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

D1.2 – Data Management Plan

WP1 Project Set-Up, Quality Procedures and Metric Definitions

Dissemination Level: Public
Document type: Open Research Data Pilot
Version: 1.0.0
Date: June 30, 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	Data Management Plan
Work package	WP1
Due Date	30/06/2021
Submission Date	30/06/2021
Start Date of Project	January 1, 2021
Duration of project	36 months
Main Responsible Partner	University of Piraeus Research Centre (UPRC)
Deliverable nature	Report
Author name(s)	George Manias (UPRC), George Marinos (UPRC), Stavroula Meimetea (UPRC), Dimosthenis Kyriazis (UPRC), Kenneth Muir (UNIMAN), Shwetambara Malwade (TMU), Rostislav Kostadinov (MUP), Andrea Damiani (FPG), Nerea Aguado Lopez (HDM)
Reviewer name(s)	Fabio Melillo (ENG), Usman Wajid (ICE)

Document Revision History

Version History			
Version	**Date**	**Author(s)**	**Changes made**
0.1	2021-04-25	Dimosthenis Kyriazis (UPRC), George Manias (UPRC), George Marinos (UPRC)	Initial version and content
0.2	2021-06-01	Shwetambara Malwade (TMU), Kenneth Muir (UNIMAN), Nerea Lopez (HDM), Andrea Damiani (FPG), Rostislav Kostadinov (MUP)	Updates and contributions from pilot partners
0.3	2021-06-09	George Manias (UPRC), George Marinos (UPRC), Stavroula Meimetea (UPRC)	Consolidated version ready for Internal Review
0.4	2021-06-24	Fabio Melillo (ENG), Usman Wajid (ICE), George Manias (URPC)	Updated version based on internal review feedback
1.0	2021-06-28	George Manias (URPC), Fabio Melillo (ENG)	Final version

Table of Contents

Executive summary 4
1 Introduction 5
2 Pilots' Data Summary 6
2.1 Pilot#1 - Study of Genomics and Epigenomics Markers for Early Risk Assessment of Pancreatic Cancer 6
2.1.1 Data Purpose 6
2.1.2 Data Collection & Description 7
2.1.3 Personal Data/Ethical Issues 12
2.1.4 Data Utility 13
2.2 Pilot#2 - Interventional Monocentric Study based on Patient Reported Outcomes 14
2.2.1 Data Purpose 14
2.2.2 Data Collection & Description 14
2.2.3 Personal Data/Ethical Issues 15
2.2.4 Data Utility 15
2.3 Pilot#3 - Study of Lifestyle Choices on Elevating the Risk Factors for Pancreatic Cancer 17
2.3.1 Data Purpose 17
2.3.2 Data Collection & Description 18
2.3.3 Personal Data/Ethical Issues 18
2.3.4 Data Utility 19
2.4 Pilot#4 - Study of Risk, Personalised Recommendations and Measures to Raise Awareness of Relevant Factors 20
2.4.1 Data Purpose 20
2.4.2 Data Collection & Description 20
2.4.3 Personal Data/Ethical Issues 24
2.4.4 Data Utility 25
2.5 Pilot#5 - Study of Improved Risk Prediction Models and Targeted Interventions that can Delay the Onset of Cancer 26
2.5.1 Data Purpose 26
2.5.2 Data Collection & Description 26
2.5.3 Personal Data/Ethical Issues 28
2.5.4 Data Utility 28
3 Data Management Policy 29

3.1 Data Management in the Consortium 30
4 Data Security 32
5 Conclusions 33
Bibliography 34
List of Acronyms 35

Executive summary

This Deliverable defines and describes the overall Data Management Plan (DMP) of the iHelp project. The DMP seeks to identify the best practices and specific standards for the generated data and assess their suitability for sharing and reuse in accordance with official guidelines. To this end, it aims to support the data management lifecycle for all data that will be collected, processed or generated by the project in order to maximise their accessibility and usage, according to the H2020 Pilot on Open Research Data (ORDP) in which the project participates.

The iHelp DMP will comply with the European Commission's (EC) Data Management Plan template and will specify how the generated data will be easily discovered and accessed, ensuring open access by adopting the adequate licensing scheme (e.g. Creative Commons License). The latter reflects the status of the data that are collected, processed, or generated, and the respective methodology and standards, whether and how these data will be shared and made open, and how they will be curated and preserved.

The DMP defines the general policy and approach to data management in the iHelp project, covering the data management related issues both on the administrative and on the technical level. To this end, this deliverable includes topics like application reconfiguration logs and monitoring metrics collection, publication and deposition of open data, the data repository infrastructure and compliance with the Open Access Infrastructure for Research in Europe (OpenAIRE). The current document also defines the foreseen resources needed for the data openness, and the ethical aspects that are taken into consideration in the context of iHelp.

1 Introduction

This document summarizes the iHelp's Data Management Plan (DMP), which has been conceived to support the data management life cycle for all data that will be collected, generated, and processed by the project. The iHelp's DMP identifies best practices for gathering information about the variety of data to be used in the project that will optimise the development, specific processes and repositories for the generated data and assess their suitability for the sharing and reuse in accordance with official guidelines.

The structure of this document complies with the guidelines and recommendations specified in the European Commission's Data Management Template and will offer open access to its scientific results reported in publications, to the relevant scientific data and to data generated during the course of the project. The rest of this deliverable is organised and structured into three main sections. Section 2 has a two-fold objective. In one hand it seeks to deliver and detail pilot's datasets and collection actions, while in the other hand it also summarizes the purpose of the collected data and ethical issues that may derive from the collection of personal and sensitive data. To this end, this section is being further separated into pilot-oriented subsections in which each pilot describes a summary of the data to be handled and analysed. Moreover, Section 3 provides the data management policy to be followed about the search, retrieval and access to data, while Section 4 address the security of handled datasets. Finally, Section 5 concludes this deliverable.

2 Pilots' Data Summary

In this section are being described and listed the pilots' datasets that will be collected and generated in the scopes of the iHelp project. These datasets, their collection procedures, and their purpose have been initially identified through the requirements collection, reported in the scopes of D2.1 State of the art & Requirements Analysis I that will also be released during M06, and are input datasets for each of the project's pilot. Moreover, in the below subsections are being described ethical, legal, and regulatory issues related to the protection of personal and sensitive data that may arise during the iHelp project as they have been reported in previous ethics related deliverables of the project (G., G., 21), (S., S., G., 21), (Sar, 21).

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

The pilot focuses on *Genomics and Epigenomics Markers for Early Risk Assessment of Pancreatic Cancer*.

2.1.1 Data Purpose

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

- 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 the implementation of an interactive iHelp platform.

The risk prediction functionality in the UNIMAN pilot will be developed using available large datasets such as the UKBiobank[1]. The pilot partner 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 consenting participants within in a range of community health check settings including the National Health Service - Health Check (NHS-HC).

On top of this, the UNIMAN pilot seeks (i) to provide individuals with information about risk factors for future disease is an important public health approach; (ii) to raise awareness of health conditions and educate individuals on how to prevent life threatening diseases like Pancreatic Cancer. To this end, UNIMAN will recruit pilot subjects via the UK National Health Service-Health Check and other community engagement opportunities. Subject criteria include age over 50 and no history of any type of cancer. After consent, UNIMAN will run cancer risk prediction and only subjects with high risk will be invited to provide blood/DNA sample to assess their predisposition, biological age and potential cancer markers. All subjects will then be guided towards iHelp platform which aims to support risk mitigation activities. UNIMAN will evaluate the compliance in all groups and in particular, high-risk group to see if by adding "omics" information will enhance behavioral change.

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

2.1.2 Data Collection & Description

The UNIMAN pilot will assemble 2 types of data as also stated in the scopes of D2.1 State of the art & Requirements Analysis I of the project that will also be released during M06.

- Primary data

The primary data will be acquired from large studies including the UKBiobank, the lifeline data and the Albertas' Tomorrow Project (ATP) data. UNIMAN pilot has permission to use these datasets in order to develop its risk models and validate them.

- Secondary data

In this pilot, secondary data will be collected from healthy individuals who are eligible to attend NHS-HC. Data will be collected using 1) the questionnaire (general characteristic data) 2) for cancer risk assessment, data will be collected through UNIMAN's online cancer risk prediction platforms 3) in high-risk groups, biological samples will be collected to process and assess genetics and epigenetic markers, 4) all individuals will provide data from agreed set of target activities via iHelp platform.

Datasets that will be utilized, examined, and analysed in the scopes of UNIMAN pilot datasets are being described below.

Table 1: 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 are 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)	Data on disease outcomes will be updated periodically. The 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 2: 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 are owned 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 3: UNIMAN Pilot – 3rd Primary dataset detail

Section	Description
ID	DS-P1-03
Title	The Albertas' Tomorrow Project (ATP).
Description	A large scale Canadian population based data with comprehensive medical, lifestyle data.

Owner	The data are owned 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 4: UNIMAN Pilot – 4th Primary dataset detail

Section	Description
ID	DS-P1-04
Title	Survey data
Description	These 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 5: UNIMAN Pilot – 5th Primary dataset detail

Section	Description
ID	DS-P1-05
Title	Risk assessment data
Description	These 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 6: UNIMAN Pilot – 6th Primary dataset detail

Section	Description
ID	DS-P1-06
Title	Biological data
Description	These 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 biological data will be not be sharing as a raw data but instead will be shared as an information of predisposition to any partners. For example, participant identification number and summary genetic 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.

Table 7: UNIMAN Pilot – 7th Primary dataset detail

Section	Description
ID	DS-P1-07
Title	Risk mitigation data
Description	These 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 (interactive with periodically data collection at 3, 6 and 12 months)
Type of process (Stream or static data)	Dynamic stream 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.

2.1.3 Personal Data/Ethical Issues

UNIMAN pilot will account for personal data protection and privacy and hence data from participants will be identified by ID and the linkage file will only be held by the core team. UNIMAN will abide by the ethics of data protection and privacy and only consented data will be shared with pseudonymous identification number. Any linkage that could lead to person identifier will not be carry out to protect privacy and if participants has withdrawn, data will be deleted promptly.

Moreover, data will be encrypted and securely stored in a password protected computer (University of Manchester managed machine).

A sample consent form is cited in deliverable D1.9 (Sar, 21).

Participants who engage with the NHS-Health Check or other community-based health checks will be assessed for:

- No prior cancer diagnosis
- Consent to the study
- No history of psychiatric problems
- English speaking as there will be no interpreter providing.

Once ethical approval is obtained a master file will be created for the pilot and all relevant documents will be submitted. Details will be made available through relevant deliverables D6.3 and D6.4 (Pilot setup and implementation of digital trials – I & II) at M24 and M32 respectively.

2.1.4 Data Utility

The collected data content will be of interest to both the healthcare sector, as well as to a wider community of data scientists, or practitioners of healthcare data science, to carry out machine learning research. DevOps operators will also be interested in the application and infrastructure monitoring data as a means of performance testing against the end users' practices. The usefulness of other derivative data which will be published, as well as platform performance data included in scientific journal publications, conference paper submissions and other scientific dissemination activities. These will be useful to researchers and hospitals looking to expand their understanding of challenges of pancreatic cancer and other types of cancer.

Moreover, UNIMAN pilot will generate various types of data and each will offer utility to various users and/or stakeholders.

- Data generated using the questionnaires to monitor behavioural changes. These data can be useful for machine learning or AI community data scientists for future use of AI to predict factors or input that can help enable behavioural changes.
- Data of risk factors collected to enabling risk prediction. These data could be of interest to Public Health sector as to prevalence of specific risk factors in the community. These data can also be used to help target public message in the UK NHS-Health Check.
- Data on biological markers. These data can be added to the literature related to pancreatic cancer. Summary data can also be shared with the peers. This is useful for further knowledge and validation exercise for future studies.
- Data on factors contributing to the risk mitigation together with identification of risk mitigation based on individual persona can also be shared with the wider communities including social care, marketing research and health authorities.

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

The pilot focuses on *Interventional Monocentric Analysis based on Patient Reported Outcomes*.

2.2.1 Data Purpose

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. The FPG pilot study in iHelp aims at the definition of personalised therapeutic strategies that can maximize their efficacy and reduce the risks offside effect in patients affected by Pancreatic Cancer and bringing significant improvements in terms of Quality of Life (QoL) of high-risk individuals. Patients affected by pancreatic cancer with indication to neo-adjuvant, exclusive or adjuvant radiotherapy or chemoradiation will be enrolled according to the inclusion criteria (see below). A written consent will be obtained by the patient to be enrolled. Patients will be able to withdraw consent in each moment of the study without any changes in the diagnostic and therapeutic process.

2.2.2 Data Collection & Description

FPG pilot's prospective acquisition of secondary data is conducted from the first clinical evaluation, until the end of the treatment and the first follow-up period. On top of this, secondary data will be collected from mobile application, IoT device, and clinical evaluation. In particular, data will be extracted from the Hospital health records and integrated with lifestyle data coming from IOT/wearables/apps that will be utilized in the scopes of iHelp project.

On top of this, primary retrospective data come from a cohort of patients treated by adjuvant radiotherapy (with or without chemotherapy) are already available in pilot's institution. This dataset includes data regarding staging, markers, surgery, histological examination, radiotherapy, chemotherapy, toxicity, clinical outcomes.

Table 8: 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	Pseudonymisation
Regulatory Constraint Requirements	Driven by Ethical Committee feedback

2.2.3 Personal Data/Ethical Issues

Technical and organisational measures are already in place at FPG to ensure safeguard of all rights and freedoms of data subjects / research participants, according to the laws and regulations both at the E.U. and the national level.

Lawful basis for the reuse of previously collected patient data will be obtained from DPO and Ethical Committee, with all due and appropriate technical and organisational measures suggested by said entities, in total respect of the laws and regulations both at the E.U. and the national level. Moreover, the rights and freedoms of the data subjects/research participants will abide by the principles of the Declaration of Helsinki (WMA, 01).

Furthermore, pseudonymisation is an established policy at Policlinico Gemelli that is achieved via a combination of standard algorithms (hashing, rotations, key-based encryption) and proprietary solutions and will be used during the development of FPG pilot.

Also, the evaluation of the ethical risks related to the data processing activities and more in general with the pilot will be performed by the Ethical Committee as a preliminary step in the execution of the project.

No activity on patient data will be possible prior to explicit approval from the Ethical Committee.

If so decided by the FPG DPO, a data protection impact assessment will be conducted internally during year 2021.

2.2.4 Data Utility

Clinical researchers and policymakers at large will obtain information on how oncological patients accept wearables and apps during their care path; data about the lifestyle of oncological patients will be available,

making it possible to measure the impact of lifestyle on outcomes, with special attention to undesired effects and toxicities.

On top of this, clinical researchers will also benefit from the dissemination activities on provided data, e.g. by reading scientific papers describing the study, they will be able to replicate it on the same pathology or translate it into a different setting.

Information on wearable and app acceptance will also be useful for designers and producers, to guide them toward an improved design.

Moreover, local clinicians will have a more complete view of their patients and will be able to know them better under different points of view, thus improving their care path. This information will also be important to specialized psychologists, who will be able to study a new kind of data, more related to everyday life and the social environment of the oncological patient.

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

The pilot focuses on *Lifestyle Choices on Elevating the Risk Factors for Pancreatic Cancer*.

2.3.1 Data Purpose

The Hospital de Denia-MarinaSalud (HDM) pilot main objective 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. The HDM pilot will be targeted on the studying the effects of lifestyle choices on the risks associated with Pancreatic Cancer. Human beings will be involved in HDM pilot through the treatment of their health data. The two main lines and types of data that will be examined and utilized will be:

- Treatment of primary data: In this activity bulk historical data should be extracted from central Electronic Medical Reports (EMRs) and will be used to train Artificial Intelligence (AI) algorithms. Clinical staff of HDM will select those patients and their data that would be of interest for the project. These data will be anonymized in a way that it will not be possible to identify the human being it refers to and send to iHelp data sources to train AI algorithms.
- Treatment of secondary data: In this activity current patients that will accept to participate in the project will sign consent for the use of their data. These data will be pseudonymized and only the technical staff of HDM will have access to the pseudonymization keys to revert the process. The signed consent, so, will be the base legitimation for use of the health data, what will be aligned with GDPR rules. These data will be treated from European and Non-European stakeholders and will not be held or further used unless this is essential for reasons that were clearly stated in advance to support data privacy.

It is currently known that a risk factor is anything that increases individual's chance of getting a disease such as cancer. Different cancers have different risk factors. Some risk factors, like smoking, can be changed. Others, like a person's age or family history, can not be changed.

In some cases, there might be a factor that may decrease individual's risk of developing cancer or has an unclear effect. That is not considered a risk factor, but it can be noted clearly on this page as well.

Having a risk factor, or even many, does not mean that someone will get cancer. And some people who get cancer may have few or no known risk factors.

In the Marina Salud 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.

2.3.2 Data Collection & Description

The HDM pilot will assemble 2 types of data.

- Primary data based on the Electronic Health Records (HER) of Marina Salud hospital.
- Acquisition of secondary data from the trial of patients selected. These data will be collected from mobile application, IoT device, clinical evaluation and self-evaluation.

Table 9: 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	Data are stored in a SQL Server, hence they can be exported and provided in any format that is needed (e.g. json, csv).
Data store	Marina Salud DWH
Data Security	Data will be moved to a pseudonymized data store before being extracted.
Regulatory Constraint Requirements	Driven by Ethical Committee feedback

2.3.3 Personal Data/Ethical Issues

The pilot will involve the use of health data, which are considered as personal data. The protection and privacy of these data will be ensured by applying the GDPR requirements, that is, the appliance of the appropriate organizational and technical measures. Only the required data will be used, and it will be both anonymized in the case of primary data or pseudonymized in the case of secondary data. In case of

pseudonymization only selected persons from HDM staff involved in the project will be able to revert the process. Data will not be held or further used unless this is essential for reasons that were clearly stated in advance. Also, technical measures will be applied to the databases used.

2.3.4 Data Utility

Data that will be collected and analysed can facilitate the procedures and mechanisms of providing recommendations to the health authorities to design a public health programme, identify lifestyle habits that increase the risk of developing the disease, and select the people most at risk of developing pancreatic cancer throughout their lives, in order to guide an early detection programme for pancreatic cancer at a European level.

The technological development that this project will produce is pioneering in healthcare. The digitisation of information, analysis and results will hopefully have an impact on how the European population becomes aware of their health, has access to guiding technologies to support them and help create a virtual environment for the promotion of health and healthy habits.

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

The pilot focuses on *Risk, Personalised Recommendations and Measures to Raise Awareness of Relevant Factors*.

2.4.1 Data Purpose

The main objective of the Medical University Plovdiv (MUP) pilot is to implement the AI (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 two 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 and etc. The MUP pilot will analyse and measure the healthcare awareness regarding the risk factors that increase the probability of malignant Pancreatic processes development. On top of this, an important aspect will be to analyse the existing Pancreatic Cancer related datasets to understand key risk factors in the local context e.g. local diet and other lifestyle aspects. Based on the identification of key risk factors through the application of AI learning models, the next step will be to understand the likelihood and significance of identified risks. Throughout the pilot, surveys will be carried out on medical specialists who are involved with the diagnostic and treatment processes of patients with risk of developing or who have already developed Pancreatic Cancer. Medical specialists will assess and evaluate the iHelp outcomes and generated data - in terms of their accessibility, significance for the diagnosis and recommendations, as well for the improvement of the ongoing treatment. In the scopes of the iHelp project participants from the pilot side will be medical specialists - medical doctors and registered nurses.

Both MUP participants in the project will have access to and will process data throughout the project, thus they are both responsible for the data protection in accordance with EU and Bulgarian regulations.

2.4.2 Data Collection & Description

The first data collection will be conducted 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 platform in order an evaluation of the risk to be performed. Based on the results and outcomes 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).

Furthermore, secondary data that will be utilized under the scopes of iHelp will be obtained and collected from the patients and will be related to measures according to quality of life, body temperature and weight, pain level, alcohol, meat, vegetables, grains consumption, pain killers and other medicine intake, social exclusion. The patients from whom these secondary data will be collected are to be included into the survey after been selected from the professionals and have given their written consent as also stated in ethics related deliverables of the project (G., G., 21), (S., S., G., 21), (Sar, 21). The results produced from the iHelp

will be used for proving the reliability of the product and its value for medical specialists. Healthcare providers 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 seeks 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 these data, the healthcare provider will be informed regarding the risk and recommended program to be discussed between him/her and patient at risk.

Table 10: 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 Format (?) – Most probably they will be in excel/csv format but since this deliverable will be submitted by M6 there may be some undefined topics regarding the data by the time of submission of this deliverable.
Data store	All obtained data will be stored into the MUP
Data Security	Access to the data will be granted only to the researchers working for the pilot
Regulatory Constraint Requirements	In accordance with the Bulgarian regulations

Table 11: 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 Format (?) – Most probably they will be in excel/csv format but since this deliverable will be submitted by M6 there may be some undefined topics regarding the data by the time of submission of this deliverable.
Data store	All obtained data will be stored into the MUP
Data Security	Access to the data will be granted only to the researchers working for the pilot
Regulatory Constraint Requirements	In accordance with the Bulgarian regulations

Table 12: 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 Format (?) – Most probably they will be in excel/csv format but since this deliverable will be submitted by M6 there may be some undefined topics regarding the data by the time of submission of this deliverable.
Data store	All obtained data will be stored into the MUP
Data Security	Access to the data will be granted only to the researchers working for the pilot
Regulatory Constraint Requirements	In accordance with the Bulgarian regulations

Table 13: 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 Format (?) – Most probably they will be in excel/csv format but since this deliverable will be submitted by M6 there may be some undefined topics regarding the data by the time of submission of this deliverable.
Data store	All obtained data will be stored into the MUP
Data Security	Access to the data will be granted only to the researchers working for the pilot
Regulatory Constraint Requirements	In accordance with the Bulgarian regulations

Table 14: 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 Format (?) – Most probably they will be in excel/csv format but since this deliverable will be submitted by M6 there may be some undefined topics

	regarding the data by the time of submission of this deliverable.
Data store	All obtained data will be stored into the MUP
Data Security	Access to the data will be granted only to the researchers working for the pilot
Regulatory Constraint Requirements	In accordance with the Bulgarian regulations

Table 15: 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 Format (?) – Most probably they will be in excel/csv format but since this deliverable will be submitted by M6 there may be some undefined topics regarding the data by the time of submission of this deliverable.
Data store	All obtained data will be stored into the MUP
Data Security	Access to the data will be granted only to the researchers working for the pilot
Regulatory Constraint Requirements	In accordance with the Bulgarian regulations

2.4.3 Personal Data/Ethical Issues

Protection of the data sources will be achieved through the exclusion of any personal data (name, address, phone number, contact details) by the medical specialists involved into the pilot. They will be identified by their specialty, years of medical practice and working place - prehospital or hospital healthcare providers. The data extracted by the hospital databases will be anonymised and the patients will be presented as 'patient 1, patient 2,' etc - again without names, gender, addresses. Thus, the personal data will be fully protected. The only personal data of the patients that must be included are data according to their age,

since age is one of the risk factors related to the development of the Pancreatic cancer. The data will be extracted on the above-described mode by the hospital records.

At the second stage of the prospective study medical specialists from various specialty will select patients at risk of Pancreatic cancer development. Medical specialists have to obtain patient consent to be included into the pilot monitoring, as a prerequisite for further pilot inclusion. After patient agrees to be included, then his/her data will be anonymised as those of the hospital records.

Every patient could terminate his/her participation into the pilot study by informing the medical specialist.

2.4.4 Data Utility

Data generated and analyzed in the scopes of this pilot can be used by various medical professionals:

- From the Medical Faculties and Colleges professors in order to teach the medical students on pancreatic cancer clinical features and risk factors for its development, as well as the required preventive measures to be imposed in certain patients that could be assessed with moderate or high risk.
- From the public health managers in order to establish a program for achieving general population awareness regarding those customs, environmental and behavioral factors that could increase the risk of Pancreatic cancer development.
- From Healthcare systems Managerial bodies - based on iHelp's generated data a screening and preventive program could be established and implemented into the routine out-of-hospital medical care for decreasing the risk level of the individual and population for Pancreatic cancer development.
- Healthcare providers from different levels - general practitioners, specialist, medical technicians, physician assistants, registered nurses for better screening of the patients at risk.

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

The pilot focuses on *Improved Risk Prediction Models and Targeted Interventions that can Delay the Onset of Cancer*.

2.5.1 Data Purpose

The Taiwan Medical University (TMU) pilot aims to develop machine learning algorithms that can be tuned to predict high risk individuals for pancreatic as well as liver cancer for early management of modifiable risk factors (lifestyle, behavior) among these individuals. TMU will also be conducting digital trials involving the use of mobile applications for personalized healthcare and monitoring of risks and recommendations related to pancreatic and liver cancer.

Thus, the goals include:

- Applying data analytics, AI based algorithms or deep learning technologies to analyze 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.

2.5.2 Data Collection & Description

In the first phase of risk prediction, the required 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, symptoms, subjective sleep assessment etc.

Table 16: TMU Pilot – 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 17: TMU Pilot – Secondary 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.

2.5.3 Personal Data/Ethical Issues

The rights and freedoms of the data subjects'/research participants will abide by the principles of the Declaration of Helsinki (WMA, 01). Participants will be recruited only after they have signed the patient consent form (from the TMU-Joint Institutional Review Board: TMU ethical committee). Patient's identification will be confidential, and all the patient's test results and diagnoses made in the trial/research will be labelled with a project serial number and patient's name will be taken off from all labels. All documents containing evidence to study eligibility, history and physical findings, laboratory data, results of consultations, etc., as well as Institutional Review Board (IRB) records and other regulatory documentation will be retained by the Protocol Lead Investigator in a secure storage facility in compliance with Taiwan regulations and guidance.

The TMU pilot will undergo pseudonymisation. During the research period a research number will replace personal ID for labelling, and data will be stored in a locker for restricted access and for privacy protection.

To this end, ethical issues to be considered:

- Exposing identity and sensitive data (privacy breach): Personal ID information and trial/research-related records will be kept confidential.
- Security/safety risks for the research participants: Data in the database is already anonymized. During the research period a research number will replace personal ID for labelling. This allows for the data to be safe and secured.
- Potential for misuse of data: Data will be secured in a storage facility with restricted access for privacy protection. TMU will require the data scientist to sign a Non-disclosure agreement (NDA) with TMU for the use of the dataset in Taiwan.

2.5.4 Data Utility

The database in the first phase (TMU-CDR) may be of interest to risk model-builders, for instance health care data scientists/machine learning developers/researchers/technicians, etc. In addition, it could also be of interest to students of healthcare informatics research for conducting data analytics related research activities. The performance of the developed models will be considered for scientific dissemination activities including publications in scientific journals, conference paper presentations and posters and other similar communication activities.

Furthermore, medical professionals or model-users (clinicians /consultants/ specialists / oncologist or general practitioners/ nurses or paramedical), could be interested to expand their understanding of risk factors towards development of pancreatic and liver cancer.

3 Data Management Policy

iHelp project involves the processing of data from different sources to provide evidence-based policy making in the full lifecycle of policy management. In this section it is described the Data Management Policy of iHelp, based on the guidelines and on the template for FAIR Data Management in H2020, as part of making research data findable, accessible, interoperable and re-usable (FAIR).

The overall Data Management Policy in iHelp Project has been set according to the following principles:

- Identification of datasets collected or generated by the project, as reported in subsections of Section 2.
- Definition of the principles for exploitation, availability, access rights and re-use of data managed by the project.
- Definition of the principles for data archiving and preservation.
- Definition of the principles for ethical and legal compliance.

A dedicated task in the iHelp work program, more specific Task 6.1 “Scientific Coordination of Pilot Scenarios”, focuses on the definition of pilot protocol i.e. how the pilots will be conducted in a uniform way while complying with relevant ethical, security and privacy constraints. Moreover, a relevant Deliverable, D6.1 Coordination of pilot scenarios for personalized healthcare – early risk identification, prevention and intervention measures I, aims at the definition and circulation of this protocol. The protocol will highlight the objective(s) of the pilot, purpose, methodology for data collection and analysis, selection of individuals, the use of AI, design of relevant interventions, the use of technologies for monitoring and evaluation, the techniques for the analysis of impact (of personalised recommendations) and the general organisation of a user studies that ensures the safety of the subjects and integrity of the data collected. The pilot protocol will also define the procedures that can facilitate exchange of data/knowledge between different stakeholders in during the pilot phase.

The pilot protocol will be supported by the data management guidelines and the implementation of the Big Data Platform, provided by Task 4.4 “Big Data Platform and Knowledge Management System”, of the project. This integration between different tasks and components will support the establishment of a standardised real-world data framework that supports the highest degree of protection for research data, in accordance with all GDPR, privacy and security requirements. In this respect, the pilot protocol will adopt “protection by design” in each step of the process.

On top of this, one of the main objectives of the iHelp project is to deliver the iHelp Data Management Framework (DMF) that will be a cornerstone of any technical effort aiming at personalized healthcare solutions. This framework will be the basis for the integration of heterogeneous datasets and management of big health data in a standardised format. In the meanwhile, big data management mechanisms are being integrated while preserving the data ownership aspects to realise the complete data path: from acquisition, cleaning, to data harmonisation, interoperability, integration, modelling, analysis, information extraction, interpretation and decision support (through appropriate alerts, visualisation and reports).

Furthermore, as many iHelp results and outcomes as possible will be made openly accessible according to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020[2]. These will be made available through an Open Access repository that will be identified in collaboration with all partners and especially communication task leaders of the iHelp consortium. To this end, Zenodo[3] repository is the first identified research data repository that can be utilized and is fully OpenAIRE compliant. On top of this, the Zenodo metadata can be used, which is compliant with DataCite's Metadata Schema minimum and recommended terms, with a few additional enrichments[4].

Furthermore, open access to peer-reviewed scientific publications will be performed by publishing either in green or gold open access journals, and announced on the iHelp website[5], the OpenAIRE portal[6] and in the R&I Participants Portal[7].

As for the re-use and interoperability of the data, reports and open data produced in the course of the project will be made available in commonly used formats, as ODF or PDF for documents and JSON, XML or RDFs for data. In addition, the following standard vocabularies will be used in the default metadata schema for all types of open data:

- License: Open Definition[8]
- Funders: FundRef[9]
- Grants: OpenAIRE[10]
- Citation & Search: DataCite Metadata Schema[11]

3.1 Data Management in the Consortium

Data management activities are responsibility of the whole consortium, from creation of data, publication and dissemination, following the Data Management Plan, the Project Management Handbook (delivered in M3), the Communication and Collaboration Plan and Activities (to be delivered in M12), and the Exploitation plan (to be delivered in M12) guidelines. Moreover, specific individuals and groups of partners of the iHelp consortium have corresponding responsibilities.

The **Project Coordinator (PC)** is responsible to:

- Develop the data management plan in collaboration with the consortium partners.
- Supervise of the data management activities (for collection and publication) and associated milestones in coordination with Work package Leaders.
- Write the Data Management Plan (D1.2) release (in M6).
- Provide support to arising issues.

[2] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf. Retrieved 2020-05-21
[3] https://zenodo.org/
[4] https://about.zenodo.org/principles/
[5] https://ihelp-project.eu/
[6] https://www.OpenAIRE.eu/
[7] https://ec.europa.eu/research/participants
[8] http://opendefinition.org/
[9] https://www.crossref.org/services/funder-registry
[10] http://api.OpenAIRE.eu/
[11] https://schema.datacite.org/

The **Work Package Leaders** are responsible for:

- Implement the Data Management guidelines in their respective WPs.
- Monitor the Data Management activities and milestones with partners.
- Provide input to the data management plan by analysing and summarising the WP-specific related data.
- Provide support and guidance for the publication of open data and documents.
- Monitor that the open results (data and software) are deposited in the default repository or a complementary OpenAIRE-compliant repository and sending reminders to partners.
- Monitor that the open results available in OpenAIRE are properly linked with iHelp (see https://www.OpenAIRE.eu/participate/claim).
- Coordinate with the Project Coordinator (PC) for any ethical or privacy issues that may prevent the publication of data.

The **Communication task leader** is responsible for:

- Provide support on the best publication path (green or gold open access).
- Provide support for publication on scientific publications.
- Monitor that green access (self-archiving) publications are deposited in repositories.
- Monitor that metadata about publications is made available in the R&I Participant Portal (preferably automatically through OpenAIRE) and on the iHelp website.
- Ensure that research data related to a publication is made available in repositories and linked to respective publication.
- Monitor possible embargo periods and remind to partners.
- Ensure that publications available in OpenAIRE are properly linked with iHelp (see https://www.OpenAIRE.eu/participate/claim).

The **Data Provider / Scientist** is responsible for:

- Informing data & dissemination leaders when new open data / papers ready for publication are available.
- Describe the data (through appropriate metadata) or scientific publication in accordance with the iHelp data management policy (e.g. according to the chosen metadata standard).
- Depositing (publishing into a repository) the data or scientific publication in accordance with the iHelp data management policy and with help of the tools (catalogue, repository, etc.) provided by the project.

4 Data Security

Ethics and the advancement of Ethical AI stand at the forefront of the iHelp mission. The comprehensive use and management of data in the project and its solutions make the ethical, legal, socio-economic, cultural (ELSEC) and data protection considerations a cornerstone of this project. Therefore, iHelp consortium has a strong commitment to guarantee and foster the human-centric orientation that characterizes the European AI community. The iHelp consortium seeks to implement the Trustworthy AI principles during the whole project lifetime, both in design and piloting phases, while it also will make use of the tools provided by the High-Level Expert Group on Artificial Intelligence (HLEG) to implement their requirements. The security framework of the iHelp project will provide identification, authentication and authorization for business-to-business and business-to-consumer models. In addition, the framework ensures data confidentiality and integrity based on appropriate encryption mechanisms and methodologies. The latter takes into consideration the management of personal data and the solutions to minimize the need for such data in the data to be analyzed, not only for project results to be shared but also to ensure compliance with data protection regulations of the platform being developed. To this end, in Section 2 appropriate techniques, mechanisms and procedures are being identified and analyzed in terms of protection and security of the examined data.

Moreover, the ethical and legal principles that will guide iHelp project are declared in Section 5 “Ethics and Security” of the iHelp GA (EC, 2020), including privacy, for the protection of personal data, and security enforcement within the project. To this end, ethical, legal, and regulatory issues related to the protection of personal and sensitive data that may arise during the iHelp project and appropriate guidelines and principles for addressing these issues have been identified and reported in previous ethics related deliverables of the project (G., G., 21), (S., S., G., 21), (Sar, 21).

On top of this, iHelp seeks to deliver mechanisms that will standardize data management and knowledge sharing within security and privacy constraints in order to allow the use and reuse of sensitive health related data for bringing improvements in the advice and interventions provided to patients. The latter will be achieved by ensuring that the data gathered, and knowledge generated in the project is managed through security and privacy by design approach and based on the combination of appropriate standards (i.e. HL7 FHIR[12]) with ontology alignment and semantic annotation techniques. Moreover, data sharing across the whole data lifecycle of the project is supported by SSL/TLS encryption for any external connection and ZFS based encryption techniques to ensure security and privacy of data.

[12] https://www.hl7.org/fhir/

5 Conclusions

The Data Management Plan (DMP) for the iHelp project, reports the provisions for the data management strategy to be applied to the project datasets and across the whole data lifecycle to make project's data compliant with the FAIR guidelines.

iHelp DMP procedures make use of solutions and standards provided by the High-Level Expert Group on Artificial Intelligence (HLEG) for their implementation. To this end, iHelp's results, including public reports, open data, open access publications and open-source software can be accessible and available once the project is finished.

Furthermore, the DMP provides the first overview on the data that is collected, processed, or generated following the methodology and standards set out in the data management policy.

Bibliography

European Commission, “iHELP Grant agreement ID: 101017441”, December 2020.

S. Fairhurst, “D1.9 Ethical issues related to the involvement of Humans in iHelp”, iHelp, March 2021.

S. Malwade, S. Syed-Abdul, G. Marinos, and G. Manias, “D1.11 Ethical Issues Related to the Involvement of Non-European Countries”, iHelp, March 2021.

G. Marinos and G. Manias, “D1.10 Ethical issues related to the involvement of Humans in iHelp”, iHelp, March 2021.

World Medical Association, “World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects”, Bulletin of the World Health Organization 79, no. 4, pp. 373 – 374, 2001.

List of Acronyms

AI	Artificial Intelligence
ATP	Albertas' Tomorrow Project
CA	Consortium Agreement
CSV	Comma-Separated Values
D	Deliverable
DMF	Data Management Framework
DMP	Data Management Plan
DoA	Description of Action
DPO	Data Protection Officer
DWH	Data Warehouse
EC	European Commission
EHR	Electronic Health Record
EMRs	Electronic Medical Records
EU	European Union
FPG	Agostino Gemelli University Policlinic
GDPR	General Data Protection Regulation
HDM	Hospital de Denia-Marina Salud
HHR	Holistic Health Records
HLEG	High-Level Expert Group
ID	Identity Document
IoT	Internet of Things
IRB	Institutional Review Board
M	Month
MUP	Medical University Plovdiv
NDA	Non-Disclosure Agreement
NHS-HC	National Health Service - Health Check
ORDP	Pilot on Open Research Data
PREMs	Patient Reported Experience Measures
PROMs	Patient-Reported Outcome Measures
REST	Representational State Transfer
RWD	Real-World Data
SSL	Secure Sockets Layer
TLS	Transport Layer Security
TMU	Taipei Medical University
UNIMAN	University of Manchester
ZFS	Zettabyte File System