 (juvenile) diabetes have a higher risk.	▪ Race African Americans are slightly more likely to develop pancreatic cancer than whites. The reasons for this are not clear, but it may be due in part to having higher rates of some other risk factors for pancreatic cancer, such as diabetes, smoking, and being overweight.	▪ Coffee Some older studies have suggested that drinking coffee might increase the risk of pancreatic cancer, but more recent studies have not confirmed this.
▪ Chronic pancreatitis Chronic pancreatitis, a long-term inflammation of the pancreas, is linked with an increased risk of pancreatic cancer. Chronic pancreatitis is often seen with heavy alcohol use and smoking.	▪ Family history Pancreatic cancer seems to run in some families. In some of these families, the high risk is due to an inherited syndrome (explained below). In other families, the gene causing the increased risk is not known. Although family history is a risk factor, most people who get pancreatic cancer do not have a family history of it.	▪ Infections Some research suggests that infection of the stomach with the ulcer-causing bacteria Helicobacter pylori (H. pylori) or infection with Hepatitis B may increase the risk of getting pancreatic cancer. More studies are needed.
▪ Workplace exposure to certain chemicals Heavy exposure at work to certain chemicals used in the dry cleaning and metal working industries may raise a person's risk of pancreatic cancer.	▪ Inherited genetic syndromes. Inherited gene changes (mutations) can be passed from parent to child. These gene changes may cause as many as 10% of pancreatic cancers. Sometimes these changes result in syndromes that include increased risks of other cancers (or other	

	health problems). Examples of genetic syndromes that can cause pancreatic cancer include: - Hereditary breast and ovarian cancer syndrome, caused by mutations in the BRCA1 or BRCA2 genes - Hereditary breast cancer, caused by mutations in the PALB2 gene. - Familial atypical multiple mole melanoma (FAMMM) syndrome, caused by mutations in the p16/CDKN2A gene and associated with skin and eye melanomas. - Familial pancreatitis, usually caused by mutations in the PRSS1 gene. - Lynch syndrome, also known as hereditary non-polyposis colorectal cancer (HNPCC), most often caused by a defect in the MLH1 or MSH2 genes - Peutz-Jeghers syndrome, caused by defects in the STK11 gene. This syndrome is also linked with polyps in the digestive tract and several other cancers.	
- Alcohol Some studies have shown a link between heavy alcohol use and pancreatic cancer. Heavy alcohol use can also lead to conditions such as chronic pancreatitis, which is known to increase pancreatic cancer risk.	- Chronic pancreatitis (due to a gene change) Chronic pancreatitis is sometimes due to an inherited gene mutation. People with this inherited (familial) form of pancreatitis have a high lifetime risk of pancreatic cancer.	

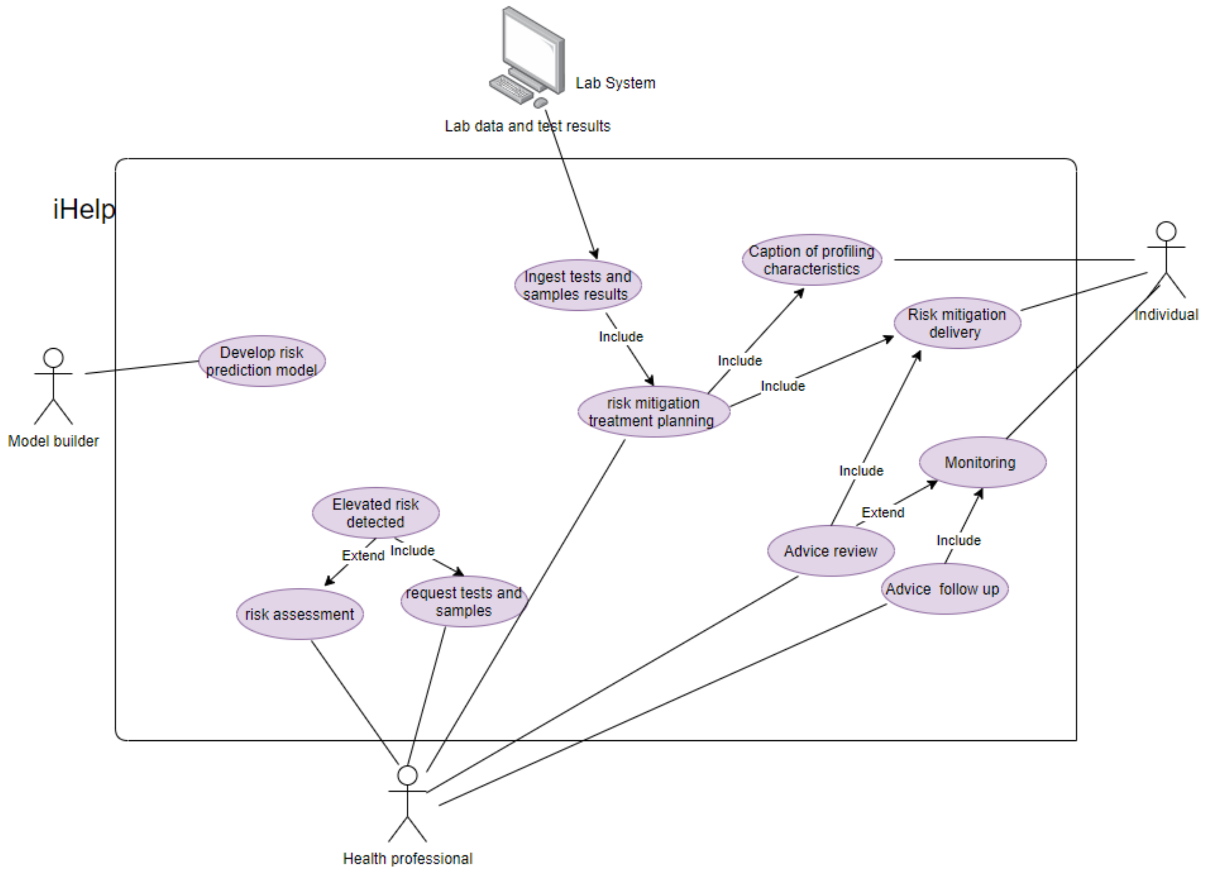


Figure 3: HDM Pilot use case diagram

2.3.2 "As-Is vs To-Be" Scenario

Marina Salud started in 2009 serving a population of 180,000 inhabitants with an ambitious commitment to the implementation of an Electronic Health Record (EHR) that configures a paperless environment and that provides a clinical database with a level of detail that has not been exploited from the point of view of clinical research.

Therefore, a primary database is available, where it is planned to be applied artificial intelligence algorithms to search for patterns that identify risk factors for pancreatic cancer. Those primary data will also be complemented with a secondary data set from the implementation of a trial on a set of diagnosed, risk and healthy patients.

These combined datasets and types of data, coupled with the indications on the modification of habits that the doctors will be able to recommend to the patients will incorporate information into the database on which the algorithms will have to search and identify the relationships.

Not only is the case of success in the study of pancreatic cancer sought, but the definition of a model that can be extrapolated to any other pathology from the technological and organizational models implemented, so that small organizations such as Marina Salud can carry out research, which is currently only possible in large University Hospitals.

2.3.3 Description of scenarios

Table 18: HDM Pilot – 1st Scenario

Section	Description
ID	SCE-P3-01
Title	Study of Lifestyle Choices on Elevating the Risk Factors for Pancreatic Cancer in non-cancer patients.
Description	Life habits have an influence on the suffering of certain pathologies, but the relationship between these life habits and diseases are not correlated due to the complexity that this entails. In this pilot we are going to correlate life habits and their effect as risk factors
Actors	Model Builder, Primary Care Clinician, Cancer Free Patient
Objectives	Introduce lifestyle habits in the HHR model that allows an analysis of the correlation between these and the risk of suffering from the disease and / or its evolution if it is already diagnosed.
Pre-conditions	Technical specifications of App and IoT's, with CE certification, Ethical Committee approval, Patient informed consensus, infrastructure, App and IoT's
Process Dialog	Incorporate patients, train them in technology, follow evolution, apply AI algorithms to the HHR database to find correlation, suggest changes in habits, evaluate evolution.
Variations	-
Post-conditions	-
Diagrams	This scenario is represented in the **Figure 3**
Issues and Notes	-

Table 19: HDM Pilot – 2nd Scenario

Section	Description
ID	SCE-P3-02
Title	Study of Lifestyle Choices on Elevating the Risk Factors for Pancreatic Cancer in diagnosed patients.
Description	Life habits have an influence on the suffering of certain pathologies, but the relationship between these life habits and diseases are not correlated due to the complexity that this entails. In this pilot we are going to correlate life habits and their effect as risk factors
Actors	Model Builder, Oncologists, Cancer Patient Diagnosed

Objectives	Introduce lifestyle habits in the HHR model that allows an analysis of the correlation between these and the risk of suffering from the disease and / or its evolution if it is already diagnosed.
Pre-conditions	Technical specifications of App and IoT's, with CE certification, Ethical Committee approval, Patient informed consensus, infrastructure, App and IoT's
Process Dialog	Incorporate patients, train them in technology, follow evolution, apply AI algorithms to the HHR database to find correlation, suggest changes in habits, evaluate evolution.
Variations	-
Post-conditions	-
Diagrams	This scenario is represented in the **Figure 3**
Issues and Notes	-

2.3.4 Scenarios to requirements

The following tables contain the initial list of the requirements for the scenarios of this pilot that were described in the previous subsection.

Table 20: HDM Pilot – 1st Requirement

Section	Description
ID	REQ-P3-01
Type	DATA
Short Name	iHelp Monitoring (IoT)
Description & quantification	It represents the monitoring carried out towards the patient through all the information followed by the same (tests, questionnaires, activity data) through graphs and analytics of different types. A wearable device to acquire: ▪ Steps [MAN] ▪ Climbs [MAN] ▪ energy expenditure [MAN] ▪ exercise sessions autodetected by the device [OPT] The physical activity data should be collected daily and only once at the end of the day, at night. ▪ heart rate Min - Max - Resting heart rate - Average daily heart rate, automatically acquired from IoT [OPT/ENH]; or, if it is not possible, sporadic automatic heart rate acquisition and storage [MAN]

	▪ sleep metrics [MAN] Standard sleep information (Start, End, Sleep Stages) will be collected daily. ▪ blood oxygen saturation [DES] Automatic periodic (daily) measurement of blood oxygen saturation [DES]; or, if it is not possible, a manual acquisition and reporting by the patient [MAN] ▪ body temperature [OPT] ▪ blood pressure [OPT] ▪ weight NB: according to the need to allows patient compliance this feature is interesting only if it is possible to automatically acquire without any patient active action. Data about the body's signals should be collected weekly. Weight, blood pressure and body temperature are collected by manual entries and blood oxygen saturation automatically, if possible. 30 devices are required; 15 for healthy patients and 15 for diagnosed patients.
Additional information	For ethical committee also the following documents are needed: ▪ CE certificate: IoT ▪ Technical docs: IoT
Priority	Previously reported point by point ▪ MAN: Mandatory requirement ▪ DES: Desirable requirement ▪ OPT: Optional requirement ▪ ENH: Possible future enhancement
Reference Scenarios	SCE-P3-01, SCE-P3-02
Success Criteria	Incorporation of information in the HHR and application of AI algorithms.

Table 21: HDM Pilot – 2nd Requirement

Section	**Description**
ID	REQ-P3-02
Type	DATA
Short Name	iHelp Questionnaire collection (Capture of Profiling characteristics)
Description & quantification	Since the study will be carried out in two / three groups of patients, it will be necessary to have different types of questions since the starting situation and the objectives will be different.

Undiagnosed patients, it is proposed to assess lifestyle habits, while for diagnosed patients it is proposed to add questions that relate lifestyle habits to the subjective evolution of the disease.

For the collection of diet information, the frequency in which the data are collected and the food groups to be reported by users are proposed:

Frequency: Data on the patient's diet should be collected daily, only once at the end of the day, at night.

Proposed food groups and measures:

- Fruits - In pieces
- Meats (Chicken, Veal, Pork, Lamb ...) - In pieces
- Eggs - In pieces
- Fish - In pieces
- Pasta, rice, potatoes ... - In approximate grams
- Bread, cereals ... - In approximate grams
- Vegetables - In pieces
- Legumes - In approximate grams
- Cold Cuts - In pieces
- Dairy Products (Milk, Cheese, Yogurt ...) - In pieces
- Sweets (Cookies, Pastries, Jam ...) - In pieces
- Sugary Soft Drinks - Approximate Cups

Data about the lifestyle of patients should be collected weekly. The following parameters should be monitored through a weekly questionnaire:

- Tobacco Consumption - Packs / Week
- Consumption of alcoholic beverages - Cups / Week
- Consumption of other drugs

In the case of the collection of information on symptoms, only the need to incorporate a questionnaire on a specific scenario is reported (SCE-P3-02). Chemotherapy symptoms will be reported in each cycle for diagnosed patients.

For patient enrolment (initial questionnaire), only the information of the SIP number will be needed. The SIP number is a patient identification number that we are interested in having stored in the patient's information. This will be used to make requests for information to the API that will provide the secondary data information. This number already exists for each patient, they only have to enter it in their registration to the APP, it is a number that can always contain a maximum of 8 numerical digits.

For scheduled questionnaires, they must collect the information weekly. Standard questionnaires:

- Dietary reporting (Type of food and amount eaten daily)
- Lifestyle aspects (Packs of tobacco and cups of alcoholic beverages weekly)
- Psychological Information. For our pilot it is proposed to collect the state of mind, in a general way. It can be collected by means of a widget with faces that represent moods for example.

Additional information	For ethical committee submission and consequent approval, the following documents are needed: ▪ CE certificate: Healthentia app ▪ Technical docs: Healthentia app The application must be in Spanish, like all interfaces, questionnaires, etc
Priority	Previously specified point by point ▪ MAN: Mandatory requirement ▪ DES: Desirable requirement ▪ OPT: Optional requirement ▪ ENH: Possible future enhancement
Reference Scenarios	SCE-P3-01, SCE-P3-02
Success Criteria	Incorporation of information in the HHR and application of AI algorithms.

Table 22: HDM Pilot – 3rd Requirement

Section	Description
ID	REQ-P3-03
Type	FUNC
Short Name	iHelp Advice (Risk Mitigation Delivery)
Description & quantification	Both personal dashboard and app-advice are sensitive issues for ▪ Personal dashboard of activities [DES] NB: no data regarding other patients (also in aggregate form) have to be proposed because no benchmark of ideal activities have been previously reported in scientific literature for patients who will undergo radiotherapy. Global inter-patient statistics or gamification features must be made available only for MDs. ▪ Possibility to communicate to the patient and to see if the patient read the content of the message [DES] NB good clinical practice advice will be proposed by clinicians to the patient. AI based advice will be proposed first to the clinician and, if coherent with a good clinical practice recommendation, will be proposed to the patient.
Additional information	For ethical committee submission and consequent approval, the following documents are needed: ▪ CE certificate: Healthentia app ▪ Technical docs: Healthentia app
Priority	Previously specified point by point ▪ MAN: Mandatory requirement

	▪ DES: Desirable requirement ▪ OPT: Optional requirement ▪ ENH: Possible future enhancement
Reference Scenarios	SCE-P3-01, SCE-P3-02
Success Criteria	Communication of recommendations to clinicians for use with patients, the device is not considered

Table 23: HDM Pilot – 4th Requirement

Section	**Description**
ID	REQ-P3-04
Type	FUNC
Short Name	Secondary data integration (Ingest Test and Samples)
Description & quantification	Laboratory tests must be generated. Once here, all the analyses or tests that have been carried out must be inserted into the system. Several types of laboratory tests must be able to be inserted. The clinical information collected by the platform must be able to be transferred to the HCE of Marina Salud in parallel to the incorporation into the HHR. The information will be requested through the REST API that is provided to retrieve the secondary data information. This will be done through HTTP requests and the agreement of a patient identifier that can serve us once the data travels outside the platform (which in the case of our pilot will be the SIP number, for our patients).
Additional information	-
Priority	[MAN]
Reference Scenarios	SCE-P3-01, SCE-P3-02
Success Criteria	Integration of data from Healthentia to Marina Salud in a standard and secure form.

Table 24: HDM Pilot – 5th Requirement

Section	**Description**
ID	REQ-P3-05
Type	FUNC
Short Name	Develop Risk Prediction Model

Description & quantification	It represents the entire set of information that will be stored in the database and that can also be accessed, applying filters and it will be displayed in the form of data graphics. Artificial intelligence algorithms may be used to visualize the data with graphs.
Additional information	-
Priority	[MAN]
Reference Scenarios	SCE-P3-01, SCE-P3-02
Success Criteria	To be able to visualize all the necessary information of the data model, to be able to search or filter information of interest and also represent this information in different types of graphs. Type of non-analytical information that must be supported and stored about the patient: Personal-family history. Types: ▪ Diabetes ▪ Hypertension ▪ Rotor syndrome ▪ Dubin-Johnson syndrome ▪ Gilbert's disease ▪ Peutz-Jeghers syndrome ▪ Familial pancreatitis. ▪ Sdr. Lynch, ▪ Mutations in the BRCA-1 or BRCA-2 gene. Life habits: ▪ Smoker (packs / year) ▪ Alcoholic intake (grams / day) ▪ Type of diet (vegetarian, vegan, everything) ▪ Exercise (days per week) ▪ Level of studies ▪ Work environment Physical examination parameters: ▪ Body mass index ▪ Abdominal circumference ▪ Weight ▪ Blood Pressure

Table 25: HDM Pilot – 6th Requirement

Section	Description
ID	REQ-P3-06
Type	FUNC

Short Name	Elevated Risk Detected
Description & quantification	It represents the entire set of artificial intelligence algorithms that will be able to process the entire data set acquired (Model Data, Primary Data, Secondary Data and medical evaluations) in order to detect high levels of risk of suffering from cancer.
Additional information	-
Priority	[MAN]
Reference Scenarios	SCE-P3-01
Success Criteria	These algorithms must be trained and able to find patterns that suggest that patients have a high probability of suffering from cancer.

Table 26: HDM Pilot – 7th Requirement

Section	**Description**
ID	REQ-P3-07
Type	FUNC
Short Name	Clinical risk assessment.
Description & quantification	It represents the advice by the clinician with whom the patient is treated, and a risk assessment is contemplated, once a high risk is detected from artificial intelligence patterns.
Additional information	-
Priority	[MAN]
Reference Scenarios	SCE-P3-01
Success Criteria	-

Table 27: HDM Pilot – 8th Requirement

Section	**Description**
ID	REQ-P3-08
Type	ENV
Short Name	Requests Tests and Samples

Description & quantification	Once these abnormal risk levels are detected for a patient, the clinician must order the entire set of samples or tests necessary for the patient.
Additional information	-
Priority	[MAN]
Reference Scenarios	SCE-P3-01
Success Criteria	The following types of analytical tests must be able to be integrated and ingested by the system: ▪ Hemogram ▪ Glycosylated Hemoglobin ▪ Bilirubin (Direct / indirect) ▪ Transaminases (AST; ALT; GGT; FA; LDH) ▪ Kidney function (Cr; Glomerular filtration) ▪ Cholesterol (Total, LDL, HDL) ▪ Urea ▪ Tumor marker CA19.9

Table 28: HDM Pilot – 9th Requirement

Section	Description
ID	REQ-P3-09
Type	ENV
Short Name	Risk Mitigation/ Treatment Planning
Description & quantification	Once all the data referring to laboratory tests have been processed, as well as primary data on activity or information acquired from the questionnaires, the mitigation of risks for the patient and the definition of a specific treatment plan must be considered.
Additional information	-
Priority	[MAN]
Reference Scenarios	SCE-P3-01, SCE-P3-02
Success Criteria	-

Table 29: HDM Pilot – 10th Requirement

Section	Description
ID	REQ-P3-10

Type	ENV
Short Name	Advice Follow-up
Description & quantification	Represents: once it is detected that the patient is at risk, the clinician should initiate a follow-up of the same, in order to advise the patient through monitoring.
Additional information	-
Priority	[MAN]
Reference Scenarios	SCE-P3-01, SCE-P3-02
Success Criteria	-

Table 30: HDM Pilot – 11th Requirement

Section	**Description**
ID	REQ-P3-11
Type	ENV
Short Name	Advice Review
Description & quantification	It represents when the patient is already being offered and performing the entire set of advice by the clinicians, and even so abnormalities are detected and reported to the clinician. The clinician must contact the patient either for a review or to know what is happening, in order to know what measures should be taken.
Additional information	-
Priority	[MAN]
Reference Scenarios	SCE-P3-01, SCE-P3-02
Success Criteria	-

2.4 Pilot #4 - Study of Risk, Personalised Recommendations and Measures to Raise Awareness of Relevant Factors

2.4.1 Goals and objectives

The main objective of the Medical University Plovdiv (MUP) pilot is to incorporate the AI aspects provided by iHelp into the daily medical practice. The iHelp seeks to facilitate the decision-making process into the early diagnosis of those conditions that are elevating the health risk level for Pancreatic cancer development, as well as diagnosing the Pancreatic cancer in the stage 1 or at least stage 2 of its development. The early detection of the above-mentioned conditions will support the healthcare providers into building-up a personalised approach to every individual at risk - recommendation regarding required changes into lifestyle, diet, consultation, and medication etc. The individuals into the pilot will be notified by the iHelp platform only with a notice like “there is a need to visit your healthcare provider”. The clinical doctor is the one to be notified by the iHelp about the elevated risk level for Pancreatic cancer development. Afterwards the clinicians are to inform the patient at risk or with diagnose regarding the required procedures to be followed. The system will issue advice to the clinician about the required changes and proposed measures regarding the lifestyle, behaviour, diet, regime, tests, treatment etc. The clinical doctor is the one to decide to what extent the system recommendation is useful for every case.

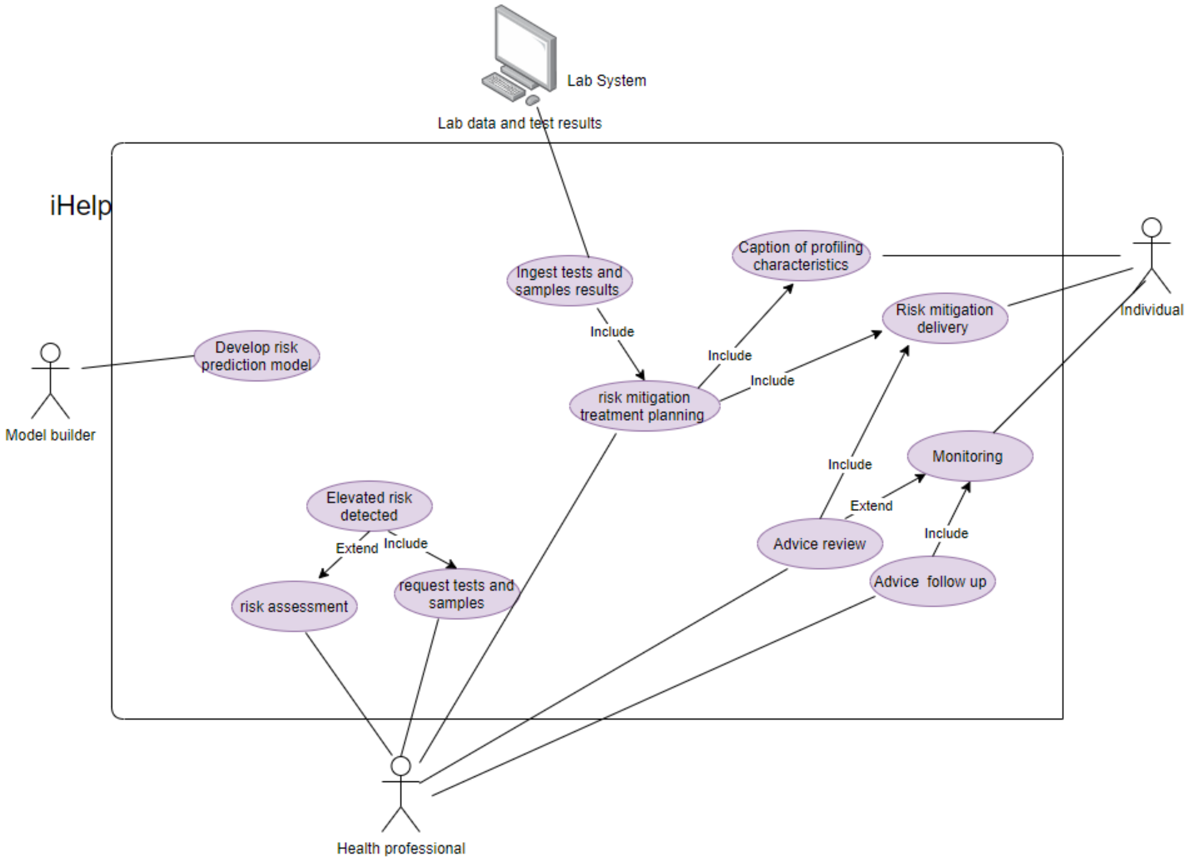


Figure 4: MUP Pilot use case diagram

2.4.2 “As-Is vs To-Be” Scenario

The capabilities of the system to collect, analyse and process the data extracted from the various sources - patients complains, test results, family history, comorbidities etc and to compare the findings to the

Pancreatic cancer patterns in order to assess the cancer development risk level is to increase immensely the Pancreatic cancer prevention. The patients' data storage in a particular file is a prerequisite for constant follow-up of the patients' status, as well to monitor every single patient at risk adherence to the prescribed program.

The system has to add every new data to the previous stored and to reassess the health risk for Pancreatic cancer development. The latter introduces a novel capacity speeding-up the patients at risk monitoring and clinical assessment.

2.4.3 Description of scenarios

Different healthcare providers related to the diagnosing, monitoring and treatment of the patients at risk and with developed Pancreatic cancer - general practitioners, specialists in internal medicine, gastroenterology, endocrinology, general surgery, oncology, registered nurses will be included as end-users. These medical professionals will be asked to quantify the weight of the prognostic risk factors associated the elevated risk for cancer development, as well the factors related to clinical staging and development of the cancer itself. Their recommendations and feed-back will be valuable regarding the structure and impact of iHelp on clinical practice.

Table 31: MUP Pilot – 1st Scenario

Section	Description
ID	SCE-P4-01
Title	Pancreatic cancer risk assessment
Description	Identification and assessment of Pancreatic cancer risk factors based on the study of complains, comorbidities, clinical findings, history, tests' results evaluation, and cancer risk level assessment
Actors	Medical Professionals ▪ Risk model-builders (Specialists in general practitioners, specialists in internal medicine, gastroenterology, endocrinology, general surgery, oncology,) ▪ Model -users (Clinician's consultants in gastroenterology, oncology specialists and General Practitioners) ▪ Model-users (Nurses and Paramedical)
Objectives	To facilitate the decision-making process into the early diagnosis of those conditions that are elevating the health risk level for Pancreatic cancer development, as well as diagnosing the Pancreatic cancer at its early stages.
Pre-conditions	Access to various data - from patients, from healthcare provider, from testing
Process Dialog	Accessing the diverse data sources, extracting them for analyses and processing for risk level assessment. Finally evaluate and explore the results in the visual analytic tool to be presented in an appropriate manner to the clinicians.
Variations	Risk levels for development of the Pancreatic cancer or for deterioration of already developed cancer
Post-conditions	

Diagrams	This scenario is covered in 4 & Figure 5
Issues and Notes	

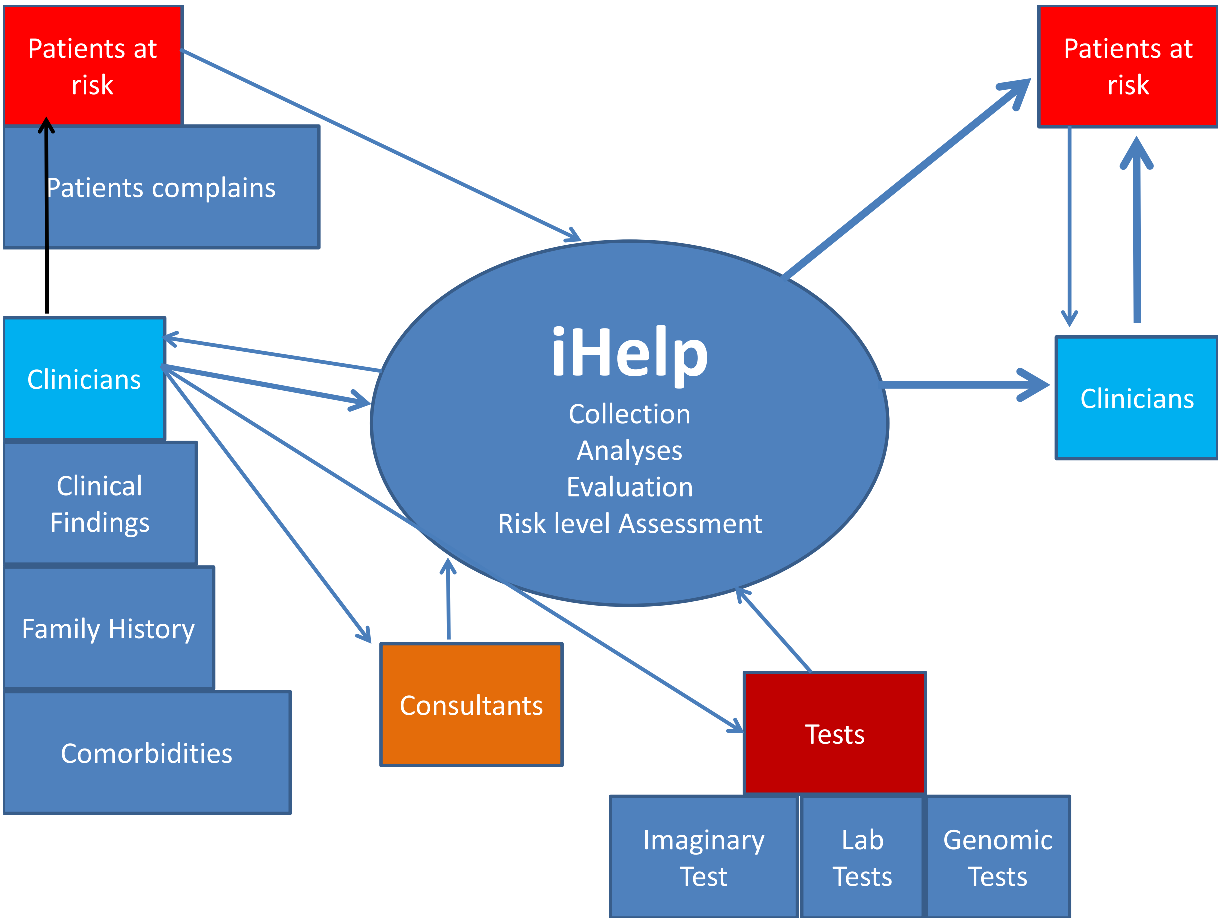


Figure 5: MUP Pilot Scenario diagram

2.4.4 Scenarios to requirements

The following tables contain the initial list of the requirements for the scenarios of this pilot that were described in the previous subsection.

Table 32: MUP Pilot – 1st Requirement

Section	**Description**
ID	REQ-P4-01
Type	DATA
Short Name	Data assessment

Description & quantification	Access to and evaluation of large amount data from patients' complains is a prerequisite for Pancreatic cancer risk assessment
Additional information	The data will be stored at Medical University Plovdiv and send after anonymization
Priority	MAN
Reference Scenarios	SCE-P4-01
Success Criteria	Collection of the specific data from patients complains and patients' records

Table 33: MUP Pilot – 2nd Requirement

Section	Description
ID	REQ-P4-02
Type	DATA
Short Name	Data assessment
Description & quantification	Access to and evaluation of large amount data from the clinical findings during the examinations of the patients as a prerequisite for Pancreatic cancer risk assessment
Additional information	
Priority	MAN
Reference Scenarios	SCE-P4-01
Success Criteria	Collection of the specific data from examinations

Table 34: MUP Pilot – 3rd Requirement

Section	Description
ID	REQ-P4-03
Type	DATA
Short Name	Data evaluation
Description & quantification	Access to and evaluation of large amount data from consultations and laboratory and imaginary tests as a prerequisite for Pancreatic cancer risk assessment

Additional information	
Priority	MAN
Reference Scenarios	SCE-P4-01
Success Criteria	Collection of the specific data from consultations and testing

Table 35: MUP Pilot – 4th Requirement

Section	**Description**
ID	REQ-P4-04
Type	DATA
Short Name	Questionnaire’s data evaluation
Description & quantification	Access to and evaluation of large amount data from patients regarding prescribed regime outcomes: ▪ Quality of life questionnaire ▪ Pain level ▪ Body weight and temperature ▪ Alcohol consumption ▪ Red and processed meat consumption ▪ Vegetables and grains consumption ▪ Daily Medicines
Additional information	
Priority	MAN (Mandatory)
Reference Scenarios	SCE-P4-01
Success Criteria	Collection of the specific data from the patients

Table 36: MUP Pilot – 5th Requirement

Section	**Description**
ID	REQ-P4-05
Type	DATA
Short Name	Recommendations and mitigation

Description & quantification	Access to and evaluation of the proposed by the iHelp program for risk reduction measures to be recommended to the patient to follow in order to prevent development or deterioration of the Pancreatic cancer
Additional information	
Priority	MAN
Reference Scenarios	SCE-P4-01
Success Criteria	Collection of the specific data from iHelp

2.5 Pilot #5 - Study of Improved Risk Prediction Models and Targeted Interventions that can Delay the Onset of Cancer

2.5.1 Goals and objectives

The objective of the Taipei Medical University (TMU) pilot is to predict high risk individuals towards pancreatic and liver cancer for early-stage management of the disease and further explore the effect of digital therapeutics solutions among these individuals.

Thus, the goals include:

- Applying data analytics, AI based algorithms or deep learning technologies to analyse electronic health record data and predict high risk individuals towards pancreatic and liver cancer for early-stage management of the disease.
- Conducting a digital trial, an observational study, for mobile applications for personalized healthcare and improved quality of life, among the high-risk individuals.

The digital trial will be focused to comply with the unmet needs for supportive care among high-risk individuals in our case- liver and pancreatic cancer.

In TMU pilot individual users will not be informed directly. However, clinicians will notify the individual users about elevated risk of developing a disease. TMU pilot digital solution (mobile app) will provide dashboard of symptom status to the clinicians to keep them informed regarding the elevated risk for the users and further monitor or investigate them for the diagnosis of the disease.

The system, which is in the form of a dashboard with visualization on the mobile phones of users, will be used by the clinicians in order to monitor them. Users would further be encouraged to make an appointment with the doctors in case of high-risk indication.

For this, a self-service Decision Support Solution (DSS) is planned to be created during the iHelp project. The established strategy will not be limited to clinicians providing standardized health advice to users, rather, it will be focused on continuous examination of the person’s behaviour and current condition in order to tailor input to the specific situation. A consistent interaction with the mobile app (which the user is constantly providing feedback) in order to provide the user with multiple input tailored to his or her healthcare needs will be established.

TMU pilot does not plan to use AI for decision making. However, AI will be used for triggering appropriate messages in the mobile app.

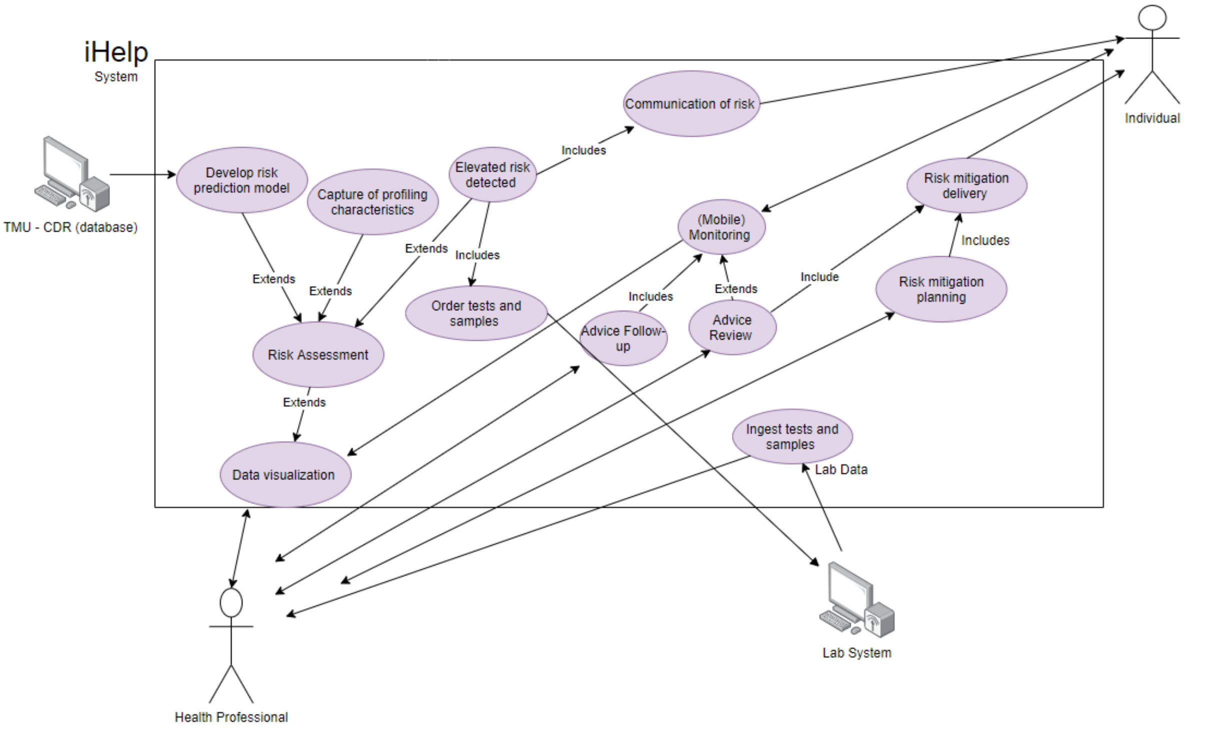


Figure 6: TMU Pilot use case diagram

2.5.2 “As-Is vs To-Be” Scenario

In the TMU pilot, data is stored in TMU-Clinical Data Repository (TMU-CDR) database that includes, historic patient data. The clinical database includes data from three affiliated hospitals, namely TMU Hospital, Wanfang Hospital and Shuang Ho Hospital. The data collected must be stored in Taiwan, and the server and other database specifications must be specified in compliance with TMU's ethical requirements and regulations. For this purpose, we have already applied for ethical approval from the Institutional Review Board of TMU.

Regarding the data acquisition, data in TMU pilot will be collected from about 1000 patient records for those diagnosed with pancreatic or liver cancer over a ten-year period. Patients aged 20 to 90 years old with at least 3 years of records in TMU hospitals with more than 1 outpatient visit or admission claim between 1998 and 2008 will be identified.

This research would consider hospital appointments, diagnoses, and medications, as well as reports of pancreatic and liver cancer diagnostic tests and comorbidities. TMU will collect data from people who have not been diagnosed with pancreatic or liver cancer in the last ten years for the research dataset. The propensity score of the patients in the training dataset for age, comorbidities, visits, and drugs will be used to obtain this information. TMU may also provide algorithms and techniques for detecting and correcting (or removing) corrupt or incomplete records from all the data collected. The collected data will be combined and transformed to the iHelp-specified Holistic Health Record (HHR) format before being stored in the TMU data storage facility. On the training dataset, AI models will be developed to predict whether these patients will be diagnosed with liver or pancreatic cancer in the next one, two, or three years.

TMU's objective is to provide scientific evidence that shows how technical interventions can enhance people's quality of life. During follow-ups, the patients' quality of life will be assessed using a paper-based European Organization for the Research and Treatment of Cancer QLQ-C30 (EORTC QLQ-C30) questionnaire. The mobile app will capture patient-reported results related to cancer symptoms, facilitating subjective data collection. These patient outcomes include diarrhea, breathing difficulties, emotional state, fatigue, loss of smell/ taste, headache, muscle pain, nausea, etc. The data will be restructured so that AI algorithms can explore it for new correlations, patterns, and dependencies. The dashboard, which displays visualizations for both healthcare professionals and patients, can also be used to track this data. As a result, we will obtain subjective data via questionnaires and PROs.

In order to provide decision support to the relevant people/patients, mobile apps will be designed with a user-centred approach that incorporates a health recommender framework. The app will be tailored to Taiwanese cultural and linguistic needs, and it will be available on both Android and iOS platforms. For high-risk people, the user-centric mobile app can provide tailored preventive and intervention initiatives. It is based on artificial intelligence-based tailored messages provided via health recommender systems to enable patients to stop smoking, engage in physical activity, consume a healthier diet, and enhance their overall well-being.

Table 37: Below are listed the relevant previous studies where the Taipei Medical University has participated, their contribution in each study and the improvements that are planned to be done from TMU through iHelp

	CrowdHEALTH Study	SmokeFreeBrain Study
What TMU contributions were in relevant studies?	▪ Developing mobile app that was used to select citizens with the potential risk of having Obesity or being Over-weight and giving healthy tips. ▪ Use of machine learning techniques to predict the outcome of CKD in this project. ▪ An active partner in the dissemination of the project and results in all possible prestigious medical informatics conferences in the region of Asia- Pacific (SYE, 20)	A study “Mobile Motivational Messages for Change (3M4Chan) intervention in TMU” was done. It involved development of health recommendation system based mobile App that is programmed to push tailored messages for health concern and readiness to quit, tips for sustaining abstinence, use of interactive self-assessments, and helpful cessation information (P., M., S., 19), (F., M., P., 19), (F., M., S., 18).
What Improvements are planned to be done from TMU through iHelp?	Provide analytical tools for prediction of high risk pancreatic and liver cancer patients.	In addition to using patient’s basic information to tailor the messages, iHelp will also utilise patient reported outcomes for different symptoms, in order to personalize the messages.

2.5.3 Description of scenarios

Table 38: TMU Pilot – 1st Scenario

Section	Description
ID	SCE-P5-01
Title	Prediction of High-risk patients (Pancreatic and Liver)
Description	Analysis of all of our patients' EHRs over the last 20 years in order to produce risk measures for individuals. EHR data includes age, sex, diagnostic codes, laboratory test reports, medications, comorbidities, family history.
Actors	Analysts (Risk model-builders), System administrators
Objectives	Early detection of people who are at high risk for cancer that will help to postpone the progression of the disease for early-stage management of the disease.
Pre-conditions	Access to TMU-CDR database.
Process Dialog	Accessing the data and pre-processing it so that analysts can use machine learning techniques to stratify risks.
Variations	Risk levels for development of the Pancreatic and Liver cancer. Risk prevention for three types of pancreatic cancer: ▪ Adenocarcinoma prevention ▪ Squamous cell carcinoma ▪ Colloid carcinoma Risk prevention for four types of liver cancer: ▪ Hepatocellular carcinoma ▪ Cholangiocarcinoma ▪ Liver angiosarcoma ▪ Hepatoblastoma
Post-conditions	Informing clinicians about patients with high risk, and suggestions to contact them.
Diagrams	This scenario is covered in Figure 6 & Figure 7
Issues and Notes	SCE-P5-01

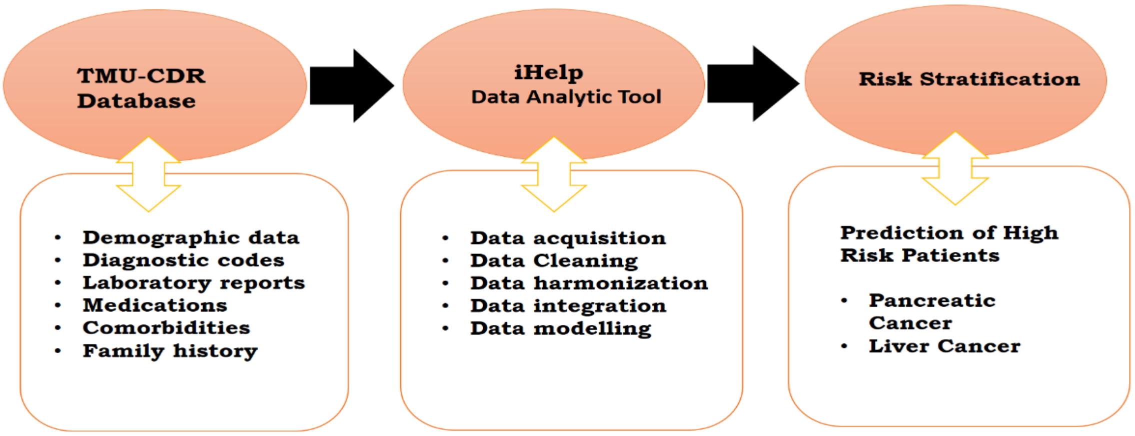


Figure 7: TMU Pilot 1st Scenario diagram

Table 39: TMU Pilot – 2nd Scenario

Section	Description
ID	SCE-P5-02
Title	Digital trial among high-risk individuals (Pancreatic and Liver)
Description	Conducting digital trial using mobile apps to evaluate the effect of digital healthcare solutions on quality of life. An observational study would be conducted among the high-risk individuals susceptible towards pancreatic and liver cancer.
Actors	▪ Individuals ○ Individuals profiled as having (high or medium) risk for developing a disease. ○ Individuals/(Patients) with one or two pre-cursory diseases to Pancreatic Cancer (i.e., diabetes) ▪ Medical Professionals ○ Risk model-builders (Specialists (in which disease) ○ Model -users (Clinicians /consultants/ specialists / oncologist or General Practitioners) ○ Model-users (Nurses and Paramedical)
Objectives	To explore the effects of digital therapeutic solutions on high-risk individuals in order to reduce the chances of disease risks, to improve their quality of life and overall well-being.
Pre-conditions	**Inclusion Criteria:** ▪ Participants identified with a high risk to develop pancreatic or liver cancer. ▪ Willing to share EHR information, including laboratory reports with the researchers so that they can receive personalized care.

	▪ Participants who have an Android / iOS phone **Exclusion Criteria:** ▪ Less than 20 years of age ▪ Participants who are having other serious illness, are bed ridden, or suffering from a mental condition.
Process Dialog	High risk group: Patients with high risk to develop pancreatic/ liver cancer will be contacted by a research assistant/study nurse and explained the study objectives. After considering the eligibility criteria, a research nurse will explain the participants the intent of the study and enrol them if they agree to follow the study. Participants will be recruited only after they fill in the patient consent form (from the TMU-JIRB ethical board).
Variations	High risk intervention group: ▪ Mobile app: Personalized health recommender system based mobile app providing interactive tailored prompts, to assess the change in behaviour as reflected in the Quality of Life of these individuals. In addition, patient reported outcomes will also be collected through the app. ▪ Usual care (treatment received from the hospitals)
Post-conditions	Follow –up (every 2 months) at 2, 4 and 6 months. Outcome assessment: **Clinical assessment of the patients and User Experience assessment of the Digital solutions**
Diagrams	This scenario is covered in 7 & Figure 8
Issues and Notes	

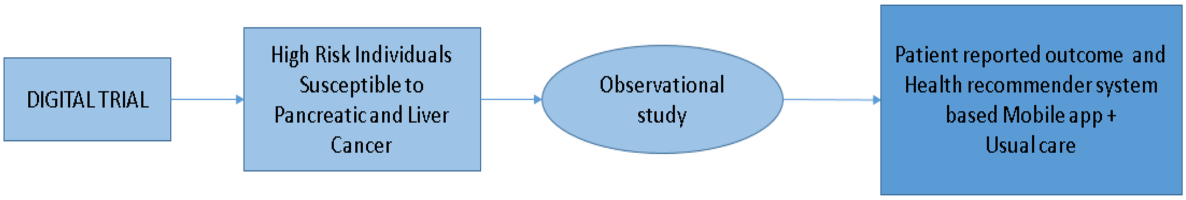

DIGITAL TRIAL
High Risk Individuals Susceptible to Pancreatic and Liver Cancer
Observational study
Patient reported outcome and Health recommender system based Mobile app + Usual care

Figure 8: TMU 2nd Scenario diagram

2.5.4 Scenarios to requirements

The following tables contain the initial list of the requirements for the scenarios of this pilot that were described in the previous subsection.

Table 40: TMU Pilot – 1st Requirement

Section	Description
ID	REQ-P5-01
Type	Data
Short Name	Assessment of Data
Description & quantification	Access to and evaluation of large amount data from various sources as a prerequisite for Liver and Pancreatic cancer risk assessment. The data encompasses variables like age, sex, diagnostic codes, medication, family history and laboratory results.
Additional information	The dataset cannot leave Taiwan
Priority	MAN
Reference Scenarios	SCE-P5-01
Success Criteria	Successful access to the TMU-CDR database

Table 41: TMU Pilot – 2nd Requirement

Section	Description
ID	REQ-P5-02
Type	FUNC
Short Name	Data analysis
Description & quantification	It represents the entire set of artificial intelligence algorithms may be used to analyse the data for risk prediction. The following variables from the TMU-CDR will be considered: ▪ Basic demographic information ▪ Comorbidities ▪ Medications ▪ Lab results
Additional information	-
Priority	[MAN]
Reference Scenarios	SCE-P5-01

Success Criteria	Successful risk evaluation using the risk prediction model from the EHR, for determining patients with high risk towards pancreatic cancer and liver cancer.

Table 42: TMU Pilot –3rd Requirement

Section	Description
ID	REQ-P5-03
Type	FUNC
Short Name	Risk assessment and communication
Description & quantification	Identification and communication of high risk to the individuals and advise towards the use of iHelp digital solutions to enhance behavioural change.
Additional information	-
Priority	MAN
Reference Scenarios	SCE-P5-01, SCE-P5-02
Success Criteria	Successful enrolment of high-risk individuals susceptible to pancreatic and liver cancer. Successful application of iHelp digital solutions to enhance behavioural change.

Table 43: TMU Pilot –4th Requirement

Section	Description
ID	REQ-P5-04
Type	FUNC
Short Name	Targeted Recommendation
Description & quantification	Usage of co-designed iHelp platform to provide feedback on targeted recommendations and level of activity.
Additional information	-
Priority	MAN

Reference Scenarios	SCE-P5-02
Success Criteria	Successful implementation of iHelp digital solution (Mobile app) and looping the usage data for feedback at defined time interval.

Table 44: TMU Pilot – 5th Requirement

Section	**Description**
ID	REQ-P5-02
Type	DATA
Short Name	Subjective data collection
Description & quantification	The main objective is to gather qualitative data from the subjects, such as basic user details, family history, habits, stress, subjective sleep assessment, etc. The mobile application will be embedded with health recommender system that will enable health related awareness, healthcare advice, behavioural nudges, education, phycological support personalized motivational messages. Assessment of patient reported outcomes from the Mobile App for -- pain, happiness, depression, anxiety, mood, physical activity, sleep behaviour and fatigue. **User Experience assessment of the Digital solutions** User engagement, usefulness of personalised information, impact of recommendations etc. will be performed based on the responses gathered through the mobile application.
Additional information	
Priority	MAN
Reference Scenarios	SCE-P5-02
Success Criteria	▪ Delivery of personalized recommendations/decision-support. ▪ Timely received user feedback for the personalized motivational messages.

Table 45: TMU Pilot – 6th Requirement

Section	Description
ID	REQ-P5-03
Type	FUNC
Short Name	Data storage
Description & quantification	Role of server: ▪ Anonymization and integration of real time and AI analytic data. ▪ Keeping the data secure. ▪ Additional information
Additional information	
Priority	MAN
Reference Scenarios	SCE-P5-01, SCE-P5-02
Success Criteria	Creation and storage of anonymized secured database

Table 46: TMU Pilot – 7th Requirement

Section	Description
ID	REQ-P5-04
Type	FUNC
Short Name	Data visualization
Description & quantification	Decision Support Solution (DSS) will allow clinicians/medical experts to analyse the collected data. The system will be presented in the form of a dashboard with visualizations to the clinicians as well as patients.
Additional information	
Priority	DES
Reference Scenarios	SCE-P5-02
Success Criteria	Improves clinical decision processes for patients

3 Pilot datasets and data regulatory constraints

This section contains the definition of all available datasets that will be used in the scope of the iHelp project, along with potential data regulatory constraints that might be needed to be enforced when these datasets are being accessed by third parties or are being collected and stored in cloud environments, such as the deployment of the iHelp platform outside of the proprietary 's premises.

3.1 Pilot #1 Study of Genomics and Epigenomics Makers for Early Risk Assessment of Pancreatic Cancer

3.1.1 Data Collection

The UNIMAN pilot will assemble 2 types of data.

- Primary data

The primary data will be acquired from large studies including the UKBiobank, the lifeline data and the ATP data. We have permission to use these datasets in order to develop our risk models and validate them. For cancer risk assessment, data will be collected through our online cancer risk prediction platforms. Furthermore, in high-risk groups, biological samples will be collected to process and assess genetics and epigenetic markers.

- Secondary data

Secondary data will be collected from healthy individuals who are eligible to attend NHS-HC. Secondary data will also be collected using questionnaires (of general characteristic data). All individuals will provide data from agreed set of target activities via iHelp platform and associated Healthentia application.

3.1.2 Dataset Description & Requirements

3.1.2.1 Primary Data

Table 47: UNIMAN Pilot – 1st Primary dataset detail

Section	Description
ID	DS-P1-01
Title	UKBiobank
Description	A large-scale UK population-based data with comprehensive medical, lifestyle, genetics data. The data contains 500k individuals.
Owner	The data is owned by the UKBiobank and the UNIMAN partner is a third-party recipient.
Licence / Privacy	The data must stay on premises and be accessed externally to the platform.
Data type	Structural
Type of process (Stream or static	Data on disease outcomes will be updated periodically. The

data)	UKBiobank will send e-mail to inform of the updates (annually).
Data format	Direct connection to the datastore and transform into STATA format for the analysis purpose using MDCHECKSUM and key file to unlock the access.
Data store	Accessed by a third-party REST interface
Data Security	Data was anonymised and only designated names on the usage of the data are allowed to access and process the data.
Regulatory Constraint Requirements	Data can only be used within the scope of the application made to the UKBiobank.

Table 48: UNIMAN Pilot – 2nd Primary dataset detail

Section	**Description**
ID	DS-P1-02
Title	Lifelines
Description	A large-scale Dutch population-based data with comprehensive medical, lifestyle, genetics and epigenetic data.
Owner	The data is own by the Lifelines and the UNIMAN partner is a third-party recipient.
Licence / Privacy	The data must stay on premise and be accessed externally to the platform.
Data type	Structural
Type of process (Stream or static data)	Linkage data (data that linked to the Official National Statistic and other parties that the Lifelines deems to be useful for the research purposes) will be updated periodically.
Data format	Direct connection to the datastore and transform into STATA format for the analysis purpose.
Data store	Accessed by a third-party REST interface
Data Security	Data was anonymised and only designated names on the usage of the data are allowed to access and process the data.
Regulatory Constraint Requirements	Data can only be used within the scope of the application made to the Lifelines.

Table 49: UNIMAN Pilot – 3rd Primary dataset detail

Section	**Description**

ID	DS-P1-03
Title	The Alberta’s Tomorrow Project (ATP).
Description	A large-scale Canadian population-based data with comprehensive medical, lifestyle data.
Owner	The data is own by the ATP and the UNIMAN partner is a third-party recipient.
Licence / Privacy	The data must stay on premises and be accessed externally to the platform.
Data type	Structural
Type of process (Stream or static data)	Linkage data (data that linked to the Official National Statistic) will be updated periodically.
Data format	Data will be sent in CSV format with password protected and only one designated person from the UNIMAN will be assigned to call to the ATP staff team to redeem code for data access.
Data store	Accessed by a third-party REST interface
Data Security	Data was anonymised and only designated names on the usage of the data are allowed to access and process the data.
Regulatory Constraint Requirements	Data can only be used within the scope of the application made to the ATP study.

Table 50: UNIMAN Pilot – 4th Primary dataset detail

Section	Description
ID	DS-P1-04
Title	Survey data
Description	This data will be collected from participants by the UNIMAN team.
Owner	The UNIMAN partner will be a primary owner.
Licence / Privacy	The data will be anonymised prior to sharing.
Data type	Structural
Type of process (Stream or static data)	Static data
Data format	Data will be shared in CSV format with password protected.
Data store	At the UNIMAN assigned premise and university managed

	computer.
Data Security	Data will be anonymised.
Regulatory Constraint Requirements	Data can only be used for the iHelp research purpose and DTA will be obtained from any partners who will be using the data prior to any data sharing.
Section	Description
ID	DS-P1-05
Title	Risk assessment data
Description	This data will be collected from participants by the UNIMAN team.
Owner	The UNIMAN partner will be a primary owner.
Licence / Privacy	The data will be anonymised prior to sharing.
Data type	Structural
Type of process (Stream or static data)	Static data
Data format	Data will be shared in CSV format with password protected.
Data store	At the UNIMAN assigned premise and university managed computer. Data will be stored on the P-drive which is a secure storage for personal files and documents on the University's network. University managed computer is encrypted and maintained by IT Services. Furthermore, as part of the University's Cyber Security Programme, the UNIMAN has introduced 2-factor authentication so access to data can be safe and secured.
Data Security	Data will be anonymised.
Regulatory Constraint Requirements	Data can only be used for the iHelp research purpose and Non-disclosure Agreement (NDA) will be obtained from any partners who will be using the data prior to any data sharing.

Table 51: UNIMAN Pilot – 5th Primary dataset detail

Section	**Description**
ID	DS-P1-05
Title	Risk assessment data
Description	This data will be collected from participants by the UNIMAN team.

Owner	The UNIMAN partner will be a primary owner.
Licence / Privacy	The data will be anonymised prior to sharing.
Data type	Structural
Type of process (Stream or static data)	Static data
Data format	Data will be shared in CSV format with password protected.
Data store	At the UNIMAN assigned premise and university managed computer. Data will be stored on the P-drive which is a secure storage for personal files and documents on the University's network. University managed computer is encrypted and maintained by IT Services. Furthermore, as part of the University's Cyber Security Programme, the UNIMAN has introduced 2-factor authentication so access to data can be safe and secured.
Data Security	Data will be anonymised.
Regulatory Constraint Requirements	Data can only be used for the iHelp research purpose and NDA will be obtained from any partners who will be using the data prior to any data sharing.

Table 52: UNIMAN Pilot – 6th Primary dataset detail

Section	Description
ID	DS-P1-06
Title	Biomarker data
Description	This data will be collected from only high-risk participants by the UNIMAN team.
Owner	The UNIMAN partner will be a primary owner.
Licence / Privacy	The biomarker data will not be sharing as a raw data but instead will be shared as an information of predisposition scores to any partners. For example, participant identification number and summary genetic score and methylation score data.
Data type	Structural
Type of process (Stream or static data)	Static data
Data format	Data will be shared in CSV format with password protected.

Data store	At the UNIMAN assigned premise and university managed computer. Data will be stored on the P-drive which is a secure storage for personal files and documents on the University's network. University managed computer is encrypted and maintained by IT Services. Furthermore, as part of the University's Cyber Security Programme, the UNIMAN has introduced 2-factor authentication so access to data can be safe and secured.
Data Security	Data will be anonymised.
Regulatory Constraint Requirements	Data can only be used for the iHelp research purpose and DTA will be obtained from any partners who will be using the data prior to any information sharing.

3.1.2.2 Secondary Data

Table 53: UNIMAN Pilot – Secondary data measurement requirements

Requirement	Description
REQ-MEAS-P1-001	"Self-reported baseline general characteristics"
REQ-MEAS-P1-002	"Data derived from risk mitigation including stream data related to lifestyle including physical activity, diet, alcohol, well-being and smoking habits"

Table 54: UNIMAN Pilot – Secondary data application requirements

Requirement	Description
REQ-APP-P1-001	"Individuals must be able to read and write English and consent to participate the study"
REQ-APP-P1-002	"Individual in high-risk group must consent to have blood sample collection"
REQ-APP-P1-003	"Individual must be able to use wearable device over the trail period and receive regular feedback in a form of "chatbot"

3.2 Pilot #2 Interventional Monocentric Study based on Patient Reported Outcomes

3.2.1 Data Collection

The FPG pilot seeks to implement retrospective extrapolation of available primary data, along with the prospective acquisition of secondary data from the first clinical evaluation, until the end of the treatment and the first follow-up period. Data will be collected from mobile application, IoT device, clinical evaluation.

3.2.2 Data Description

3.2.2.1 Primary Data

Primary data involve a retrospective cohort of patients treated by adjuvant radiotherapy (with or without chemotherapy) in Agostino Gemelli University Policlinic institution. This dataset includes data regarding staging, markers, surgery, histological examination, radiotherapy, chemotherapy, toxicity, clinical outcomes.

Table 55: FPG Pilot – 1ST Primary dataset detail

Section	Description
ID	DS-P2-01
Title	Dataset for adjuvant radiotherapy
Description	Retrospective monocentric dataset for adjuvant radiotherapy in pancreatic cancer. n>100 patients. All underwent surgery.
Owner	Fondazione Policlinico Universitario Agostino Gemelli IRCCS
Licence / Privacy	Data can be hosted at Gemelli Generator (FPG) and queries can be run in a sandbox environment; results will be made available online to the project as needed, with license to use them for the project goals. If data are to be used in a learning effort, a federated learning environment can be deployed at Gemelli Generator.
Data type	Structured data
Type of process (Stream or static data)	Static data
Data format	Excel file
Data store	At Gemelli Generator, FPG
Data Security	Pseudonymization
Regulatory Constraint	Driven by Ethical Committee feedback

Requirements	

3.2.2.2 Secondary Data

Table 56: FPG Pilot – Secondary data measurement requirements

Requirement	Description
REQ- MEAS-P2-001	Behavioural ▪ Steps [MAN] ▪ Climbs [MAN] ▪ energy expenditure [MAN] ▪ heart rate Min - Max - Resting heart rate - Average daily heart rate, automatically acquired from IoT [OPT/ENH]; or, if it is not possible, sporadic automatic heart rate acquisition and storage [MAN]
REQ- MEAS-P2-002	Medical ▪ BMI The patient periodically enters his/her weight and having entered the height at registration, BMI can be computed. ▪ blood pressure [OPT] ▪ blood oxygen saturation [DES] Automatic periodic (daily) measurement of blood oxygen saturation [DES]; or, if it is not possible, a manual acquisition and reporting by the patient [MAN]
REQ- MEAS-P2-003	Symptoms ▪ body temperature [OPT]
REQ- MEAS-P2-004	▪ sleep metrics [MAN]
REQ- MEAS-P2-005	▪ exercise sessions autodetected by the device [OPT]
REQ- MEAS-P2-006	Regulatory Requirements For ethical committee also the following documents are needed: ▪ CE certificate: IoT ▪ Technical docs: IoT
REQ- MEAS-P2-007	Questionnaires ▪ EORTC QLQ-C30 questionnaire [MAN] Authorization already acquired from EORTC

	▪ EORTC QLQ-PAN26 questionnaire [MAN] Authorization already acquired from EORTC ▪ COVID-19 prevention daily questionnaire [MAN] ▪ eventual psychological condition extracted from social media [OPT]

Table 57: FPG Pilot – Secondary data application requirements

Requirement	Description
REQ- APP-P2-001	Both personal dashboard and app-advice are sensitive issues for ▪ Personal dashboard of activities [DES] NB: no data regarding other patients (also in aggregate form) should be proposed because no benchmark of ideal activities have been previously reported in scientific literature for patients who will undergo radiotherapy. Global inter-patient statistics or gamification features must be made available only for MDs.
REQ- APP-P2-002	▪ Possibility to communicate to the patient and to see if the patient read the content of the message [DES] NB good clinical practice advice will be proposed by clinicians to the patient. AI based advice will be proposed first to the clinician and, if coherent with a good clinical practice recommendation, will be proposed to the patient.

3.3 Pilot #3 Study of the Lifestyle choices on Elevating the Risk Factors for Pancreatic Cancer

3.3.1 Data Collection

Primary data are based on the Electronic Health Records (EHRs) of Hospital de Dénia-Marina Salud. Acquisition of secondary data will be collected from the trial of patients selected, from mobile application, IoT devices, clinical evaluation, and self-evaluation.

3.3.2 Data Description

3.3.2.1 Primary Data

Table 58: HDM Pilot – 1st Primary dataset detail

Section	Description
ID	DS-P3-01
Title	Marina Salud Atenea
Description	The Marina Salud EHR contain about 300.000 EMR's. The dimensions of information are Person, Encounter, Orders, Clinical Event, Laboratory, Radiology, Diagnosis and Procedures.
Owner	Marina Salud
Licence / Privacy	The data will be anonymised prior to sharing.
Data type	Structured data
Type of process (Stream or static data)	Static data
Data format	MS SQL Server
Data store	Marina Salud DWH
Data Security	Pseudonymization
Regulatory Constraint Requirements	Driven by Ethical Committee feedback

3.3.2.2 Secondary Data

Table 59: HDM Pilot – Secondary data measurement requirements

Requirement	Description
REQ- MEAS-P3-001	Behavioural ▪ Steps [MAN] ▪ Climbs [MAN]

	▪ energy expenditure [MAN] ▪ heart rate Min - Max - Resting heart rate - Average daily heart rate, automatically acquired from IoT [OPT/ENH]; or, if it is not possible, sporadic automatic heart rate acquisition and storage [MAN]
REQ- MEAS-P3-002	Medical ▪ BMI The patient periodically enters his/her weight and having entered the height at registration, BMI can be computed. ▪ blood pressure [OPT] ▪ blood oxygen saturation
REQ- MEAS-P3-003	Symptoms ▪ body temperature [OPT]
REQ- MEAS-P3-004	▪ sleep metrics [MAN]
REQ- MEAS-P3-005	▪ exercise sessions autodetected by the device [OPT]
REQ- MEAS-P3-006	Regulatory Requirements For ethical committee also the following documents are needed: ▪ CE certificate: IoT ▪ Technical docs: IoT
REQ- MEAS-P3-007	Questionnaires ▪ EORTC QLQ-C30 questionnaire [MAN] Authorization already acquired from EORTC ▪ EORTC QLQ-PAN26 questionnaire [MAN] Authorization already acquired from EORTC ▪ COVID-19 prevention daily questionnaire [MAN] ▪ eventual psychological condition extracted from social media [OPT]

Table 60: HDM Pilot – Secondary data application requirements

Requirement	Description
REQ- APP-P3-001	Both personal dashboard and app-advice are sensitive issues for ▪ Personal dashboard of activities [DES]

REQ- APP-P3-002	▪ Possibility to communicate to the patient and to see if the patient read the content of the message [DES] NB good clinical practice advice will be proposed by clinicians to the patient. AI based advice will be proposed first to the clinician and, if coherent with a good clinical practice recommendation, will be proposed to the patient.

3.4 Pilot #4 Study of Risk, Personalised Recommendations and Measures to Raise Awareness of Relevant Factors

3.4.1 Data Collection

The first data collection will be done by extracting retrospective data from the patients' files into the 4 MUP hospitals and Centre for Oncology. The data acquired will be grouped in accordance with the source - patients complains, examinations, family and comorbidities history, laboratory and imaginary tests, consultations. These data will be inserted to the iHelp AI in order evaluation of the risk to be performed and based on the results a preventative program including treatment, dietary and physical regime, consultation, examination, laboratory and imaginary checks schedule and behavioural changes suggestions will be extracted by the iHelp and after consideration will be recommended to the patient at risk (depending on the assessed risk level).

Secondary data will be collected by the iHelp related to the measured and obtained from the patient - quality of life, body temperature and weight, pain level, alcohol, meat, vegetables, grains consumption, pain killers and other medicine intake, social exclusion. Healthcare provider will be alerted by iHelp for changes into these parameters, in order to actively enter in contact with patients at risk for assessing the sources for the program violation.

On the second layer the iHelp AI is to extract and group into the above-mentioned categories the data entered by the physicians, healthcare providers into the patient's file during the examination. Once extracted and evaluated the data will inform the healthcare provider regarding the risk and recommended program to be discussed between him/her and patient at risk.

3.4.2 Data Description

3.4.2.1 Primary Data

Table 61: MUP Pilot – 1st Primary dataset detail

Section	Description
ID	DS-P4-01
Title	Patient complaints
Description	Large amount data of patient complaints
Owner	Medical University Plovdiv
Licence / Privacy	Medical University Plovdiv

Data type	Structural
Type of process (Stream or static data)	Static data
Data format	Excel Workbook (.xlsx) Format
Data store	Medical University of Plovdiv – MS SQL database
Data Security	Pseudonymization
Regulatory Constraint Requirements	Driven by Ethical Committee feedback

Table 62: MUP Pilot –2nd Primary dataset detail

Section	**Description**
ID	DS-P4-02
Title	Family history
Description	Large amount data of patient's family history
Owner	Medical University Plovdiv
Licence / Privacy	Medical University Plovdiv
Data type	Structural
Type of process (Stream or static data)	Static data
Data format	Excel Workbook (.xlsx) Format
Data store	Medical University of Plovdiv – MS SQL database

Data Security	Pseudonymization
Regulatory Constraint Requirements	Driven by Ethical Committee feedback

Table 63: MUP Pilot – 3rd Primary dataset detail

Section	**Description**
ID	DS-P4-03
Title	co-morbidities
Description	Large amount data of patient co-morbidities
Owner	Medical University Plovdiv
Licence / Privacy	Medical University Plovdiv
Data type	Structural
Type of process (Stream or static data)	Static data
Data format	Excel Workbook (.xlsx) Format
Data store	Medical University of Plovdiv – MS SQL database
Data Security	Pseudonymization
Regulatory Constraint Requirements	Driven by Ethical Committee feedback

Table 64: MUP Pilot – 4th Primary dataset detail

Section	**Description**

ID	DS-P4-04
Title	Medication
Description	Large amount data of medication data (e.g.: what kind of medicines are taken)
Owner	Medical University Plovdiv
Licence / Privacy	Medical University Plovdiv
Data type	Structural
Type of process (Stream or static data)	Static data
Data format	Excel Workbook (.xlsx) Format
Data store	Medical University of Plovdiv – MS SQL database
Data Security	Pseudonymization
Regulatory Constraint Requirements	Driven by Ethical Committee feedback

Table 65: MUP Pilot – 5th Primary dataset detail

Section	**Description**
ID	DS-P4-05
Title	Laboratory & imaginary tests
Description	Large amount data of laboratory tests & imaginary tests
Owner	Medical University Plovdiv

Licence / Privacy	Medical University Plovdiv
Data type	Structural
Type of process (Stream or static data)	Static data
Data format	Excel Workbook (.xlsx) Format
Data store	Medical University of Plovdiv – MS SQL database
Data Security	Pseudonymization
Regulatory Constraint Requirements	Driven by Ethical Committee feedback

Table 66: MUP Pilot – 6th Primary dataset detail

Section	**Description**
ID	DS-P4-06
Title	Physical examination
Description	Large amount data of physical examination findings (e.g.: the status of the patient)
Owner	Medical University Plovdiv
Licence / Privacy	Medical University Plovdiv
Data type	Structural
Type of process (Stream or static data)	Static data
Data format	Excel Workbook (.xlsx) Format

Data store	Medical University of Plovdiv – MS SQL database
Data Security	Pseudonymization
Regulatory Constraint Requirements	Driven by Ethical Committee feedback

3.4.2.2 Secondary Data

Table 67: MUP Pilot – Secondary data measurement requirements

Requirement	Description
REQ-MEAS-P4-001	Questionnaires: ▪ Quality of life questionnaire
REQ-MEAS-P4-002	Medical: ▪ Pain level ▪ Body weight and temperature ▪ Daily Medicines
REQ-MEAS-P4-003	Symptoms: ▪ Temperature ▪ Pain level
REQ-MEAS-P4-004	Behavioural: ▪ Alcohol consumption ▪ Red and processed meat consumption ▪ Vegetables and grains consumption

Table 68: MUP Pilot – Secondary data application requirements

Requirement	Description
REQ- APP-P4-001	The clinicians must be able to view the data in a dashboard display

3.5 Pilot #5 Study of Improved Risk Prediction Models and Targeted Interventions that can Delay the Onset of Cancer

3.5.1 Data Collection

In the first phase of risk prediction, the required data variables for AI based big data analytics would be considered from the available TMU-CDR database. These variables include clinical visits, diagnoses, and medications in this study along with the records of pancreatic and liver cancer diagnostic tests and comorbidities.

In the second phase which includes digital trial, a mobile app would be used for collecting subjective data from the subjects, such as basic user details, family history, habits, stress, symptoms, subjective sleep assessment etc.

3.5.2 Data Description

3.5.2.1 Primary Data

Table 69: TMU Pilot – 1st Primary dataset detail

Section	Description
ID	DS-P5-01
Title	TMU Clinical Data Repository (TMU-CDR) database
Description	Contains Electronic Health Records (EHR) of all our patients from last 20 years (1998-2018) in three university affiliated hospitals. (Laboratory Test reports, medications, family history, comorbidities.)
Owner	Taipei Medical University- available for use only for TMU staff, students, researchers, and faculty members
Licence / Privacy	It must stay on premise and can be accessed externally to the platform in Taiwan, after acquiring required permissions from the authorities.
Data type	Unstructured
Type of process (Stream or static data)	Static
Data format	Data storage infrastructure ▪ CSV format
Data store	Data will be stored in a secure storage facility at TMU
Data Security	Access control
Regulatory Constraint Requirements	Needs to be in compliance with Taiwan regulations and guidance.

Table 70: TMU Pilot – 2nd Primary dataset detail

Section	Description
ID	DS-P5-02
Title	Data collected via mobile app
Description	Data collected during digital trial using mobile apps to evaluate the effect of digital healthcare solutions in order to reduce the chances of disease risks, to improve their quality of life and overall well-being. An observational study would be conducted among the high-risk individuals susceptible towards pancreatic and liver cancer.
Owner	Taipei Medical University
Licence / Privacy	Anonymised data will be available for iHelp project's partners.
Data type	Structured
Type of process (Stream or static data)	Dynamic stream data
Data format	Data storage infrastructure ▪ CSV format with password protected
Data store	Data will be stored in a secure storage facility at TMU as additionally in internal iHelp's components and Big Data Platform.
Data Security	Access control
Regulatory Constraint Requirements	Needs to be in compliance with Taiwan regulations and guidance.

3.5.2.2 Secondary Data

Table 71: TMU Pilot – Secondary data measurement requirements

Requirement	Description
REQ- MEAS-P5-001	Behavioural ▪ Mobile app: basic information, message ratings

REQ- MEAS-P5-002	Behavioural ▪ Patient reported outcomes through mobile app: diarrhea, breathing difficulties, emotional state, fatigue, loss of smell/ taste, headache, muscle pain, nausea, etc.
REQ- MEAS-P5-003	Behavioural ▪ Questionnaires: European Organization for the Research and Treatment of Cancer QLQ-C30 (EORTC QLQ-C30). ▪ Pittsburgh Sleep Quality Index (PSQI)
REQ- MEAS-P5-004	Regulatory requirements: ▪ TMU-IRB Ethical approval

Table 72: TMU Pilot – Secondary data application requirements

Requirement	**Description**
REQ- APP-P5-001	▪ Information presentation Measures would be presented with dashboard-like display. Patient information to be presented individually and may include comparison with statistics of other patients belonging to the same group.
REQ- APP-P5-002	▪ Advice delivery The patient must be able to receive advice (motivational text/ alert/suggestion) in the form of push notifications. The messages that would be delivered to patients would be automatic, that is, originating from AI-based health recommender system integrated in the mobile application.
REQ- APP-P5-003	▪ Means of subjective, patient-generated data collection. Questionnaires, as well as patient-reported outcome measures, would be collected by the mobile application**.**

4 Modules

This section provides a list of the requirements of the different modules that will be implemented during the iHelp project. These requirements are related with specific component portions, which can be either a program, a software component, an existing product that will be used as part of the overall platform, or a set of combinations of all the above, that implements a specific functionality and provides a set of capabilities via well-defined interfaces.

Each module is described in its own Section. For each module there is a sub-section where are explained the goals and objectives of this module, a "state-of-the-art" sub-section which is dedicated to the description of the related work and the state-of-the-art approaches and a "background technology" sub-section where is described the already existed "baseline technology" of each module and partners plan to further develop them in order to fulfil the requirements of the iHelp platform. Furthermore the "module to user requirements" sub-section is a mapping between the module and the relevant user requirements identified in Section 2. Finally, in the "module to technical requirements" sub-section the technical requirements of each module are reported in detail.

The table below describes the different types of the requirements and each technical requirement identified in 4.X.5 sections is characterized using one of the available types.

Table 73: Types of the requirements

FUNCTIONAL	*Functional*	**FUNC**
	Data	**DATA**
NON-FUNCTIONAL	*Look and Feel Requirements*	**L&F**
	Usability Requirements	**USE**
	Performance Requirements	**PERF**
	Operational / Environment Requirements	**ENV**
	Maintainability and Support Requirements	**SUP**

4.1 Data Modelling and Integrated Health Records

4.1.1 Goals and Objectives

The goal of "Data Modelling and Integrated Health Records" task is to create new structures using the current health records revealing personalized information. An important aspect in this task is the modelling of primary and secondary data that will come from WP3. These data involve health associated properties such as lifestyle, fitness, wellbeing etc. as well as social-related parameters such as behaviors, relationships, interactions, etc., both broad categories of sensitive data. More specifically, the main goals of this task include techniques to:

- Address how data is organized/modelled following the HL7 FHIR structure (and then organized in the new HHR structure before it is stored in the big data platform.
- Establish a reliable approach that allows modelling of primary and secondary data.
- Develop rules to automatically establish correlations between different parameters in the HHR structure.
- Extent the structures to allow new data parameters to be added with time.
- Improve scalability and extensibility is to ease the sharing of specific aspects of HHR between different health platforms.

4.1.2 State of the Art

Several attempts have been made to model the concepts in health domain. Most of these efforts were made with the help of health ontologies. An ontology is representation of concepts that describe a certain area and define the properties of these concepts as well as the relations between them. An ontology in the end can be seen as a graph where the nodes are the concepts, and the relations are the arcs between the concepts. The main language that is used for defining an ontology is the OWL 2 Web Ontology Language[2]. In the health domain today exist ontologies that define medical information, concepts about of food, social and physical activities etc. Some indicative ontologies are:

- SNOMED CT3, which defines concepts related to diseases in terms of possible causes, effect in the body and symptoms.
- The International Classification of Functioning, Disability and Health (ICF) 4 which is an ontology classifying health and health-related domains from a body, personal activities, and a societal perspective.
- The USDA Food Composition Database5 is the standard reference for nutrients, food, and food products, including classification with respect to manufacturers.
- The SMASH6 ontology (Semantic Mining of Activity, Social, and Health data). It describes concepts used in describing the semantic features of healthcare data and social networks and also includes categories for physical and social activities.

[2] https://www.w3.org/TR/owl2-overview/

[3] https://bioportal.bioontology.org/ontologies/SCTO

[4] https://www.who.int/classifications/international-classification-of-functioning-disability-and-health

[5] https://ndb.nal.usda.gov/ndb/

[6] https://bioportal.bioontology.org/ontologies/SMASHPHYSICAL

This work on the ontologies despite significant, was not coordinated by a central organization and as a result there is not necessarily a link between them, thus overlapping or differences in the definition of the concepts and their relationships exist.

In parallel or in conjunction with these efforts regarding the ontologies, several standards were emerged with the goal to make easier the standardization of the Electronic Health Records and the increase of their interoperability. The most important and most widely adopted such standard is HL7 FHIR. Health Level Seven or HL7 refers to a set of international standards for transfer of clinical and administrative data between software applications used by various healthcare providers[7]. FHIR stands for Fast Healthcare Interoperable Resource and combines the best features of HL7 v2, v3 and CDA. FHIR is based on called "resources" which can be jointed, extended, and adapted to provide a more manageable solution for optionality and customization.

In the context of the iHelp project, the HL7 FHIR standard will be adopted in a way that it covers the pilot health organizations' needs in the case of the Pancreatic Cancer.

4.1.3 Background technology

Table 74: T3.1 – Background technology

Technology Name	Technology Description	Advancements / Usage
OWL	A language to define ontologies in a specific domain.	Semantic definition of the pilot data sources and the HHR model
HL7 FHIR	A standard for the definition and exchange of EHR	Definition of the pilot data sources and the HHR model.
Apache Jena[8]	A java library for handling OWL resources	Handling of the OWL/RDF resources and transform them to JSON-LD

4.1.4 Module to User Requirements

Table 75: T3.1 - 1stFunctionality

Section	Description
ID	U-REQ-T3.1-01
Title	HHR Specifications
Functionality Description	Definition and transformation of patient data to HHR. The HHRs will be used in all the pilot functionalities related to analysis and thus the module is a prerequisite for the realization of most of the user requirements.

[7] https://en.wikipedia.org/wiki/Health_Level_7

[8] https://jena.apache.org/index.html

Source User Requirement	▪ All except the ones related to the IoT and mobile devices
Use case Quote	▪ Pilot#1 – "be able to analyze the data..." ▪ Pilot#2 – "to predict outcomes and toxicity." ▪ Pilot#3 – "machine learning methods for risk stratification…" ▪ Pilot#4 – "The system will issue some advice to the clinician..." ▪ Pilot#5 -- "to analyze electronic health record data and predict high risk individuals..."
Generic / Specific	Generic (because it satisfies the requirements of more than one pilot)
Task / Component	T3.1 / Data Modelling and integrated Health Records
Lead Partner	ATC
Notes	-

4.1.5 Module to Technical Requirements

Table 76: T3.1 – 1st Technical requirement

Section	Description
ID	T-REQ-T3.1-01
Type	FUNC
Short Name	Data feed
Functionality ID	U-REQ-T3.1-01
Description & quantification	Cleaned and HL7 compatible primary and secondary data need to be fed into this module
Additional information	Data source and data schema should be described.
Priority	▪ MAN
Reference Pilot	P1, P2, P3, P4, P5
Success Criteria	Successful transformation to HHR schema

Table 77: T3.1 – 2nd Technical requirement

Section	Description
ID	T-REQ-T3.1-02
Type	PERF
Short Name	Scalability
Functionality ID	U-REQ-T3.1-01
Description & quantification	The module must be able to handle more than one data streams in parallel
Additional information	-
Priority	▪ MAN
Reference Pilot	P1, P2, P3, P4, P5
Success Criteria	Two or more instances of the module can co-exist handling different data streams each

Table 78: T3.1 – 3rd Technical requirement

Section	Description
ID	T-REQ-T3.1-03
Type	ENV
Short Name	Connectivity
Functionality ID	U-REQ-T3.1-01
Description & quantification	The module must be able to listen to events and send events from / to a message bus
Additional information	-
Priority	▪ MAN
Reference Pilot	P1, P2, P3, P4, P5
Success Criteria	The module receives and sends events to the project's event bus as part of the data processing pipeline

4.2 Primary Data Capture and Ingestion

4.2.1 Goals and Objectives

The main goal of the Primary Data Capture and Ingestion group of modules is to capture data coming from primary data sources (i.e., electronic patient health records, results of research studies regarding socio-demographic data, individual social habits such as smoking or drinking etc.) by establishing connectivity mechanisms with the heterogeneous data sources and setup dynamically the data ingestion pipeline that may consist of various serverless functions. The latter need to perform a pre-processing on the incoming data, transform them to the common HHR model and finally store them into the Big Data Platform of iHelp. It is worth to mention that the objective of this group of modules is not to provide the serverless functions that perform data transformations or applying quality assurance regulations, rather than provide the means for the dynamic deployment of such functions in a data pipeline, ensure their internal data exchange and establish the data connectivity with the primary data sources.

4.2.2 State of the Art

Serverless computing and the use of the Function-as-a-Service (FaaS) paradigm has been emerged during the recent years, following the uprising of the cloud computing and the swift of modern enterprises to the microservices paradigm. Modern ICT world has been changed a lot compared to the situation it was in the beginning of the last decade. At that time, cloud computing was in an infant situation and most organizations hesitated to move their software to such vendors. However, during the global economic crisis of the 2010s, new start-ups emerged bringing their own ICT applications but could not cover the initial cost of buying expensive hardware to host three solutions. Another key factor was the demand for quick growth, bringing the requirement to frequently scale out the applications to cover the requested user traffic. As cloud computing became more mature and extended with Backend-as-a-Service support, more and more organizations started to decide to move their own solutions to cloud vendor and benefit from the pay-as-you-go pricing model, while at the same time outsourcing basic operations to the cloud vendors, such as the maintenance of the hardware, the monitoring of the ICT solutions etc.

At the same time, virtualization of the software became a dominant paradigm that helped the automation of deployments via tools like docker-compose or Kubernetes that were beneficial for the maintenance of software and its portability. This provoked the use of the microservices approach that started to be favourable compared to traditional monolithic applications. The latter started to be broken down to smaller pieces, where each piece was not a single microservice, moving to a truly *separation of concern* between the different building blocks of an integrated solution. However, during the recent years, virtualization has moved as step beyond hardware to virtualized containers in what we now call *serverless computing* and *Function-as-a-Service*.

Serverless computing describes the paradigm that software does not have to run on dedicated machines, but it can make use of cloud services instead. Using virtualization combined with the microservice pattern, applications can be fully deployed in a cloud vendor who is getting the responsibility of managing the underlying infrastructure resources. A step beyond is the FaaS, where the application can be broken down to specific stateless functions, and the developer only needs to take control of the business logic implemented within the function itself. The code is packaged, and it is the responsibility of the cloud provider to deploy it in a running container. This is ideal for short-living processes that usually need to respond to a specific request following an event-driven approach, and there is no demand to be continuously deploy, thus consuming resources when it is idle. It can scale out very quickly handling unexpected traffic bursts when they occur, rather than having to maintain and pay unused resources. This event-driven approach is suitable in situations like the Primary Data Capture and Ingestion set of technologies, where the services and functions deployed need to respond to specific events (i.e. the periodically ingestion of a dataset) rather than being deployed all time. The FaaS tend to use service choreography instead of service orchestration, as they do not require a central component to orchestrate the interaction between the functions. Instead, once the functions have been dynamically deployed, they can respond to events and produce events for other functions to respond. On the other hand, FaaS comes with the drawback of not being suitable for long-running process doing heavily batch jobs and they suffer for a cold start upon their initial invocation. Bu this is not the case in the Data Capture and Ingestion, where each function has a small job to do, and the data needs to be ingested eventually, thus an additional overall delay of some seconds is acceptable.

As the goal of the Primary Data Capture and Ingestion is to provide the means for deploying such functionality dynamically and setup the overall data ingestion pipeline, we focused in FaaS related platforms. AWS Lambda[9] can be considered as the first platform that introduced such concepts, along with Microsoft Azure Functions[10]. AWS Lambda requires that the services run in AWS machines, which is not suitable for iHelp that deals with sensitive data that might not be able to leave the healthcare organization, while Microsoft Azure requires the use of the Azure application platform which is a vendor lock-in that the project is not willing to take. Google Cloud Functions[11] also suffer from the same constraint due the handling of sensitive data and the fact that iHelp tends the use of a private cloud. Other solutions are Iron Functions[12], Auth0 WebTask[13] and Galactic Fog Gestal Laser[14]. However, as the IBM's OpenWhisk[15] has been currently moved to the Apache Foundation, the project tends to be more favourable of this solution due to its wider community support and the experience of many of the technical partners of the consortium with this platform.

4.2.3 Background technology

The main technology the Primary Data Capture and Ingestion will make use of is Openwhisk that will allow for the dynamic deployment of serverless functions to establish he data ingestion pipeline.

Table 79: T3.2 – Background Technology

Technology Name	Technology Description	Advancements / Usage
Apache OpenWhisk	A distributed Serverless platform that executes functions in response to events, managing the infrastructure, severs and scaling using Docker	In iHelp we will extend the OpenWhisk automating the deployment of a set of functions to establish a data pipeline communicating using Kafka queues

[9] https://aws.amazon.com/lambda/
[10] https://azure.microsoft.com/en-us/services/functions/
[11] https://cloud.google.com/functions/
[12] https://open.iron.io/
[13] https://webtask.io/
[14] https://hub.docker.com/_/galacticfog
[15] https://openwhisk.apache.org/

	containers	

4.2.4 Module to User Requirements

Table 80: T3.2 – 1st Functionality

Section	Description
ID	U-RED-T3.2-01
Title	Primary Data Ingestion
Functionality Description	This module will be the basis for the establishing the data ingestion pipelines and deploy and interconnect the corresponding functions for cleansing, transformation, and storage of both primary and secondary data to the big data platform.
Source User Requirement	This module is primarily related with the following user requirements: REQ-P1-01, REQ-P1-02, REQ-P3-04, REQ-P4-01, REQ-P4-02, REQ-P4-03, REQ-P4-04, REQ-P5-01, REQ-P5-02 More, it must also address secondary requirements that can be further derived from the ones listed in section 2 and are related with the ingestion of both primary and secondary data.
Use case Quote	Access to and evaluation of large amount data …
Generic / Specific	Generic
Task / Component	T3.2 Primary Data Capture and Ingestion
Lead Partner	LXS
Notes	This module should make use of the same underlying technology as of the Analytic Workbench.

4.2.5 Module to Technical Requirements

Table 81: T3.2 – 1st Technical requirement

Section	Description
ID	T-REQ-T3.2-01
Type	ENV
Short Name	Dynamic deployment of functions

Functionality ID	U-RED-T3.2-01
Description & quantification	Once functions have been packaged, they should be able to be deployed automatically, without the manual creation of virtualized resources
Additional information	
Priority	MAN
Reference Pilot	All
Success Criteria	Each function can be deployed dynamically without the manually intervention of a system administrator

Table 82: T3.2 – 2nd Technical requirement

Section	Description
ID	T-REQ-T3.2-02
Type	FUN
Short Name	Transparent communication of functions
Functionality ID	U-RED-T3.2-01
Description & quantification	The deployed functions must communicate transparently via a mean, without having to take care the responsibility of establishing the means of communication between them
Additional information	
Priority	MAN
Reference Pilot	All
Success Criteria	Each function must be able to consume and produce results via a standard method infrastructure. Functions must not implement their own ways of communication

Table 83: T3.2 – 3rd Technical requirement

Section	Description
ID	T-REQ-T3.2-03
Type	ENV
Short Name	Automatic Orchestration of the Choreography

Functionality ID	U-RED-T3.2-01
Description & quantification	There must be an API that allows for the definition of which services take part in each data pipeline, so that the deployment can be automated.
Additional information	We call this requirement as the *orchestration of service choreography*. Instead of manually deploying each service that takes part in the choreography of the data pipeline, there should be an API that orchestrates this deployment. This will allow the descriptive definition of which services are part of each instantiated data pipeline, and deploy them while establishing the communication means among them
Priority	DES
Reference Pilot	All
Success Criteria	A data pipeline can be deployed with a single instructor

Table 84: T3.2 – 4th Technical requirement

Section	**Description**
ID	T-REQ-T3.2-04
Type	L&F
Short Name	Data Ingestion UI Dashboard
Functionality ID	U-RED-T3.2-01
Description & quantification	A UI dashboard should allow the data provider to define the data pipeline, and once this is defined, this should be translated to a REST call to the *orchestrator of choreography* that will deploy the data pipeline
Additional information	
Priority	OPT
Reference Pilot	All
Success Criteria	We can use a web UI to deploy a data pipeline, instead of console commands

Table 85: T3.2 – 5th Technical requirement

Section	**Description**

ID	T-REQ-T3.2-05
Type	FUNC
Short Name	Data source connectivity
Functionality ID	U-RED-T3.2-01
Description & quantification	Different mechanisms for data connectivity (REST pull, DB connections, FTP connections) must be provided by the module to allow retrieving data from external sources and put them as an input to the data ingestion pipelines
Additional information	
Priority	MAN
Reference Pilot	All
Success Criteria	Data can be captured from external sources by all required means for connectivity

4.3 Secondary Data Extraction and Interoperability

4.3.1 Goals and Objectives

The goal of secondary data extraction is to enrich the HHRs that traditionally contain medical/clinical data (primary data), with data about patients'/peoples' lifestyle. These we term Real-World Data, to discriminate data captured continuously and ubiquitously from the everyday life of our participants to data collected sporadically in a clinical setting.

Raw data are collected, measured by wearable or other devices, or reported by our participants using our questionnaires. All these data are processed at the edge, to reason about the habits, behaviours, and relationships of our participants.

4.3.2 State of the Art

Activity tracking is becoming widespread, with an increased amount of people using their smartphones or dedicated activity trackers do understand their physical activity, sleep, or heart. As technology progresses, more measurements are made available, and the trackers' form factor changes: (largish) watches, compact bands and screen-less rings can be found as commercial products. IoT devices are also used in sporadically collecting medical signals at home.

Questionnaires have been used in collecting data from participants since the early days of studies. Traditionally questionnaires are in paper, the data being collected in interviews, from participants' memory or diaries. In recent years electronic systems from collecting Patient Reported Outcomes (ePRO) emerged: Castor EDC, Medidata Rave, Encapsia, Medable, Science 37, uMotif and Healthentia from Innovation Sprint.

4.3.3 Background technology

Healthentia is an eCLinical platform that facilitates RWD collection and analytics. The platform is built around three pillars:

- Secure Real-World Data collection involving measurements and questionnaires. The measurements are done either by a sensing service implemented in Healthentia, or via data ingestion built in Healthentia interfacing different activity tracking services, like Fitbit, Garmin or Apple Health.
- Analytics provision via dashboards addressing the patient or the healthcare professional.
- Smart services for participants' understanding (phenotyping and prediction) and advice delivery.

The platform's users interact with Healthentia via the two apps:

- The mobile app is what the study participants use to measure and report as well as view data. It is available on Google Play and Apple Store. The app is also used for advice delivery, as discussed in section 4.11.
- The web portal app is where healthcare professionals view data and machine learning results about the participants in individual or collective modes.

All pilots have agreed to use Healthentia for secondary data collection. Currently a demo study is ongoing for partners to get used to the current version of the mobile and portal apps. In the project, an extended version of the app will be used for secondary data collection. The extensions planned are spanning two areas:

- RWD collection extensions: Integrate different/more devices, both wearables and sporadically used for measurements (scales, blood pressure monitors, etc.). Also, implement the necessary questionnaires.
- Machine learning extensions: Implement the edge processing for the extraction of the habits, behaviours and relationships of our participants.

Table 86 T3.3 Background Technology

Technology Name	Technology Description	Advancements / Usage
Healthentia Edge Component	Secondary data collection app & platform, including reasoning module for higher –level secondary data	▪ More measurements (devices, questionnaires) ▪ Ml for the extraction of habits, behaviours and relationships

4.3.4 Module to User Requirements

The tables below describe a specific functionality that needs to be supported by the secondary data extraction component. For each of these functionalities, we refer to the source user requirement (as defined in Section 2) and/or the source secondary data user requirement (as defined in Section 3).

Table 87: T3.3 – 1st Functionality

Section	Description
ID	U-REQ-T3.3-01
Title	Physical activity
Functionality description	Collecting different aspects of physical activity
Source User Requirement	REQ-APP-P1-003, REQ-MEAS-P1-004, REQ-MEAS-P1-006, REQ-MEAS-P1-005 REQ-MEAS-P2-001 REQ-MEAS-P3-001, REQ-MEAS-P3-005, REQ-MEAS-P3-006, REQ-P3-02 REQ-P5-02
Use case Quote	▪ Physical activity (REQ-MEAS-P1-004, REQ-P5-02) ▪ Use of a wearable (REQ-APP-P1-003) ▪ Regulatory compliance of wearable: CE certificate, technical documentation (REQ-MEAS-P1-006, REQ-MEAS-P3-006, REQ-P3-02) ▪ Steps, floors climbed, energy expenditure – MAN (REQ-MEAS-P2-001, REQ-MEAS-P3-001) ▪ Auto-detected physical activities (exercise sessions) – OPT (REQ-MEAS-P1-005, REQ-MEAS-P3-005)
Generic / Specific	Generic
Task / Component	T3.3, Secondary Data Extraction, and Interoperability
Lead Partner	Innovation Sprint
Notes	More information is needed on the measurement's frequency. Is intraday info useful? No needs reported by pilot 4.

Table 88: T3.3 – 2nd Functionality

Section	Description
ID	U-REQ-T3.3-02
Title	Diet
Functionality description	Collecting dietary information
Source User Requirement	REQ-MEAS-P1-004 REQ-MEAS-P4-004
Use case Quote	▪ Diet: REQ-MEAS-P1-004 ▪ Red & processed meat (REQ-MEAS-P4-004) ▪ Vegetables & grain (REQ-MEAS-P4-004)
Generic / Specific	Generic
Task / Component	T3.3, Secondary Data Extraction, and Interoperability
Lead Partner	Innovation Sprint
Notes	Pilot 1 needs to elaborate their needs. Pilots 2,3 and 5 have not reported any dietary reporting needs

Table 89: T3.3 – 3rd Functionality

Section	Description
ID	U-REQ-T3.3-03
Title	Lifestyle aspects
Functionality description	Collecting lifestyle information
Source User Requirement	REQ-MEAS-P1-004 REQ-MEAS-P4-004 REQ-P5-02
Use case Quote	▪ Alcohol (REQ-MEAS-P1-004, REQ-MEAS-P4-004) ▪ Smoking: REQ-MEAS-P1-004 ▪ Habits (REQ-P5-02)
Generic / Specific	Generic
Task / Component	T3.3, Secondary Data Extraction and Interoperability

Lead Partner	Innovation Sprint
Notes	Pilot 5 needs to elaborate their needs. Pilots 2 and 3 have not reported any lifestyle reporting needs

Table 90: T3.3 – 4th Functionality

Section	**Description**
ID	U-REQ-T3.3-04
Title	Heart information
Functionality description	Collecting heart information
Source User Requirement	REQ-MEAS-P2-001 REQ-MEAS-P3-001
Use case Quote	▪ Sporadic heart rate entry as a symptom - MAN (REQ-MEAS-P2-001, REQ-MEAS-P3-001) ▪ Min, max, resting heart rate per day automatically measured – OPT/ENH (REQ-MEAS-P2-001, REQ-MEAS-P3-001)
Generic / Specific	Generic
Task / Component	T3.3, Secondary Data Extraction and Interoperability
Lead Partner	Innovation Sprint
Notes	Pilots 2 and 3 only need heart rate, preferably supported by the activity tracker. Worst case sporadic measurements should be manually entered.

Table 91: T3.3 - 5th Functionality

Section	**Description**
ID	U-REQ-T3.3-05
Title	Sleep information
Functionality description	Collecting sleep information
Source User Requirement	REQ-MEAS-P2-001 REQ-MEAS-P3-001 REQ-P5-02
Use case Quote	▪ Sleep metrics – MAN (REQ-MEAS-P2-004, REQ-MEAS-P3-004) ▪ Subjective sleep assessment, sleep behaviour (REQ-P5-02)

Generic / Specific	Generic
Task / Component	T3.3, Secondary Data Extraction and Interoperability
Lead Partner	Innovation Sprint
Notes	Pilots 2,3 and 5 are interested in sleep information. All three pilots need to give details on what they need.

Table 92: T3.3 - 6th Functionality

Section	**Description**
ID	U-REQ-T3.3-06
Title	Body signals
Functionality description	Collecting information about different body measurements
Source User Requirement	REQ-MEAS-P1-003 REQ-MEAS-P2-002 REQ-MEAS-P3-002 REQ-MEAS-P4-002
Use case Quote	▪ Weight manual entry, BMI given height (REQ-MEAS-P2-002, REQ-MEAS-P3-002, REQ-MEAS-P4-002) ▪ Blood pressure manual entry - OPT (REQ-MEAS-P2-002, REQ-MEAS-P3-002) ▪ Blood oxygen saturation automatic periodic measurement - DES (REQ-MEAS-P2-002, REQ-MEAS-P3-002) ▪ Blood oxygen saturation manual entry - MAN (REQ-MEAS-P2-002, REQ-MEAS-P3-002) ▪ Body temperature manual entry - OPT (REQ-MEAS-P2-003, REQ-MEAS-P3-003, REQ-MEAS-P4-002) ▪ Blood sample test (REQ-MEAS-P1-003)
Generic / Specific	Generic
Task / Component	T3.3, Secondary Data Extraction and Interoperability
Lead Partner	Innovation Sprint
Notes	Pilot 1 needs blood pressure test. This is clinical info, not part of secondary data. Pilots 2, 3 and 4 are interested in weight, blood pressure and body temperature manual entry, as well as blood oxygen saturation, if possible, automatically acquired. Pilot 5 is not interested in body signal measurements.

Table 93: T3.3 - 7th Functionality

Section	Description
ID	U-REQ-T3.3-07
Title	Symptoms
Functionality description	Collecting information about symptoms the patients experience
Source User Requirement	REQ-MEAS-P4-002 REQ-P5-02
Use case Quote	▪ Pain level (REQ-MEAS-P4-002, REQ-P5-02) ▪ Diarrhoea, breathing difficulties, emotional state, fatigue, loss of smell/ taste, headache, muscle pain, nausea, etc. (REQ-P5-02)
Generic / Specific	Generic
Task / Component	T3.3, Secondary Data Extraction, and Interoperability
Lead Partner	Innovation Sprint
Notes	Pilots 1, 2 and 3 are not interested in symptoms collection. Is location of pain important? Is intensity of symptoms important or only occurrence?

Table 94: T3.3 - 8th Functionality

Section	Description
ID	U-REQ-T3.3-08
Title	Medication
Functionality description	Collecting information about medication
Source User Requirement	REQ-MEAS-P4-002
Use case Quote	Daily medication (REQ-MEAS-P4-002)
Generic / Specific	Specific
Task / Component	T3.3, Secondary Data Extraction and Interoperability
Lead Partner	Innovation Sprint

Notes	Only pilot 4 is interested in medication. What is needed though is unclear. Is it a medication diary? A questionnaire?

Table 95: T3.3 - 9th Functionality

Section	Description
ID	U-REQ-T3.3-09
Title	Patient enrolment
Functionality description	Collecting information about patients at enrolment (initial questionnaire)
Source User Requirement	REQ-MEAS-P1-001 REQ-MEAS-P2-002 REQ-MEAS-P3-002 REQ-MEAS-P4-002 REQ-P5-02
Use case Quote	▪ Baseline general characteristics (REQ-MEAS-P1-001) ▪ Height (REQ-MEAS-P2-002, REQ-MEAS-P3-002, REQ-MEAS-P4-002) ▪ Basic user details, family history (REQ-P5-02)
Generic / Specific	Generic
Task / Component	T3.3, Secondary Data Extraction and Interoperability
Lead Partner	Innovation Sprint
Notes	All pilots need initial info about patients but must be more specific on what will be asked of them.

Table 96: T3.3 - 10th Functionality

Section	Description
ID	U-REQ-T3.3-10
Title	Scheduled questionnaires
Functionality description	Forwarding questionnaires to patients (in general or to specific ones)
Source User Requirement	REQ-MEAS-P1-002 REQ-MEAS-P2-007 REQ-MEAS-P3-007, REQ-P3-02 REQ-MEAS-P4-001 REQ-MEAS-P5-003
Use case Quote	▪ EORTC QLQ-C30 questionnaire - MAN (REQ-MEAS-P2-007, REQ-MEAS-P3-007, REQ-MEAS-P5-003)

	- EORTC QLQ-PAN26 questionnaire – MAN (REQ-MEAS-P2-007, REQ-MEAS-P3-007) - COVID-19 prevention daily questionnaire – MAN (REQ-MEAS-P2-007, REQ-MEAS-P3-007) - Questionnaires about lifestyle & habits (REQ-P3-02) - Questionnaires about lifestyle related to cancer risk assessment (REQ-MEAS-P1-002) - Questionnaires about evolution of the disease (REQ-P3-02) - Questionnaire about quality of life (REQ-MEAS-P4-001) - Ability to schedule which questionnaires are addressed to which participants within a pilot (REQ-P3-02)
Generic / Specific	Generic
Task / Component	T3.3, Secondary Data Extraction and Interoperability
Lead Partner	Innovation Sprint
Notes	All pilots need to forward questionnaires to patients. The specific needs must be provided in place of general characterization of a questionnaire.

Table 97: T3.3 - 11th Functionality

Section	Description
ID	U-REQ-T3.3-11
Title	Psychological information
Functionality description	Collecting psychological information
Source User Requirement	REQ-MEAS-P2-007 REQ-MEAS-P3-007 REQ-P5-02
Use case Quote	- Eventual psychological condition extracted from social media – OPT (REQ-MEAS-P2-007, REQ-MEAS-P3-007) - Stress, happiness, depression, anxiety, mood (REQ-P5-02)
Generic / Specific	Generic
Task / Component	T3.3, Secondary Data Extraction and Interoperability
Lead Partner	Innovation Sprint
Notes	Pilots 1 and 4 are not interested in psychological information. Pilot 5 should specify the exact question to be asked, and if they consider psychological information entry similar to symptom entry.

Table 98: T3.3 - 12th Functionality

Section	Description
ID	U-REQ-T3.3-12
Title	Secondary data ingestion
Functionality description	Ingesting secondary data
Source User Requirement	REQ-P3-04 REQ-APP-P5-003
Use case Quote	Information should be transferred to the pilot healthcare institution (REQ-P3-04) Questionnaires handled by nurses, patient reports via the app (REQ-APP-P5-003)
Generic / Specific	Generic
Task / Component	T3.3, Secondary Data Extraction and Interoperability
Lead Partner	Innovation Sprint
Notes	Questionnaires are also handled via the app.

Table 99: T3.3 - 13th Functionality

Section	Description
ID	U-REQ-T3.3-13
Title	User experience
Functionality description	Information about the experience of the patients and their engagement
Source User Requirement	REQ-P5-02
Use case Quote	▪ User experience assessment of the digital solution (REQ-P5-02) ▪ User engagement automatically measured (REQ-P5-02)
Generic / Specific	Specific
Task / Component	T3.3, Secondary Data Extraction and Interoperability
Lead Partner	Innovation Sprint

Notes	User engagement with the data collection app is automatically inferred by the reported data. The app also has the feedback aspects, that should be considered in T5.3. User aspects are of interest only to pilot 5.

4.3.5 Module to Technical Requirements

Table 100: T3.3 - 1st Technical requirement

Section	Description
ID	T-REQ-T3.3-01
Type	FUNC
Short Name	Activity tracking
Functionality ID	U-REQ-T3.3-01, U-REQ-T3.3-04, U-REQ-T3.3-05, U-REQ-T3.3-06
Description & quantification	Physical activity, sleep, heart and SPO2 information must be collected with an integrated, CE certified and technical documented activity tracker.
Additional information	It is still to be determined if daily summaries or intra-day data are needed.
Priority	MAN
Reference Pilot	P1, P2, P3, P5
Success Criteria	The 3rd party activity tracker is integrated, providing the required data.

Table 101: T3.3 - 2nd Technical requirement

Section	Description
ID	T-REQ-T3.3-02
Type	FUNC
Short Name	Exercise sessions
Functionality ID	U-REQ-T3.3-01
Description & quantification	Auto-detected exercise sessions should be registered.
Additional information	It is still to be determined which data apart from activity type, time and duration are needed.

Priority	DES
Reference Pilot	P1, P3
Success Criteria	The 3rd party activity tracker supporting exercise session auto-detection is integrated, providing the required data.

Table 102: T3.3 - 3rd Technical requirement

Section	**Description**
ID	T-REQ-T3.3-03
Type	FUNC
Short Name	Lifestyle information
Functionality ID	U-REQ-T3.3-02, U-REQ-T3.3-03
Description & quantification	Patients must be given the possibility to report on their diet, smoking and alcohol intake.
Additional information	Mode of information collection has to be identified: Daily questionnaires or widgets the patient manipulates to add/remove stuff at any time (even for past days)
Priority	MAN
Reference Pilot	P1, P4, P5
Success Criteria	The mobile application allows patients to report daily habit data.

Table 103: T3.3 - 4th Technical requirement

Section	**Description**
ID	T-REQ-T3.3-04
Type	FUNC
Short Name	Manual entry of body measurements
Functionality ID	U-REQ-T3.3-06
Description & quantification	Patients must be given the possibility to log their weight, blood pressure, body temperature (and possibly SPO2) at any time.
Additional information	Log (date/time, value) pairs. Need to finalize measurements' list.

Priority	MAN
Reference Pilot	P1, P2, P3, P4
Success Criteria	The mobile application allows patients to report body metrics and other measurements from non-connected devices at will.

Table 104: T3.3 - 5th Technical requirement

Section	Description
ID	T-REQ-T3.3-05
Type	FUNC
Short Name	Symptoms' logging
Functionality ID	U-REQ-T3.3-07
Description & quantification	Patients must be given the possibility to log a set of symptoms.
Additional information	Just date/time or also intensity? Need to finalize symptoms' list.
Priority	MAN
Reference Pilot	P4, P5
Success Criteria	The mobile application allows patients to report symptoms at will.

Table 105: T3.3 - 6th Technical requirement

Section	Description
ID	T-REQ-T3.3-06
Type	FUNC
Short Name	Logging medication
Functionality ID	U-REQ-T3.3-08
Description & quantification	Patients must be given the possibility to log their medication intake.
Additional information	List of medication to select from, or free text?
Priority	MAN

Reference Pilot	P4
Success Criteria	The mobile application allows patients to log medication.

Table 106: T3.3 - 7th Technical requirement

Section	Description
ID	T-REQ-T3.3-07
Type	FUNC
Short Name	Enrolment data collection
Functionality ID	U-REQ-T3.3-09
Description & quantification	Patients must be given an enrolment questionnaire to answer.
Additional information	Information collected needs to be finalised.
Priority	MAN
Reference Pilot	All
Success Criteria	The mobile application allows patients to fill in and submit one-time questionnaires.

Table 107: T3.3 - 8th Technical requirement

Section	Description
ID	T-REQ-T3.3-08
Type	FUNC
Short Name	Scheduled questionnaires
Functionality ID	U-REQ-T3.3-10
Description & quantification	Patients must be given questionnaires at different intervals, or conditional to their state.
Additional information	The exact questionnaires to be used are not always determined.
Priority	MAN
Reference Pilot	All

Success Criteria	The mobile application allows patients to fill in and submit scheduled questionnaires.

Table 108: T3.3 - 9th Technical requirement

Section	**Description**
ID	T-REQ-T3.3-09
Type	FUNC
Short Name	Scheduled questionnaires
Functionality ID	U-REQ-T3.3-10
Description & quantification	Patients must be given questionnaires at different intervals, or conditional to their state.
Additional information	The exact questionnaires to be used are not always determined.
Priority	MAN
Reference Pilot	All
Success Criteria	The iHelp system is able to conditionally send patients notifications to fill in and submit questionnaires.

Table 109: T3.3 - 10th Technical requirement

Section	**Description**
ID	T-REQ-T3.3-10
Type	FUNC
Short Name	Reported psychological information
Functionality ID	U-REQ-T3.3-11
Description & quantification	Patients must be able to report their psychological state.
Additional information	The psychological info needs to be determined. The mode of information collection is the same as symptoms.
Priority	MAN
Reference Pilot	P5

Success Criteria	The mobile application allows patients to submit information related to their psychological state.

Table 110: T3.3 - 11th Technical requirement

Section	**Description**
ID	T-REQ-T3.3-11
Type	FUNC
Short Name	Extracted psychological information
Functionality ID	U-REQ-T3.3-11
Description & quantification	Psychological condition of the patients should be extracted from their social media.
Additional information	This could be a requirement for T5.4, Social Analytics. O be discussed with WP5/T5.4.
Priority	OPT
Reference Pilot	P2, P3
Success Criteria	The iHelp platform is able to determine patients' psychological condition from their social media.

Table 111: T3.3 - 12th Technical requirement

Section	**Description**
ID	T-REQ-T3.3-12
Type	ENV
Short Name	Secondary data availability
Functionality ID	U-REQ-T3.3-12
Description & quantification	The organization hosting the study must have access to the data.
Additional information	The organization hosting the study is the data controller in GDPR. Healthentia/Innovation Sprint are data processors. The data belongs to the hosting organization, and they should formally authorize Innovation Sprint to offer the data to the iHelp platform and/or monitor the study on their behalf through a Data Processing Agreement. They can obtain the data at any time using the data exporting features of Healthentia and/or the Healthentia API.

Priority	MAN
Reference Pilot	All
Success Criteria	The pilot partners can get the data exports they require by having their SW systems utilising the provided API endpoints.

Table 112: T3.3 - 13th Technical requirement

Section	Description
ID	T-REQ-T3.3-13
Type	USE
Short Name	User experience reporting
Functionality ID	U-REQ-T3.3-13
Description & quantification	The patients should report on their experience with the data collection app at some milestone instances
Additional information	The reporting must be done using means external to the data collection app.
Priority	MAN
Reference Pilot	All
Success Criteria	There is an independent feedback mechanism the patients are asked to use to periodically report their experience with the data collection system.

4.4 Standardisation and Quality Assurance of Heterogeneous Data

4.4.1 Goals and Objectives

Nowadays, healthcare domain faces various challenges due to the diversity and variety of data, their huge volume, and their distribution, thus processing and analysis of these data become more and more complex and difficult procedures. Hence, approaches, applications and solutions that seek to address issues that derive from the wealth of Big Data are vitally important. The collection, the quality estimation, as well as the interpretation and the harmonization of the data, that derive from the existing huge amounts of heterogeneous medical devices and data sources, face a dramatic increase of interest in the healthcare domain. To this end, to address all these issues this specific task of the iHelp project has three-fold objectives and targets. In one hand it seeks to assure the incoming data accuracy, integrity, and quality, while in the meanwhile it seeks to provide an automated structure mapping mechanism between data resources and widely known and approved in the healthcare domain HL7 FHIR resources (D., K., 13). On top of this, this task aims through the utilization of specific measures and rules to ensure provide a predictive selection mechanism for achieving data sources' reliability during runtime and for providing the decision whether a connected data source will be considered as reliable or not because of the execution. Based on the above three (3) discrete, but integrated with each other, subcomponents have been identified, as presented in the below figure (Fig. 9).

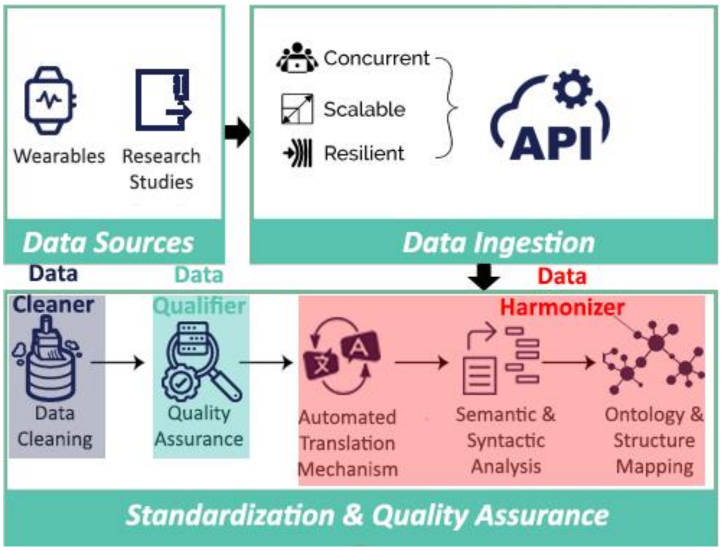


Figure 9: Standardization & Quality Assurance mechanism

4.4.2 State of the Art

Under the scopes of this task state-of-the-art approaches and techniques from the fields of AI and Semantic Web technologies will be utilized in order to enhance the quality of the incoming data, the interoperability and harmonization of them and the extraction of valuable information and knowledge out of them. To this end, three different subcomponents have been identified, as, also, mentioned in the previous section (Section 4.4.1). These specific three subcomponents are the Data Cleaner, the Data Qualifier and the Data Harmonizer, each of them with the aim to enhance the value of the incoming/raw data.

- ***Data Cleaner***: Especially in the healthcare domain, clean data can lead to better decision making, care of high value, and reduced inefficiencies. To this end, this subcomponent seeks to deliver the software implementation that will provide the assurance that the provided data coming from several heterogeneous data sources will be clean and complete, to the extent possible. This microservice will be designed to minimize and filter the non-important data, thus improving the data quality and importance. On top of this, the Data Cleaner service will be comprised of a multi-fold process consisted of three (3) sub mechanisms which are listed below and are also presented in the below figure (Fig. 10).

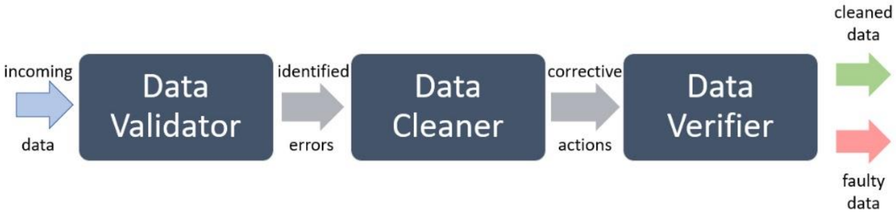


Figure 10: Data Cleaner Microservice

  - *Data Validator*: This mechanism will ensure that the rest of the iHelp microservices will operate on clean, correct and useful data. Therefore, the Data Validation service will perform data validation of the incoming information data with the purpose of identifying errors associated with the conformance to specific set of constraints.
  - *Data Cleaner*: Entails the main sub mechanism of the Data Cleaner subcomponent. Its main goal is to correct or remove all the data elements for which validation errors were raised, considering missing, irregular, unnecessary, and inconsistent data. Thus, the Data Cleaning sub mechanism will perform the necessary corrections or removals of errors identified by the Data Validation Service.
  - *Data Verifier*: The main objective of this sub mechanism is to check the data elements of a dataset for accuracy and inconsistencies after the steps of data validation and cleaning are performed. To this end, it will ensure that all the corrective actions performed by the Data Cleaning service will be executed in compliance to the data models' design of the iHelp platform. To this end, this service seeks to ensure that the data will accurately be corrected or completed, and the dataset will eventually be error free.
- ***Data Qualifier***: The Data Qualifier subcomponent will be utilized for every new registered data source and every incoming dataset in the platform since it seeks to enhance and adapt the selection of reliable sources. This specific microservice will also be integrated with the message bus mechanism provided in the scope of iHelp project. To this end, the whole dataset and information about the data source should be the input to this microservice in order to provide the reliability levels of the connected data sources into a message (i.e. String) format, which will be annotated to the dataset as metadata.
- ***Data Harmonizer***: In this final step high-quality data are kept and translated into a common format, being able to be used for further utilization. Achieving true interoperability of data entails different representations, purposes, and syntaxes and will enable improved access to assets, records, datasets, and policies. The European Commission, through their program ISA[2] has defined the European Interoperability Framework (EIF) which defines interoperability across four layers: (i) organizational interoperability, (ii) semantic interoperability, (iii) technical interoperability and (iv)

legal interoperability. Semantic Interoperability is the aspect of interoperability which is concerned with ensuring that the precise format and meaning of exchanged data and information is preserved and understood throughout by any other application that was not initially developed for this purpose. Therefore, it enables systems to combine received information with other information resources and to process it in a meaningful manner, hence it is a prerequisite for the front-end multilingual delivery of services to the user. To this end, achieving meaningful Semantic Interoperability of data from heterogenous sources is a challenging issue for policy makers, as it is a complex procedure since it covers both semantic and syntactic aspects. A research in the biological sector highlights the usage of a JSON-LD system, which provides a standard way to add semantic context to the existing JSON data structure, for the purpose of enhancing the interoperability between APIs and data. iHelp project will enhance interoperability based on data driven-design, coupled with linked data technologies (e.g. JSON-LD and RDF) and standards-based ontologies and vocabularies from the HL7 FHIR format in order to improve both semantic and syntactic interoperability of the incoming. Moreover, a data modelling by standard metadata schemas will be defined in order to specify the metadata elements that should accompany a dataset within a domain. In deeper details, the Data Harmonizer mechanism incorporates the utilization of three (3) integrated subcomponents, as also presented in the below figure (Fig. 11).

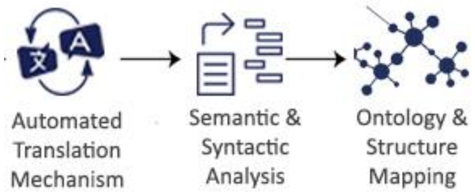

Automated Translation Mechanism
Semantic & Syntactic Analysis
Ontology & Structure Mapping

Figure 11: Data Harmonizer Microservice

4.4.3 Background technology

Table 113: T3.4 - Background Technology

Technology Name	Technology Description	Advancements / Usage
Scikit-learn	Incorporates tools and techniques for data processing and cleaning.	It contains implementations of machine learning models and techniques to be evaluated, i.e. text preprocessing, tokenization, categorization algorithms, clustering, and deep neural network.
Pandas	A Python tool that provides fast, flexible, and expressive data structures designed to make working with structured and time series data.	It contains tools and techniques that facilitate the munging and cleaning of data, while also the analysis, modelling and transformation of the data.
Numpy	A Python library for performing vectorization, indexing, and array	It offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier

	procedures.	transforms, and more.
Keras	A powerful and easy-to-use free open-source Python library for developing and evaluating deep learning models.	Keras is a deep learning API running on top of the machine learning platform TensorFlow. Incorporates and facilitates the utilization of powerful Neural Networks and state-of-the-art Transformers models.
Cerberus	A Python validation library which provides powerful yet simple and lightweight data validation functionality.	Cerberus works by defining a validation schema for data to be processed and analyzed.
spaCy	A library for advanced Natural Language Processing in Python.	It contains pretrained pipelines and supports tokenization and training for 60+ languages. It features neural network models for tagging, parsing, named entity recognition, text classification and more.
TensorFlow	An open-source library that supports traditional machine learning tools as well deep learning applications.	With TensorFlow it will be feasible to train and run deep neural networks for image recognition, word embeddings, recurrent neural networks, sequence-to-sequence models for machine translation, and Natural Language Processing (NLP) simulations. It also supports production prediction at scale.

4.4.4 Module to User Requirements

Table 114: T3.4 - 1st Functionality

Section	Description
ID	U-REQ-T3.4-01
Title	Data Standardization
Functionality Description	Facilitates the combination and integration of received information with other information resources and to process all incoming data in a meaningful manner.
Source User Requirement	REQ-P3-04, REQ-P4-01, REQ-P4-02, REQ-P4-03, REQ-P4-04
Use case Quote	“Integration of the information collected in the HCE of Marina Salud” and “Data assessment”
Generic / Specific	Generic

Task / Component	Data Harmonizer
Lead Partner	UPRC
Notes	Further to the User Requirements that have been identified, this functionality will be utilized under the scopes of every Pilot partner and scenario and across all incoming data.

Table 115: T3.4 – 2nd Functionality

Section	**Description**
ID	U-REQ-T3.4-02
Title	Data Processing & Data Cleaning
Functionality Description	This functionality seeks to provide the assurance that the provided data coming from several heterogeneous data sources will be cleaned and processed in order to ensure the correct analysis of them.
Source User Requirement	REQ-P1-03, REQ-P1-01
Use case Quote	“Successful processing the sample and data return” & “Data access”
Generic / Specific	Generic
Task / Component	Data Cleaner
Lead Partner	UPRC
Notes	Further to the User Requirements that have been identified, this functionality will be utilized under the scopes of every Pilot partner and scenario and across all incoming data.

Table 116: T3.4 – 3rd Functionality

Section	**Description**
ID	U-REQ-T3.4-03
Title	Data Annotation
Functionality Description	This functionality seeks to annotate incoming data and datasets with appropriate metadata in order to enhance the interoperability and add extra value and knowledge to them.
Source User Requirement	REQ-P3-05

Use case Quote	“Develop Risk Prediction Model”
Generic / Specific	Generic
Task / Component	Data Harmonizer
Lead Partner	UPRC
Notes	Further to the User Requirements that have been identified, this functionality will be utilized under the scopes of every Pilot partner and scenario and across all incoming data.

Table 117: T3.4 – 4th Functionality

Section	Description
ID	U-REQ-T3.4-04
Title	Knowledge Graphs
Functionality Description	This functionality seeks correlate and interlink high-quality data to common knowledge graphs. Through this approach internal and external data can be automatically linked and can be used as rich datasets for any Machine Learning (ML) and AI task.
Source User Requirement	REQ-P3-05
Use case Quote	“Develop Risk Prediction Model”
Generic / Specific	Generic
Task / Component	Data Harmonizer
Lead Partner	UPRC
Notes	N/A

4.4.5 Module to Technical Requirements

Table 118: T3.4 – 1st Technical Requirement

Section	Description
ID	T-REQ-T3.4-01
Type	FUN

Short Name	Standardised Interface to other internal iHelp components
Functionality ID	U-REQ-T3.4-01
Description & quantification	It should facilitate the standardised connection to other internal components of the iHelp platform, such as the Data Gateway. The standardisation of the messages should follow a well-defined and structured format, such us XML or JSON.
Additional information	N/A
Priority	MAN
Reference Pilot	All
Success Criteria	Proper specification of message structure

Table 119: T3.4 – 2nd Technical Requirement

Section	Description
ID	T-REQ-T3.4-02
Type	FUN
Short Name	Error Identification
Functionality ID	U-REQ-T3.4-02
Description & quantification	The Data Cleaner subcomponent should facilitate the identification of errors associated with conformance to specific constraints, safeguarding that the data measures compare to defined business rules or constraints.
Additional information	Python libraries of Pandas, NumPy, Scikit-learn, Keras will be utilized under this scope.
Priority	MAN
Reference Pilot	All
Success Criteria	Identification of on-purpose included errors

Table 120: T3.4 – 3rd Technical Requirement

Section	Description
ID	T-REQ-T3.4-03
Type	FUN

Short Name	Conformance to specific data type
Functionality ID	U-REQ-T3.4-02
Description & quantification	The Data Cleaner subcomponent should facilitate the conformance to specific and needed data types (integers, strings etc.).
Additional information	Python libraries of Pandas, NumPy, Scikit-learn, Keras will be utilized under this scope.
Priority	MAN
Reference Pilot	All
Success Criteria	Identification of correct and appropriate data types.

Table 121: T3.4 – 4th Technical Requirement

Section	Description
ID	T-REQ-T3.4-04
Type	FUN
Short Name	Conformance to specific and predefined data values
Functionality ID	U-REQ-T3.4-02
Description & quantification	The Data Cleaner subcomponent should safeguard the conformance to specific and predefined data values (e.g. specific values from a dropdown list).
Additional information	Python libraries of Pandas, NumPy, Scikit-learn, Keras will be utilized under this scope.
Priority	MAN
Reference Pilot	All
Success Criteria	Identification of correct and needed data values.

Table 122: T3.4 – 5th Technical Requirement

Section	Description
ID	T-REQ-T3.4-05
Type	FUN

Short Name	Conformance to specific and predefined regular expression patterns
Functionality ID	U-REQ-T3.4-02
Description & quantification	The Data Cleaner subcomponent should safeguard the conformance to specific and predefined regular expression patterns (e.g. data that has a certain pattern in the way it is displayed, such as emails, ICD10 codes etc.).
Additional information	Python libraries of Pandas, NumPy, Scikit-learn, Keras will be utilized under this scope.
Priority	OPT
Reference Pilot	All
Success Criteria	Identification of specific and predefined regular expression patterns.

Table 123: T3.4 – 6th Technical Requirement

Section	**Description**
ID	T-REQ-T3.4-06
Type	FUN
Short Name	Conformance to unique values
Functionality ID	U-REQ-T3.4-02
Description & quantification	The Data Cleaner subcomponent should safeguard the conformance to unique values (e.g. patients' ID must be unique).
Additional information	Python libraries of Pandas, NumPy, Scikit-learn, Keras will be utilized under this scope.
Priority	OPT
Reference Pilot	All
Success Criteria	Identification of non-conformance to uniqueness

Table 124: T3.4 – 7th Technical Requirement

Section	**Description**
ID	T-REQ-T3.4-07
Type	FUN

Short Name	Conformance to mandatory fields
Functionality ID	U-REQ-T3.4-02
Description & quantification	The Data Cleaner subcomponent should safeguard that all the mandatory fields of a dataset are filled in.
Additional information	Python libraries of Pandas, NumPy, Scikit-learn, Keras will be utilized under this scope.
Priority	OPT
Reference Pilot	All
Success Criteria	Identification of missing mandatory fields.

Table 125: T3.4 – 8th Technical Requirement

Section	**Description**
ID	T-REQ-T3.4-08
Type	FUN
Short Name	Identification of duplications
Functionality ID	U-REQ-T3.4-02
Description & quantification	The Data Cleaner subcomponent should safeguard the identification of duplications that could then be removed facilitating easier and more efficient record management and maintenance.
Additional information	Python libraries of Pandas, NumPy, Scikit-learn, Keras will be utilized under this scope.
Priority	MAN
Reference Pilot	All
Success Criteria	Identification of duplicate records and values.

Table 126: T3.4 – 9th Technical Requirement

Section	**Description**
ID	T-REQ-T3.4-09
Type	FUN

Short Name	Automatic field completion
Functionality ID	U-REQ-T3.4-02
Description & quantification	The Data Cleaner subcomponent should safeguard that the data set provided is fully complete and should empower the automatic filling in of information based on interpolation / extrapolation techniques.
Additional information	Python libraries of Pandas, NumPy, Scikit-learn, Keras will be utilized under this scope.
Priority	MAN
Reference Pilot	All
Success Criteria	Automatic completion of on-purpose excluded values.

Table 127: T3.4 – 10th Technical Requirement

Section	Description
ID	T-REQ-T3.4-10
Type	FUN
Short Name	Automatic error correction
Functionality ID	U-REQ-T3.4-02
Description & quantification	The Data Cleaner subcomponent should safeguard that inconsistencies and errors identified are corrected.
Additional information	Python libraries of Pandas, NumPy, Scikit-learn, Keras will be utilized under this scope.
Priority	MAN
Reference Pilot	All
Success Criteria	Automatic correction of erroneous and incorrect values.

Table 128: T3.4 – 11th Technical Requirement

Section	Description
ID	T-REQ-T3.4-11
Type	FUN

Short Name	Data verification
Functionality ID	U-REQ-T3.4-02
Description & quantification	The Data Cleaner subcomponent should safeguard that data provided is accurate, especially referring to erroneous inliers, i.e., data points generated by error but falling within the expected range (erroneous inliers often escape detection).
Additional information	Python libraries of Cerberus will be utilized under this scope.
Priority	MAN
Reference Pilot	All
Success Criteria	Automatic verification of data for accuracy and inconsistencies.

Table 129: T3.4 – 12th Technical Requirement

Section	Description
ID	T-REQ-T3.4-12
Type	FUN
Short Name	Ontology Mapping
Functionality ID	U-REQ-T3.4-04
Description & quantification	The Data Harmonizer should define the appropriate techniques and tools to map concepts, classes, and semantics defined in different ontologies and datasets and to achieve transformation compatibility through extracted metadata.
Additional information	N/A
Priority	MAN
Reference Pilot	All
Success Criteria	Successful annotation, transformation and mapping of data and corresponding ontologies in terms of semantic and syntactic interoperability of data.

Table 130: T3.4 – 13th Technical Requirement

Section	Description
ID	T-REQ-T3.4-13

Type	FUN
Short Name	Data Schemas & HL7 FHIR Data Models
Functionality ID	U-REQ-T3.4-01, U-REQ-T3.4-03
Description & quantification	Define the exact data schemas and HL7 FHIR models that will be used for the transformation of incoming data and. Incoming and cleaned data will be modelled and transformed according to the defined schemas and models.
Additional information	N/A
Priority	MAN
Reference Pilot	All
Success Criteria	Transformation of data to the wanted HL7 FHIR data models and appropriate data schemas.

Table 131: T3.4 – 14th Technical Requirement

Section	**Description**
ID	T-REQ-T3.4-14
Type	FUN
Short Name	Named Entity Recognition (NER)
Functionality ID	U-REQ-T3.4-01, U-REQ-T3.4-02, U-REQ-T3.4-03, U-REQ-T3.4-04
Description & quantification	The Data Harmonizer subcomponent should facilitate the semantic and syntactic interoperability of data through the correct identification of proper entities from raw texts.
Additional information	N/A
Priority	MAN
Reference Pilot	All
Success Criteria	Identification of proper entities.

Table 132: T3.4 – 15th Technical Requirement

Section	**Description**
ID	T-REQ-T3.4-15

Type	FUN
Short Name	Translation of data
Functionality ID	U-REQ-T3.4-01
Description & quantification	The Data Harmonizer subcomponent should safeguard the language independence of incoming data and facilitate the processing and analysis of data in different languages.
Additional information	N/A
Priority	MAN
Reference Pilot	All
Success Criteria	Proper translation of incoming data in English language.

Table 133: T3.4 – 16th Technical Requirement

Section	Description
ID	T-REQ-T3.4-16
Type	FUN
Short Name	Deep Learning and Neural Networks procedures
Functionality ID	U-REQ-T3.4-01, U-REQ-T3.4-02, U-REQ-T3.4-03
Description & quantification	The utilization of Deep Learning techniques and especially of Neural Networks is mandatory, therefore it should be supported by the provided infrastructure.
Additional information	N/A
Priority	MAN
Reference Pilot	All
Success Criteria	Unhindered utilization of Deep Learning and Neural Network models.

Table 134: T3.4 – 17th Technical Requirement

Section	Description
ID	T-REQ-T3.4-17
Type	ENV

Short Name	Triplestore Database
Functionality ID	U-REQ-T3.4-04
Description & quantification	Triplestore is needed in order to store correlated, annotated and interoperable data as linked ontologies. Hence, it will be feasible the storage of semantic facts and the support of the corresponding data schema models.
Additional information	N/A
Priority	MAN
Reference Pilot	All
Success Criteria	Successful storage of semantic interoperable data as linked ontologies.

Table 135: T3.4 – 18th Technical Requirement

Section	**Description**
ID	T-REQ-T3.4-18
Type	ENV
Short Name	Communication with iHelp's Database
Functionality ID	U-REQ-T3.4-01, U-REQ-T3.4-02
Description & quantification	The correct communication and integration with iHelp's Database should be safeguarded.
Additional information	N/A
Priority	MAN
Reference Pilot	All
Success Criteria	Successful storage of cleaned, transformed, and harmonized data.

4.5 Personalised Health Modelling and Predictions

4.5.1 Goals and Objectives

Athens Technology Center (ATC) in Task 4.5 ‘Personalized health modeling and predictions’ aims to will deliver a data modelling tool that will assist in the development of personalized and disease specific health models using AI learning techniques. This tool will assist in the identification of specific diseases, along with their contributing factors and it will support the scalability and security needs of personalized health data. The data modelling techniques will enable the development of prediction algorithms for a certain risk, according to the analysis of relevant trends and patterns across groups of individuals.

4.5.2 State of the Art

Various ML methods like the random forest, decision tree, and logistic regression models had been developed to model disease risk predictions and had notably poor performance in comparison with Deep Neural Network and gradient boosting methods. The poor performance of the decision tree and logistic regression models could be attributed to the fact that these methods are not designed to model the very complex nonlinear relationships that exist in the data. DNN model describes the target as a nonlinear function of the input features and seems appealing to be evaluated in describing the complex underlying patterns in the health related and the habit/social related data.

Data from diseased people will be used to feed Deep Neural Networks that their aim will be to find rule associations between the data and the disease. Association Rule Learning techniques lag behind in terms of computational and time complexity, while deep learning techniques gain pace. We suggest using models such as Node2vec and by using methods displaying the most similar embeddings, finding correlations between the data. Hidden patterns will be discovered and knowledge about the disease will be extracted. Data from a new user will be fed to this model, in order to predict the risk of developing the disease and different symptoms.

4.5.3 Background technology

Table 136: T4.1 – Background Technology

Technology Name	Technology Description	Advancements / Usage
Node2Vec	Scalable feature learning.	Node2vec is proposed as an algorithmic framework for learning continuous feature representations on graphs.
Scikit-learn	Tools for predictive data analysis.	It contains implementations of machine learning models to be evaluated, i.e. ridge logistic regression, decision tree, random forest, gradient boosting, and deep neural network.
Gensim	Library for representing text as semantic vectors.	It can help modeling the textual features identified in the datasets acquired in the context of iHelp.

4.5.4 Module to User Requirements

Table 137: T4.1 – 1st Functionality

Section	Description
ID	U-REQ-T4.1-01
Title	Risk prediction
Functionality Description	Apply risk prediction to 700 participants
Source User Requirement	SCE-P1-01 / REQ-P1-01
Use case Quote	"To develop a comprehensive cancer risk prediction model"
Generic / Specific	Generic
Task / Component	T4.1: Personalized Health Modelling and Predictions
Lead Partner	ATC
Notes	-

Table 138: T4.1 – 2nd Functionality

Section	Description
ID	U-REQ-T4.1-02
Title	Health models from EHR
Functionality Description	Develop/enrich a health model based on EHR data to support risk level assessment. Data includes age, sex, diagnostic codes, laboratory test reports, medications, comorbidities, family history.
Source User Requirement	SCE-P4-01 / REQ-P4-01 SCE-P4-01 / REQ-P4-02 SCE-P4-01 / REQ-P4-03 SCE-P5-01 / REQ-P5-01
Use case Quote	▪ P4: "Access to and evaluation of large amount data from the clinical findings during the examinations of the patients as a prerequisite for Pancreatic cancer risk assessment" ▪ P4: "Access to and evaluation of large amount data from patients complains a prerequisite for Pancreatic cancer risk assessment"

	▪ P4: "Access to and evaluation of large amount data from consultations and laboratory and imaginary tests as a prerequisite for Pancreatic cancer risk assessment" ▪ P5: "Analysis of all patients' electronic health records (EHR) over the last 20 years in order to produce risk measures for individuals."
Generic / Specific	Generic
Task / Component	T4.1: Personalized Health Modelling and Predictions
Lead Partner	ATC
Notes	-

Table 139: T4.1 – 3rd Functionality

Section	Description
ID	U-REQ-T4.1-03
Title	Health models from secondary data
Functionality Description	Develop/enrich a health model based on secondary data to support risk level assessment.
Source User Requirement	SCE-P4-01 / REQ-P4-04
Use case Quote	"Access to and evaluation of large amount data from patients regarding prescribed regime outcomes"
Generic / Specific	Generic
Task / Component	T4.1: Personalized Health Modelling and Predictions
Lead Partner	ATC
Notes	-

4.5.5 Module to Technical Requirements

Table 140: T4.1 – 1st Technical Requirement

Section	Description
ID	T-REQ-T4.1-01
Type	FUNC

Short Name	Build health modes based on genomics and epidemiological data.
Functionality ID	U-REQ-T4.1-01
Description & quantification	Model the genomics and epidemiological data as vector representations to feed into neural networks for training.
Additional information	Feature extraction methods will be applied.
Priority	MAN
Reference Pilot	P1
Success Criteria	Health models based on genomics and epidemiological data.

Table 141: T4.1 – 2nd Technical Requirement

Section	**Description**
ID	T-REQ-T4.1-02
Type	FUNC
Short Name	Build health modes based on various types of data.
Functionality ID	U-REQ-T4.1-01
Description & quantification	Model the various types of data (data from patients complains and patients records, data from examinations, data from consultations and testing and secondary data) as vector representations to feed into neural networks for training.
Additional information	Feature extraction methods will be applied.
Priority	MAN
Reference Pilot	P4
Success Criteria	Health models based on various types of data.

Table 142: T4.1 – 3rd Technical Requirement

Section	**Description**
ID	T-REQ-T4.1-03
Type	FUNC
Short Name	Build health modes based on EHR data.

Functionality ID	U-REQ-T4.1-02, U-REQ-T4.1-03
Description & quantification	Model the EHR data (includes age, sex, diagnostic codes, laboratory test reports, medications, comorbidities, family history) as vector representations to feed into neural networks for training.
Additional information	Feature extraction methods will be applied.
Priority	MAN
Reference Pilot	P5
Success Criteria	Health models based on EHR data.

4.6 Model Library: Implementation and Recalibration of Adaptive Models

4.6.1 Goals and Objectives

The implementation and recalibration of adaptive models is realised through an analytic workbench in iHelp platform. The analytic workbench provides a complete information processing framework in iHelp that will incorporate declarative methods for the specification of targeted analysis tasks, as well as declarative analytics to include adaptive and predictive data services. The analytic workbench also facilitates openness and usability by allowing any actor (e.g. researcher, healthcare professional, data provider, etc) to develop on-demand adaptive learning models for different risks based on the application of advance AI analytic techniques. The analytic workbench is composed of several sub-components, as elaborated in the following figure:

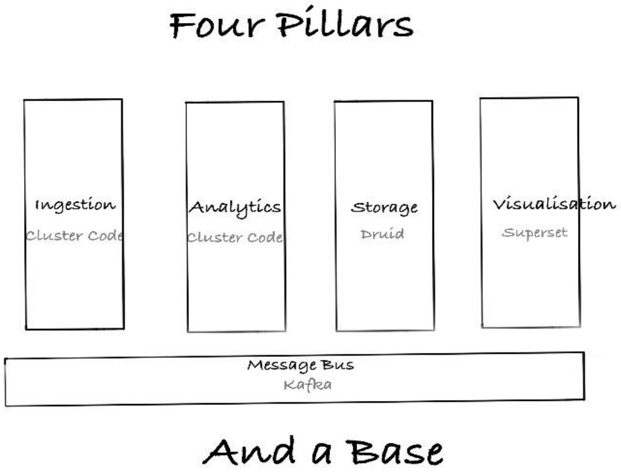


Figure 12: High-level architecture of the Analytic Workbench

4.6.2 State of the Art

The Analytic Workbench in the iHelp platform will use the following open-source technologies:

- Apache Druid
- Apache Superset
- Apache Kafka

The use of the open-source technologies/sub-components will enable the Analytic Workbench to easily connect multiple functionalities in an extensible approach. In essence these sub-components will be packaged in a series of containers making it easily deployable in Cloud.

4.6.3 Background technology

Table 143: T4.2 – Background Technology

Technology Name	Technology Description	Advancements / Usage
Apache Druid	Druid is an open-source, distributed data store	Druid streams data from message buses such as Kafka and support the

	designed to quickly ingest massive quantities of event data and provide low-latency queries on top of the data.	design of workflows that provide fast ad-hoc analytics, instant data visibility and concurrency to power UIs where an interactive, consistent user experience is desired.
Apache Superset	Apache Superset is an open-source software cloud-native application for data exploration and data visualization able to handle data at petabyte scale (big data)	Superset is used as a fast, lightweight and intuitive solution that make it easy for users of all skill sets to explore and visualize their data, from simple line charts to highly detailed geospatial charts.
Apache Kafka	Apache Kafka is a framework implementation of a software bus using stream-processing	Apache Kafka is used as a distributed event streaming platform that is used to support high-performance data pipelines, streaming analytics and data integrations

4.6.4 Module to User Requirements

Table 144: T4.2 – 1st Functionality

Section	Description
ID	U-REQ-T4.2-01
Title	Analytic Workbench
Functionality Description	The Analytic Workbench should provide the deployment and execution support to the AI analytic models in the iHelp platform
Source User Requirement	REQ-P4-02, REQ-P4-03, REQ-P4-04, REQ-P5-01 *(All above requirements relate to the same technical requirement)*
Use case Quote	*"Access to and evaluation of large amount data …"*
Generic / Specific	Generic
Task / Component	Analytic Workbench
Lead Partner	ICE
Notes	The Analytic Workbench should support the deployment and execution of analytic services or adaptive models in a way that allows them to perform their (analytic/evaluation) operations in an efficient way

4.6.5 Module to Technical Requirements

Table 145: T4.2 – 1st Technical Requirement

Section	Description
ID	T-REQ-T4.2-01
Type	FUN
Short Name	Deployment of Analytic Functions
Functionality Id	U-REQ-T4.2-01
Description & quantification	The Analytic Workbench should provide the necessary infrastructure to support the deployment of software services (e.g. AI learning and predictive models) that can address specific data analytic needs in the iHelp platform
Additional information	The Analytic Workbench should provide the relevant infrastructure that supports the deployment needs of different types of analytic services (or AI models) developed in the iHelp project.
Priority	MAN
Reference Pilot	All
Success Criteria	Successful deployment of iHelp analytic services

Table 146: T4.2 – 2nd Technical Requirement

Section	Description
ID	T-REQ-T4.2-02
Type	FUN
Short Name	Execution and Orchestration of Analytic Functions
Functionality Id	U-REQ-T4.2-01
Description & quantification	The Analytic Workbench should support the execution and orchestration of software services (e.g. AI learning and predictive models) that can address specific data analytic needs in the iHelp platform
Additional information	The Analytic Workbench should provide necessary infrastructure to support the execution needs of the analytic services deployed in the iHelp platform. Moreover, relevant API management techniques should be supported to fulfil the orchestration needs of different types of analytic services (or AI models) developed in the iHelp project.

Priority	MAN
Reference Pilot	All
Success Criteria	Successful execution and orchestration of iHelp analytic services

Table 147: T4.2 – 3rd Technical Requirement

Section	Description
ID	T-REQ-T4.2-03
Type	FUN
Short Name	Adaptation of Analytic Functions
Functionality Id	U-REQ-T4.2-01
Description & quantification	The Analytic Workbench should provide the necessary infrastructure to support the adaptation of AI learning and predictive models that can address specific data analytic needs in the iHelp platform
Additional information	The Analytic Workbench should provide the relevant support to the runtime adaptation needs of different types of analytic services (or AI models) developed in the iHelp project.
Priority	MAN
Reference Pilot	All
Success Criteria	Runtime adaptation of iHelp analytic services

4.7 Clinical DSS Suite with Visual Analytic Tools

4.7.1 Goals and Objectives

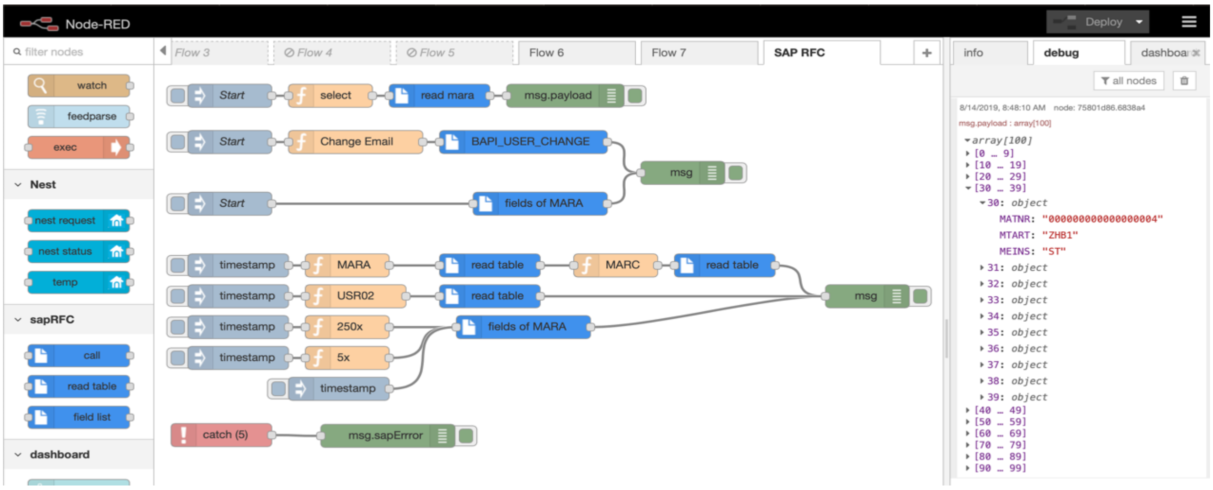
The Clinical Decision Support System (DSS) is a visualization tool which will allow Clinicians and policy makers to analyse data and take decisions. The DSS is going to apport query building, analytic and visualization capabilities by means of different interfaces. The query builder interface will allow to create queries workflow and execute federated queries. The dashboard interface will allow to show queries and analytic results in different chart types such as line, bar, pie, world maps or heatmaps and tables.

4.7.2 State of the Art

Workflow: Many tools for service, data and analytic workflows allow the composition of distributed artifacts in modular applications. An analysis of some of the popular workflow solution is provided below:

Node-Red

Node-Red[16] is a flow-based development tool, released under the Apache 2.0 license. Allows to wire together APIs, online services and hardware devices as part of the IoT. Node-Red has a web browser editor interface that allows to create flows of drag and dropping nodes. Nodes have different functionalities such as inject, function, MQTT requests, HTTP request, among others. There is a big community contributing with the development of new nodes with new functionalities. One example are the nodes that allows to connect to MySQL and PostgreSQL databases or execute Python scripts.



```
8/14/2019, 8:48:10 AM   node: 75801d86.6838a4
msg.payload : array[100]
▾array[100]
▸[0 … 9]
▸[10 … 19]
▸[20 … 29]
▾[30 … 39]
  ▾30: object
     MATNR: "000000000000000004"
     MTART: "ZHB1"
     MEINS: "ST"
  ▸31: object
  ▸32: object
  ▸33: object
  ▸34: object
  ▸35: object
  ▸36: object
  ▸37: object
  ▸38: object
  ▸39: object
▸[40 … 49]
▸[50 … 59]
▸[60 … 69]
▸[70 … 79]
▸[80 … 89]
▸[90 … 99]
```

Figure 13 – Node Red example

Rete

Rete[17] is a visual programming modular framework released under the MIT License. It allows to create node-base flows in the web browser interface. Node's functionality must be defined to allow users to create instructions for processing data without coding. In order to work with Rete it has to be integrated with

[16] Node-Red. (s.f.). *Securing Node-Red* . Recuperado el 08 de 03 de 2021, de https://nodered.org/docs/user-guide/runtime/securing-node-red

[17] Retejs. (s.f.). *Retejs*. Recuperado el 26 de 02 de 2021, de https://rete.js.org/#/

Angular. All nodes have to be programmed in order to give some functionality, there are not previously coded nodes as it can be found in Node-Red.

Figure 14 – Rete example

Apache Airflow

Apache Airflow[18] is a platform that allows create workflows as directed acyclic graphs (DAGs) of tasks, released under the Apache 2.0 license. It was joined to the Apache Software Foundation's incubation program in 2016. It allows to execute, schedule and distribute tasks across worker hosts. Apache Airflow is integrated with a lot of architectures such as Hive, AWS and Google cloud among others. Nodes can be configured to react to data sensors that can trigger a DAG when data arrives. It is possible to monitor system status and can be configured to send email alerts.

Apache Airflow architecture is composed by 4 different components: web server, scheduler, executor, metadata database. The initial deployment and the learning curve of this system are much more expensive compared to the systems mentioned above.

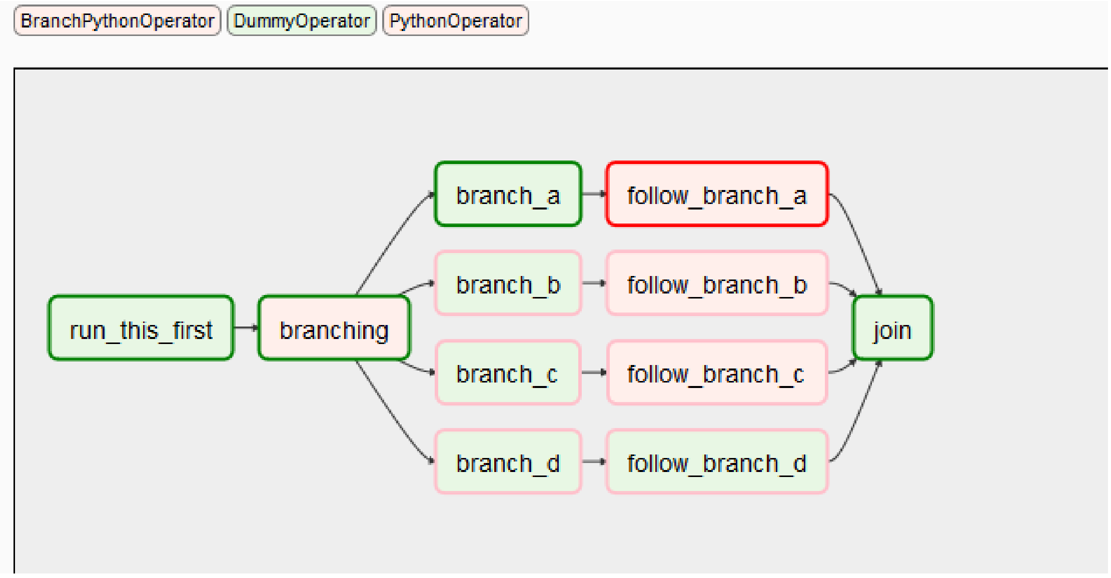


Figure 15 – Apache Airflow example

Visualisation: Similar to workflows, several tools for data and analytic visualisations exist. Most of the them are open source and therefore provide good options for integration in any analytic application.

Node-Red Dashboard

[18] Apache. (s.f.). *Apache Airflow*. Recuperado el 26 de 02 de 2021, de https://airflow.apache.org/

Node-Red Dashboard[19] is a module included in the Node-Red framework distributed under the Apache 2.0 license. This module allows to create live data dashboards by means of a set of nodes. Moreover, this module contains different widgets that can be installed and that provide new functionalities. Some of the widgets that can be included are the *node-red-contrib-web-worldmap*, *node-red-node-ui-table* and *node-red-contrib-ui-heatmap*, among others. Specifically, these widgets allow to prompt world map, heat maps and tables in the designed dashboards. If it is required new widgets can be created in order to fulfill dashboards' requirements.

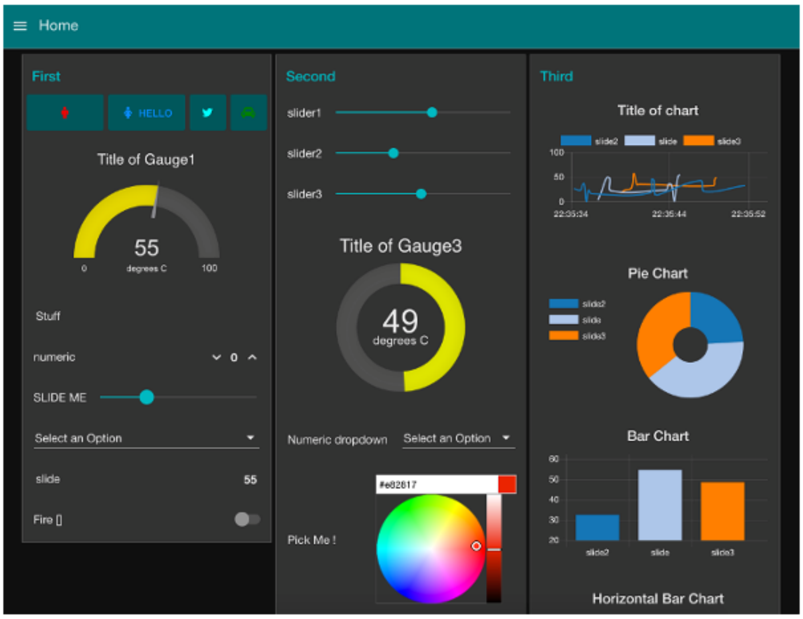


Figure 16 – Node Red Dashboard example

Tableau Public

Tableau Public[20] is one of the tools available at the Tableau ecosystem. Tableau is a tool for data visualization and business analytics that contains some add-ons for data management. Particularly, Tableau Public consists of a desktop tool available for windows and mac. This version allows the creation of dashboards that can be published at the Tableau website.

[19] Node-Red. (s.f.). *Node-Red Dashboard*. Recuperado el 26 de 02 de 2021, de https://flows.nodered.org/node/node-red-dashboard

[20] Tableau. (s.f.). *Tableau Public*. Recuperado el 18 de 02 de 2021, de https://public.tableau.com/en-us/s/

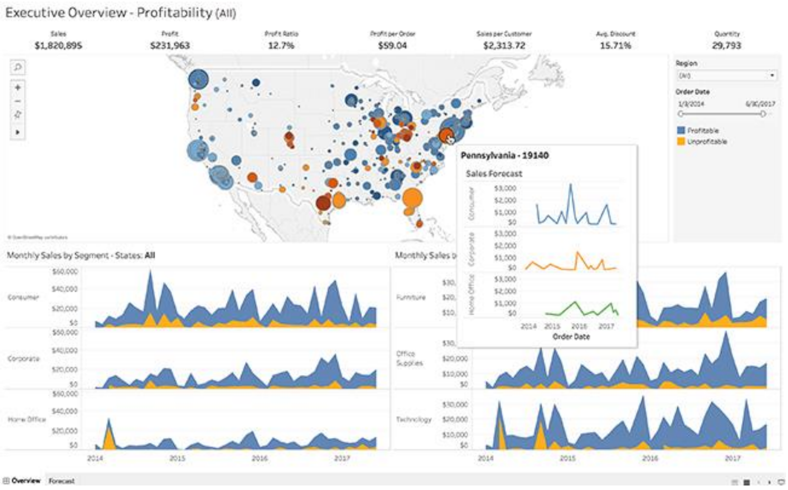


Figure 17 – Tableau example

Tableau Public is easy to use and has different types of visualizations that can be added to each dashboard. Between the different types of visualizations, we can find the heatmap. However, this tool free tool has some inconveniences:

- Does not allow JDBC database connection.
- Must pay in order to make your data private.
- In order to share the dashboards, the user must have a Tableau account.
- Once the dashboard has been published on the tableau website, it cannot be modified.
- Only email alerting is allowed.
- Not so many add-ons.
- One data source for dashboard.

Grafana

Grafana[21] is a popular tool for dashboards creation, available under the Apache 2.0 license. Grafana can be deployed in all operative systems (Windows, mac and Linux) and can be dockerized. Main component is the dashboard where the user can add as many panels as desired. Data can be obtained from several data sources such as Prometheus, Graphite, InfluxDB and PostgreSQL among more than 30 different, all of them native integrated. Moreover, it is possible to create data source plugins to connect to other databases.

Grafana's dashboards are dynamic, users can modify the previously created dashboard as many times as desired. Data can be show in panels using different visualization such as heatmaps, histograms and geomaps. Grafana contains a lot of different plugins that allow to use different data visualizations. In each panel is possible to set different alarms and send the notifications to other systems like email and slack.

[21] Grafana. (s.f.). *Grafana*. Recuperado el 18 de 02 de 2021, de https://grafana.com/

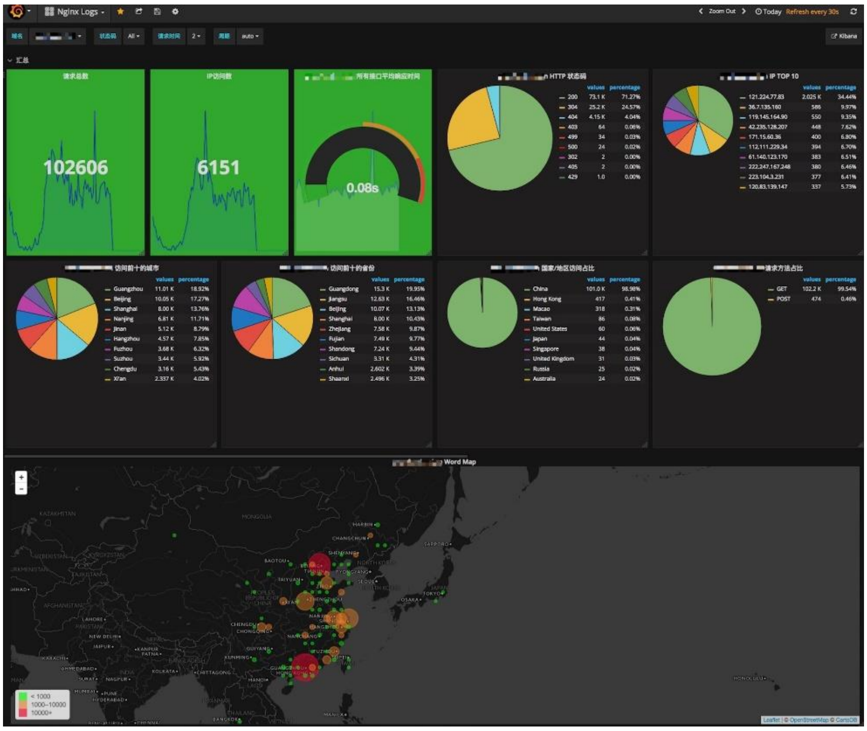
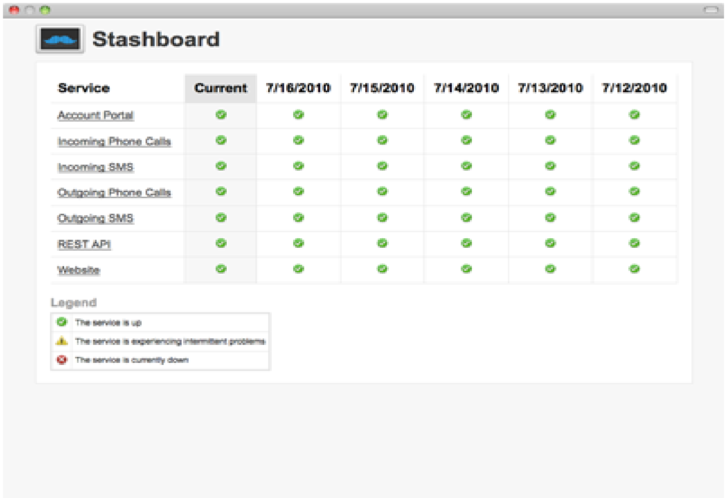


Figure 18 – Grafana example

Stashboard

Stashboard[22] is an open-source tool, distributed under the MIT license, that allows to create dashboards to show the status information about cloud services and APIs. Data is obtained and the dashboards are configured by means of a REST API. It is easy to use, however it does not contain so many visualizations formats.



Stashboard

Service	Current	7/16/2010	7/15/2010	7/14/2010	7/13/2010	7/12/2010
Account Portal	✓	✓	✓	✓	✓	✓
Incoming Phone Calls	✓	✓	✓	✓	✓	✓
Incoming SMS	✓	✓	✓	✓	✓	✓
Outgoing Phone Calls	✓	✓	✓	✓	✓	✓
Outgoing SMS	✓	✓	✓	✓	✓	✓
REST API	✓	✓	✓	✓	✓	✓
Website	✓	✓	✓	✓	✓	✓

Legend

✓	The service is up
⚠	The service is experiencing intermittent problems
✖	The service is currently down

Figure 19 – Stashboard example

Freeboard

[22] Stashboard. (s.f.). *Stashboard*. Recuperado el 18 de 02 de 2021, de https://www.stashboard.org/

Freeboard[23] is a data and visualization opensource tool distributed under the MIT license. It contains several widgets that can be added to the interactive dashboards. It was originally designed to interact with IoT devices, and it is possible to add widgets and create data sources. Dashboards are public and is required to get the premium edition in order to have private dashboards and unlimited widgets.

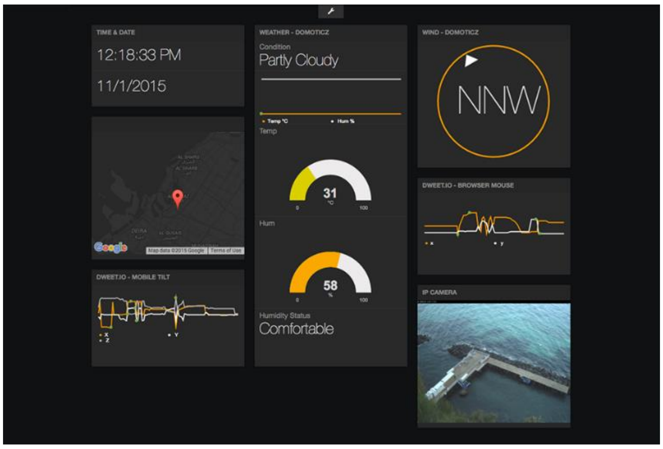
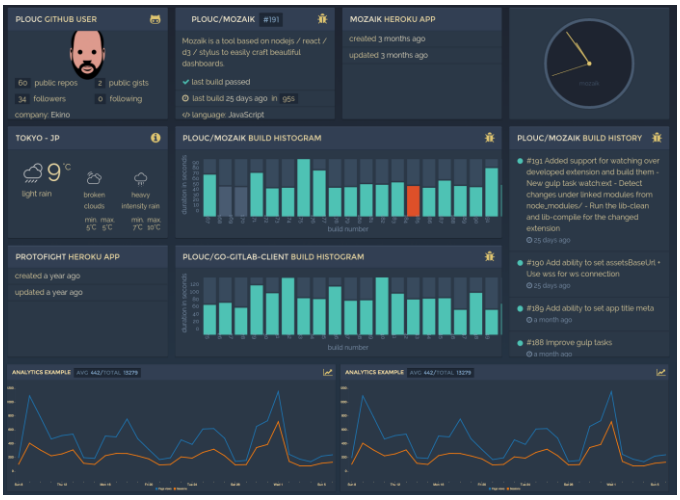


Figure 20 – Freeboard example

Mozaïk

Mozaïk[24] is a tool that allows to easily create dashboard. It is opensource, distributed under the MIT license. It has several widgets to be used and new ones can be created coding. Nevertheless, the creation of new widgets is a challenging task and there is a lack of data manipulation.



Mozaik is a tool based on nodejs / react / d3 / stylus to easily craft beautiful dashboards.

- #191 Added support for watching over developed extension and build them - New gulp task watch:ext - Detect changes under linked modules from node_modules/ - Run the lib-clean and lib-compile for the changed extension
  25 days ago
- #190 Add ability to set assetsBaseUrl + Use wss for ws connection
  25 days ago
- #189 Add ability to set app title meta
  a month ago
- #188 Improve gulp tasks
  a month ago

Figure 21 – Mozaik example

Dashbuilder

[23] Freeboard. (s.f.). *Freeboard*. Recuperado el 18 de 02 de 2021, de https://freeboard.io/

[24] Mozaïk. (s.f.). *Mozaïk*. Recuperado el 18 de 02 de 2021, de http://mozaik.rocks

Dashbuilder[25] is a web application that allows users to create business dashboards. It is an opensource software, released under the Apache 2.0 license. Data can be extracted from different data sources such as JDBC databases and regular text files. Dashboards are created with drag and drop and can be modified online. The main constraint is the lack of customization and personalization of the dashboards. Only contains three types of charts displayers: bar, pie and line.

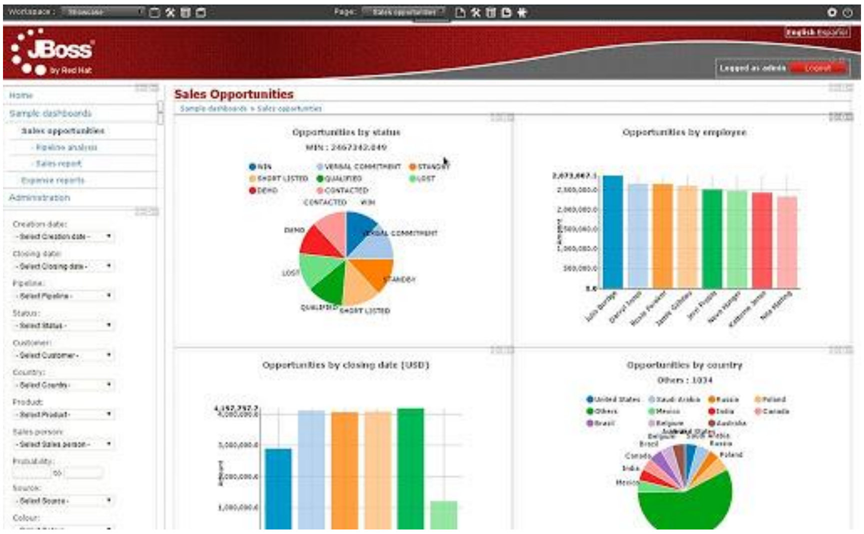


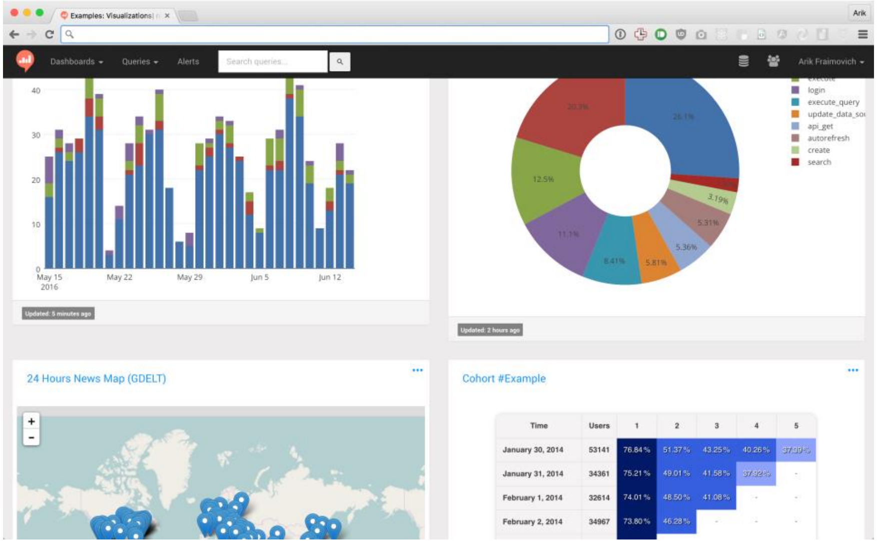
Figure 22 – Dashbuilder example

Redash

Redash[26] is a query builder tool that allows to visualize queried data. Data can be prompted throw forms such as chart, boxplot, map and word cloud, among others. Whoever, it is not an opensource solution and has 30 days free trial. It requires a considerable amount of time to properly learn how to use the dashboard. And allow to share the dashboards and to set alerts on certain events on the data.

[25] Dashbuilder. (s.f.). *Dashbuilder*. Recuperado el 18 de 02 de 2021, de http://dashbuilder.org/index.html

[26] Redash. (s.f.). *Redash*. Recuperado el 18 de 02 de 2021, de https://redash.io/



Cohort #Example

Time	Users	1	2	3	4	5
January 30, 2014	53141	76.84 %	51.37 %	43.25 %	40.26 %	37.39 %
January 31, 2014	34361	75.21 %	49.01 %	41.58 %	37.92 %	-
February 1, 2014	32614	74.01 %	48.50 %	41.08 %	-	-
February 2, 2014	34967	73.80 %	46.28 %	-	-	-

Figure 23 – Redash example

Visualization components

Other solution proposed for the data visualization part of the DSS, is the usage of different libraries that uses Angular as render. Angular[27] is a framework for the development of web applications. In order to create the visualization components there are two different libraries such as ngx-chart[28] for the different types of charts and amCharts[29] for the maps.

User Management: Similar workflow and analytics, user management can also be performed through several open source solutions that implement standardised protocols for security and privacy of user data as well as for access control, authentication and authorisation.

Node-Red security

Node-Red is not secured by default, but it can be configured to make it secure. Node-Red has a guide (Node-Red, s.f.) to show how to make Node-Red secure. In the guide shows how to add users and passwords and to configure different level permissions.

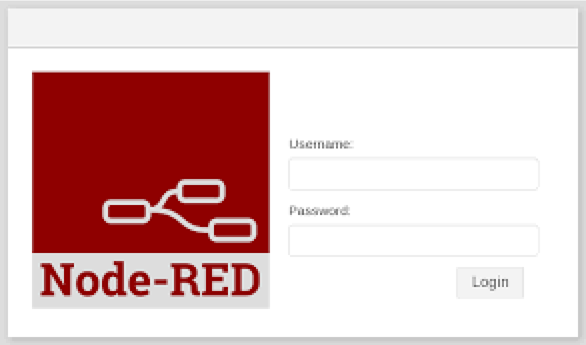


Figure 24 – Node Red for user management

[27] Google. (s.f.). *Angular*. Recuperado el 18 de 02 de 2021, de https://angular.io

[28] https://swimlane.github.io/ngx-charts

[29] https://www.amcharts.com/

Keycloak

Keycloak[30] is an opensource product released under the Apache 2.0 license that user's authentication. Keycloak can be integrated with other dashboard tools presented previously such us Node-Red and Angular.

Figure 25 – Keycloak for user management

4.7.3 Background technology

Table 148: T4.3 – Background Technology

Technology Name	Technology Description	Advancements / Usage
Visual Query Builder	The Visual Query Builder provides a set of operators that allow the user to build queries using drag and drop visual elements for relational databases.	In the scope of *iHelp*, the Visual Query Builder component allows the Physicians to compose SQL queries that will execute federated queries internally. By this way the Physicians will take the result of the data stored in one or more different data stores and locations.
Data Analytic Workflow	The Data Analytic Workflow provides the outcome of different isolated analytics presenting additional insights.	In the scope of *iHelp*, the Data Analytic Workflow component allows to indicate among a set of existing analytical models, the model or models to be executed and it will show the combined and individual results obtained by each of the models.
Physician's Dashboard	The Physician's Dashboard provides different visualization panels which will allow physicians to have an overview of a patient, or a population set to assist in decision making.	In the scope of *iHelp*, the Physician's Dashboard component consist of a set of preconfigured dynamic panels that will show the information of an individual patient or a set. The patient or set will be specified by the physician in the dashboard it shelf,

[30] Hat, R. (s.f.). *Keycloak*. Recuperado el 5 de 03 de 2021, de https://keycloak.org

		and all panels will show the information dynamically.

4.7.4 Module to User Requirements

Table 149: T4.3 – 1st Functionality

Section	Description
ID	U-REQ-T4.3-01
Source User Requirement	▪ REQ-P3-05 ▪ SCE-P4-01 ▪ REQ-P5-04
Use case Quote	▪ Pilot#3 – "… visualize all the necessary information of the data model…", "… different types of graphs." ▪ Pilot#4 – "…explore the results in the visual analytic tool…" ▪ Pilot#5 – "allow clinicians/medical-experts to analyse the collected data. The system will be presented in the form of a dashboard with visualizations to the clinicians as well as patients."
Generic / Specific	Generic
Task / Component	T4.3 Decision Support System (DSS)
Lead Partner	UPM
Notes	-

Table 150: T4.3 – 2nd Functionality

Section	Description
ID	U-REQ-T4.3-02
Source User Requirement	▪ REQ-P3-05 ▪ SCE-P5-01
Use case Quote	▪ Pilot#3 – "Artificial intelligence algorithms may be used to visualize the data with graphs." ▪ Pilot#5 – "…analysts can use machine learning techniques …"
Generic / Specific	Generic
Task / Component	T4.3 Decision Support System (DSS)
Lead Partner	UPM

Notes	-

Table 151: T4.3 – 3rd Functionality

Section	Description
ID	U-REQ-T4.3-03
Source User Requirement	▪ REQ-P3-05
Use case Quote	▪ Pilot#3 – "…information that will be stored in the database and that can also be accessed, applying filters…"
Generic / Specific	Generic
Task / Component	T4.3 Decision Support System (DSS)
Lead Partner	UPM
Notes	-

4.7.5 Module to Technical Requirements

Table 152: T4.3 – 1st Technical Requirement

Section	Description
ID	T-REQ-T4.3-01
Type	FUNC
Short Name	Analytics workflow
Functionality ID	U-REQ-T4.3-02
Description & quantification	The administrator must be able to create a DAG workflow that allows data to be read from one or more databases and perform analytics on this data.
Additional information	-
Priority	MAN
Reference Pilot	P1, P2, P3, P4, P5
Success Criteria	Successful execution of analytics

Table 153: T4.3 – 2nd Technical Requirement

Section	Description
ID	T-REQ-T4.3-02
Type	FUNC
Short Name	Federated queries builder workflow
Functionality ID	U-REQ-T4.3-03
Description & quantification	The administrator must be able to create a DAG workflow that allows to compose SQL statements that will be later executed against several databases.
Additional information	-
Priority	MAN
Reference Pilot	P1, P2, P3, P4, P5
Success Criteria	Successful execution of federated queries

Table 154: T4.3 – 3rd Technical Requirement

Section	Description
ID	T-REQ-T4.3-03
Type	FUNC
Short Name	Dashboard creation
Functionality ID	U-REQ-T4.3-01
Description & quantification	The administrator must be able to create different dashboards that will be used by physicians and policy makers will use to make decisions about patients.
Additional information	-
Priority	MAN
Reference Pilot	P1, P2, P3, P4, P5
Success Criteria	Physicians and policy makers must have all the relevant information requested in the different dashboards.

Table 155: T4.3 – 4th Technical Requirement

Section	Description
ID	T-REQ-T4.3-04
Type	FUNC
Short Name	Data visualizations
Functionality ID	U-REQ-T4.3-01
Description & quantification	Data obtained from the federated queries and the analytics must be prompted on different types of charts, such as bar, linear, radar and world maps.
Additional information	-
Priority	MAN
Reference Pilot	P1, P2, P3, P4, P5
Success Criteria	Successful data representation in the different chart types.

4.8 Big Data Platform and Knowledge Management System

4.8.1 Goals and Objectives

The main goal of the Big Data Platform and Knowledge Management System is to serve as the main data repository of the iHelp platform. Towards this, it will need to support data ingestion of external sources in very high rates, ensuring at the same time data consistency in terms of database transactions. Moreover, as it will be used for data retrieval by the analytical tools, it needs to offer a rich query processing mechanism in order for the data processing to be pushed down to the storage level, thus making the execution of the analytical algorithms more efficient. Finally, it has been identified a potential need for execution of queries over federated datastores that might exists externally to the deployment of the iHelp platform, thus a target objective from this component is to allow for polyglot query processing.

4.8.2 State of the Art

It has been a long discussion whether the use of traditional relational database management systems supporting SQL query processing is valid for Big Data applications, where huge amounts of data need to be ingested, while the query processing involves big data. Traditional relational databases ensure ACID properties and provide transactional semantics, whose main drawback is their lack of scalability of the transactions. Due to this, new data management technologies have been emerged, that can be categorized as NoSQL datastores and lack the support for ACID transactions (delegating any consistency check and complex transaction concept on the applications) trading this off for scalability. Their drawback however is that they also lack support for rich query processing mechanisms so even if they can support a highly rated data ingestion flow by downgrading the needs for data consistency (which is not the first priority in such scenarios), they are incapable of performing analytics. To this end, they has often been used by popular analytical frameworks to delegate this work to them. This introduces the need for those frameworks to retrieve a vast amount of data from the NoSQL datastores and perform the analysis in memory, which requires vast amount of computational and memory resources. In other words, they cannot use a database management system to push down these types of operations as closer to the storage as possible, which is mandatory for performing analytical processing efficiently.

To solve this problem, the most popular approach is the use of an additional data warehouse. By doing this, the NoSQL datastores can be used for primary data ingestion of the data, as their key-value nature and their lack of preserving transactions allows them to scale out horizontally as much as it is needed to serve the incoming workload. Therefore, they can be used as the primary storage of the raw data coming from various resources at any rate. However, as they are not capable of performing sophisticated query processing, the system integrators and architectures take the benefit of a data warehouse that can do this instead. In these cases, data is continuously being migrated from the NoSQL store to the data warehouse, by using expensive ETLs for data extraction, transformation and finally loading. To make things worse, the execution of the ETL procedures happen periodically in batches (usually during the offload of the system by nights) and the data warehouses are now capable of being used for analytical purposes. The drawback with these types of lambda architectures is one hand that they are difficult to be implemented and maintained, as they are rather complex and require additional effort to maintain two different database management systems at the same time, while on the other hand, data that are being kept in the data warehouse are always outdates and does not reflect to the current situation. In simpler words, the data analysts are performing their analysis over data that had been collected the previous day and cannot extract knowledge from the current view of the newly inserted data. This is a significant obstacle for real-time Business Intelligence (BI).

In order to deal with this inherit problem, hybrid solutions have been emerged during the recent years. The provide Hybrid Transactional and Analytical Processing (HTAP). They need to provide both a scalable transactional processing mechanism in order to allow the service of incoming data loads at very high rates, and analytical processing, in order to reduce the need for a data warehouse. The important thing to be mentioned is their ability to perform such types of analytics over the operational data.

The de facto atomic commit protocol that has been dominant in the operational datastores is the two-phase-commit (GRA., 78), (L., S., 79) Two-phase commit is heavy on time because it requires two rounds of messages between the coordinator and the participants of the transaction. In the last decade however, there have been quite a few solutions, which provide scalable transactional support without sacrificing consistency, such as Percolator (P., D., 2010) or Spanner (COR +, 13). However, most of them are either based on variations of two-phase commit, which is inherently limited because of its high overload; make use of expensive hardware to time events, like Spanner, Deuteronomy, and LEAP; or have a centralized transaction manager, like Omid (F., J., K., +, 14) and Apache Tephra[31].

As the need for real-time Business intelligence has been emerged the past few years, there is the case where organizations increasingly require analytics on fresh operational data to derive timely insights. This cannot be addressed by the use of complex lamba architectures using NoSQL solutions on one hand and periodically migrating data to a data warehouse on the other. To meet these requirements, database engines have been evolved to efficiently support hybrid transactional and analytical workloads (HTAP). However, this is challenging because OLTP workloads require ACID semantics, high throughput, and performance isolation, while OLAP workloads require interactive response times for very complex analytical queries that can span a big set of data. HTAP database engines are now a growing trend among which probably the most known database engine is SAP HANA[32].

Another important point for a big data platform aimed at a diverse potential public is polyglot support and the capability of accessing and providing different datastores that can fit the needs of different institutions. Accessing heterogeneous data sources has been addressed by multi-database systems (O., V, 11) and data integration systems (D., H., +, 2012). The typical solution is to provide a common data model and query language to seamlessly access data sources, hiding datastore details and enabling joins across datastores. The dominant state-of-the-art architectural model is the mediator/wrapper architecture in which each data source has an associated wrapper (or integration layer) that gives visibility to the source schema and data. The mediator centralizes the information provided in a unified view of all available data.

4.8.3 Background technology

The main background technology that will be used for the development of the BigData Platform and Knowledge Management System will be the LeanXcale ultra scalable datastore, with its ability for Hybrid Analytical and Transactional Processing, its dual SQL and NoSQL interface and its extensions for polyglot query processing.

[31] Apache, "Apache Tephra," [Online]. Available: http://tephra.incubator.apache.org. [Accessed May 2017]

[32] SAP, "Hana," [Online]. Available: https://www.sap.com/products/hana.html. [Accessed May 2017].

Table 156: T4.4 – Background Technology

Technology Name	Technology Description	Advancements / Usage
LeanXcale datastore	An ultra-scalable transactional relational datastore with support for Hybrid Analytical and Transactional Processing, providing a dual SQL/NoSQL interface and allows for polyglot query processing	In the scope of iHELP, the LeanXcale datastore will be further extended in order to provide parallel analytical query processing while its dual SQL/NoSQL interface will be validated against real life scenarios in order to advance its existing TRL. Moreover, its support for polyglot query processing will be used in scenarios where a query needs to be executed against federated datastores, being able to retrieve information from external instances of the iHelp platform.

4.8.4 Module to User Requirements

Table 157: T4.4 – 1st Functionality

Section	Description
ID	U-RED-T4.4-01
Title	Big Data Platform
Functionality Description	The Big Data Platform will provide all means to effectively store data being ingested by the corresponding data pipelines, even coming in high rates. It will also provide the query processing mechanism to allow to data analytics to retrieve data, using a rich query processing interface based on well-known standards, accessing live data that has been ingested.
Source User Requirement	This module is primarily related with the following user requirements: REQ-P1-01, REQ-P2-04, REQ-P4-01, REQ-P4-02, REQ-P4-03, REQ-P4-04, REQ-P5-01, REQ-P5-03 More, it must also address secondary requirements that can be further derived from the ones listed in section **Error! Reference source not found.** and are related with the data management aspects of the AI algorithms and data ingestion pipelines for primary and secondary data.
Use case Quote	“Access to and evaluation of large amount data …”
Generic / Specific	Generic
Task / Component	T4.4 Big Data Platform and Knowledge Management System
Lead Partner	LXS

Notes	The Big Data Platform will rely on LXS baseline technology that will be further extended in order to cope with the needs of the identified requirements.

4.8.5 Module to Technical Requirements

Table 158: T4.4 – 1st Technical Requirement

Section	Description
ID	T-REQ-T4.4-01
Type	FUNC
Short Name	HTAP Provision
Functionality ID	U-RED-T4.4-01
Description & quantification	Big Data Platform and Knowledge Management System must withstand operational workloads (OLTP) as the ingestion of patient HHRs or the streaming records coming from the secondary data ingestion process and - at the same time - allow analytical queries to do analytics over that data.
Additional information	
Priority	MAN
Reference Pilot	All
Success Criteria	Data is stored at the pace of input and the analytical tools can execute complex analytical queries over it.

Table 159: T4.4 – 2nd Technical Requirement

Section	Description
ID	T-REQ-T4.4-02
Type	PERF
Short Name	Scalable transactional mechanism
Functionality ID	U-RED-T4.4-01
Description & quantification	Big Data Platform must have ACID properties to guarantee the ability of transaction, while being able at the same time to scale out without sacrificing the performance in order to handle diverse high rated ingestion workloads

Additional information	
Priority	DES
Reference Pilot	All
Success Criteria	Data is stored consistently and transactionally while the performance can be kept steady when the incoming load is being increased

Table 160: T4.4 –3rd Technical Requirement

Section	Description
ID	T-REQ-T4.4-03
Type	DATA
Short Name	Local-Global Deployment
Functionality ID	U-RED-T4.4-01
Description & quantification	Because of data protection regulations among the Countries, the iHelp platform has to be able to deal with a deployment where some activities and data is stored at local level, while others can be done at a global level.
Additional information	
Priority	MAN
Reference Pilot	All
Success Criteria	Deployments fit UCs local data protection needs, while there can be a global view of the data easily integrated.

Table 161: T4.4 – 4th Technical Requirement

Section	Description
ID	T-REQ-T4.4-04
Type	FUNC
Short Name	Polyglot Query Processing
Functionality ID	U-RED-T4.4-01
Description & quantification	To have a wide range of options to cover all needs, the platform should have Polyglot integration of different data-stores. However, the focus should be on polyglot integration with other instances of the BigData

	Platform deployed in other pilots, so that the data analyst can benefit from the data federation among different deployments.
Additional information	
Priority	DES
Reference Pilot	All
Success Criteria	Queries can retrieve data stored in different deployments from a single-entry point.

4.9 Techniques for Early Risk Identification, Predictions and Assessment

4.9.1 Goals and Objectives

ATC seeks to utilise AI –based detection algorithms in the scopes of iHelp in order to deliver mechanisms that reveal the risks related to a specific disease. The data that will be used, will come from the integrated HHR platform and will be anonymized. Mechanisms will be developed, in order to detect the risks robustly and develop prediction models, using anonymized records and real-time data, not necessarily related to health, that correspond to each patient. The secondary data that will be used, include measured data from wearables, medical devices, questionnaires or reports etc. The predictive models that will be developed, will be applied to the collected and aggregated datasets, controlling over-fitting with various techniques that ensure that the learned models are able to generalize to unseen data.

4.9.2 State of the Art

The iHELP project aims to automate the process of disease detection and risk prediction for patients, as well as knowledge gain –related to the disease- and optimization in decision-making. A number of different and heterogeneous, historic or real-time data sources will be acknowledged (lifestyle and health records), in order to take into consideration all factors that could play a role in the development of a disease.

Frugal AI models are possibly needed since the information might be abundant or expensive to label (such as health records and images). Frugal AI has met a considerable development lately, with many state-of-the-art models, tackling all different types of problems, especially in the health sector. In particular, Active Learning intends to learn the characteristics of different classes by choosing which examples it will consider and, thus, reducing drastically the amount of data for training. Models such as Matching Nets might prove themselves valuable for our task.

Moreover, anomaly detection models must be trained, in order to distinguish healthy and diseased people. The diseased people will be further processed. Anomaly detection models have seen a great development lately. Generative Adversarial Nets (GANs), Autoencoders and Variational Autoencoders (VAE) are the state-of-the-art models for this task, that learn to recognize normal behavior, and then calculate anomaly scores for the data, in order to identity irregularities. Based on the characteristics of the data, the most suitable model will be chosen.

To simulate the evolution of a disease over time for a new patient, there are two path one can follow. First, if the evolution of a disease of other patients is available, traditional models such as CNNs, LSTMs and autoencoders, that look for similarities in the input data, perform well in that task. Otherwise, some novel and very promising models try to predict the evolution of a dynamic graph by using Graph Neural Networks. Such a model is the EvoNet for example.

4.9.3 Background technology

Table 162: T5.1 – Background Technology

Technology Name	Technology Description	Advancements / Usage
TensorFlow	Open-source machine learning platform.	It allows to build and train ML models easily using high-level APIs like Keras with eager execution. Predictive models implemented as neural networks will be used in medical diagnostics as well as in risk predictions.
Keras	Open-source Python library for developing and evaluating deep learning models.	Provides the interfaces to train neural networks and build predictive analytics. It offers a library of classifiers to be used in the risk prediction algorithms.

4.9.4 Module to User Requirements

Table 163: T5.1 – 1st Functionality

Section	Description
ID	U-REQ-T5.1-01
Title	Risk assessment and predictions
Functionality Description	Usage of risk prediction algorithms in order to perform risk assessment
Source User Requirement	SCE-P1-01 / REQ-P1-02
Use case Quote	"Perform risk assessment by applying risk predictions"
Generic / Specific	Generic
Task / Component	T5.1 Techniques for Early Risk Identification, Predictions and Assessment
Lead Partner	ATC
Notes	Based on extended model trained with large scale genomics and epidemiological data.

Table 164: T5.1 – 2nd Functionality

Section	Description
ID	U-REQ-T5.1-02
Title	Life habits and risks correlation
Functionality Description	Analysis of the correlation between life habits and the risk of suffering from the disease and / or its evolution if it is already diagnosed.
Source User Requirement	SCE-P3-01 / REQ-P3-01, REQ-P3-02
Use case Quote	“Correlate life habits and their effect as risk factors”
Generic / Specific	Generic
Task / Component	T5.1 Techniques for Early Risk Identification, Predictions and Assessment
Lead Partner	ATC
Notes	-

Table 165: T5.1 – 3rd Functionality

Section	Description
ID	U-REQ-T5.1-03
Title	Early risk detection
Functionality Description	Early detection of people who are at high risk for cancer. Data include complains, comorbidities, clinical findings, history, tests' results.
Source User Requirement	SCE-P4-01 / REQ-P4-01, SCE-P5-01 / REQ-P5-01
Use case Quote	P4: “Pancreatic cancer risk factors evaluation and cancer risk level assessment.”, P5: “Access to and evaluation of large amount data from various sources as a prerequisite for Liver and Pancreatic cancer risk assessment.”
Generic / Specific	Generic
Task / Component	T5.1 Techniques for Early Risk Identification, Predictions and Assessment
Lead Partner	ATC

Notes	There may be a differentiation between the two pilots based on the provided data and the type of cancer, e.g. P5 mentions also Liver Cancer risk assessment.

4.9.5 Module to Technical Requirements

Table 166: T5.1 – 1st Technical Requirement

Section	Description
ID	T-REQ-T5.1-01
Type	FUNC
Short Name	Build a risk prediction model
Functionality ID	U-REQ-T5.1-01
Description & quantification	Train a risk prediction model with genomics and epidemiological data and use it to assess a participant’s risk.
Additional information	-
Priority	MAN
Reference Pilot	P1
Success Criteria	The trained model is able to generalise and predict with high accuracy the risk of Pancreatic cancer among 3 clusters: low, medium and high risk.

Table 167: T5.1 – 2nd Technical Requirement

Section	Description
ID	T-REQ-T5.1-02
Type	FUNC
Short Name	Identify risk factors.
Functionality ID	U-REQ-T5.1-02
Description & quantification	Find patterns in correlations between life habits and their effect as risk factors.
Additional information	-
Priority	MAN

Reference Pilot	P3
Success Criteria	The trained model is able to generalise and identify patterns indicated as risk factors.

Table 168: T5.1 – 3rd Technical Requirement

Section	Description
ID	T-REQ-T5.1-03
Type	FUNC
Short Name	Predict the evolution of a diagnosed disease.
Functionality ID	U-REQ-T5.1-02
Description & quantification	Find patterns in correlations between recommended life habits and their possible effect in the evolution of a diagnosed disease.
Additional information	-
Priority	MAN
Reference Pilot	P3
Success Criteria	The trained model is able to generalise and predict with high accuracy the evolution of the disease.

Table 169: T5.1 – 4th Technical Requirement

Section	Description
ID	T-REQ-T5.1-04
Type	FUNC
Short Name	Pancreatic cancer risk assessment.
Functionality ID	U-REQ-T5.1-03
Description & quantification	Train an AI model with different types of data and use it to assess the cancer development risk level.
Additional information	Data include complains, comorbidities, clinical findings, history, tests' results.
Priority	MAN

Reference Pilot	P4
Success Criteria	The trained model is able to generalise and predict with high accuracy the risk of Pancreatic cancer among 3 clusters: low, medium and high risk.

Table 170: T5.1 – 5th Technical Requirement

Section	**Description**
ID	T-REQ-T5.1-05
Type	FUNC
Short Name	Risk prediction.
Functionality ID	U-REQ-T5.1-04
Description & quantification	Train an AI model with EHR data and use it to assess the cancer development risk level.
Additional information	Data includes age, sex, diagnostic codes, laboratory test reports, medications, comorbidities, family history.
Priority	MAN
Reference Pilot	P5
Success Criteria	The trained model is able to generalise and predict with high accuracy the Pancreatic and Liver cancer high risk individuals.

4.10 Design of Personalised Prevention and Intervention Measures

4.10.1 Goals and Objectives

The main goal of this activity is to develop conceptual models, which will enable us to plan for and deliver iHelp interventions in a personalised way. We will capture and evaluate a range of different personal factors, such as personality type, motivational factors, genetic predisposition, epigenomics, lifestyle, cancer outcome, and demographics. Through capture of this data, we will be able to profile iHelp users and deliver a specific intervention in a way that is tailored to the needs of different profiles (segments). Initial conceptual models will be later validated and translated into AI algorithms which will be deployed by iHelp in the range of processes, ranging from initial advice, to coaching.

4.10.2 State of the Art

Majority of health interventions show limited degree of personalisation of health advice. Our approach assumes personalisation based on a range of factors, including demographic, psychological traits, also in relation to genetics, epigenomics, and cancer. Such personalisation models, if executed by AI (e.g., personal characteristics could be elicited by iHelp chatbot), in a way that does not interfere with individual's everyday lives, could enable inducing and maintaining motivation for individuals to engage with recommended healthy behaviours. This use of AI technology for health engagement could help to advance AI in healthcare and enable (1) connection between a device and individual, (2) creation of feedback loop for an individual (enabled by the device), (3) greater uptake of healthy behaviours and healthy habit creation, and (4) help individuals decrease their risk of pancreatic cancer in the future.

4.10.3 Background technology

This section does not describe a technology component and therefore there is no relevant background technology to describe.

4.10.4 Module to User Requirements

Table 171: T5.2 – 1st Functionality

Section	Description
ID	U-REQ-T5.2-01
Title	Collection of user characteristics (profiling variables)
Functionality Description	Data collection
Source User Requirement	SCE-P1-02 REQ-P1-02

Use case Quote	(1) "Fill in a questionnaire to tell us something about you" or (2) "Hi, I am Rob, your personal assistant, I would love to find out more about you!"
Generic / Specific	Generic
Task / Component	User characteristics can be collected in two ways (1) by allowing the user to fill in a short questionnaire, (2) by allowing user to engage in a short conversation with a chatbot – where user characteristics are collected by a chatbot during a casual conversation
Lead Partner	UNIMAN
Notes	(1) or (2) point to two different ways this could be done. (1) using a traditional questionnaire, (2) using chatbot as a personal assistant

Table 172: T5.2 – 2nd Functionality

Section	Description
ID	U-REQ-T5.2-02
Title	Applying profiling formula/model/algorithm
Functionality Description	Data analysis resulting in profiling
Source User Requirement	SCE-P1-02 - Capture of profiling (Personalisation) characteristics REQ-P1-02 - Capture personality types
Use case Quote	"So, we have a promoter here, he is more interested in longevity and will probably not like to hear about cancer"
Generic / Specific	Generic
Task / Component	Data analysis
Lead Partner	UNIMAN can provide formulas/ algorithms, execution will rest with a technical partner enabling data collection and analysis
Notes	NA

Table 173: T5.2 – 3rd Functionality

Section	Description
ID	U-REQ-T5.2-03
Title	Communicating risk to people
Functionality Description	Matching the way risk is communicated to personal needs and characteristics, as indicated by one's profile.
Source User Requirement	SCE-P1-04 REQ-P1-04
Use case Quote	"Your body ages a bit faster than it should. To slow down that aging process, would you consider doing x, y, z" or "Do you know you can actually be healthier by making some small adjustments to your lifestyle by doing x, y, z" or "Do you know that you could reduce your future risk of being ill by doing x, y, z"…..
Generic / Specific	Specific
Task / Component	Communicating own disease risk in a way that it motivates behaviour change
Lead Partner	UNIMAN (design of the messages), execution will rest with a technical partner enabling communicating with users
Notes	NA

Table 174: T5.2 – 4th Functionality

Section	Description
ID	U-REQ-T5.2-04
Title	Provision of personalised feedback supporting habit formation/maintenance
Functionality Description	Providing feedback to users in a way it is consistent with their needs (as dictated by their profile characteristics).
Source User Requirement	SCE-P1-05 REQ-P1-05
Use case Quote	"You are doing great! By doing x,y, z you managed to slow down the aging process of your body!" or "Great job! You did x, y, z and your body is healthier now. It will thank you for that effort later😉" or "Well done you! You did x, y, z and this helped you to reduce your risks of future diseases!" …..

Generic / Specific	Specific
Task / Component	Communicating feedback in relation to behaviour change in a way that it motivates behaviour change and habit formation
Lead Partner	UNIMAN (design of the messages), execution will rest with a technical partner enabling communicating with users
Notes	This is a continuous effort. This activity will also interact with Delivery Mechanisms for Personalised Healthcare and Real-time Feedback (section 4.11, delivered by INS)

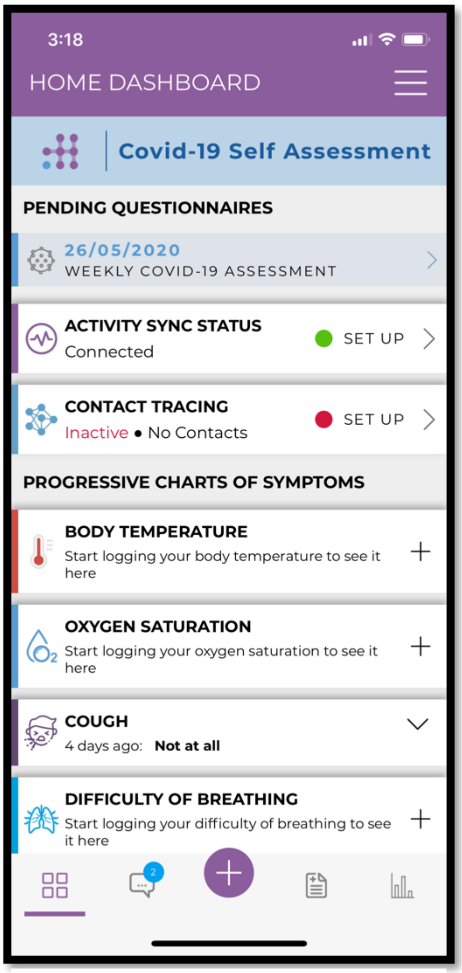
4.10.5 Module to Technical Requirements

This section does not describe a technology component and therefore there is no relevant technical requirements to describe.

4.11 Delivery Mechanisms for Personalised Healthcare and Real-time Feedback

4.11.1 Goals and Objectives

The delivery mechanisms for personalised healthcare and real-time feedback deal with the direct communication between platform and end-user. More specifically, in iHelp such delivery mechanisms are embedded in a mobile application that communicates information about collected data, as well as health- or wellbeing advice to end-users (who may be considered *patients*, or *citizens* depending on the pilot). The delivery mechanisms can be seen as a module, or a set of modules that will be designed and developed within the iHelp project, and integrated in the existing Healthentia mobile application (see Figure 26), that is brought into the project as background by Innovation Sprint (INS).The Healthentia mobile application is a production-ready app that runs on Android or iOS devices. The application is classified as a Class I Medical Device. End-users can register and be assigned to various "trial" configurations. Specific configurations determine the availability of various features (e.g., which questionnaires may be answered periodically, which types of information may be provided by the end-user, or whether or not the user is able to contact his/her assigned medical professional). This configurability will be used in the context of the iHelp project to allow different "versions" of the Healthentia application to serve the different pilots in the project. In a similar fashion, the *delivery mechanisms for personalised healthcare and real-time feedback* will consist of various features and functionalities that can be enabled or disabled depending on the specific app configuration. The Healthentia mobile application additionally serves as a tool for the *extraction of secondary data* – or the collection of lifestyle information in order to enrich iHelp's Holistic Health Records (HHRs). Additional information about the Healthentia mobile application, and specifically its measuring or data collection capabilities can be found in Section 3.



3:18

HOME DASHBOARD

Covid-19 Self Assessment

PENDING QUESTIONNAIRES

26/05/2020
WEEKLY COVID-19 ASSESSMENT

ACTIVITY SYNC STATUS
Connected ● SET UP

CONTACT TRACING
Inactive • No Contacts ● SET UP

PROGRESSIVE CHARTS OF SYMPTOMS

BODY TEMPERATURE
Start logging your body temperature to see it here

OXYGEN SATURATION
Start logging your oxygen saturation to see it here

COUGH
4 days ago: **Not at all**

DIFFICULTY OF BREATHING
Start logging your difficulty of breathing to see it here

Figure 26: Screenshot of the existing Healthentia mobile application.

The specific objective of the delivery mechanisms for personalised healthcare and real-time feedback is to provide the means necessary to deliver communicative acts to the end-user of the mobile application. These communicative acts can be e.g., measures, alerts, feedback, recommendations or other types of actions, the specifics of which will be designed within Task 5.2 ("Design of Personalised Prevention and Intervention Measures"), Task 5.5 ("Monitoring, Altering, Feedback and Evaluation Mechanisms"), and other work areas of the iHelp project.

4.11.2 State of the Art

Delivery of personalised healthcare and real-time feedback can be achieved in many ways. In order to limit the scope of this state-of-the-art analysis, we logically limit ourselves to the domain of smartphone applications – as those will be the medium used within the iHelp project. Yet, the scope still remains enormous, as delivering healthcare and feedback through a mobile phone application is a research field in and of itself: *mHealth* – a field that was already vast 8 years ago as shown by a historical overview article by Fiordelli et al., published in 2013 (F., D., S, 13).

Although a complete state-of-the-art analysis is thus out of scope, it may also not be useful for the purpose of this deliverable document. Instead, we provide a quick overview of common methods used for delivering personalised health information in the field of mHealth, specifically we look at data visualisation, feedback messages or nudges, and conversational coaching.

Data visualisation

The first and most obvious form of feedback when you are measuring health related data of users is to simply show that information to the user in the form of graphs, pie charts or other visualisations. Figure 27 below shows examples of visualisations of different data types by three popular health/wellbeing apps on the market today: Fitbit (left), Apple Health (middle), and WiThings (right).

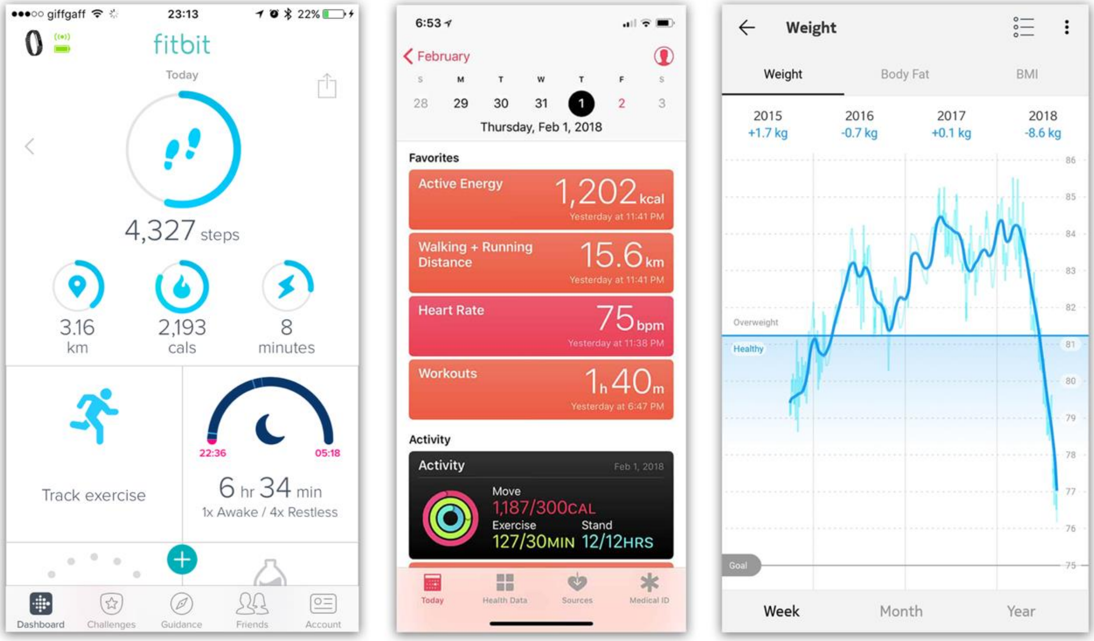


Figure 27: Examples of data visualisation in popular health and wellbeing apps, fltr: Fitbit, Apple Health, and WiThings.

As shown above, data visualisation can be simple like showing the number of calories burned in a block-element (centre), it can include information about progress such as the "steps circle" that fills up as the user approaches their daily step goal (left), or it can include derived information, such as a smoothed average plotted over "raw" values as shown in the weight graph on the right.

The state of the art is visually appealing and has certainly gone beyond simply plotting information in a graph; measured information is compared against certain goal values, and derived values are calculated to improve the usefulness of the plotted data. Yet, the potential of visualising how different types of measured data correlates to each other seems to be still relatively unexplored – a finding supported by the conclusions of research done on the creation of a taxonomy of physical activity data visualisation by Alrehiely et al. in 2018 (A., 18).

Feedback messages and nudges

Feedback messages or nudges are short, sometimes time-critical messages that are sent to the user to either inform them about their behaviour or to suggest some type of action to be taken on their behalf in order to improve their behaviour. Simply informing a user about their behaviour can often better be achieved through data visualisation, but otherwise might be achieved through a message like “You did 3420 steps so far today”. A message that aims to improve the user’s behaviour implies that it has been defined what “good behaviour” is – often achieved through a process called goal setting (L., L., 02). An example of such a message might be, “It looks like you only did 3420 steps so far today, why don’t you go for a walk?”. This type of feedback can take a simpler form, as exemplified by e.g., Fitbit devices that vibrate shortly if you haven’t done at least 250 steps in the current hour, and the hour is almost up. The simple “bzzt bzzt” that is felt on the user’s arm implies a message “You haven’t been active this hour, get up and move now!”. But feedback through natural language is much richer and has much greater potential in informing and educating users.

Natural language feedback messages are a well-studied phenomenon in and of itself, and we don’t aim to provide an exhaustive overview of this field here. In our earlier work (A., C., J., + 15) we have studied the literature on generating motivational messages in the area of physical activity and created a model on how such messages might be generated automatically, specifically focusing on ways to tailor such messages to individual users. Figure 28 below shows the outcome of this work – a model that shows how you might sequentially define a message *timing*, its *intention*, its *content*, and finally its *representation*, and how you can apply various tailoring techniques (A., G., H., 14) in each of those steps.

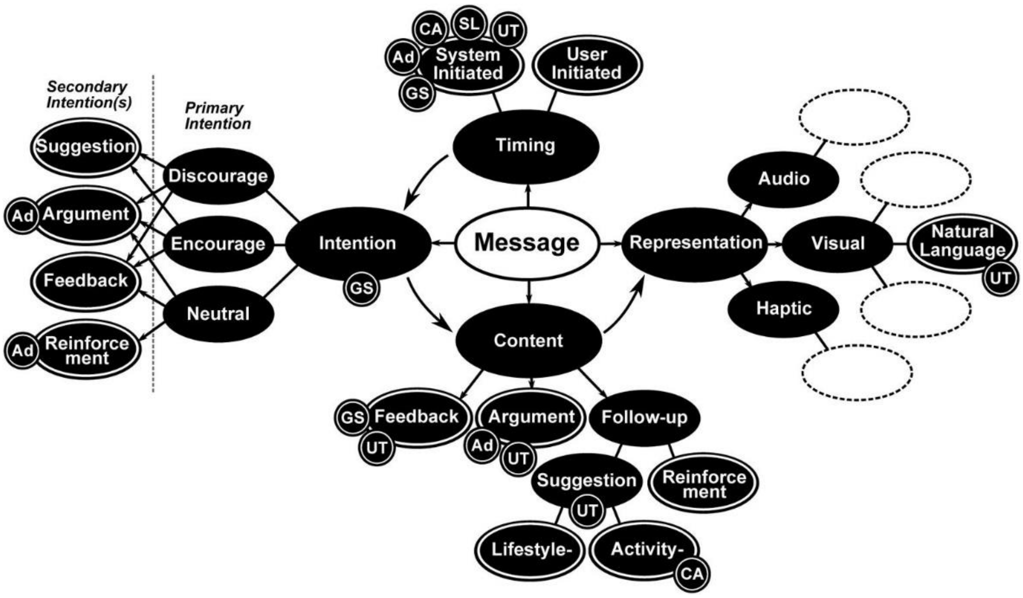


Figure 28: The model for tailoring motivational messages for real-time physical activity coaching from (op den Akker et al., 2015).

Conversational coaching

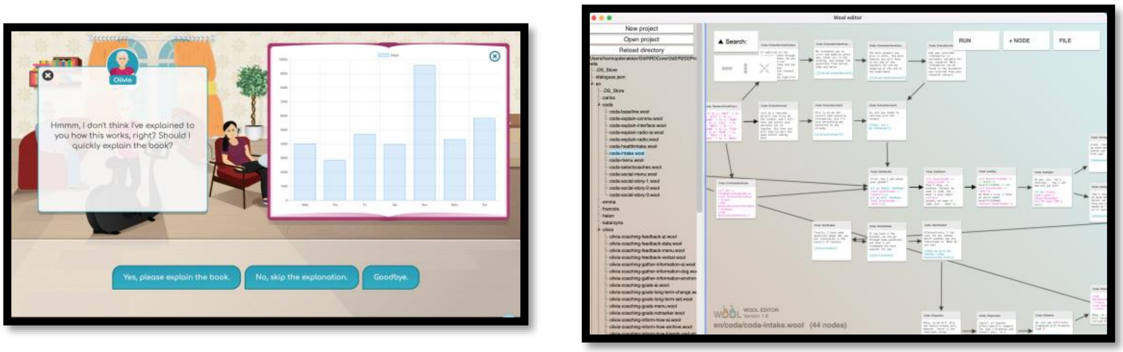
A natural enrichment of the personalized natural language feedback messages as described above, would be to give the user the ability to respond to the message – effectively starting a *conversation* with the user. Building on the example provided earlier, let's consider the following interaction:

- **System:** It looks like you only did 3420 steps so far today, why don't you go for a walk?
- **User:** I'm in a meeting.
- **System:** Okay, I see in your agenda that you have some time at 15:00, I will remind you then.
- **User:** Great, thanks!

By adding the ability for the user to reply to a feedback message, we have created a natural dialogue between a *virtual health coach* and the user. Supporting such conversations opens a host of possibilities in delivering feedback in a personalised, engaging and natural way.

Unfortunately, the state of the art in natural language processing, understanding and generation is not capable of providing fully automated voice interactions between a virtual coach and a user. As such, current applications in the serious domain of eHealth tend to focus on implementing *chatbots* or scripted virtual coaching dialogue.

Embodied virtual coaches with scripted health dialogue has been the focus of the recently finished EU H2020 project Council of Coaches (A., A., B., + 18). In Council of Coaches, a mature demonstrator was built in which users could engage in coaching dialogue with a set of 7 different virtual coaches, covering domains like physical activity, diet, cognition and more. An important outcome of the project was the open-source release of the underlying dialogue platform – the WOOL Platform (www.woolplatform.eu). The WOOL Platform defines a simple, but powerful scripting language for coaching dialogue, and includes an intuitive GUI Tool that allows domain experts to author coaching dialogue, as well as software libraries for executing those dialogues in web and mobile applications. Figure 29 below shows the Council of Coaches demonstrator (left), and the WOOL Editor that was used to author the dialogues (right).



Hmmm, I don't think I've explained to you how this works, right? Should I quickly explain the book?

Figure 29: The Council of Coaches demonstrator with scripted coaching dialogue (left), and the WOOL Editor that is used to author such coaching dialogue (right).

4.11.3 Background technology

Table 175: T5.3 – Background Technology

Technology Name	Technology Description	Advancements / Usage
Healthentia Edge Component (www.healthentia.com)	The Healthentia Edge Component is a platform consisting of mobile app, back-end and professional web portal that is used for capturing Real World Data and patient reported outcomes (see also §4.11.1).	The Healthentia platform will receive major upgrades, focusing on the delivery of personalised coaching to its end-users of the mobile (and possible web portal). Various back-end functionalities need to be added to support the delivery of coaching actions to the mobile app, while a major UI overhaul of the mobile app is foreseen to include the mentioned modalities of feedback: data visualisations, messages, and dialogue.
WOOL Platform (www.woolplatform.eu)	An open source (MIT License) platform for authoring, and executing dialogue, specifically designed, and developed to support eHealth use cases.	We expect to build on the WOOL Platform to better support the iHelp use cases in several ways. First, web-based (micro-)services need to be developed to deliver coaching dialogues. Second, the WOOL Editor is likely to require improvements in usability to allow iHelp domain experts to author serious coaching content. Third, we may require feature updates to the WOOL language as needed.

4.11.4 Module to User Requirements

The tables below (Table 174 through Table 178) each describe a specific functionality that needs to be supported by this component – *the delivery mechanism for personalised healthcare and real-time feedback*. For each of these functionalities, we refer to the Source User Requirement and provide a Scenario Quote to link these functionalities to the Scenarios and Requirements as defined in Section 2.

Table 176: T5.3 – 1st Functionality

Section	Description
ID	U-REQ-T5.3-01
Title	Feedback on relevant risk factors
Functionality Description	Showing the measured information that is related to identified risk factors (e.g. physical activity) to the users of the mobile app.
Source User Requirement	REQ-P1-04, REQ-P5-02

Scenario Quote	*"Participants will be using the platform to inform their risk mitigation activities"* (SCE-P1-04) *"The mobile application will be embedded with health recommender system that will enable health related awareness"* (REQ-P1-04)
Generic / Specific	Generic
Task / Component	T5.3: Delivery Mechanisms for Personalised Healthcare and Real-time Feedback / Mobile App
Lead Partner	Innovation Sprint
Notes	-

Table 177: T5.3 – 2nd Functionality

Section	Description
ID	U-REQ-T5.3-02
Title	Goal setting for relevant risk factors
Functionality Description	Showing target values for the measured information that is related to identified risk factors (e.g. physical activity) and allowing users a level of control over those target values based on the configuration of the specific pilot.
Source User Requirement	REQ-P1-04
Scenario Quote	*"Participants will be using the platform to inform their risk mitigation activities **and set targets**"* (SCE-P1-04)
Generic / Specific	Generic
Task / Component	T5.3: Delivery Mechanisms for Personalised Healthcare and Real-time Feedback / Mobile App
Lead Partner	Innovation Sprint
Notes	-

Table 178: T5.3 – 3rd Functionality

Section	Description
ID	U-REQ-T5.3-03
Title	Communicating with mobile app users
Functionality Description	Receiving communication from healthcare professionals to be displayed in the mobile app to the user.
Source User Requirement	REQ-P2-03, REQ-P3-04
Scenario Quote	*"Possibility to communicate to the patient…"* (REQ-P2-03)

Generic / Specific	Generic
Task / Component	T5.3: Delivery Mechanisms for Personalised Healthcare and Real-time Feedback / Mobile App
Lead Partner	Innovation Sprint
Notes	-

Table 179: T5.3 – 4th Functionality

Section	Description
ID	U-REQ-T5.3-04
Title	Providing messages and conversational coaching
Functionality Description	Receiving various types of coaching messages from the personalised advisor module. Simple messages cover short to the point health related awareness information, healthcare advice and behavioural nudges, while more in-depth conversational coaching cover the need for education, psychological support and more in-depth personalized coaching.
Source User Requirement	REQ-P5-02
Scenario Quote	*"The mobile application will be embedded with health recommender system that will enable health related awareness, healthcare advice, behavioural nudges, education, phycological support personalized motivational messages.* *Assessment of patient reported outcomes from the Mobile App for -- pain, happiness, depression, anxiety, mood, physical activity, sleep behavior and fatigue."* (REQ-P5-02)
Generic / Specific	Generic
Task / Component	T5.3: Delivery Mechanisms for Personalised Healthcare and Real-time Feedback / Mobile App
Lead Partner	Innovation Sprint
Notes	-

Table 180: T5.3 – 5th Functionality

Section	Description
ID	U-REQ-T5.3-05
Title	Tracking user interaction with messages and advice
Functionality Description	Any interaction that the mobile app user has with messages / notifications / reminders / conversations through the app, should be logged. This information must be made available to the back end in order for it to be displayed to healthcare professionals.

Source User Requirement	REQ-P2-03, REQ-P3-04
Scenario Quote	*"Possibility to communicate to the patient **and to see if the patient read the content of the message**"* (REQ-P2-03)
Generic / Specific	Generic
Task / Component	T5.3: Delivery Mechanisms for Personalised Healthcare and Real-time Feedback / Mobile App
Lead Partner	Innovation Sprint
Notes	-

4.11.5 Module to Technical Requirements

The following tables (Table 178 through 183), list the requirements for implementing the functionalities described in the previous paragraph (§4.11.4). Each requirement references the specific functionality that it relates to.

Table 181: T5.3 – 1st Technical Requirement

Section	**Description**
ID	T-REQ-T5.3-01
Type	FUNC
Short Name	Mobile App Data Visualisation
Functionality ID	U-REQ-T5.3-01
Description & quantification	The mobile application shall visualize the data for all measured information that is related to identified risk factors (e.g. physical activity) to the users of the mobile app.
Additional information	The exact data types that need to be visualized must be further specified.
Priority	MAN
Reference Pilot	All Pilots
Success Criteria	Being able to show a data visualisation for all the identified relevant data types.

Table 182: T5.3 – 2nd Technical Requirement

Section	**Description**
ID	T-REQ-T5.3-02
Type	FUNC
Short Name	Mobile App Goal Visualisation
Functionality ID	U-REQ-T5.3-02

Description & quantification	The user of the mobile app shall be able to see what the goal value is for any data related to relevant risk factors (to be determined, but e.g. physical activity) on a daily/weekly basis or the time frame as deemed relevant for the specified data type.
Additional information	For each data type identified as requiring data visualisation (see T-REQ-T5.3-01), it must be determined whether a goal value is needed, and what the time frame(s) for this goal should be (e.g. daily, weekly).
Priority	MAN
Reference Pilot	All Pilots
Success Criteria	Ability to see goal values for each relevant data type.

Table 183: T5.3 – 3rd Technical Requirement

Section	**Description**
ID	T-REQ-T5.3-03
Type	FUNC
Short Name	Mobile App Goal Setting
Functionality ID	U-REQ-T5.3-02
Description & quantification	The user of the mobile app shall be able to change any of the goals as supported in T-REQ-T5.3-02 under the conditions as defined by the health care professional (e.g. not allowed/allowed, and possible goal-value bounds).
Additional information	Further clarification on whether or not a goal should be changeable by the end-users for each specific target data type, and if so, under which conditions is needed.
Priority	DES
Reference Pilot	All Pilots
Success Criteria	Ability to change goal-values (if allowed) for each relevant data type.

Table 184: T5.3 – 4th Technical Requirement

Section	**Description**
ID	T-REQ-T5.3-04
Type	FUNC
Short Name	Mobile App Receiving Messages
Functionality ID	U-REQ-T5.3-03
Description & quantification	End-users of the mobile app shall be able to receive communication messages from healthcare professionals and shall be notified when a new message is available.

Additional information	-
Priority	MAN
Reference Pilot	All Pilots
Success Criteria	For every message sent to the user, the mobile app shall provide a notification and display the message.

Table 185: T5.3 – 5th Technical Requirement

Section	Description
ID	T-REQ-T5.3-05
Type	FUNC
Short Name	Mobile App Coaching
Functionality ID	U-REQ-T5.3-04
Description & quantification	End users of the mobile app shall be able to receive (a) coaching messages, and (b) conversational coaching triggers. Both message type shall trigger a notification in the mobile app. Coaching messages (a) shall consists of a single text field that the user can read and confirm reading. Conversational Coaching Triggers (b) shall trigger an interactive dialogue with the system's virtual coach, the content of which is defined by the specific conversation to be triggered.
Additional information	Coaching Conversations can be defined using the Open-Source WOOL Platform.
Priority	MAN
Reference Pilot	All Pilots
Success Criteria	Every coaching trigger received by the app shall display either a single text message (a) or trigger an interactive dialogue with the virtual coach (b).

Table 186: T5.3 – 6th Technical Requirement

Section	Description
ID	T-REQ-T5.3-06
Type	FUNC
Short Name	Mobile App Read Confirmation
Functionality ID	U-REQ-T5.3-05
Description & quantification	For every communication interacted with by the end-user of the mobile app (e.g., read/dismissed/entered the conversation), the mobile app shall send a "read confirmation" to the iHelp back-end to indicate that the communication has been received by the end-user.
Additional information	-

Priority	MAN
Reference Pilot	All Pilots
Success Criteria	Every communication received by the end-user that the end-user interacts with must lead to a notification ("read confirmation") to be received by the iHelp back-end.

4.12 Social Analytics for the Study of Societal Factors and Policy Making

4.12.1 Goals and Objectives

The Social Analytics component in the iHelp platform provides the ability to gather and analyse information from different social media platforms to get an insight into the social interaction concerning specific topics of interest. A key feature of this component is a Complex Event Processing (CEP) engine that allows the capture and analysis of streaming data from multiple social media platforms. The CEP engine incorporate state of the art Natural Language Processing, Sentiment Network Analysis and AI technologies to provide a complete toolkit for clinicians and policy makers; allowing them to collect, process, visualize and store online data related to Cancer. The CEP engine will be used to analyse the lifestyle trends, mental models, emotional intelligence, social interactions, and societal influences on the individuals who either have developed the Cancer or are in the early stage of risk identification. The engine will be extended to enable the inference of probabilistic events suitable for early risk identification and evaluation of targeted recommendations. The development of AI-based CEP functionality will support the real-time statistical analysis and processing of large social media datasets that complement the real-time data gathered from mobile/wearable devices; to find the dynamics, interactions, feedback loops, causal connections and trends that are of relevance to address the Cancer related challenges on the social and physio-social fronts.

4.12.2 State of the Art

In the EC funded Red-Alert project (https://redalertproject.eu/), ICE has developed a social media analysis solution that used Complex Event Processing (CEP) functionality to analyse the terrorism related content on the social media platforms (such as Twitter). The CEP functionality in the Red-Alert solution is used in conjunction with Natural Language Processing (NLP) and Social Network Analysis (SNA) techniques to generate alerts for the law-enforcement agencies (partners in the project).

The Red-Alert solution is able to process data that is directly extracted from social networks by means of social media APIs. Each social media platform has their own API to get data from it, usually by query searches (searches of content based on keywords). The Red-Alert solution is also able to process batch (social media) data uploaded by the users, which may come from one of the various social networks. Several techniques for data analysis were implemented, including NLP (extraction of concepts, sentiment, topics, etc.) and SNA (relations between users) to obtain interesting insight into the social media data e.g. community of users that discuss a specific topic, trends and patterns of specific topics, the most influential users with regards to a specific topic, etc. Based on the analysis, the solution is able to generate alerts related to the patterns of suspicious messages, message contents or the authors. The alerts are generated on a visual dashboard that is tuned to provide necessary information to the target uses.

It is important to note that some of the components (NLP and SNA) associated with the Red-Alert CEP functionality were developed by other Red-Alert partners under specific IPR conditions, which puts restrictions on their use after the project.

4.12.3 Background technology

Table 187: T5.4. – Background Technology

Technology Name	Technology Description	Advancements / Usage
MySQL	MySQL is an open-source relational database management system (RDBMS). It organizes data into one or more data tables in which data types may be related to each other	MySQL is used to store data about users of the social media analysis solution
MongoDB	MongoDB is a document-oriented NoSQL database used for high volume data storage	MongoDB is used to store the social media data
Apache Flink	Apache Flink is a framework and distributed processing engine for stateful computations over unbounded and bounded data streams. Flink can be used for stream and batch processing, sophisticated state management, event-time processing semantics, and exactly once consistency guarantees for state	Flink is used as the CEP engine in the social media analysis component in iHelp. The open-source nature of the Flink engine allows customisations and enhancements according to the needs of the CEP functionality in the iHelp project. In the iHelp project, the CEP functionality will be extended to enable the inference of probabilistic events suitable for early risk identification and evaluation of targeted recommendations. Moreover, NLP, SNA and AI techniques will be developed and integrated to deliver an integrated CEP engine capable of performing effective social media analysis
Apache Kafka	Apache Kafka is a system to manage the events' logs. It is a durable messaging system that send messages between processes, applications, and servers	Apache Kafka is used as a message brokering system in the social media analysis component. It is where the messages will be published for processing by different components in the social media analysis solution
Apache ZooKeeper	ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services	Apache ZooKeeper is used to coordinate and manage different components in the social media analysis solution
Java	Java is a class-based, object-oriented programming language. It is a general-purpose programming language intended to let application developers write and run code that can run on all platforms that support Java without the need for recompilation	Java is used as a server-side language for the development of the social media analysis solution

4.12.4 Module to User Requirements

There are no specific user requirements concerning the social analysis and support for policy making. However, social analysis is considered a fundamental aspect of the iHelp platform and therefore technical requirements for the social analysis solution are described in the next section.

Table 188: T5.4 - 1st Functionality

Section	Description
ID	U-REQ-T5.4-01
Title	Social analyzer for risk assessment
Functionality Description	Using machine learning based techniques from data collected from social media for risk assessment
Source User Requirement	REQ-P1-03
Scenario Quote	*"Using innovative risk analysis algorithms to evaluate pancreatic cancer risk in 700 participants"* (REQ-P1-03) *"To be able to communicate the assessed cancer risk in the most motivational way to the target participant"* (SCE-P1-05)
Generic / Specific	Specific
Task / Component	T5.4: Social analytics for the study of societal factor and policy making
Lead Partner	ICE
Notes	Although it fulfils no direct user requirement, the social media component fulfils the goal of REQ-P1-03 by providing novel combinations of user data and analysis

4.12.5 Module to Technical Requirements

Table 189: T5.4 – 1st Technical Requirement

Section	Description
ID	T-REQ-T5.4-01
Type	FUNC
Short Name	SMC - Social Media Collector
Functionality ID	U-REQ-T5.4-01
Description & quantification	SMC is a component that collects posts and updates from the APIs of the social media platform(s)

Additional information	Each social media platform exposes an API that allows the collection of real-time data (posts/messages). This component interfaces with those APIs to collect latest social media data
Priority	MAN
Reference Pilot	N/A
Success Criteria	Data from social media data is extracted through the provided APIs and is made available for use (processing, analysis etc)

Table 190: T5.4 – 2nd Technical Requirement

Section	**Description**
ID	T-REQ-T5.4-02
Type	FUNC
Short Name	SMA - Social Media Aggregator
Functionality ID	U-REQ-T5.4-01
Description & quantification	SMA is a component that allows to collate social media posts and updates from many different social media feeds.
Additional information	There may be a need to collect data from multiple social media platforms. This component enables the aggregation of data from multiple social media platform. The aggregated data gives a broader picture of the discussions/interactions taking place in the social media space
Priority	OPT
Reference Pilot	N/A
Success Criteria	Data from multiple social media platforms is aggregated and made available for use (processing, analysis etc)

Table 191: T5.4 – 3rd Technical Requirement

Section	**Description**
ID	T-REQ-T5.4-03
Type	FUNC
Short Name	NLP - Natural Language Processing
Functionality ID	U-REQ-T5.4-01

Description & quantification	Natural language processing (NLP) is a branch of artificial intelligence or AI, concerned with giving computers the ability to understand text and spoken words. The NLP component is responsible for the extraction and interpretation of concepts, sentiment, topics, etc. from the social media data
Additional information	NLP is an important technique to understand the data in natural language. The NLP component will enable the interpretation of social media messages/posts and extraction of meaningful information
Priority	MAN
Reference Pilot	U-REQ-T5.4-01
Success Criteria	Key concepts and topics are identified from the processing of social media data

Table 192: T5.4 – 4th Technical Requirement

Section	Description
ID	T-REQ-T5.4-04
Type	FUNC
Short Name	SNA - Sentiment Analysis
Functionality ID	U-REQ-T5.4-01
Description & quantification	The SNA component is responsible for the analysis of the social networks on the social media platforms and to extract relevant information about the structure of the networks, identify clusters or communities and predict new links between users
Additional information	SNA is essential for the analysis of networks and communities in social networks. The identification of communities can help with the analysis of which topics are popular among which type of communities
Priority	MAN
Reference Pilot	N/A
Success Criteria	Identification of user networks and communities around specific topics

Table 193: T5.4 – 5th Technical Requirement

Section	Description

ID	T-REQ-T5.4-05
Type	FUNC
Short Name	CEP - Complex Event Processing
Functionality ID	U-REQ-T5.4-01
Description & quantification	Complex event processing is an organizational tool that helps to aggregate a lot of information and identify cause-and-effect relationships among events in real time. The CEP component in the social media analysis solution will be responsible for converting social streams in readable event stream that can be processed in real-time to discover useful patterns and trends. Moreover, CEP functionality will be extended to include inference of probabilistic events suitable for early risk identification and evaluation of targeted recommendations
Additional information	The role of CEP in the social media analysis solution is to apply patterns in real time to disparate sources and formats of data (eg text, authors.) to help to identify specific messages, behaviour or content in social media.
Priority	MAN
Reference Pilot	N/A
Success Criteria	Identification of key events as well as trends and patterns of specific types of events through the analysis real-time social media feeds

Table 194: T5.4 – 6th Technical Requirement

Section	Description
ID	T-REQ-T5.4-06
Type	FUNC
Short Name	Dashboard – Alert Creation
Functionality ID	U-REQ-T5.4-01
Description & quantification	The social media analysis solution in Help should provide an administrative dashboard to the users, allowing users to create alerts by specifying which topics (keywords and phrases) they are interested in monitoring on specific social media platforms. These topics can relate to the lifestyle trends, mental models, emotional intelligence, social interactions, and societal influences on the individuals.
Additional information	This administrative interface of the social media analysis solution will allow the registration of users and creation of alerts.
Priority	MAN

Reference Pilot	N/A
Success Criteria	Creation of alerts on specific keywords/phrases

Table 195: T5.4 – 7th Technical Requirement

Section	**Description**
ID	T-REQ-T5.4-07
Type	FUNC
Short Name	Dashboard – Monitoring
Functionality ID	U-REQ-T5.4-01
Description & quantification	The Monitoring Dashboard of the social media analysis solution provides real-time monitoring updates on the pre-specified alerts through a set of intuitive visualisations. The alerts will show e.g. how many social media posts are being posted on specific topic(s), from whom, when and the pattern of posts.
Additional information	Monitoring will be the second UI of the Dashboard
Priority	MAN
Reference Pilot	N/A
Success Criteria	The real-time Monitoring of pre-specified alerts is presented to the users through a set of intuitive visualisations

Table 196: T5.4 – 8th Technical Requirement

Section	**Description**
ID	T-REQ-T5.4-08
Type	L&F
Short Name	Dashboard – Monitoring Look and Feel
Functionality ID	U-REQ-T5.4-01
Description & quantification	The Monitoring Dashboard should be designed in such a way that makes it usable by target users i.e. policy makers. This means the monitoring data is presented to the users in such a way that makes is easily interpretable, explainable, and usable in their decision making
Additional information	Specific usability and end-user development guidelines should be consulted in the design of Monitoring Dashboard

Priority	DES
Reference Pilot	N/A
Success Criteria	The target users provide 100% satisfaction on the way social media analysis outcomes are presented to them through the Monitoring Dashboard

4.13 Monitoring, Alerting, Feedback and Evaluation Mechanisms

4.13.1 Goals and Objectives

The main goal of the Monitoring and Alerting module is to track and/or monitor the effect of the recommendations that are prescribed to the individuals by the clinician/medical experts. The module (and its sub-modules) will run as a separate service and will observe, and check and aggregate data coming from primary (patient dossier and medical records) or secondary data sources (mobile and or wearable devices) and compare them with personalized targets issued by the clinician/medical experts.

4.13.2 State of the Art

Monitoring Alerting and feedback module will utilize the scalability, flexibility, modularity, separation of concerns, containerization, virtualization, and automated deployments offered by Function-as-a-Service (FaaS) paradigm, microservices architecture and serverless computing.

Those paradigms encourage loosely coupling of system modules and use service choreography that allow building effective complex system based on clear and well-defined elements.

Demonstrated effectiveness of such approach is one of the reasons for its increasing use in modern application demanding big data, high volume complex processing.

The monitoring mechanism will be expose a rule engine executed on certain configurable intervals (daily, weekly, monthly etc.) that will make a quick assessment of the goals' progress and will decide whether to activate the alerting submodule. The decisions will be done based on comparing certain monitored parameter target values defined in the personalized advice against real achieved values. Different experts defined escalation policies and threshold values will be considered when generating the proper alert and alert recipient (monitored individual or the advisor) in order to achieve optimal results.

Examples of such rules will be that monitored person have made recommended by his clinician 10000 steps per week, while maintaining blood pressure within 80- 140 range and limiting the calorie intake to max 3000 calories per day

A gamification and user behavioral profiling will be utilized to ensure continues involvement, creating and maintaining a desire for achieving desired results – introduction of healthier lifestyle and avoiding risk factor for cancer diseases.

4.13.3 Background technology

Table 197: T5.5 – Background Technology

Technology Name	Technology Description	Advancements / Usage
REDIS	Data structure store, used as a distributed, in-memory key–value database, cache, and message broker, with optional durability that supports different kinds of abstract data structures	Redis cache allow to store run time artifacts for a highly scalable distributed system that are needed in order to calculate fast and effectively various targets set for individual by personalized advice given to him

Spring Boot	Spring Boot makes it easy to create stand-alone, production-grade Spring based Applications that you can "just run". Most Spring Boot applications need minimal Spring configuration.	Spring Boot Applications Embed Tomcat directly (no need to deploy WAR files) Provide production-ready features such as metrics, health checks, and externalized configuration. Ideally fits to the separation of concerns paradigm as well as FaaS. Traditional and battle tested framework for microservices and docker images.
spring-kafka	The Spring for Apache Kafka (spring-kafka) applies core Spring concepts to the development of Kafka-based messaging solutions. It provides a "template" as a high-level abstraction for sending messages. In all these cases, you will see similarities to the JMS sup in the Spring.	Spring-kafka provides an easy-to-use interface to kafka streams, thus enhancing us for easier, robust and fast way to implement the communication of the monitoring and alerting modules to the kafka data bus, and the other related modules.

4.13.4 Module to User Requirements

Table 198: 1st Functionality

Section	Description
ID	U-REQ-T5.5-01
Title	Analytical data
Functionality Description	Collecting the measured information that is related to identified risk factors (e.g., physical activity) and provide it for analytical processing and advice fine tuning
Source User Requirement	REQ-P1-05
Scenario Quote	*"Usage of co-designed iHelp platform to provide feedback on targeted recommendations and level of activity"* (REQ-P1-05) *"Each mitigation package will be tailored to fit the suite of target sets for each participant based on their starting profile (pre-defined by pilot team)."* (SCE-P1-05)
Generic / Specific	Specific
Task / Component	T5.5: Monitoring, alerting, feedback, and evaluation mechanisms
Lead Partner	KODAR
Notes	-

Table 199: 2nd Functionality

Section	Description

ID	U-REQ-T5.5-02
Title	Monitoring effects of personalized advices
Functionality Description	Communicate the personalized advice to the end user and keep him informed about progress towards set goals via the mobile app
Source User Requirement	REQ-P3-10, REQ-P5-03
Scenario Quote	*"Once it is detected that the patient is at risk, the clinician should initiate a follow-up of the same, in order to advise the patient through monitoring"* (REQ-P3-10) The pilot will look for compliance rate and level of achievement compared to goals set for each participant. (P1) "Identification and communication of high risk to the individuals and advise towards the use of iHelp digital solutions to enhance behavioral change." (REQ-P5-03)
Generic / Specific	Specific
Task / Component	T5.5: Monitoring, alerting, feedback, and evaluation mechanisms
Lead Partner	KODAR
Notes	-

Table 200: 3rd Functionality

Section	**Description**
ID	U-REQ-T5.5-03
Title	Advice Follow up, feedback, alerting and periodic review
Functionality Description	Constantly monitor the effects of personalized advises given to the user and in case of major abnormalities involve the health care practitioner in the feedback loop.
Source User Requirement	REQ-P3-11, REQ-P5-06
Scenario Quote	*"When the patient is already being offered and performing the entire set of advice by the clinicians, and even so abnormalities are detected and reported to the clinician. The clinician must contact the patient either for a review or to know what is happening, in order to know what measures should be taken. "*(REQ-P3-11) to allow participants to monitor changes in their personalized risk before and after and their progress with their risk mitigation activities. (P3) "Health recommender system that will enable health related awareness, healthcare advice, behavioural nudges, education, phycological support personalized motivational messages."(REQ-P5-06)

Generic / Specific	Specific
Task / Component	T5.5: Monitoring, alerting, feedback, and evaluation mechanisms
Lead Partner	KODAR
Notes	-

4.13.5 Module to Technical Requirements

Table 201: T5.5 – 1st Technical Requirement

Section	Description
ID	T-REQ-T5.5-01
Type	FUNC
Short Name	Monitoring of personalised advice
Functionality ID	U-REQ-T5.5-02
Description & quantification	Provide monitoring mechanisms to track and/or monitor the impact of targeted recommendations (interventions-prevention measures) that are prescribed to individuals by the clinician/medical-experts.
Additional information	The availability of the monitoring data allows better and early identification of risks, recalibration of the prevention-intervention models and fine tuning of the targeted recommendations. The availability of the monitoring data allows better and early identification of risks, recalibration of the prevention-intervention models and fine tuning of the targeted recommendations.
Priority	MAN
Reference Pilot	All Pilots
Success Criteria	This mechanism allows better and early identification of risks, recalibration of the prevention-intervention models and fine tuning of the targeted recommendation

Table 202: T5.5 – 2nd Technical Requirement

Section	Description
ID	T-REQ-T5.5-02

Type	FUNC
Short Name	Alerting mechanism for recommendations
Functionality ID	U-REQ-T5.5-03
Description & quantification	Multiple health and behaviour related parameters must be assessed in real time and alerts raised when anomalies in expected behaviour or risks are identified
Additional information	The alerts include behavioural nudges, temporal feedback, therapeutic guidelines etc.
Priority	MAN
Reference Pilot	All Pilots
Success Criteria	Used as an escalation mechanism for delivering updates to clinician/medical experts in case of deviations or new risks identifications

Table 203: T5.5 – 3rd Technical Requirement

Section	**Description**
ID	T-REQ-T5.5-03
Type	FUNC
Short Name	Personalised treatment plan
Functionality ID	U-REQ-T5.5-02
Description & quantification	The alerting mechanism operate on top of the monitoring data to provide personalised alerts (specific to nature of risks associated with the individuals) that allow individuals to make on time improvements in the lifestyle, behaviours, and social interactions.
Additional information	
Priority	DES
Reference Pilot	All Pilots
Success Criteria	Ensure following a personalised health plan by observing change of parameters and their targets and motivating the users with nutrition tips, physical activity encouragement and suggestions for healthier lifestyle (reducing smoking and alcohol intake)

Table 204: T5.5 – 4th Technical Requirement

Section	**Description**

ID	T-REQ-T5.5-04
Type	FUNC
Short Name	Individual Lifestyle and health monitoring
Functionality ID	U-REQ-T5.5-02
Description & quantification	Monitoring alerting and feedback mechanism replace ‘one size fits all’ approach to the treatment and care of patients with a particular condition, to one which convert AI based models to uses individualised help plans broken down to achievable and closely monitored targets
Additional information	Currently, the decision making of healthcare professionals is generally based on population averages rather than individual characteristics. The individualised targets and feedback approach will provide genuine insight into the thinking and actions of the patients and enable us to understand the behaviour and the sentiments behind certain lifestyle choices.
Priority	DES
Reference Pilot	All Pilots
Success Criteria	Obtain personalised targets for various health parameters that AI models produce (and being validated or corrected by clinicians) so they are presented to users along with reference information to support increase in health literacy as well as better monitoring

Table 205: T5.5 – 5th Technical Requirement

Section	**Description**
ID	T-REQ-T5.5-05
Type	FUNC
Short Name	Identification of early symptoms in healthy individuals
Functionality ID	U-REQ-T5.5-03
Description & quantification	Allow usage of monitoring alerting and feedback mechanism not only for monitoring advice provided to individuals with confirmed cancer but also for early identification of warning signs.
Additional information	Early warning signs relay to either healthy person monitoring several body parameters e.g., to conduct healthy lifestyles and increase physical activity levels or to the detection of the deterioration of the condition of already diseased patients.

	Prediction models from aggregated patient data of certain health events/complications are used to set a combination of parameters and target levels monitored and alerted
Priority	DES
Reference Pilot	P1
Success Criteria	Detection of health and wellbeing issues at earlier and more treatable stages

Table 206: 6th Technical Requirement

Section	**Description**
ID	T-REQ-T5.5-06
Type	FUNC
Short Name	Provide information to analytical workbench for aggregated activities and alerts
Functionality ID	U-REQ-T5.5-01
Description & quantification	Store collected aggregated monitoring data and generated alerts for the particular user or user group
Additional information	Collected information from the monitoring to be visualised in graphical user interfaces and used by person itself and by clinicians in order to monitor trends and direction of changes This raw data should be stored in DB and exposed to T4.3 “DSS system” for historical data , trend monitoring , personal motivation , individual progress monitoring and therapy decision or analytical processing for individual or group of users
Priority	DES
Reference Pilot	All Pilots
Success Criteria	Monitored data aggregation requested by DSS and analytical workbench modules is being calculated for monitored individuals

5 Conclusions

This document firstly summarizes the methodology that was agreed in the scope of Task 2.1 of the project for collecting the user and technical requirements of the project. Based on this methodology, a list of concrete scenarios and user requirements for each of the pilot was specified, along with the initial version of their relevant user requirements. What is more, the technical partners of the consortium also provided the initial set of technical requirements as they were foreseen at this starting phase of the project. Additionally, it provides the state-of-the-art analysis of the base technology sectors that the iHelp project is involved, and could possibly exploit, along with a list of baseline technological tools and solutions that are planned to be incorporated in the overall platform. At this initial phase of the project, the outcomes of this deliverable have created valuable input for the progress of the task that is related with the design of the overall architecture of the platform.

This is the second of a series of versions that are planned to be released through the project, following the first version – i.e. D2.1 – State of the art & Requirement Analysis I. As it has been published on M12, this second version updates the current list of the user and technical requirements, taking into consideration that the use cases, data and technical requirements have been more mature. At that point, the overall architecture will need to be further refined and extended, in order to cover more advanced scenarios that were not initially considered. Finally, a third version is planned to be delivered on M18, in order to cover or remaining aspects and to correct potential erroneous decisions or unnecessary requirements that might have been identified earlier, so that it can drive the final definition of the requirements that will drive the overall architecture of the project, as the latter will be heading towards to its conclusion.

Bibliography

A. Doan, A. Halevy, and Z. G. Ives, Principles of Data Integration. Elsevier, 2012.

D. Bender and K. Sartipi, “HL7 FHIR: An Agile and RESTful approach to healthcare information exchange”, *In Proceedings of the 26th IEEE international symposium on computer-based medical systems*, p. 326-331, June, 2013.

D. Peng and F. Dabek, "Large-scale incremental processing using distributed transactions and notifications," 2010.

D. G. Ferro, F. Junqueira, I. Kelly, B. Reed and M. Yabandeh, "Omid: Lock-free transactional support for distributed data stores,", Apr. 2014, DOI**:** 10.1109/ICDE.2014.6816691

D. Cusumano *et al.*, “Delta Radiomics Analysis for Local Control Prediction in Pancreatic Cancer Patients Treated Using Magnetic Resonance Guided Radiotherapy,” *Diagnostics*, no. 1, p. 72, Jan. 2021, doi: 10.3390/diagnostics11010072.

E. A. Locke and G. P. Latham, “Building a practically useful theory of goal setting and task motivation: A 35-year odyssey.,” *American Psychologist*, no. 9, pp. 705–717, Sep. 2002, doi: 10.1037/0003-066x.57.9.705.

F. Luna-Perejon *et al.*, “Evaluation of user satisfaction and usability of a mobile app for smoking cessation,” *Computer Methods and Programs in Biomedicine*, p. 105042, Dec. 2019, doi: 10.1016/j.cmpb.2019.105042.

F. Cellini *et al.*, “Basics and Frontiers on Pancreatic Cancer for Radiation Oncology: Target Delineation, SBRT, SIB Technique, MRgRT, Particle Therapy, Immunotherapy and Clinical Guidelines,” *Cancers*, no. 7, p. 1729, Jun. 2020, doi: 10.3390/cancers12071729.

G. Macchia *et al.*, “Quality of Life and Toxicity of Stereotactic Radiotherapy in Pancreatic Tumors: A Case Series,” *Cancer Investigation*, no. 2, pp. 149–155, Feb. 2012, doi: 10.3109/07357907.2011.640649.

G. C. Mattiucci *et al.*, “External Beam Radiotherapy Plus 24-Hour Continuous Infusion of Gemcitabine in Unresectable Pancreatic Carcinoma: Long-Term Results of a Phase II Study,” *International Journal of Radiation Oncology*Biology*Physics*, no. 3, pp. 831–838, Mar. 2010, doi: 10.1016/j.ijrobp.2009.02.013.

G. C. Mattiucci *et al.*, “Prognostic Impact of Presurgical CA19-9 Level in Pancreatic Adenocarcinoma: A Pooled Analysis,” *Translational Oncology*, no. 1, pp. 1–7, Jan. 2019, doi: 10.1016/j.tranon.2018.08.017.

G. Macchia *et al.*, “Preoperative Chemoradiation and Intra-Operative Radiotherapy for Pancreatic Carcinoma,” *Tumori Journal*, no. 1, pp. 53–60, Jan. 2007, doi: 10.1177/030089160709300110.

G. C. Mattiucci *et al.*, “Hypofractionated sequential radiotherapy boost: a promising strategy in inoperable locally advanced pancreatic cancer patients,” *Journal of Cancer Research and Clinical Oncology*, no. 3, pp. 661–667, Oct. 2020, doi: 10.1007/s00432-020-03411-7.

H. op den Akker, M. Cabrita, R. op den Akker, V. M. Jones, and H. J. Hermens, “Tailored motivational message generation: A model and practical framework for real-time physical activity coaching,” *Journal of Biomedical Informatics*, pp. 104–115, Jun. 2015, doi: 10.1016/j.jbi.2015.03.005.

H. op den Akker, V. M. Jones, and H. J. Hermens, “Tailoring real-time physical activity coaching systems: a literature survey and model,” *User Modeling and User-Adapted Interaction*, no. 5, pp. 351–392, Jun. 2014, doi: 10.1007/s11257-014-9146-y.

Harm op den Akker ,Rieks op den Akker, Tessa Beinema, Oresti Banos, Dirk Heylen,Björn Bedsted, Alison Pease, Catherine Pelachaud, Vicente TraverSalcedo, Sofoklis Kyriazakos, and Hermie Hermens, *“Council of Coaches - A Novel Holistic Behavior Change Coaching Approach”,* DOI: 10.5220/000678770219022

J. N. Gray, “Notes on data base operating systems,” in *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 1978, pp. 393–481.

J. C. Corbett *et al.*, “Spanner,” *ACM Transactions on Computer Systems*, no. 3, pp. 1–22, Aug. 2013, doi: 10.1145/2491245.

L. Boldrini, D. Cusumano, F. Cellini, L. Azario, G. C. Mattiucci, and V. Valentini, “Online adaptive magnetic resonance guided radiotherapy for pancreatic cancer: state of the art, pearls and pitfalls,” *Radiation Oncology*, no. 1, Apr. 2019, doi: 10.1186/s13014-019-1275-3.

L. Placidi *et al.*, “On-line adaptive MR guided radiotherapy for locally advanced pancreatic cancer: Clinical and dosimetric considerations,” *Technical Innovations & Patient Support in Radiation Oncology*, pp. 15–21, Sep. 2020, doi: 10.1016/j.tipsro.2020.06.001.

M. Fiordelli, N. Diviani, and P. J. Schulz, “Mapping mHealth Research: A Decade of Evolution,” *Journal of Medical Internet Research*, no. 5, p. e95, May 2013, doi: 10.2196/jmir.2430.

M. Alrehiely, “A Taxonomy for Visualisations of Personal Physical Activity Data on Self-Tracking Devices and their Applications – ScienceOpen,” *ScienceOpen*. https://dx.doi.org/10.14236/ewic/HCI2018.17

M. T. Özsu and P. Valduriez, Principles of Distributed Database Systems. Springer International Publishing, 2020.

S. Hors-Fraile *et al.*, “A recommender system to quit smoking with mobile motivational messages: study protocol for a randomized controlled trial,” *Trials*, no. 1, Nov. 2018, doi: 10.1186/s13063-018-3000-1.

S. Hors-Fraile *et al.*, “Opening the Black Box: Explaining the Process of Basing a Health Recommender System on the I-Change Behavioral Change Model,” *IEEE Access*, pp. 176525–176540, 2019, doi: 10.1109/access.2019.2957696.

S. Syed-Abdul *et al.*, “Artificial Intelligence based Models for Screening of Hematologic Malignancies using Cell Population Data,” *Scientific Reports*, no. 1, Mar. 2020, doi: 10.1038/s41598-020-61247-0.

Syed-Abdul, S., Firdani, R.P., Chung, H.J., Uddin, M., Hur, M., Park, J.H., Kim, H.W., Gradišek, A. and Dovgan, E., 2020. Artificial Intelligence based Models for Screening of Hematologic Malignancies using Cell Population Data. Scientific Reports, 10(1), pp.1-8

V. Valentini *et al.*, “Intraoperative Radiation Therapy in Resected Pancreatic Carcinoma: Long-Term Analysis,” *International Journal of Radiation Oncology*Biology*Physics*, no. 4, pp. 1094–1099, Mar. 2008, doi: 10.1016/j.ijrobp.2007.07.2346.

Xerox Corporation. Palo Alto Research Center, B. W. Lampson, and H. E. Sturgis, *Crash Recovery in a Distributed Data Storage System*. 1979.

List of Acronyms

EU	European Union
DoA	Description of Action
RwD	Real World Data
PREMs	Patient Reported Experience Measures
PC	Pancreatic Cancer
PROMs	Patient Reported Outcome Measures
CA	Consortium Agreement
D	Deliverable
IoT	Internet of Things
M	Month
DPO	Data Protection Officer
HHR	Holistic Health Records
AI	Artificial Intelligence
UPRC	University of Piraeus Research Center
ATC	Athens Technology Centre
LXS	LeanXcale
INS	Innovation Sprint
KOD	KODAR Systems
ENG	Engineering Ingegneria Informatics SpA
SIE	Siemens
ICE	Information Catalyst for Enterprise
UNIMAN	University of Manchester
UPM	Universidad Politecnica de Madrid
FPG	Agostino Gemelli University Policlinic
HDM	Hospital de Denia-MarinaSalud
KI	Karolinska Institutet
MUP	Medical University Plovdiv
TMU	Taipei Medical University
ATP	Alberta’s Tomorrow Project
NHS-HC	National Health Service cardiovascular health check
EORTC	European Organization for Research and Treatment of Cancer
EHR	Electronic Health Records
EMR	Electronic Medical Records
FaaS	Function as a Service
HTAP	Hybrid Transactional and Analytical Processing
OLAP	Online Analytical Processing

OLTP	Online Transactional Processing
BI	Business Intelligence
ACID	Atomicity, Consistency, Isolation, Durability
P	Pilot
SQL	Structured Query Language
CEP	Complex Event Processing
OWL	Web Ontology Language
RDF	Resource Description Framework