PROJECT
i-PROGNOSIS: Intelligent Parkinson early detection guiding novel supportive interventions

GRANT AGREEMENT No.
690494

D5.1 - Open research data management plan

CONTRACTUAL SUBMISSION DATE
July 2016

ACTUAL SUBMISSION DATE
July 2016

REVISION DATE
August 2017

DELIVERABLE VERSION
11.0

MAIN AUTHOR(S)
Antonis Billis (AUTH)
Evdokimos Konstantinidis (AUTH)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 690494.

GRANT AGREEMENT No.	690494
PROJECT ACRONYM	i-PROGNOSIS
PROJECT FULL TITLE	Intelligent Parkinson early detection guiding novel supportive interventions
Type Of Action	Research & Innovation Action (RIA)
Topic	H2020-PHC-21-2015 - Advancing active and healthy ageing with ICT: Early risk detection and intervention
Start Of Project	1 February 2016
Duration	48 months
Project URL	www.i-prognosis.eu
EU Project Officer	Ramón Sanmartín Sola

DELIVERABLE TITLE	Open research data management plan
DELIVERABLE No.	D5.1
Deliverable Version	11.0
Deliverable Filename	i-PROGNOSIS-690494_D5.1_UpdY2.docx
Nature Of Deliverable	R (Report)
Dissemination Level	PU (Public)
Number Of Pages	70
Work Package	WP5 - i-PROGNOSIS Integration and Orchestration
Partner Responsible	AUTH
Author(s)	Evdokimos Konstantinidis (AUTH), Antonis Billis (AUTH) Konstantinos Kyritsis (AUTH), Stelios Hadjidimitriou (AUTH), Vasileios Charisis (AUTH), Leontios Hadjileontiadis (AUTH)Christos Frantzidis (AUTH), Ioannis Ioakeimidis (KI), Michael Stadtschnitzer (FRAUNHOFER), Nikos Grammalidis (CERTH), Hugo Silva (PLUX), Dhaval Trivedi (KCL), Lisa Klingelhoefer (TUD)
Editor	Panagiotis Bamidis (AUTH)

ABSTRACT

Deliverable D5.1 aims to present the plan concerning the open research data management according to the guidelines for data management in the H2020 Online Manual. This deliverable will evolve during the lifetime of the project in order to present the status of the project's reflections on data management. The deliverable contains provisional information about the data that will be produced and collected within the project, whether and how it will be made

	accessible for re-use and further exploitation, and how it will be curated and preserved.
KEYWORDS	Data management plan (DMP); Datasets; Open access; Open research

SIGNATURES

WRITTEN BY	RESPONSIBILITY - COMPANY	DATE
Antonis Billis	Main author 1 - AUTH	22/7/2017
Evdokimos Konstantinidis	Main author 2 - AUTH	22/7/2017
REVIEWED BY		
Fotis Karayiannis	Internal Reviewer 1 - MICROSOFT	26/7/2016
Michael Stadtschnitzer	Internal Reviewer 2 - FRAUNHOFER	26/7/2016
APPROVED BY		
Leontios Hadjileontiadis	Project Coordinator - AUTH	28/7/2016
REVISED BY		
Antonis Billis	Revisions author - AUTH	31/08/2017

TABLE OF CONTENTS

LIST OF MAIN ABBREVIATIONS

API	Application Programming Interface
BDI	Beck Depression Inventory
CSV	Comma Separated Values
DMP	Data Management Plan
DoA	Description of the Action
EC	European Commission
EDF	European Data Format
GData	Generic Data
HDF5	Hierarchical Data Format 5
HRA	Health Research Authority
JSON	JavaScript Object Notation
MoCA	Montreal Cognitive Assessment
NMSQuest	Parkinson's Disease Non Motor Symptoms Questionnaire
PDSS	Parkinson's Disease Sleep Scale
PGS	Personalised Game Suite
RBD-SQ	REM Sleep Behaviour Questionnaire
SData	Specific Data
UPDRS	Unified Parkinson Disease Rating Scale
XML	Extensible Markup Language

1 EXECUTIVE SUMMARY

This deliverable is the initial version of the Data Management Plan (DMP) of the i-PROGNOSIS project, in accordance to the regulations of the Pilot action on Open Access to Research Data of the Horizon 2020 programme (H2020). It contains provisional information about the data that will be produced and collected within the project, whether and how it will be made accessible for re-use and further exploitation, and how it will be curated and preserved.

Based on the guidelines on Data Management in Horizon 2020[1], a dataset description template was initially drafted to provide the main pillar for the dataset descriptions (see Appendix I). All relevant datasets were recognized and a detailed list was produced, based on the datasets that have been described within the DoA to be produced within the life span of the project. The present deliverable consolidates all the partners' feedback and provision for the datasets they contribute to.

The project at its present stage is foreseen to develop a series of datasets, related to issues ranging from user requirements to the intervention and sensor captured data stemming from patients with PD and also healthy participants. Specifically, datasets are planned to be collected in two ways: the development data collection and the data that will be collected during the deployment of i-PROGNOSIS system. The former dataset will help the development and improvement of the algorithms and systems of the i-PROGNOSIS, while the latter will constitute the actual datasets that are foreseen to be the main input of the i-PROGNOSIS system.

Given that the majority of the i-PROGNOSIS datasets involve data collection from human participants, the respective data produced either raw or processed, should be carefully handled, under thorough consideration of ethical and privacy issues involved in such datasets. For all the identified i-PROGNOSIS datasets, specific parts that can be made publicly available have been identified in the current first version of the project's DMP. The public datasets of the i-PROGNOSIS project will become available through a common repository that will be formulated on the basis of the i-PROGNOSIS "data management portal" based on theZenodo repository service for open datasets[2].

This is an initial version of the Data Management Plan of the i-PROGNOSIS system. Having said that, the datasets described at this stage, represent an early reflection on the data that we foresee to be collected. During the evolution of the project, we expect that there will some changes either to the content of the datasets or the information classification. However, main principles - as described within this deliverable - is expected to remain intact until the end of the project, thus forming the main strategic axes of the overall Data Management Plan.

The overall DMP will be delivered at the end of the i-PROGNOSIS project (M48) within the deliverable "D5.7 Report on open research data management".

[1] Guidelines on Data Management in Horizon 2020
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] Zenodo https://zenodo.org/

2 INTRODUCTION

During the lifetime of i-PROGNOSIS, data of different nature will be generated and collected. These data are user-related and thus require a clear plan on how they are to be managed, i.e., stored, accessed, protected against unauthorized or improper use, etc. Thus, the main goals of i-PROGNOSIS Data Management Plan (DMP) are:

1. Outline of the types of data foreseen for generation at this stage of the project, including the context and procedures of this generation, as well as the degree of privacy and confidentiality of the data.
2. Outline of the protocols that will be followed to assess the generated/collected data with respect to their sensitiveness.
3. Outline of the data acquisition plan for the duration of the project.
4. Outline of the measures that are foreseen for the adequate management of the data from the ethical and security points of view.

The remainder of the deliverable is structured as follows. In Section 3, the common data collection procedures are outlined. A short description about specifications of the data management portal is provided, in addition to different types of classified information. Finally, Section 4 elaborates on each dataset, based on the provided template of Appendix I.

3 PRINCIPLES

3.1 PARTICIPATION IN THE PILOT ON OPEN RESEARCH DATA

The Data Management Plan will guide all activities regarding the anonymisation, exchange and release of data gathered during the project, as required for participating in the Open Research Data Pilot of the H2020 framework. Datasets that are to be produced within the i-PROGNOSIS project span from users' demographic and clinical data to sensor data, intervention data and outcome measures of several interventions. These data will allow the research community to benchmark algorithms with respect to several aspects of Parkinson's disease (PD) detection and treatment, providing a common basis for augmented policy decision making. Since i-PROGNOSIS data collection phases involve human participants, data collected will contain sensitive, personal information, and, as a result, the focus is also placed on possible ethical issues and access restrictions regarding personal data, so that no regulations on sensitive information are violated (see also D1.2 "Ethics and safety manual").

In this scope, this version of the deliverable describes in detail the datasets that are to be collected in each data collection phase, by each different technical or medical partner. Individual data collection mechanisms are treated as individual cases and therefore individual data management plans are foreseen, including the following information - as described in the available data management plan template and guidelines of the European Commission (EC)[1]: i) dataset reference and name, ii) dataset description, iii) standards and metadata, iv) data sharing, v) archiving and preservation (including storage and backup).

As part of the Data Management Plan, the i-PROGNOSIS consortium aims at developing a central database station, where data and evaluation outcomes of all relevant data collection phases will be deposited. Furthermore, a central Data Management Portal will be developed, in order to serve for the main access point to the publicly available datasets. The i-PROGNOSIS Data Management Portal will be based on open-source data portal platforms, such as Zenodo, allowing for each partner to publish its datasets to the research community.

3.2 THE i-PROGNOSIS DATA MANAGEMENT PORTAL

3.2.1 Publicly available service

i-PROGNOSIS will publish a subset of the data within the Zenodo repository service. Zenodo builds and operates a simple and innovative service that enables researchers, scientists, EU projects and institutions to share, preserve and showcase multidisciplinary research results (data and publications) that are not part of the existing institutional or subject-based repositories of the research communities.

Zenodo enables researchers, scientists, EU projects and institutions to:

- Easily share the long tail of small research results in a wide variety of formats including text, spreadsheets, audio, video, and images across all fields of science.
- Display their research results and receive credit by making the research results citable and integrating them into existing reporting lines to funding agencies like the EU.
- Easily access and reuse shared research results.

Some of the features provided by the Zenodo service are:

- An open digital repository for everyone and everything not served by a dedicated service; the so-called "long tail" of research results.
- A modern look and feel in line with current trends in state-of-the-art online services.
- Integration with OpenAIRE infrastructure and assured inclusion in OpenAIRE corpus.
- Easy upload and semi-automatic metadata completion by communication with existing online services such as DropBox for upload, Mendeley/ORCID/CrossRef/OpenAIRE for upload and pre-filling metadata.
- Easy access to research results via an innovative viewing option, open APIs, integration with existing online services, and the preservation of community independent data formats.
- A safe and trusted service by combining community-based curation with short- and long-term archival and digital preservation strategies in accordance with best practices.
- Persistent identifiers, Digital Object Identifiers (DOIs), for sharing research results.

- Service hosting according to industry best practices in CERN's professional data centres.
- An easy way to link research results with other results and products, funding sources, institutions, and licenses.
- Supports Dublin Core, MARC and MARCXML for metadata exporting.
- Complies with OAI-PMH for data dissemination.

3.3 COMMON PROCEDURES

3.3.1 Data Collection

3.3.1.1 Development Data Collection

The newly introduced development data acquisition period is the first and the earliest collection period and it is highly correlated with software development and implementation. The existence of non-artificial data is mandatory for the effective development of software-related, and not limited, to minor tremor or dysphonia detection. Moreover, the need for data is further augmented due to the fact that only limited publicly available corpora exist, and the existing ones incorporate different data capturing methods than i-PROGNOSIS, constituting them not suitable for our needs.

Furthermore, it is crucial for software implementation to capture and have access to raw data (e.g., unprocessed speech signals). The above is one of the main differences with the rest of the data collection periods, where sensitive data will be manipulated with proper methods.

Goals of the development data collection period:

- ❖ Aid the implementation of the processing algorithms (machine learning approaches require data during their training phase).
- ❖ Evaluate the effectiveness of various data descriptors. The latter is essential in order to successfully analyse, understand and translate the raw signals to meaningful information, as well as, to initially train the machine learning algorithms.

Development datasets will only be publicly available in the form of processed data. Raw signal data will remain confidential and will be available within the consortium, only after the data collectors provide their approval to share the data.

3.3.1.2 Data to be collected during the deployment of i-PROGNOSIS

The goal of the data collection at this stage of the project is to build the predictive algorithm for PD detection and evaluate it against the medical golden standard. In addition, the efficacy of the interventions regimen will be measured by medical partners, in terms of the specified medical evaluation protocol (see D2.2 "Data collection and medical evaluation protocol").

Data collected during the deployment of i-PROGNOSIS will be fully anonymised and will be offered to third parties in a processed format. Opening data during this stage of the project in a raw format will not be an option, due to privacy issues. Clinical data

will be used to label the processed data, thus forming the required "semantic ground truth" for other researchers to apply their own machine learning techniques and statistical analyses and to compare their findings.

3.3.2 Data Access Procedures

3.3.2.1 Public

For the sections of the dataset that will be made publicly available, a respective Web page will be created on the i-PROGNOSIS site, that will provide a description of the dataset, along with a download url. External researchers will have unrestricted access, following the steps as indicated by the Zenodo service. In addition, any third party stakeholder will be explicitly informed about the publication that need to be cited if part of the dataset has been used for publications. All public datasets should be strictly anonymised.

3.3.2.2 Protected

Data denoted as protected can be shared out of the consortium, as long as interested parties *a priori* request access from the consortium, explaining how these datasets will be used, e.g., research or commercial purposes. Appropriate forms will be created and be available through the i-PROGNOSIS website, from where interested parties can use them and request for explicit access to certain datasets. Upon the approval by the i-PROGNOSIS partners, interested parties will be provided with credentials, in order to download the requested datasets.

3.3.2.3 Confidential/Private

The private part of the datasets will be stored at a specific and designated private space of the partner responsible for the dataset, namely the data collector, on dedicated hard disk drives, to which only members of the data collector, whose work directly relates to these data will have access. In order for other i-PROGNOSIS partners to obtain access to these data, each partner must provide a proper written request to the data collector, including a justification over the need of access to the particular data. Once deemed necessary, the data collector will provide the respective data to the partner.

4 i-PROGNOSIS DATASETS

4.1 DATASETS NAMING

Concerning the convention followed for naming the project datasets, it should be noted that the name of each dataset comprises:

1. A prefix "**DS**" indicating a dataset.
2. Its unique identification number depending on the dataset category (see next Section), e.g., "**DS1**" for datasets belonging to the category of personal and clinical data, "**DS2**" for sensed and captured GData or SData, "**DS3**" for interventions data and "**DS4**" for requirements data.
3. Since sensed and captured GData or SData or interventions data contain different types of data, such as physiological recordings, audio, and visual

features, further distinction takes place as follows: "**DS2.1**" for physiological features, "**DS2.2**" for visual features, "**DS3.1**" for serious games' metrics, etc.

4. A short name indicative of its content and purpose.

For example, body and gesture-related features collected during the interventions data collection phase form a dataset named:

Prefix	*Category identification number*	*Type of data discriminator*	*Indicative Name*
"**DS**"+	"**3.**"+	"**1-**"+	"**BodyAndGesture**"+

i.e., "**DS3.1-BodyAndGesture**"

4.2 SUMMARY OF THE I-PROGNOSIS DATASETS

The i-PROGNOSIS datasets are divided in the following categories:

1. Personal and clinical data.
2. Sensed and captured GData or SData.
3. Interventions data.
4. Requirements data.

Personal and clinical data will accompany and annotate Sensed and captured GData or SData and Interventions data, as essential metadata, thus allowing for the semantic annotation of the collected datasets.

TABLE 1 presents cumulatively all datasets that are initially planned to be collected within the i-PROGNOSIS project with respect to each aforementioned category.

TABLE 1 Indicative Datasets to be collected within the i-PROGNOSIS project

Dataset #	Dataset Name	Description
PERSONAL & CLINICAL DATA		
1.1	**DS1.1-ElectronicHealthRecordData**	It contains information related to medical evaluation protocols, such as health screening results and interventions' outcome measures.
GDATA / SDATA		
2.1	**DS2.1-VoiceQualityAnalysis**	It contains vocal features extracted from recordings captured using the smartphone microphone.
2.2	**DS2.2-PhotosFacialAnalysis**	It contains facial features extracted from photos taken with the mobile phone camera (selfies for masked face etc.).
2.3	**DS2.3-ActivityAnalysis**	It contains features derived from data collected by the inertial sensors on

		the smartwatch regarding the physical activity of the wearer over time.
2.4	**DS2.4-PhysioSignalAnalysis**	It contains features derived from the physiological data sources acquired using the Smart Belt, namely bowel sounds and the newly introduced Electrogastography (EGG) sensor.
2.5	**DS2.5-TypingPatternAnalysis**	It contains keystroke dynamics-related features collected during typing on a virtual keyboard of a touch screen-enabled smartphone.
2.6	**DS2.6-ExploratoryWalkabilityAnalysis**	It contains de-personalised location information and derived features from GPS, Wi-Fi, Mobile Network and IMU originating from the users' smartphone and smartwatch.
2.7	**DS2.7-TextSentimentAnalysis**	It contains SMS text and/or tweets collected from the users' smartphone along with associated sentiment classification
2.8	**DS2.8-FoodIntakeAnalysis**	The dataset is composed of objective quantification of meal mechanics as well as derivative information about the user's eating behavioural elements (duration of meal, number of bites, eating rate and eating rate changes across the meal).
2.9	**DS2.9-BowelSoundsAnalysis**	It contains pre-processed bowel sound signals captured from the smart belt.
2.10	**DS.10-TremorAnalysis**	It contains IMU based (accelerometer, gyroscope and magnetometer) derived features, originating from the user's smartphone and smartwatch.
INTERVENTIONS		
3.1	**DS3.1-SeriousGamesMetrics**	It considers in-game metrics and performance such as scores, achievements, difficulty level, etc. of all the gaming sessions (exergames, dietary, handwriting, emotion and voice games).
3.2	**DS3.2-BodyGestureAnalysis**	The dataset consists of postures and gestures tracking experiments, as well as balance and gait features, especially during the Exergames where Kinect will be the main sensor.

3.3	**DS3.3-SleepStageAnalysis**	It comprises sleep stage data based on accelerometer, heart rate, and skin temperature measurements of users, captured by the smartwatch during their sleep (and provided they are wearing the smartwatch on their wrist).
3.4	**DS3.4-GaitAnalysis**	It contains accelerometer, gyroscope and pedometer data captured from the smart watch/band with respect to rhythm guidance in terms of sound cues
3.5	**DS3.5-VoiceEnhancementAnalysis**	It contains voice data captured using the smartphone microphone.
REQUIREMENTS		
4.1	**DS4.1-FocusGroupsDataset**	It contains data gathered within focus groups during the requirements elicitation phase.
4.2	**DS4.2-WebSurveyDataset**	It contains the questions and the qualitative answers to the questions of the i-PROGNOSIS Web survey, conducted within the context of the identification of user requirements and system specifications, from ~2000 anonymous survey participants.

4.3 DATASETS BREAKDOWN

In the following, datasets are described based on the template of Appendix I.

4.3.1 Personal & Clinical Data

DATA SET REFERENCE NAME	**DS1.1-ElectronicHealthRecordData**
DATA SET DESCRIPTION	
Generic description	
This dataset will contain the i-PROGNOSIS users' personal information, demographics and health related data (assessment batteries' scores, comorbidities, medication). This dataset will facilitate subsequent analysis with respect to the wide GData and SData collection, as well as to clinical assessment tests towards research on the stealth assessment and screening of early signs of PD. In addition, results of the medical evaluation as represented within these datasets, will allow to evaluate the effectiveness of the user participation in the interventions (i.e., their involvement in the Serious Games).	

Origin of data
This dataset a) will be collected by i-PROGNOSIS App, and b) will be entered into the i-PROGNOSIS system by the clinical sites. The data collected through the mobile App will be taken during the first use of the mobile phone app; the user will be asked to provide short personal information, demographics and health related data by multiple choice questions. During the SData collection and intervention phases, the relevant clinical data for each user will be entered into the system from the recruiting clinical site (AUTH, KCL or TUD). In case of participants who will follow all the projects' phases, multiple instances of follow up assessment tests' scores will be available.
Nature and scale of data
The dataset will be collected during all phases of the project: the GData/SData collection and the intervention phases. The questions delivered through the i-PROGNOSIS App will be designed in a multiple choice format, so that the obtained information will be in a numeric format. Health related data at this stage will include: 1. Diagnosis of Parkinson`s disease: yes / no, 2. Physical handicap: yes / no, 3. Family history of Parkinson`s disease: yes / no. The process of clinical assessment will take place in three countries (UK, Germany, and Greece), where health-centric partners are based, in order to facilitate the medical monitoring of the users and the subsequent clinical validation of PD risk. The clinical data will include all the relevant medical information collected by the recruiting clinical centres. Some of the clinical examinations include among others: Unified Parkinson Disease Rating Scale (UPDRS), Parkinson's Disease Non Motor Symptoms Questionnaire (NMSQuest), Montreal Cognitive Assessment (MoCA), Parkinson's Disease Sleep Scale (PDSS), REM Sleep Behaviour Questionnaire (RBD-SQ), Beck Depression Inventory (BDI), Parkinson Fatigue Scale, Senior fitness test and BERG balance scale. During the initial pilot phase more than 5000 older adults (above 55 years of age) is expected to participate, whereas some 80 and 60 will take part in the SData phase and the intervention phases. The data will be stored as records of a profile database, part of the i-PROGNOSIS repository.
To whom the dataset could be useful
The dataset will be used to classify and rate the obtained sensed and captured further GData and SData, intervention data and requirements data in respect to age, gender and health state dependent values. In addition, clinical data will be used for ground truth purposes, to train the machine learning algorithms responsible for the second stage early PD detection. Also, part of this dataset will be each intervention user's baseline and follow up data, for the evaluation of their health state progress through the use of the serious games Personalised Game Suite.
Related scientific publication(s)

This kind of dataset is collected in every type of clinical trial in the field of PD or otherwise. Example: Martinez-Martin et al. EuroInf: A Multicenter Comparative Observational Study of Apomorphine and Levodopa Infusion in Parkinson's Disease. Mov Disord. 2015 Apr;30(4):510-6.
Indicative existing similar data sets (including possibilities for integration and reuse)
Not applicable within the application context of the novel technologies used in i-PROGNOSIS.
STANDARDS AND METADATA
Clinical data will not contain any metadata, instead the name of the examination and the measurement score will be provided. There would be a possibility to support electronic health data exchange, via partial implementation of the CEN/ISO 13606 standard as declared the base standard of the European Interoperability Framework for Health.
DATA SHARING
Access type
Personal identification as stated in the consent forms will be kept separate from any research and health-related data, which will be pseudonymised (only initials and date of birth will be kept). Furthermore no personal data that could identify an individual person e.g., name, date of birth is obtained. The data for personal identification of the users will be kept locally at each recruitment clinical centre, outside this dataset. This dataset will only include personal and clinical data relevant to the i-PROGNOSIS protocols. Due to ethical reasons, only averaged group data could become publicly available, while individual data will be private to serve the i-PROGNOSIS R&D objectives.
Access Procedures
For the parts of the dataset that will be made publicly available, a respective Web page will be created that will provide a description of the dataset and links to the data management portal. The private part of this dataset will be stored at a specifically designated private space of clinical partners, in dedicated hard disk drives, on which only members of the clinical research teams will have access.
Embargo periods (if any)
The applicable datasets will be publicly available 2 years after the end of the project to allow the consortium prepare and submit the scientific publications.
Technical mechanisms for dissemination
For the public part of the dataset, a link will be provided from the i-PROGNOSIS site to the Zenodo service. The link will be provided in all relevant i-PROGNOSIS publications. A technical publication describing the dataset and acquisition procedure will be published.
Necessary S/W and other tools for enabling re-use

The dataset will be in a numeric manner and therefore designed to allow easy reuse with commonly available tools.
Repository where data will be stored
The public part of this dataset will be accommodated within the Zenodo service., hosted by CERN infrastructure.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The dataset will be preserved as long as there are regular downloads. After that it would be made accessible by request and preserved by AUTH at least until the end of the project. Locally, the personal and clinical data will be preserved based on the national and departmental practises about scientific data handling.
Approximated end volume of data
Some MBs.
Indicative associated costs for data archiving and preservation
There are no costs associated with data preservation in institutional servers or the Zenodo service.
Indicative plan for covering the above costs
There are no relevant costs.
PARTNERS ACTIVITIES AND RESPONSIBILITIES
Partner Owner / Data Collector — **TUD, KCL, AUTH, MICROSOFT**
Partner in charge of the data analysis — **Both clinical and technical partners**
Partner in charge of the data storage — **TUD, KCL, AUTH, MICROSOFT**
WPs and Tasks
The data are going to be collected within activities of WP4, WP6 and WP7, to mainly serve the research efforts of T4.1 - T4.5, T6.1 - T6.5, T7.1, T7.3 and T7.4.

4.3.2 Sensed and Captured GData/SData

DATA SET REFERENCE NAME	**DS2.1-VoiceQualityAnalysis**
DATA SET DESCRIPTION	
Generic description	
Speech dataset for the detection and identification of voice-related PD symptoms for early PD detection. The database contains voice features extracted from speech data collected from phone calls recorded from the i-PROGNOSIS smartphone dialler	

application. The data set includes extracted voice features and several annotations about the speech data and the underlying speakers.
Origin of data
The dataset is collected by recording conversations and subsequent extraction of voice features using the i-PROGNOSIS smartphone dialler application during GData and SData collection phase.
Nature and scale of data
The goal of this dataset is to enable the detection and identification of voice-related PD symptoms for early PD detection. The dataset will contains voice features extracted from the speech signals recorded by the i-PROGNOSIS smartphone application during the GData and SData collection phase. Annotations including date and time of the recordings along with speaker information (e.g., ID) will provide the ground truth and will facilitate the subsequent analysis of the speech recordings. The voice features are extracted from raw speech data (16 kHz sampling rate, wav unencoded, 16 bit depth) and are stored in JSON format. Data Format: JSON. The dataset is in order of 50 kB per recording.
To whom the dataset could be useful
The dataset will be used within the project for the development and evaluation of automatic signal processing and pattern recognition algorithms for the extraction of discriminative features for the detection and identification of voice-related PD symptoms for the early detection of PD. This data will be used in WP3 and WP6, supporting the early PD symptoms detection and for the overall assessment of the i-PROGNOSIS system. This dataset may also be useful in the future for other researchers who want to explore voice-related PD symptoms.
Related scientific publication(s)
This database will build the foundation of our research and development of algorithms towards the identification and detection of voice-related PD symptoms for early PD detection. We plan to propose our findings on ICASSP, Interspeech or other voice and biomedical-related conferences.
Indicative existing similar data sets (including possibilities for integration and reuse)
To the best of our knowledge no long-term speech database covering voice recordings of healthy and PD subjects is (publicly) available.
STANDARDS AND METADATA
The dataset will be accompanied with detailed documentation of its contents. Indicative metadata include: (a) description of the experimental setup and procedure that led to the generation of the dataset, and (b) documentation of the variables recorded in the dataset.
DATA SHARING
Access type

The extracted features and anonymized annotations of the collected datasets could become publicly available, while the rest of them will be private to serve the i-PROGNOSIS R&D objectives. The inclusion of a subject's data in the public part of this dataset will be done on the basis of appropriate informed consent to data publication.
Access Procedures
For the parts of the dataset that will be made publicly available, a respective web page will provide a description of the dataset and links to the data management portal. The private part of this dataset will be stored at a specifically designated private space of FRAUNHOFER, in dedicated hard disk drives, on which only members of the FRAUNHOFER research team will have access.
Embargo periods (if any)
The applicable datasets will be publicly available two years after the end of the project to allow the consortium prepare and submit the scientific publications.
Technical mechanisms for dissemination
For the public part of the dataset, a link will be provided from the i-PROGNOSIS site to the Zenodo service. The link will be provided in all relevant i-PROGNOSIS publications. A technical publication describing the dataset and acquisition procedure will be published.
Necessary S/W and other tools for enabling re-use
The dataset is designed to allow easy reuse and access with commonly available tools (e.g., Matlab, Python, GVIM) and software libraries (e.g., Tensorflow, C++ STL), because the data is stored primarily in a common file format (JSON).
Repository where data will be stored
The public data will be hosted within the Zenodo service that will serve the needs of the Data Management Portal of the i-PROGNOSIS project.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The public part of the dataset will be preserved as long as there are regular downloads. After that it would be made accessible by request and preserved by AUTH at least until the end of the project. The private part of the dataset will be preserved by FRAUNHOFER at least until the end of the project.
Approximated end volume of data
The dataset is expected to consume several GB depending on the size of the extracted features for each recording and the length and quantity of the speech signals (e.g., single channel audio waveform 16 bit, 44.1 kHz is of size 5.3 MB/min, single channel mp3 192 Kbps is of size 1.44 MB/min).analysed recordings.
Indicative associated costs for data archiving and preservation

There are no costs associated with data preservation in institutional servers or the Zenodo service.
Indicative plan for covering the above costs
There are no relevant costs.
PARTNERS ACTIVITIES AND RESPONSIBILITIES

Partner Owner / Data Collector	**TUD, KCL, AUTH, KI, FRAUNHOFER**
Partner in charge of the data analysis	**FRAUNHOFER**
Partner in charge of the data storage	**FRAUNHOFER, MICROSOFT**

WPs and Tasks
The data is collected within activities of WP3 and WP6, to mainly serve the research efforts of T3.3, T6.1, T6.2 and T6.3.

DATA SET REFERENCE NAME	**DS2.2-FacialAnalysis**
DATA SET DESCRIPTION	
Generic description	
Facial image dataset for the detection and identification of masked face symptoms for early PD detection. The database will contain selfie images collected from the frontal camera of the smartphone of the user. The data set will include annotations about a) the bounding box of the recognized face of the user, and b) self-assessment of the emotional state of the user.	
Origin of data	
The dataset will be collected from the user smartphone by collecting the (selfie) images obtained during GData and SData collection phases and detecting/recognizing the face of the user.	
Nature and scale of data	
The goal of this dataset is to enable the detection and identification of masked face symptoms for early PD detection. The dataset will contain the images (selfie) images obtained during GData and SData collection phases. Annotations including date and time of the recordings along with bounding box of the identified user face in the selfie will provide the ground truth and will facilitate the subsequent face expression analysis task. The dataset will contain JPEG selfie images and XML (or text) files for the annotations. Data Format: JPEG files for selfie images from the frontal camera of the smartphone, XML or plain text files for annotations/metadata.	
To whom the dataset could be useful	
The dataset will be used within the project for the development and evaluation of automatic facial expression recognition algorithms which will then be used to detect	

masked face symptoms for the early detection of PD. This data will be used in WP3 and WP6, supporting the early PD symptoms detection and for the overall assessment of the i-PROGNOSIS system. This dataset may also be useful in the future for other researchers who want to explore masked face PD symptoms.
Related scientific publication(s)
This database will build the foundation of our research and development of algorithms towards the masked face PD symptoms for early PD detection. We plan to propose our findings on computer vision/image processing- and biomedical-related conferences and journals (e.g., ICIP conference).
Indicative existing similar data sets (including possibilities for integration and reuse)
To the best of our knowledge no image database covering face expressions of healthy and PD subjects is (publicly) available.
STANDARDS AND METADATA
The dataset will be accompanied with detailed documentation of its contents. Indicative metadata include: a) description of the procedures that led to the generation of the dataset, and b) documentation of the contents of the dataset.
DATA SHARING
Access type
Due to ethical reasons, only the data captured by a subset of the patients, during the GData and SData collection phases, as well as normal healthy control subjects could become publicly available, while the rest of them will be private to serve the i-PROGNOSIS R&D objectives. The inclusion of a (normal healthy control) subject's data in the public part of this dataset will be done on the basis of appropriate informed consent to data publication.
Access Procedures
For the portions of the dataset that will be made publicly available, a respective web page will provide a description of the dataset and links to the data management portal. The private part of this dataset will be stored at a specifically designated private space of CERTH, in dedicated hard disk drives, on which only members of the CERTH research team will have access.
Embargo periods (if any)
The applicable datasets will be publicly available two years after the end of the project to allow the consortium prepare and submit the scientific publications.
Technical mechanisms for dissemination
For the public part of the dataset, a link will be provided from the i-PROGNOSIS site to the Zenodo service. The link will be provided in all relevant i-PROGNOSIS publications. A technical publication describing the dataset and its annotations will be produced.
Necessary S/W and other tools for enabling re-use

The dataset will be designed to allow easy reuse and access with commonly available tools and software libraries.
Repository where data will be stored
The public part of this dataset will be accommodated within the Zenodo service, hosted by CERN infrastructure.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The public part of the dataset will be preserved on the Data Management Portal of the i-PROGNOSIS project at least until the end of the project. The private part of the dataset will be preserved by CERTH at least until the end of the project.
Approximated end volume of data
Each selfie image in JPEG format has size 1-2 MBs, so a dataset of e.g., 1000 images is expected to consume up to 2 GB.
Indicative associated costs for data archiving and preservation
There are no costs associated with data preservation in institutional servers or the Zenodo service.
Indicative plan for covering the above costs
There are no relevant costs.

PARTNERS ACTIVITIES AND RESPONSIBILITIES	
Partner Owner / Data Collector	**CERTH**
Partner in charge of the data analysis	**CERTH**
Partner in charge of the data storage	**CERTH, AUTH**
WPs and Tasks	
The data is going to be collected within activities of WP3 and WP6, to mainly serve the research efforts of T3.4, T6.1 and T6.2.	

DATA SET REFERENCE NAME	**DS2.3-ActivityAnalysis**
DATA SET DESCRIPTION	
Generic description	
This dataset will include raw sensor data, such as accelerometer, gyroscope, heart rate and barometer, as well as pedometer, estimated calories, distance travelled, motion type (walking, jogging or running), heart rate, altimeter and barometer when available to identify steps climbing. The origin of the data will be from a smart watch/band regarding activity analysis. However, the list of sensors and derived feature is subject to change, depending on the final smart watch device.	

Origin of data
The dataset will be captured by the accelerometer, gyroscope, pedometer, heart rate and skin temperature sensors of the smart watch that will be worn by the participants during all the i-PROGNOSIS data collection phases. Extracted features (e.g. estimated calories, distance travelled and motion type) would need to be estimated using the available sensors. Depending on the smart watch device that will be finally selected, some of the sensors might not be available. For development purposes we are currently using the Huawei Watch 2, which includes 6-axis A + G sensor, 3-axis Compass, Heart Rate Sensor (PPG), Barometer, Capacitive Sensor, and Ambient Light Sensor.
Nature and scale of data
The first part of the dataset is going to be collected during the development data collection phase where the subjects will follow specific activity scenarios. The second part of the dataset will be collected mainly during the SData and Interventions phases, where smart watches/bands will be provided to 80 and 60 participants, respectively. Moreover, the GData phase will contribute to the dataset, however, the contribution is expected to be limited since the smart watch/band is optional at GData. At these phases, i.e., GData, SData and Interventions, the subjects will not perform specific activity tasks, instead, the data will be captured throughout their daily routine activities. The dataset will most probably, include: i) the measures (double precision) of the acceleration force in g units (9,81 m/s^2) that is applied to the smart watch/band on all three physical axes (x, y, and z), ii) the measures (double precision) of the smart watch/band's rotation in rad/s around each of the three physical axes, and iii) the absolute number of steps taken, the steps ascended and descended (based on the smart watch's altimeter), the number of kilocalories (kcals) burned, the distance travelled (in cm) along with the speed (cm/s), the pace (ms/m) and the motion type (walking, jogging, running, etc.) and the heart rate (beats/min). The minimum sampling frequency for each sensor will be based on the outcome of the analysis of the development data (where the sampling frequency will be the highest possible). Data format: TXT or CSV file. The dataset will be in the order of ~5 MB per participant per hour of usage. Usually the smart watch/band owners use it for at least 12 hours per day.
To whom the dataset could be useful
The collected data will be used for the development and evaluation of the activity analysis of the i-PROGNOSIS project. The dataset could be useful in analysing the activity levels and patterns over the course of the day. Moreover, it may be possible to identify symptoms and signs related to Parkinson's Disease, such as tremor.
Related scientific publication(s)
The dataset will accompany the research results in the field of activity analysis through a smart watch/band of people with PD. At least one publication is intended to be made in the IEEE Biomedical Engineering journal (or similar).

The following publication will be taken into account:

Sharma, Vinod, et al. "SPARK: personalized Parkinson disease interventions through synergy between a smartphone and a smartwatch." International Conference of Design, User Experience, and Usability. Springer International Publishing, 2014.

Indicative existing similar data sets (including possibilities for integration and reuse)

To the best of our knowledge, there is no available smart watch/band-based activity dataset for PD-related. Related datasets include data captured from smartwatches (some of them along with datasets originating from smartphones). For example:

Crowdsignals.io (http://crowdsignals.io). Crowdsignals creates the largest set of rich, longitudinal mobile and sensor data recorded from smartphones and smartwatches available to the community.

Dbworld (http://permalink.gmane.org/gmane.comp.db.dbworld/54394): Dbworld dataset contains free dataset for downloads originating from smartphone+smartwatch mobile, sensor, and human activity. It is part of the CrowdSignals.io data collection campaign.

Heterogeneity Activity Recognition Dataset (https://archive.ics.uci.edu/ml/datasets/Heterogeneity+Activity+Recognition). The Heterogeneity Human Activity Recognition (HHAR) dataset from smartphones and smartwatches is a dataset devised to benchmark human activity recognition algorithms (classification, automatic data segmentation, sensor fusion, feature extraction, etc.) in real-world contexts; specifically, the dataset is gathered with a variety of different device models and use-scenarios, in order to reflect sensing heterogeneities to be expected in real deployments.

STANDARDS AND METADATA

The dataset will be accompanied with detailed documentation of its contents. Indicative metadata include: a) description of the experimental setup and procedure that led to the generation of the dataset, b) health status of the subject, and c) documentation of the variables recorded in the dataset.

DATA SHARING

Access type

Due to ethical issues, only part of the dataset will be publicly available. More specifically, the data that corresponds to a subset of the PD patients, as well as the healthy control subjects, captured at the development phase of the i-PROGNOSIS project will become publicly available. The rest of the data will be private to serve the i-PROGNOSIS R&D objectives. The inclusion of a subject's data in the public part of this dataset will be done on the basis of appropriate informed consent to data publication.

Access Procedures

The access procedures for the publicly available sub-dataset and the private sub-dataset that will serve the project's R&D objectives are described in Sections 3.3.2.1 and Section 3.3.2.3 respectively.

Embargo periods (if any)

The applicable datasets will be publicly available 2 years after the end of the project to allow the consortium to prepare and submit the scientific publications.
Technical mechanisms for dissemination
For the public part of the dataset, a respective link will be provided from i-PROGNOSIS site to the Zenodo service. The link will be provided in all relevant i-PROGNOSIS publications. A technical publication describing the dataset and acquisition procedure will be published.
Necessary S/W and other tools for enabling re-use
The dataset will be designed to allow easy reuse with commonly available tools and software libraries, such as Excel, Matlab, .NET etc.
Repository where data will be stored
The dataset will be accommodated at the data management portal of the project Website, hosted by CERN infrastructure.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The public part of this dataset will be preserved online for as long as there are regular downloads. After that it would be made accessible by request. The private part of the dataset will be preserved by AUTH at least until the end of the project.
Approximated end volume of data
The dataset is expected to be several Gigabytes, provided that each participant is expected to produce at least 20MB per day and the duration is at least 22 months (SData and Interventions phase).
Indicative associated costs for data archiving and preservation
There are no costs associated with data preservation in institutional servers or the Zenodo service.
Indicative plan for covering the above costs
There are no relevant costs.
PARTNERS ACTIVITIES AND RESPONSIBILITIES

Partner Owner / Data Collector	**AUTH, PLUX**
Partner in charge of the data analysis	**AUTH**
Partner in charge of the data storage	**AUTH, PLUX**

WPs and Tasks
The data are going to be collected within the activities of WP3, WP4 and WP6, to mainly serve the research efforts of T3.2, T4.3, T6.1, T6.2, T6.3 and T6.4.

DATA SET REFERENCE NAME	**DS2.4-PhysioSignalAnalysis**

DATA SET DESCRIPTION
Generic description
This dataset will contain features derived from physiological data acquisition (e.g., bowel sounds frequency), as well as raw data. Whenever suitable, annotations will be associated with the data to provide ground truth and facilitate the subsequent analysis of pre-recorded data. Annotations may be generated by the users in real time and recorded simultaneously with the biosignal data (e.g., by means of a manual trigger) or produced by a human expert upon revision of pre-recorded data.
Origin of data
The dataset will be collected using the selected smart watch device and also the Smart Belt. On the smart watch device currently used for development purposes we can collect 3-axis accelerometer, 3-axis accelerometer gyroscope data, 3-axis Compass, Heart Rate Sensor (PPG), Barometer, Capacitive Sensor, and Ambient Light Sensor. From the Smart Belt we can collect raw multi-channel bowel sound data and also Electrogastography (EGG) data.
Nature and scale of data
The datasets will be collected in guided and controlled testing scenarios, producing SData, and include data acquired from interaction between the elder and everyday living sensorial artefacts. The goal is to identify changes in the cardiac and motion (i.e. tremor) patterns, which may relate to changes in health status, behaviours or other aspects relevant to PD. The dataset will contain raw data, as well as temporal, spectral and non-linear features extracted from the raw time series. Examples of such features are the bowel sound frequency and related spectral indicators, EGG statistics, instant heart rate, inter-beat intervals, heart rate histogram, amongst others. Data Format: Comma-Separated Values (CSV), Hierarchical Data Format 5 (HDF5) or ASCII Text Files (TXT) The dataset is expected to include individualized records per user per device usage session, possibly segmented in files of duration compatible with easily manageable post-processing (e.g., 1 hour segments).
To whom the dataset could be useful
The collected data will be used within the project for the development and evaluation of data mining and signal processing algorithms that incorporate cardiac and motion data in the identification of early indicators of PD. Physiological data can also be useful to follow-up the interventions, and also to assess the progress and status of the disease (if possible). This data will primarily feed WP4 and WP7, supporting the design of the interventions and to the overall assessment of the i-PROGNOSIS system. Together with the annotations, it may also be useful in the future for other researchers and practitioners working not only in PD but in other medical specialties as well.
Related scientific publication(s)

The dataset will contribute to extend our research results in the field of physiological data sensing to the specific case of people with PD. We envision part of the dataset to include data collected from healthy controls. Recent publications from the project team concerning multimodal physiological datasets include:

H. P. da Silva, C. Carreiras, A. Lourenço, A. Fred, R. C. das Neves, and R. Ferreira, "Off-the-person electrocardiography: Performance assessment and clinical correlation," Health and Technology, vol. 4, no. 4, pp. 309–318, 2015. http://link.springer.com/article/10.1007/s12553-015-0098-y

H. P. da Silva, A. Lourenço, A. Fred, N. Raposo, and M. A. de Sousa, "Check Your Biosignals Here: A new dataset for off-the-person ECG biometrics," Computer Methods and Programs in Biomedicine, vol. 113, no. 2, pp. 503–514, 2014. http://www.sciencedirect.com/science/article/pii/S0169260713003891

H. Gamboa, H. P. da Silva, and A. Fred, "HiMotion: a new research resource for the study of behaviour, cognition, and emotion," Multimedia Tools and Applications, vol. 73, no. 1, pp. 345–375, 2014 http://link.springer.com/article/10.1007%2Fs11042-013-1602-x

Indicative existing similar data sets (including possibilities for integration and reuse)

To the best of our knowledge there are no datasets available (at least publicly) that simultaneously target the specific case of PD and contain the physiological data sources envisioned by the project.

STANDARDS AND METADATA

The dataset will be accompanied with detailed documentation of its contents. Indicative metadata include: a) description of the experimental setup and procedure that led to the generation of the dataset, b) type of activity performed c) documentation of the variables recorded in the dataset, d) manual annotations provided by the subjects, and e) manual annotations provided by human experts.

DATA SHARING

Access type

Due to ethical reasons, only the raw data, captured by normal healthy control subjects, (during the development data collection phase) as well as a subset of the extracted features of the collected datasets could become publicly available, while the rest of them will be private to serve the i-PROGNOSIS R&D objectives. The inclusion of a (normal healthy control) subject's data in the public part of this dataset will be done on the basis of appropriate informed consent to data publication.

Access Procedures

The recorded data will be primarily stored as CSV, HDF5 or TXT files containing the raw data streams and the features derived from the data. The files will hosted on the private i-PROGNOSIS central database that will serve the needs of the Data Management Portal of the project, or in a suitable analogous digital space, with protected access reserved only to the relevant members of the i-PROGNOSIS project team or people for which the consortium partners decide that access to the data is of relevant interest to the execution of the project. Only anonymised data will be

provided, unless otherwise deemed necessary for the adequate pursuit of the project goals.
Embargo periods (if any)
The applicable datasets will be available two years after the end of the project to allow the consortium the preparation and submission of scientific publications.
Technical mechanisms for dissemination
A technical publication describing the dataset and acquisition procedure will be published is expected. Dissemination will be mostly performed through post-processed data and result analysis by means of relevant i-PROGNOSIS publications.
Necessary S/W and other tools for enabling re-use
The dataset will be designed to allow easy reuse and access with commonly available tools and software libraries.
Repository where data will be stored
The public data will be hosted within the Zenodo service that will serve the needs of the Data Management Portal of the i-PROGNOSIS project.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The private part of the dataset will be preserved on the Data Management Portal of the i-PROGNOSIS project at least until the end of the project.
Approximated end volume of data
The dataset is expected to be several hundreds of MB, considering that the devices can produce up to 5KB of data per second.
Indicative associated costs for data archiving and preservation
There are no costs associated with data preservation in institutional servers or the Zenodo service.
Indicative plan for covering the above costs
There are no relevant costs.
PARTNERS ACTIVITIES AND RESPONSIBILITIES

Partner Owner / Data Collector	**AUTH, PLUX, MICROSOFT**
Partner in charge of the data analysis	**AUTH, PLUX**
Partner in charge of the data storage	**AUTH, PLUX, MICROSOFT**

WPs and Tasks
The data are going to be collected within activities of WP4, WP6 and WP7, to mainly serve the research efforts of T4.1, T4.5, T6.1, T6.4, T7.3 and T7.4.

DATA SET REFERENCE NAME	**DS2.5-TypingPatternAnalysis**

DATA SET DESCRIPTION
Generic description
This dataset includes typing patterns in the form of keystroke dynamics and pressure-related features, extracted during the users' typing on a virtual keyboard (of a touch screen-enabled smartphone) so as to be used as indicators towards the formulation of early PD detections tests. Features will be accompanied by metadata in order to facilitate the analysis within the i-PROGNOSIS project, as well as, by other researchers outside of the project.
Origin of data
The dataset is collected as part of the development of the i-PROGNOSIS first stage of early PD detection. It comprises a development sub-dataset and corresponding metadata collected by a small number of users (healthy controls and PD patients), in Thessaloniki, Greece, as well as, a deployment sub-dataset of ecologically-valid data and corresponding metadata collected by a larger number of users (healthy controls and PD patients, in the range of hundreds), in Greece, Germany and Portugal. Both data collection procedures have received ethical approval.
Nature and scale of data
The dataset arises from the interaction of users with the virtual keyboard of a touch screen-enabled smartphone, i.e., when they are typing on the keyboard: 1) The development sub-dataset was collected while development users [19 early PD patients (62±8.5 years of age, 26% females, 16.9±7.8 UPDRS Part III score) and 13 healthy controls (56±6 years of age, 46% females, 0.0±0.0 UPDRS Part III score)] were typing specific short (ranging from 45 to115 characters) text excerpts (up to 10) in a controlled environment using a specific smartphone (LG Nexus 5X). The dataset comprises the timestamps of the press and release actions of each key [in milliseconds (ms)], providing the ability to extract keystroke dynamics-related features such as the hold time, press latency, flight time andrelease latency, all measured in milliseconds (ms), as well as the normalised pressure applied for each key tap (in the range of 0-1), for each typing session (text excerpt). 2) The deployment sub-dataset comprises the same features as the development sub-dataset and additional typing metadata such as flags for denoting if a key press was a deliberate long press, whether sound or vibration key feedback were active or not during typing, the number of "Delete" key presses as an indicator of the number of errors, and start/end date-time of typing session. It is collected through the i-PROGNOSIS custom virtual keyboard that accompanies the "iPrognosis" Android application that is available in Greece, Germany and Portugal. As a result, this dataset is formulated by a significantly larger number of users' typing on their smartphone keyboard. In the case of the deployment dataset, keystroke dynamics are collected for each typing session, i.e., starting when the keyboard becomes active and finishing when the keyboard is suppressed. *The characters underlying each key pressed are not recorded* in order to comply with personal data privacy and protection regulations. The deployment sub-dataset scale cannot be defined because it grows continuously as the number of users downloading the iPrognosis

application and using the custom keyboard increases. Until June 2017, 7,829 records of typing sessions have been collected from 45 users of the iPrognosis application.

Data Format: For the development sub-dataset, a TXT file was generated for each typing session, including the values of features. For the deployment sub-dataset a structured JSON string is generated for each typing session, including values of features and typing metadata, as well as the coded Id of the user.

To whom the dataset could be useful

The development sub-dataset is currently used to initialise the i-PROGNOSIS machine learning algorithms as far as the typing pattern inference is concerned, that will be further trained by the deployment sub-dataset.

The deployment sub-dataset will be used for training the i-PROGNOSIS machine learning algorithms in order for the i-PROGNOSIS investigators to examine whether it is useful to employ this type of information towards the realisation of early PD detection tests based on the users' interaction with their everyday digital devices, such as the smartphone.

Moreover, the dataset could be useful for biomedical researchers that are interested in studying the relationship between the interfacing (requiring motor skills) of people with digital devices and psychomotor impairments, such as PD. The later constitutes a relatively new research field. Adequate use of these data presupposes at least basic background regarding PD symptomatology, as well as, basic knowledge in data analysis methodology and experience in the use of statistical software packages.

Related scientific publication(s)

The development sub-dataset will accompany the research results regarding the feasibility of early PD detection tests based on users' interaction with everyday digital devices. Research results are planned to be published initially in the indicative journals and/or conferences provided below:

- IEEE Transactions on Biomedical Engineering (Journal)
- Nature Scientific Reports (Journal)
- Human Computer Interaction International (Conference)

The respective research field is an emerging one. Recent publications have been made:

Giancardo, L., Sánchez-Ferro, A., Butterworth, I., Mendoza, C. S., & Hooker, J. M. (2015). Psychomotor impairment detection via finger interactions with a computer keyboard during natural typing. *Scientific reports, 5*, article no: 9678, doi:10.1038/srep09678

Giancardo, L., Sanchez-Ferro, A., Arroyo-Gallego, T., Butterworth, I., Mendoza, C. S., Montero, P., ... & Estépar, R. S. J. (2016). Computer keyboard interaction as an indicator of early Parkinson's disease. *Scientific reports, 6*, article no: 34468, doi: 10.1038/srep34468

Arroyo-Gallego, T., Ledesma-Carbayo, M. J., Sanchez-Ferro, A., Butterworth, I., Sanchez-Mendoza, C., Matarazzo, M., ... & Giancardo, L. (2017). Detection of motor impairment in Parkinson's disease via mobile touchscreen typing. *IEEE Transactions on Biomedical Engineering*. doi: 10.1109/TBME.2017.2664802

Indicative existing similar data sets (including possibilities for integration and reuse)

A similar dataset has been made available by Giancardo et al. (2015), linked to the publication:

Giancardo, L., Sánchez-Ferro, A., Butterworth, I., Mendoza, C. S., & Hooker, J. M. (2015). Psychomotor impairment detection via finger interactions with a computer keyboard during natural typing. Scientific reports, 5, article no: 9678, doi:10.1038/srep09678

The dataset includes keystroke dynamics and it is part of the supplementary information accompanying the publication and can be found through:

http://www.nature.com/article-assets/npg/srep/2015/150409/srep09678/extref/srep09678-s1.pdf

i-PROGNOSIS investigators plan to use the respective experiment protocol in order to configure the framework for the collection and analysis of the development sub-dataset.

STANDARDS AND METADATA

The development sub-dataset is accompanied by a detailed description of its contents. Metadata include: 1) description of the experiment set-up, the procedure that led to the generation of the dataset, and the ethical approval protocol and 2) an Excel file (XLSX) containing a) coded Id of participant and her/his b) age, c) gender, d) level of education completed, e) years of smartphone use, f) frequency of typing using mobile phones, g) UPDRS Part III score, and h) Hoehn & Yahr stage of PD, for all participants.

The deployment sub-dataset will be annotated using a separate JSON entry including the following user metadata: a) coded Id, b) age, c) gender, d) level of education completed, e) years of smartphone use and f) health status against PD (healthy, healthy with PD history in the family, patient).

DATA SHARING

Access type

The development sub-dataset will be publicly available (see Section 3.3.2.1), accompanied by the respective ethics approval protocol. Due to ethical compliance, the deployment sub-dataset will be confidential and only i-PROGNOSIS partners will be able to access it after following the appropriate procedure (see Section 3.3.2.3)

Access Procedures

The confidential deployment sub-dataset will be handled according the framework reported in Section 3.3.2.3. The publicly available development sub-dataset will be open for third-party stakeholders to download based on the procedure described in Section 3.3.2.1.

Embargo periods (if any)
The development sub-dataset is planned to be publicly available after month 24 of the project to allow i-PROGNOSIS investigators to prepare and submit the respective scientific publications, but allow also for third-party researchers to conduct further analysis and comment on the i-PROGNOSIS results before the end of the project.
Technical mechanisms for dissemination
The development sub-dataset will be available through the i-PROGNOSIS data management portal (see Section 3.2). Links redirecting to the portal and the available dataset will be provided through a dedicated page of the i-PROGNOSIS project website (www.i-prognosis.eu). A technical description providing information on the experiment protocol through which the development sub-dataset was captured will accompany the development sub-dataset (see also STANDARDS AND METADATA).
Necessary S/W and other tools for enabling re-use
As the publicly available compressed dataset (development sub-dataset) will comprise standard TXT and XLSX (Excel) files, no specialised software or tools will be required for the dataset to be parsed and reused, other than a (de-)compression software and a basic text editor (minimum requirement to access the content of the TXT file).
Repository where data will be stored
The development sub-dataset is stored in AUTH secure servers. The deployment sub-datasets will be stored in Microsoft Azure-based i-PROGNOSIS Cloud infrastructure (data centres are located in Ireland), but access will be restricted (Section 3.3.2.3).
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The public part of this dataset will be preserved online for as long as there are regular downloads and at least one year after the end of the project. After that it would be made accessible by direct request to the owner-data collector. The confidential part of the dataset will be preserved by the owners indefinitely and at least one year after the end of the project.
Approximated end volume of data
The development sub-dataset size is 4 MB (including metadata). The deployment dataset is expected to be approximately 40 GB (Gigabytes), based on the expected number of the iPrognosis application users (in the range of thousands), an average of 20 typing sessions per day for at least 12 months and an average of 5 KB (Kilobytes) size of JSON entry per typing session.
Indicative associated costs for data archiving and preservation
The development sub-dataset is stored in AUTH servers. The deployment sub-dataset, stored currently on the Microsoft Azure-based i-PROGNOSIS Cloud

infrastructure, will be archived and preserved in AUTH servers, after the end of the project. There are no costs associated with the latter means of archiving and preservation.
Indicative plan for covering the above costs
There are no relevant costs.

PARTNERS ACTIVITIES AND RESPONSIBILITIES	
Partner Owner / Data Collector	**AUTH, MICROSOFT**
Partner in charge of the data analysis	**AUTH**
Partner in charge of the data storage	**AUTH, MICROSOFT**

WPs and Tasks
The development sub-dataset will be collected within WP3 and WP6, to mainly serve the research efforts of T3.2 and T3.7,

DATA SET REFERENCE NAME	**DS2.6-ExploratoryWalkabilityAnalysis**

DATA SET DESCRIPTION
Generic description
Dataset for building an anonymous and depersonalised user sociability profile, based on information (derived features) originating by location sensors/modules (e.g. Wi-Fi, GPS).
Origin of data
The derived information will originate from data by the smartphone and smartwatch embedded location and IMU (GPS, GSM, Wi-Fi and accelerometer) sensors and modules during the SData collection period.
Nature and scale of data
The collected data will be captured while the user carries her/his smartphone or wears her/his smartwatch while commuting. More specifically, the dataset will contain derived de-personalised features based on the fusion of multiple sensors and modules such as: Wi-Fi, GSM, GPS and IMU from the user's smartphone and GPS and IMU from the user's smartwatch device. Since anonymisation is of outmost importance when dealing with location information, all derived features will be free of any absolute world coordinates or any way to derive them by processing. Data format: JSON/XML for data representation as well as for any relevant annotations.

It is expected that the dataset, will be composed of an equal number of PD patients and healthy subjects; however, the number of subjects is heavily related with the overall user participation. The data volume per subject is anticipated to be in the order of ~10-15 MB per subject.
To whom the dataset could be useful
The collected data will be used for the development and evaluation of the location and the physical activity service (i.e. background services, see `D2.1` for additional information) as well as the module responsible for building the user's behavioural profile. Furthermore, the dataset will be of particular interest to the researchers and health professionals that are willing to explore the underlying information contained in location data regarding the sociability of potential PD patients.
Related scientific publication(s)
A paper describing the dataset is expected to be published. Furthermore, a paper describing the approach of de-personalising location information and fusing different location data sources is also target for publication. More information on the latter can be found in D8.3 (Dissemination plan; "A paper presenting an approach for generating user profiles based on de-personalized location information").
Indicative existing similar data sets (including possibilities for integration and reuse)
To this day, publicly available location based PD datasets are non-existent.
STANDARDS AND METADATA
The dataset will be accompanied by a complete documentation containing: a. A description of the overall procedure regarding the collection of the data. b. A brief definition of how the derived location-based features were generated. c. A brief definition of the de-personalisation processes. d. An annotation file, indicating if the subject (indicated solely by a `subject ID`) is a diagnosed PD patient or a healthy individual. If deemed informative or necessary, additional non-identifiable information may be included in this section.
DATA SHARING
Access type
Since all identifiable information, like correlation to absolute world coordinates, will be stripped clean off of the dataset, access type can be classified as publicly available (i.e. can be publicly shared) with an initial embargo period.
Access Procedures

Dataset will be uploaded to the zenodo portal, along with a brief description of the dataset as well as the publication(s) that need to be cited in case any part of the dataset is used.
Embargo periods
The applicable dataset will be publicly available two years after the end of the project to allow the consortium prepare and submit the scientific publications.
Technical mechanisms for dissemination
Relevant publications will include a link to the Zenodo portal. A complete technical publication describing in-depth the dataset as well as the data acquisition procedures will be published.
Necessary S/W and other tools for enabling re-use
A typical JSON/XML (e.g., libxerces, libJSON, JSON-java) parsing library is the only requirement for accessing the content of the dataset. Such libraries are typical in most (if not all) programming languages.
Repository where data will be stored
The entirety of the dataset will be accommodated at the data management portal of the project's website, hosted by Microsoft Azure-based i-PROGNOSIS Cloud infrastructure.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The dataset will be preserved online for as long as there are regular downloads. After that it would be made accessible by request.
Approximated end volume of data
Depending on the magnitude of user participation, the end volume of data is expected to be in the range of ~15-25 GB.
Indicative associated costs for data archiving and preservation
There are no costs associated with data preservation in institutional servers or the Zenodo service.
Indicative plan for covering the above costs
There are no relevant costs.
PARTNERS ACTIVITIES AND RESPONSIBILITIES

Partner Owner / Data Collector	**AUTH**
Partner in charge of the data analysis	**AUTH**
Partner in charge of the data storage	**AUTH**
WPs and Tasks	

The entire "ExploratoryWalkabilityAnalysis" dataset will be collected during WP3, to mainly serve the research efforts of T3.2, T3.7 and T4.4.

DATA SET REFERENCE NAME	DS2.7-TextSentimentAnalysis
DATA SET DESCRIPTION	
Generic description	
SMS/Tweet(Twitter message) datasets in different languages will be collected from the users' smartphones. These messages will be manually annotated by a value indicating the sentiment classification (at minimum three classes will be used: positive, negative, neutral or n/a).	
Origin of data	
The dataset will be collected from i-Prognosis users' smartphones SMS or Twitter messages (User consent will be provided).	
Nature and scale of data	
The goal of this dataset is to enable the detection of depression symptoms for early PD detection. Data Format: plain text files of SMS/Tweets with annotations/metadata.	
To whom the dataset could be useful	
The dataset will be used within the project for the development and evaluation of depression recognition algorithms as a symptom for the early detection of PD. This data will be used in WP3 and WP6, supporting the early PD symptoms detection and for the overall assessment of the i-PROGNOSIS system. This dataset may also be useful in the future for other researchers who want to explore depression symptoms from short messages.	
Related scientific publication(s)	
This database will build the foundation of our research and development of algorithms towards detecting depression symptoms from short messages for early PD detection. We plan to publish our findings on natural language processing-related conferences and journals.	
Indicative existing similar data sets (including possibilities for integration and reuse)	
Many manually annotated databases are already available for SMS/twitter sentiment analysis. To name a few: - Stanford Twitter Corpus: http://help.sentiment140.com/for-students - HCR and OMD datasets: https://bitbucket.org/speriosu/updown - Sentiment Strength Corpora: http://sentistrength.wlv.ac.uk/ - Sanders: http://www.sananalytics.com/lab/twitter-sentiment/	

- SemEval: http://www.cs.york.ac.uk/semeval-2013/task2/
STANDARDS AND METADATA
SMS/tweet messages are ASCII text with a length of 160/140 characters respectively. The available metadata will be the available sentiment labels(classes).
DATA SHARING
Access type
Due to ethical reasons, only the SMS/Tweets collected by a subset of the patients during the initial phases by normal healthy control subjects could become publicly available, while the rest of them will be private to serve the i-PROGNOSIS R&D objectives.
Access Procedures
For the portions of the dataset that will be made publicly available, a respective web page will provide a description of the dataset and links to the data management portal. The private part of this dataset will be stored at a specifically designated private space of CERTH, in dedicated hard disk drives, on which only members of the CERTH research team will have access.
Embargo periods (if any)
The applicable datasets will be publicly available two years after the end of the project to allow the consortium prepare and submit the scientific publications.
Technical mechanisms for dissemination
For the public part of the dataset, a link will be provided from the i-PROGNOSIS site to the Zenodo service. The link will be provided in all relevant i-PROGNOSIS publications. A technical publication describing the dataset and its annotations will be produced.
Necessary S/W and other tools for enabling re-use
No specialised software or tools will be required for the dataset to be parsed and reused, other than a basic text editor.
Repository where data will be stored
The public data will be hosted within the Zenodo service that will serve the needs of the Data Management Portal of the i-PROGNOSIS project, hosted by CERN infrastructure.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The public part of the dataset will be preserved on the Data Management Portal of the i-PROGNOSIS project at least until the end of the project. The private part of the dataset will be preserved by CERTH at least until the end of the project.
Approximated end volume of data

Each text/SMS message is less than 500 bytes (assuming utf-8 encoding for tweets), so a corpus of 10000 SMS/Tweets is expected to consume up to 5 MB (per language).
Indicative associated costs for data archiving and preservation
The private part of the dataset will be stored on a dedicated hard drive, with no costs for its preservation. Data publicly archived and preserved within the Zenodo service for free.
Indicative plan for covering the above costs
There are no relevant costs.

PARTNERS ACTIVITIES AND RESPONSIBILITIES	
Partner Owner / Data Collector	**CERTH**
Partner in charge of the data analysis	**CERTH**
Partner in charge of the data storage	**CERTH, AUTH**
WPs and Tasks	
The data is going to be collected within activities of WP3 and WP6, to mainly serve the research efforts of T3.4, T6.1 and T6.2.	

DATA SET REFERENCE NAME	**DS2.8-FoodIntakeAnalysis**
DATA SET DESCRIPTION	
Generic description	
Dataset for the objective quantification of meal mechanics, based on the real-time recording of the meal progression, extrapolated from raw IMU signals (collected through smartwatch) data. This dataset will also include derivative information about the specific values of the user's eating behavioural elements, e.g.: meal duration, number of bites, eating rate and eating rate changes across the meal.	
Origin of data	
The dataset will be produced by use of smartwatches during the SData collection phase. It is possible that the dataset will be complemented by data collected during the intervention phase.	
Nature and scale of data	
It is expected that the datasets will be collected mainly during the SData phase, and potentially during the intervention phase. The users will be asked to have the smartwatch activated during their main meals (breakfast, lunch, dinner) across a pre-set number of days (e.g., 5 times/week). The smartwatch will be supposed to be used throughout the duration of a main meal.	

The dataset will contain the raw smartwatch IMU data generated by the hand movements during the meal, the raw accelerometer data will be used to extract meal behavioural indicators (e.g., the meal duration, the recorded number of bites, the total eating rate and the modelled eating rate changes across the meal). Potentially, the dataset will also include averaged values per user across multiple recorded meals. Additionally, the dataset will contain around-the-meal self-rated subjective information concerning the meal, e.g., perceived fullness before/after, food taste evaluation, etc. Data Format: CSV files for all the described measures. The dataset will be on the order of 5-10 MB per recorded meal. The size will be significantly higher if the food pictures are also included in the dataset.
To whom the dataset could be useful
The dataset will facilitate evaluation of the progress of each individual and their comparison with healthy, age-matched control populations. This comparison can be valuable for health professionals who care for patients with Parkinson patients, the patients themselves and researchers who want to explore the underlying shifts that happen in eating behaviour due to Parkinson's disease.
Related scientific publication(s)
The dataset will complement our research and clinical results in the field of eating behaviour quantification. A subset of the investigated population, participating in the SData phase, without having diagnosed Parkinson disease will be the reference control population. Corresponding publications describing and validating different components of the described methodology are: Ioakimidis I, Modjtaba Z, Eriksson-Marklund L, Bergh C, Grigoriadis A, & Södersten P. (2011). "Description of Chewing and Food Intake over the Course of a Meal." Physiology & Behavior,104 (5,: 761–769. Papapanagiotou V, Diou C, Langlet B, Ioakimidis I, & Delopoulos A. (2015). "A parametric Probabilistic Context-Free Grammar for food intake analysis based on continuous meal weight measurements." Conf Proc IEEE Eng Med Biol Soc.,7853-7856. Kilintzis V, Maramis C, Maglaveras N. Wrist sensors—An application to acquire sensory data from Android Wear® smartwatches for connected health. InBiomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on 2017 Feb 16 (pp. 125-128). IEEE. Dong Y, Hoover A, Scisco J, Muth E. A new method for measuring meal intake in humans via automated wrist motion tracking. Applied psychophysiology and biofeedback. 2012 Sep 1;37(3):205-15. Sen S, Subbaraju V, Misra A, Balan RK, Lee Y. The case for smartwatch-based diet monitoring. InPervasive Computing and Communication Workshops (PerCom Workshops), 2015 IEEE International Conference on 2015 Mar 23 (pp. 585-590). IEEE.
Indicative existing similar data sets (including possibilities for integration and reuse)

Relevant dataset for measurement of food intake analysis by use of a smartwatch can be found here: https://mug.ee.auth.gr/intake-cycle-detection/ - In this dataset, a total of 10 subjects were recorded by video while eating their launch at the university's cafeteria with s smartwatch (Microsoft band 2). The mean duration of the recordings was 13.2 minutes. The Microsoft Band 2 contains a triaxial accelerometer that provides measurements in g units and gyroscope providing measurements in degrees per second (degrees/sec).
STANDARDS AND METADATA
The dataset will be accompanied with detailed documentation of its contents. Indicative metadata include: a) description of the experimental setup and procedure that led to the generation of the dataset, b) type of consumed food (probably identified through self-reporting), c) documentation of the variables recorded in the dataset, and d) the relative "positioning" of an individual in the corresponding population distribution.
DATA SHARING
Access type
Due to ethical reasons, only the data captured by a subset of the patients and normal healthy control subjects, could become publicly available, while the rest of them will be private to serve the i-PROGNOSIS R&D objectives. The inclusion of a (normal healthy control) subject's data in the public part of this dataset will be done on the basis of appropriate informed consent to data publication. It will be investigated further whether the complementary food pictures can also become publicly available.
Access Procedures
For the portions of the dataset that will be made publicly available, a respective web page will provide a description of the dataset. The private part of this dataset will be stored at a specifically designated private space of KI, in encrypted hard disk drives, on which only members of the KI research team, whose work directly relates to these data will have access. For further i-PROGNOSIS partners to obtain access to these data, they should provide a proper request to the KI primarily responsible, including a justification over the need to have access to these data. Once deemed necessary, KI will provide the respective data portions to the partner.
Embargo periods (if any)
The applicable datasets will be publicly available 2 years after the end of the project to allow the consortium prepare and submit the scientific publications.
Technical mechanisms for dissemination
For the public part of the dataset, a link to this will be provided from i-PROGNOSIS site to the Zenodo service. The link will be provided in all relevant i-PROGNOSIS publications. A technical publication describing the dataset and acquisition procedure will be published.
Necessary S/W and other tools for enabling re-use
A typical XML parsing library is the only requirement for accessing the content of the dataset. Such libraries are typical in most (if not all) programming languages.

Repository where data will be stored	
The public part of this dataset will be accommodated at the data management portal of the project website, hosted within the Zenodo infrastructure.	
ARCHIVING AND PRESERVATION (including storage and backup)	
Data preservation period	
The public part of this dataset will be preserved online for as long as there are regular downloads. After that it would be made accessible by request. The private part of the dataset will be preserved by KI at least until 10 years after the publication of the scientific results.	
Approximated end volume of data	
The dataset will be on the order of 5-10 MB per recorded meal. A first estimation of the dataset, including food pictures would be in the range from ~2-7 GB.	
Indicative associated costs for data archiving and preservation	
Probably parts from three hard disk drives in the KI lab (primary, backup 1 and backup 2) will be allocated for the dataset. There are no costs associated with its preservation. Data publicly archived and preserved within the Zenodo service for free.	
Indicative plan for covering the above costs	
There are no relevant costs.	
PARTNERS ACTIVITIES AND RESPONSIBILITIES	
Partner Owner / Data Collector	**KI**
Partner in charge of the data analysis	**KI**
Partner in charge of the data storage	**KI**
WPs and Tasks	
The data are going to be collected within activities of WP3, WP4 and WP6, to mainly serve the research efforts of T3.2, T3.7, T4.1, T4.5, T6.1 and T6.4.	

DATA SET REFERENCE NAME	**DS2.9-BowelSoundsAnalysis**
DATA SET DESCRIPTION	
Generic description	
This dataset will include bowel sound data captured from a smart belt towards bowel immobility related to increased constipation detection, since constipation is one of the earliest non-motor symptoms in PD patients.	
Origin of data	

The dataset will be captured by a custom-made smart belt that will carry distributed microphones covering the abdomen area and will be worn either directly on the skin or above clothes.
Nature and scale of data
The first part of the dataset is expected to be collected during the development data collection phase where the subjects (healthy control and PD patients) will wear the smart belt for at least 1 hour in laboratory environment. The second part of the dataset will be collected mainly during the SData and Interventions phases, where smart belts will be given to 80 (potential PD patients and healthy controls) and 60 (identified PD patients) subjects, respectively. The recording will take place at each subject's indoor/outdoor environment. Since the bowel sounds, according to the literature, exist in the range of 100 – 1000/1500 Hz, the raw signal will be filtered (high-pass filtering at 80 Hz in order to eliminate the influence of cardiac and pulmonary sounds). Data format: TXT or CSV file Considering three recording channels, an average usage of 4 hours per day and 3 kHz sampling frequency, the dataset is estimated to be on the order of ~500 MB per subject per day.
To whom the dataset could be useful
The collected data will be used for the development and evaluation of constipation detection algorithms (through sound-based intestinal motility assessment). Moreover, the dataset may be useful to identify other gastrointestinal disorders, such as obstructions, ascites, infections and trauma. The fact that there are not any available datasets of this kind, makes this dataset valuable to the research community, especially to the gastroenterologists.
Related scientific publication(s)
The dataset will accompany the research results in the field of bowel sounds analysis for constipation detection of people with PD. One publication is intended to be made in the World Journal of Gastroenterology (or similar).
Indicative existing similar data sets (including possibilities for integration and reuse)
To the best of our knowledge, there is no available bowel sounds dataset. To this end, the dataset that will be collected during i-PROGNOSIS project is expected to have major impact.
STANDARDS AND METADATA
The dataset will be accompanied with detailed documentation of its contents. Indicative metadata include: a) description of the experimental setup and procedure that led to the generation of the dataset, b) health status of the subject, and c) documentation of the variables recorded in the dataset.
DATA SHARING
Access type

Due to ethical issues, only part of the dataset will be publicly available. More specifically, the data that correspond to a subset of the PD patients as well as the healthy control subjects, captured at the initial phase of the i-PROGNOSIS project will become publicly available. The rest of the data will be private to serve the i-PROGNOSIS R&D objectives. The inclusion of a subject's data in the public part of this dataset will be done on the basis of appropriate informed consent to data publication.
Access Procedures
The access procedures for the publicly available sub-dataset and the private sub-dataset that will serve the project's R&D objectives are described in Sections 3.3.2.1 and Section 3.3.2.3 respectively.
Embargo periods (if any)
The applicable datasets will be publicly available 2 years after the end of the project to allow the consortium prepare and submit the scientific publications.
Technical mechanisms for dissemination
For the public part of the dataset, a link to this will be provided from i-PROGNOSIS site to the Zenodo service. The link will be provided in all relevant i-PROGNOSIS publications. A technical publication describing the dataset and acquisition procedure will be published.
Necessary S/W and other tools for enabling re-use
The dataset will be designed to allow easy reuse with commonly available tools and software libraries, such as Excel, Matlab, .NET etc.
Repository where data will be stored
The public part of this dataset will be accommodated within the Zenodo service, hosted by CERN infrastructure.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The public part of this dataset will be preserved online for as long as there are regular downloads. After that it would be made accessible by request. The private part of the dataset will be preserved by AUTH at least until the end of the project.
Approximated end volume of data
The dataset is expected to be several Terabytes, provided that each participant is expected to produce at least 500MB per day and the duration is at least 22 months (SData and Interventions phase).
Indicative associated costs for data archiving and preservation
Probably four dedicated hard disk drives will be allocated for the private dataset. There are no costs associated with its preservation. Data publicly archived and preserved within the Zenodo service for free.
Indicative plan for covering the above costs

There are no relevant costs.	
PARTNERS ACTIVITIES AND RESPONSIBILITIES	
Partner Owner / Data Collector	**PLUX, AUTH**
Partner in charge of the data analysis	**PLUX, AUTH**
Partner in charge of the data storage	**PLUX, AUTH**
WPs and Tasks	
The data are going to be collected within the activities of WP3 and WP6, to mainly serve the research efforts of T3.6, T6.1, T6.3 and T6.4.	

DATA SET REFERENCE NAME	**DS2.10-TremorAnalysis**
DATA SET DESCRIPTION	
Generic description	
Dataset for detecting the presence of tremor in the upper limbs. The dataset will include information from triaxialtriaxial acceleration, orientation as well as magnetic data streams.	
Origin of data	
The dataset will consist of two parts. The first part will originate from the development collection period of i-PROGNOSIS (small population), whereas the second from the G and SData collection periods (large population).	
Nature and scale of data	
The collected data will be captured during the user's interaction with the smartphone/smartwatch devices. More specifically, smartphone generated data will be collected while the user performs certain actions while holding the device (e.g., during a voice call). On the other hand, data originating from the smartwatch device will be captured while the user performs everyday activities while wearing the smartwatch. In addition, the development part of the corpus will contain 9-dimensional smartphone sensor information (3D accelerometer, gyroscope and magnetometer) and 6-dimensional (3D accelerometer and gyroscope) for the smartwatch. Regarding the G/SData part of the corpus, only 3D acceleration data will be provided. This stems from the fact that the accelerometer sensor in contrast to the gyroscope is embedded in every smartphone designed in the last five years. Data format: JSON/XML for data representation as well as for any relevant annotations. The data volume per subject is anticipated to be in the order of ~5-10 MB per subject.	
To whom the dataset could be useful	

The collected data will be used for the development and evaluation of the minor tremor detection module (i.e. `handling service`, see D2.1 for more information) of the i-PROGNOSIS smartphone detection application. Furthermore, the dataset will be of particular interest to the researchers and health professionals that are willing to explore the underlying information contained in IMU data regarding PD tremor.
Related scientific publication(s)
A paper describing the dataset is expected to be published. Furthermore, a paper describing the approach for detecting PD minor tremor is also target for publication. More information on the latter can be found in D8.3 (Dissemination plan; "A paper presenting a method for detecting the presence of minor-tremor, based on IMU data collected by the smartphone/smartwatch").
Indicative existing similar data sets (including possibilities for integration and reuse)
To this day, publicly available IMU based PD tremor datasets are non-existent.
STANDARDS AND METADATA
The dataset will be accompanied by a complete documentation containing: a) a description of the overall procedure regarding the collection of the data, b) a brief definition of how the derived IMU-based features were generated, and c) an annotation file, indicating if the subject (indicated solely by a subject ID) is a diagnosed PD patient or a healthy individual. If deemed informative or necessary, additional non-identifiable information may be included in this section.
DATA SHARING
Access type
Since all identifiable information will be stripped clean off of the dataset, access type can be classified as publicly available (i.e., can be publicly shared) with an initial embargo period.
Access Procedures
Dataset will be uploaded to the zenodo portal, along with a brief description of the dataset as well as the publication(s) that need to be cited in case any part of the dataset is used.
Embargo periods
The applicable dataset will be publicly available two years after the end of the project to allow the consortium prepare and submit the scientific publications.
Technical mechanisms for dissemination
Relevant publications will include a link to the zenodo portal. A complete technical publication describing in-depth the dataset as well as the data acquisition procedures will be published.
Necessary S/W and other tools for enabling re-use

A typical JSON/XML (e.g., libxerces, libJSON, JSON-java) parsing library is the only requirement for accessing the content of the dataset. Such libraries are typical in most (if not all) programming languages.
Repository where data will be stored
The entirety of the dataset will be accommodated at the data management portal of the project's website, hosted by Microsoft Azure-based i-PROGNOSIS Cloud infrastructure.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The dataset will be preserved online for as long as there are regular downloads. After that it would be made accessible by request.
Approximated end volume of data
Depending on the magnitude of user participation, the end volume of data is expected to be in the range of ~15-25 GB for the G/SData collection part of the corpus. The development part will not exceed 1 GB of data.
Indicative associated costs for data archiving and preservation
There are no costs associated with data preservation in institutional servers or the Zenodo service.
Indicative plan for covering the above costs
There are no relevant costs.

PARTNERS ACTIVITIES AND RESPONSIBILITIES	
Partner Owner / Data Collector	**AUTH**
Partner in charge of the data analysis	**AUTH**
Partner in charge of the data storage	**AUTH**
WPs and Tasks	
The entire dataset will be collected during WP3, to mainly serve the research efforts of T3.2, T3.7 and T4.4.	

4.3.3 Intervention Data

DATA SET REFERENCE NAME	**DS3.1-SeriousGames**
DATA SET DESCRIPTION	
Generic description	
This dataset will include in-game metrics and games' performance of all the sessions in a semantically annotated way (by adopting proper ontologies for each specific domain) so as to facilitate subsequently analysis with respect to the clinical assessment tests towards research on stealth assessment and screening of PD early	

signs and disease progress through the suite (linear correlation of in-game metrics with clinical assessment tests).
Origin of data
The dataset will be collected using the i-PROGNOSIS Personalised Game Suite (PGS) which will accommodate different types of serious gaming interventions.
Nature and scale of data
The datasets will be collected during the serious gaming interventions of i-PROGNOSIS. The serious games will ask the user to go under specific exercises/tasks or gaming tests while mechanisms in the background will track and collect the user's performance. For instance, in the case of Exergames, the user will be asked to perform specific exercises which will be captured by the Kinect sensor. The dataset will contain metrics like reaction time, player's path / optimum path, goal time, movement range, balance, min and max angles of movements, wrong choices and much more in-game metrics that will arise during the design requirements of the serious games. Data Format: RDF triples or JSON The dataset is expected to be composed of 180 records per serious gaming session.
To whom the dataset could be useful
The collected data will be used for the development and evaluation of the Personalized Game Suite in terms of usability, acceptance, effectiveness and user's assessment. The different parts of the (semantically annotated where applicable) dataset could be useful in the benchmarking of a series of serious games, focusing either on the effectiveness axis as the primary role of the interventions, as well as in detecting and assessing, if possible, the disease's progress and status. The latter will feed the T4.5 which is intended to dynamically recommend game adaptations for personalised and optimised use as well as keeping the users in the "flow zone" which represents the feeling of being complete and energized focus in an activity with a high level of enjoyment and fulfilment towards increased adherence. Finally, this dataset will be part of the overall evaluation of the i-PROGNOSIS contributing to the validation of the pilot applications and interventions.
Related scientific publication(s)
The dataset will accompany our research results in the field of human activity monitoring of people with Parkinson's. A subset of similar dataset, including recordings from elderly people without Parkinson's going through Exergames, will be used as initial input to this dataset. Corresponding publications are: Bamparopoulos, G., Konstantinidis, E., Bratsas, C., & Bamidis, P. D. (2016). Towards exergaming commons: composing the exergame ontology for publishing open game data. Journal of Biomedical Semantics, 7(1), Article nr 4. http://doi.org/10.1186/s13326-016-0046-4 Konstantinidis, E., Bamparopoulos, G., & Bamidis, P. (2016). Moving Real Exergaming Engines on the Web: The webFitForAll case study in an active and healthy ageing living lab environment. IEEE Journal of Biomedical and Health Informatics, 1–1. http://doi.org/10.1109/JBHI.2016.2559787

Indicative existing similar data sets (including possibilities for integration and reuse)
It should be noted that there are no serious games datasets available online. To the best of our knowledge, the only open dataset regarding serious games, and more specifically Exergames performed by elderly people, is the one that AUTH has published a couple of years before described in: Bamparopoulos, G., Konstantinidis, E., Bratsas, C., & Bamidis, P. D. (2016). Towards exergaming commons: composing the exergame ontology for publishing open game data. Journal of Biomedical Semantics, 7(1), Article nr 4. http://doi.org/10.1186/s13326-016-0046-4
STANDARDS AND METADATA
The dataset will be accompanied with detailed documentation of its contents. Indicative metadata include: a) description of the experimental setup and procedure that led to the generation of the dataset, b) type of exercise or game, c) documentation of the variables recorded in the dataset, and d) semantic annotation based on existing ontologies.
DATA SHARING
Access type
Due to ethical reasons, only the data captured by a subset of the patients during the initial phases by normal healthy control subjects could become publicly available, while the rest of them will be private to serve the i-PROGNOSIS R&D objectives. The inclusion of a (normal healthy control) subject's data in the public part of this dataset will be done on the basis of appropriate informed consent to data publication.
Access Procedures
An ontology that describes Exergames using the Web Ontology Language (OWL) is available at http://purl.org/net/exergame/ns#. The acquired game results will be automatically converted to RDF triples and published on the web as open data, accessible through a SPARQL Endpoint. The data will be accessible from a SPARQL endpoint that is available at http://www.fitforall.gr/sparql, where queries can be made using the GET or POST method. In order to facilitate access, links to a download section where the datasets will be downloaded as JSON files will be provided if required. The private part of this dataset will be stored at a specifically designated private space of AUTH, in dedicated hard disk drives, on which only members of the AUTH research team whose work directly relates to these data will have access. For further i-PROGNOSIS partners to obtain access to these data, they should provide a proper request to the AUTH primarily responsible, including a justification over the need to have access to these data. Once deemed necessary, AUTH will provide the respective data portions to the partner. Another option is to provide a public link to download the dataset through the i-PROGNOSIS data management portal.
Embargo periods (if any)

The applicable datasets will be publicly available two years after the end of the project to allow the consortium prepare and submit the scientific publications.
Technical mechanisms for dissemination
For the public part of the dataset, a link to this, as well as to the SPARQL endpoint http://www.fitforall.gr/sparql, will be provided from the Data management portal. The link and the SPARQL endpoint will be provided in all relevant i-PROGNOSIS publications. A technical publication describing the dataset and acquisition procedure will be published.
Necessary S/W and other tools for enabling re-use
The dataset will be designed to allow easy reuse with commonly available tools and software libraries.
Repository where data will be stored
The public part of this dataset will be accommodated at AUTH servers, accessible by the SPARQL endpoint http://www.fitforall.gr/sparql. In addition, a subset of these data will be hosted to the Microsoft Azure-based i-PROGNOSIS Cloud infrastructure.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The public part of this dataset will be preserved online for as long as there are regular downloads. After that it would be made accessible by request. The private part of the dataset will be preserved by AUTH at least until the end of the project.
Approximated end volume of data
The dataset is expected to be several hundreds of MB, provided that each session is expected to produce a volume of ~80 KB of data.
Indicative associated costs for data archiving and preservation
The dataset will be stored in the serious gaming server hosted by AUTH. There are no costs associated with its preservation. There are no costs associated with data preservation in institutional servers or the Zenodo service.
Indicative plan for covering the above costs
There are no relevant costs.
PARTNERS ACTIVITIES AND RESPONSIBILITIES

Partner Owner / Data Collector	**AUTH**
Partner in charge of the data analysis	**AUTH**
Partner in charge of the data storage	**AUTH**

WPs and Tasks
The data are going to be collected within activities of WP4, WP6 and WP7, to mainly serve the research efforts of T4.1, T4.5, T6.1, T6.4, T7.3 and T7.4.

DATA SET REFERENCE NAME	DS3.2-BodyAndGesture
DATA SET DESCRIPTION	
Generic description	
Dataset for human identification, postures and gestures tracking experiments, especially during the Exergames where Kinect will be the main sensor. The dataset is also planned to include balance and gait features.	
Origin of data	
The dataset will be collected using a Kinect2 sensor during the exergaming intervention.	
Nature and scale of data	
It is expected that the datasets will be collected mainly during the exergaming interventions. The Exergames will ask the user to go under specific exercises which will be captured by Kinect. The approximate duration of each session is expected to be close to 45min. The dataset will contain the user's silhouette as this is provided by the Kinect SKD (Skeleton with bones and joints). It will be investigated further whether the RGB and depth image will be collected. Data Format: PNG/JPG for images (both RGB and depth), JSON for the coordinates of the joints and bones, XML or TXT for annotations. The dataset will be on the order of ~2-5 GB per recording hour.	
To whom the dataset could be useful	
The collected data will be used for the development and evaluation of the human activity monitoring and the Exergames intervention of the i-PROGNOSIS project. The different parts of the dataset could be useful in the benchmarking of a series of human tracking methods, focusing either on human identification, on posture and gesture analysis and tracking as well as in detecting, if possible, symptoms and signs of those appear at people with Parkinson's.	
Related scientific publication(s)	
The dataset will accompany our research results in the field of human activity monitoring of people with Parkinson's. A subset of similar dataset, including recordings from elderly people without Parkinson's going through Exergames, will be used as initial input to this dataset. Corresponding publications are: Konstantinidis, E. I., Antoniou, P. E., Bamparopoulos, G., & Bamidis, P. D. (2014). A lightweight framework for transparent cross platform communication of controller data in ambient assisted living environments. Information Sciences, 300, 124–139. http://doi.org/10.1016/j.ins.2014.10.070 Konstantinidis, E. I., Billis, A. S., Bratsas, C., & Bamidis, P. D. (2016). Active and Healthy Ageing Big Dataset streaming on demand. In Proceedings of the 18th International Conference on Human-Computer Interaction. Toronto, Canada.	
Indicative existing similar data sets (including possibilities for integration and reuse)	

It should be noted that although several RGB-D datasets stemming from Kinect sensor dealing with human activity analysis are publicly available (see datasets below), to the best of our knowledge, there is no available any Parkinson monitoring dataset yet.

G3D (http://dipersec.king.ac.uk/G3D/G3D.html): G3D dataset contains a range of gaming actions captured with Microsoft Kinect. The Kinect enabled us to record synchronised video, depth and skeleton data. The dataset contains 10 subjects performing 20 gaming actions: *punch right, punch left, kick right, kick left, defend, golf swing, tennis swing forehand, tennis swing backhand, tennis serve, throw bowling ball, aim and fire gun, walk, run, jump, climb, crouch, steer a car, wave, flap and clap*.

MSRC-12 Kinect gesture dataset (http://research.microsoft.com/en-us/um/cambridge/projects/msrc12/): The Microsoft Research Cambridge-12 Kinect gesture data set consists of sequences of human movements, represented as body-part locations, and the associated gesture to be recognized by the system.

RGB-D Person Re-identification Dataset (http://old.iit.it/en/datasets-and-code/datasets/rgbdid.html): A new dataset for person re-identification using depth information. The main motivation is that the standard techniques (such as SDALF) fail when the individuals change their clothing, therefore they cannot be used for long-term video surveillance. Depth information is the solution to deal with this problem because it stays constant for a longer period of time.

DGait Database (http://www.cvc.uab.es/DGaitDB/Summary.html): DGait is a new gait database acquired with a depth camera. This database contains videos from 53 subjects walking in different directions.

STANDARDS AND METADATA

The dataset will be accompanied with detailed documentation of its contents. Indicative metadata include: a) description of the experimental setup and procedure that led to the generation of the dataset, b) type of exercise in case the dataset produced during exergames, c) documentation of the variables recorded in the dataset, and d) annotated posture, action and activity.

DATA SHARING

Access type

Due to ethical reasons, only the data captured by a subset of the patients during the initial phases by normal healthy control subjects could become publicly available, while the rest of them will be private to serve the i-PROGNOSIS R&D objectives.

The inclusion of a (normal healthy control) subject's data in the public part of this dataset will be done on the basis of appropriate informed consent to data publication. It will be investigated further whether the silhouette (coordinates of joints and bones) subset of the dataset could be publicly available.

Access Procedures

For the portions of the dataset that will be made publicly available, a respective web page will (the use of the CAC-playback manager[3] will be assessed) provide a

[3] https://www.youtube.com/watch?v=XLneU8O9WU8

description of the dataset, links to the data management portal and a playback possibility in case the playback manager approach is followed. The private part of this dataset will be stored at a specifically designated private space of AUTH, in dedicated hard disk drives, on which only members of the AUTH and CERTH research team whose work directly relates to these data will have access. For further i-PROGNOSIS partners to obtain access to these data, they should provide a proper request to the AUTH/CERTH primarily responsible, including a justification over the need to have access to these data. Once deemed necessary, AUTH/CERTH will provide the respective data portions to the partner. Cac-playback manager: Konstantinidis, E. I., Billis, A. S., Bratsas, C., & Bamidis, P. D. (2016). Active and Healthy Ageing Big Dataset streaming on demand. In Proceedings of the 18th International Conference on Human-Computer Interaction. Toronto, Canada.
Embargo periods (if any)
The applicable datasets will be publicly available 2 years after the end of the project to allow the consortium prepare and submit the scientific publications.
Technical mechanisms for dissemination
For the public part of the dataset, a link to this will be provided from the i-PROGNOSIS site to the Zenodo service. The link will be provided in all relevant i-PROGNOSIS publications. A technical publication describing the dataset and acquisition procedure will be published.
Necessary S/W and other tools for enabling re-use
The dataset will be designed to allow easy reuse with commonly available tools and software libraries. In case of supporting the online playback of the datasets, libraries for a variety of programming languages will be released (e.g. http://www.cac-framework.com/)
Repository where data will be stored
The public part of this dataset will be accommodated within the Zenodo service.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The public part of this dataset will be preserved online for as long as there are regular downloads. After that it would be made accessible by request. The private part of the dataset will be preserved by AUTH at least until the end of the project.
Approximated end volume of data
The dataset is expected to be several gigabytes, provided that each recording hour is expected to be ~2-5 GB.
Indicative associated costs for data archiving and preservation
Probably two dedicated hard disk drives will be allocated for the dataset; one for the public part and one for the private. In this case the costs associated with its preservation will be according to the hardware cost (hard drives) . Azure will be also

considered as candidate data archiving means although the volume of the data (1-2 TB) would make this choice very expensive.
Indicative plan for covering the above costs
Small one-time costs covered by i-PROGNOSIS. In case of Azure, the cost will be higher.
PARTNERS ACTIVITIES AND RESPONSIBILITIES
Partner Owner / Data Collector — **AUTH, CERTH**
Partner in charge of the data analysis — **AUTH, CERTH**
Partner in charge of the data storage — **AUTH**
WPs and Tasks
The data are going to be collected within activities of WP3, WP4 and WP6, to mainly serve the research efforts of T3.2, T3.7, T4.1, T4.5, T6.1 and T6.4.

DATA SET REFERENCE NAME	**DS3.3-SleepStageAnalysis**
DATA SET DESCRIPTION	
Generic description	
This dataset will include physiological (heart rate) data and IMU (accelerometer and gyroscope) data, captured by a smartwatch (Android Wear 2.0) collected during the users' sleep, to be i) used for the development of or ii) collected during the targeted nocturnal intervention (TNI), i.e., an i-PROGNOSIS intervention that will use pacifying sounds in order to reinstate a satisfactory sleep stage when a sleep disturbance episode is detected. Features will be accompanied by metadata in order to facilitate the analysis within the i-PROGNOSIS project, as well as, by other researchers outside of the project.	
Origin of data	
The dataset will be collected as part of the development of the i-PROGNOSIS supportive interventions and the respective data collection phase. It will comprise a development sub-dataset and corresponding metadata, collected by experiment participants during a short period of time (~2 days), as well as, a deployment sub-dataset and corresponding metadata collected by users during a significantly longer period (in the range of months) during the interventions data collection period. Both data collection procedures will be accompanied by ethics approval.	
Nature and scale of data	
The data collection leading to the formulation of the dataset will take place during the users' sleep. The development sub-dataset will be collected based on an experiment protocol, i.e., the participants (PD patients) will be asked to wear the smartwatch during their sleep for ~2 consecutive nights and afterwards report on their sleep quality. Indicative features that will comprise the dataset are the heart rate (beats per minute - 1 Hz sampling frequency), accelerometer data (X, Y, and Z	

acceleration in g units or meter per second squared, 62 Hz sampling frequency) and gyroscope data (X, Y, Z angular velocity in degrees per second, 62 Hz sampling frequency) as captured and outputted by the smart watch. The deployment sub-dataset will comprise the same features as the development sub-dataset, but it will be collected through the i-PROGNOSIS PD interventions application [via the **targeted nocturnal intervention (TNI) service** (see Section 6 of D2.1)] that will be provided to a subset of interventions users to use (~10 PD patients out of 60 interventions users) based on a pre-interventions medical evaluation. The dataset is expected to be composed of 60 (20 participants × 2 nights) records for the development sub-dataset, while the deployment sub-dataset will comprise approximately 1800 (~6 months interventions period × ~30 nights × ~10 interventions users) records. Data Format: For each sleep session (development sub-dataset) four CSV files, one per feature (heart rate, skin temperature, accelerometer and gyroscope data), will be generated, including the values of features and timestamps, as well as a CSV file including metadata. For each sleep session (deployment sub-dataset) a structured JSON entry will be generated, including all the aforementioned features and metadata.
To whom the dataset could be useful
The development sub-dataset will be used to develop a sleep stage/disturbances recognition model focusing on sleep disturbances that PD patients experience. The latter model will be used as part of the targeted nocturnal intervention (TNI) and the triggering of pacifying sounds will occur based on the model outputs. The deployment sub-dataset will be used for validation and fine-tuning of the model produced during development, as well as for evaluating the effectiveness of the TNI in real life scenarios. Moreover, the dataset could be useful for biomedical researchers that are interested in studying the inference of sleep disturbances, experienced by PD patients, based on data captured by commercially available wearable sensors. Adequate use of these data presupposes at least basic background regarding PD symptomatology, as well as, basic knowledge in data analysis methodology and experience in the use of statistical software packages.
Related scientific publication(s)
The development sub-dataset will accompany the research results regarding the feasibility of a sleep stage/disturbances recognition model based on data captured by commercially available wearable sensors, such as smartwatches. Research results are planned to be published initially in the indicative journals and/or conferences provided below: - IEEE Transactions on Biomedical Engineering (Journal) - Elsevier International Journal of Medical Informatics (Journal) - Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Conference)
Indicative existing similar data sets (including possibilities for integration and reuse)

To the researchers' knowledge there are no similar datasets publicly available at the moment. However, it is expected that such dataset will be available in the near future through the Michael J. Foundation for Parkinson's Research and Intel collaborative "Fox Insight" clinical study (https://foxinsight.michaeljfox.org/) that started in 2015 and aims at collecting big data arising from users' (healthy and PD patients) interaction with smartphones and smartwatches. The latter interaction will also lead to the generation of smartwatch data during the users' sleep. i-PROGNOSIS investigators will proceed to all the necessary actions to access the latter dataset (when available) and exploit it for the development of the sleep stage/disturbance model of the TNI in conjunction with the development sub-dataset.
STANDARDS AND METADATA
The development sub-dataset will be accompanied by a detailed description of its contents. Indicative metadata include: a) description of the experiment set-up and the procedure that led to the generation of the dataset, b) anthropometrics and basic health-record data of the experiment participants, c) description of the sleep-stage-related features, d) annotation of features based on (b), and e) ground truth based on self-reported sleep quality by participants. The deployment sub-dataset will be annotated based on the following indicative metadata: a) basic health record data and anthropometrics of the interventions users, b) timestamps and duration of sleep disturbances detected, c) timestamps of starting/stopping the playback of pacifying sounds, and d) statistics on the usage of the TNI.
DATA SHARING
Access type
The development sub-dataset will be protected (see Section 3.3.2.2), accompanied by the respective ethics approval. Due to ethical compliance, the deployment sub-dataset will be confidential and only i-PROGNOSIS partners will be able to access it after following the appropriate procedure (see Section 3.3.2.3)
Access Procedures
The deployment sub-dataset that will be confidential will be handled according the framework reported in Section 3.3.2.3. The development sub-dataset that will be protected will be available for third-party stakeholders to download based on the procedure described in Section 3.3.2.1.
Embargo periods (if any)
The development sub-dataset is planned to be available as protected after month 32 of the project to allow i-PROGNOSIS investigators to prepare and submit the respective scientific publications, but enable also third-party researchers to conduct further analysis and comment on the i-PROGNOSIS results before the end of the project.
Technical mechanisms for dissemination

The development sub-dataset will be available through the i-PROGNOSIS data management portal (see Section 3.2). Links redirecting to the portal and the available dataset will be provided through a dedicated page of the i-PROGNOSIS project website (www.i-prognosis.eu). A technical description providing information on the experiment protocol through which the development sub-dataset was captured will accompany the development sub-dataset.
Necessary S/W and other tools for enabling re-use
As the publicly available compressed dataset (development sub-dataset) will comprise standard CSV files, no specialised software or tools will be required for the dataset to be parsed and reused, other than a (de-)compression software and a basic text editor (minimum requirement to access the content of the CSV file).
Repository where data will be stored
The development sub-dataset will be stored in AUTH secure servers and the Zenodo service web infrastructure. The deployment sub-datasets will be stored in the Microsoft Azure-based i-PROGNOSIS Cloud infrastructure (European data centres are located in Ireland), but access will be restricted (Section 3.3.2.3).
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The public part of this dataset will be preserved online for as long as there are regular downloads and at least one year after the end of the project. After that it would be made accessible by direct request to the owner-data collector. The confidential part of the dataset will be preserved by the owners indefinitely and at least for one year after the end of the project.
Approximated end volume of data
The development sub-dataset is expected to be approximately 600 MB (including metadata), based on the number of experiment participants (30) and the approximate size of features recorded during an average 7-hour sleep (~10 MB), for 2 nights. The deployment dataset is expected to be in the range of tens of gigabytes (GB), based on the expected number of the i-PROGNOSIS interventions users that will test the TNI (~10 users), an average 7-hour sleep per night for ~6 months (~30 days per month), and an average size of features of ~10 MB per 7-hour sleep.
Indicative associated costs for data archiving and preservation
The development sub-dataset is stored in AUTH servers. The deployment sub-dataset, stored currently on the Microsoft Azure-based i-PROGNOSIS Cloud infrastructure, will be archived and preserved in AUTH servers, after the end of the project. There are no costs associated with the latter means of archiving and preservation.
Indicative plan for covering the above costs
There are no relevant costs.

PARTNERS ACTIVITIES AND RESPONSIBILITIES	
Partner Owner / Data Collector	**AUTH, MICROSOFT**
Partner in charge of the data analysis	**AUTH**
Partner in charge of the data storage	**AUTH, MICROSOFT**
WPs and Tasks	
The development sub-dataset will be collected within WP4 and WP6, to mainly serve the research efforts of T4.2.	

DATA SET REFERENCE NAME	DS3.4-GaitAnalysis
DATA SET DESCRIPTION	
Generic description	
This dataset will include accelerometer, gyroscope and pedometer data captured from a smart watch/band regarding gait rhythm identification and analysis so as to provide personalised cueing and guidance in case of gait freezing episodes.	
Origin of data	
The dataset will be captured by the accelerometer, gyroscope and pedometer sensors of the smart watch/band that will be worn by the participants mainly during the GData, SData and Interventions phases.	
Nature and scale of data	
The first part of the dataset is going to be collected during the development data collection phase where the subjects will follow specific walking scenarios. The second part of the dataset will be collected mainly during the SData and Interventions phases, where smart watches/bands will be provided to 80 and 10 participants, respectively. Moreover, the GData phase will contribute to the dataset; however, the contribution is expected to be limited since the smart watch/band is optional at GData. At these phases, i.e., GData, SData and Interventions, the subjects will not perform specific walking tasks, instead, the data will be captured throughout their daily routine activities. The dataset will include: i) the measures (double precision) of the acceleration force in g units (9,81 m/s^2) that is applied to the smart watch/band on all three physical axes (x, y, and z), ii) the measures (double precision) of the smart watch/band's rotation in rad/s around each of the three physical axes, and iii) the absolute number of steps taken. In Android Wear, the step counter returns the number of steps taken since the last reboot and while the sensor was activated. The above data will be annotated with the health status of the subject, i.e., healthy control or Parkinson's disease patient. The minimum sampling frequency for accelerometer and gyroscope data is 8Hz. However, this value is too high for pedometer data. Consequently, pedometer data will be recorded every 15 seconds.	

Data format: TXT or CSV file. The dataset will be on the order of ~1.3 MB per participant per hour of usage and usually the smart watch/band owners use it for at least 12 hours per day.
To whom the dataset could be useful
The collected data will be used for the development and evaluation of the personalised gait rhythmic guidance intervention of the i-PROGNOSIS project. The dataset could be useful in analysing the gait patterns, speed, periodicity, complexity and habits as well as in detecting sudden freezing episodes and falls. Moreover, it may be possible to identify symptoms and signs of those appear at people with Parkinson's Disease, such as tremor.
Related scientific publication(s)
The dataset will accompany the research results in the field of rhythm gait analysis and freezing episodes' identification of people with PD. At least one publication is intended to be made in the IEEE Biomedical Engineering journal (or similar).
Indicative existing similar data sets (including possibilities for integration and reuse)
To the best of our knowledge, there is no available smart watch/band-based accelerometer/gyroscope and pedometer dataset for PD-related gait freezing detection. Related datasets include data captured from standalone inertia sensors attached to specific parts of the body, and/or accelerometer/gyroscope of a smartphone. For example: *Daphnet* (https://archive.ics.uci.edu/ml/datasets/Daphnet+Freezing+of+Gait): Daphnet dataset contains the annotated readings of 3 acceleration sensors at the hip, thigh and ankle of 10 Parkinson's disease patients that experience freezing of gait during walking tasks. Users performed three kinds of tasks: straight line walking, walking with numerous turns, and finally a more realistic activity of daily living (ADL) task, where users went into different rooms while fetching coffee, opening doors, etc. *OU-ISIR* (http://www.am.sanken.osaka-u.ac.jp/BiometricDB/InertialGait.html): OU-ISIR database contains accelerometer and gyroscope data captured from: i) 3 inertia sensors located around user's waist and ii) a smartphone which includes only a triaxial accelerometer located at the back waist of the users. There were three walking scenarios: flat level, slope up and slope down. *ZJU-GaitAcc* (http://www.cs.zju.edu.cn/~gpan/database/gaitacc.html): The ZJU-GaitAcc dataset contains the gait acceleration series of 175 subjects, who were equipped with 5 Wii remotes, acting as the inertia sensors, fastened at 5 body locations: the left upper arm, the right wrist, the right side of the pelvis, the left thigh, and the right ankle. *Gait Dataset* (http://www.cs.mcgill.ca/~jfrank8/data/gait-dataset.html): This dataset was collected at McGill University using the HumanSense open-source Android data collection platform. It contains the raw sensor data collected from a mobile phone in the pocket of 20 individuals, performing two separate 15 minute walks.
STANDARDS AND METADATA

The dataset will be accompanied with detailed documentation of its contents. Indicative metadata include: a) description of the experimental setup and procedure that led to the generation of the dataset, b) health status of the subject, and c) documentation of the variables recorded in the dataset.
DATA SHARING
Access type
Due to ethical issues, only part of the dataset will be publicly available. More specifically, the data that correspond to a subset of the Parkinson's Disease patients as well as the healthy control subjects, captured at the initial phase of the i-PROGNOSIS project will become publicly available. The rest of the data will be private to serve the i-PROGNOSIS R&D objectives. The inclusion of a subject's data in the public part of this dataset will be done on the basis of appropriate informed consent to data publication.
Access Procedures
See Section 3.3.2.1 and Section 3.3.2.3.
Embargo periods (if any)
The applicable datasets will be publicly available 2 years after the end of the project to allow the consortium prepare and submit the scientific publications.
Technical mechanisms for dissemination
For the public part of the dataset, a link to this will be provided from i-PROGNOSIS site to the Zenodo service. The link will be provided in all relevant i-PROGNOSIS publications. A technical publication describing the dataset and acquisition procedure will be published.
Necessary S/W and other tools for enabling re-use
The dataset will be designed to allow easy reuse with commonly available tools and software libraries, such as Excel, Matlab, .NET etc.
Repository where data will be stored
The public part of this dataset will be accommodated within the Zenodo service, hosted by CERN infrastructure.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The public part of this dataset will be preserved online for as long as there are regular downloads. After that it would be made accessible by request. The private part of the dataset will be preserved by AUTH at least until the end of the project.
Approximated end volume of data
The dataset is expected to be several gigabytes, provided that each participant is expected to produce at least 6 MB per day and the duration is at least 22 months (SData and Interventions phase).
Indicative associated costs for data archiving and preservation

Probably two dedicated hard disk drives will be allocated for the dataset; one for the public part and one for the private. There are no costs associated with its preservation. Azure will be also considered as candidate data archiving means.
Indicative plan for covering the above costs
The costs associated with the data archiving and preservation will be covered by the i-PROGNOSIS project budget.

PARTNERS ACTIVITIES AND RESPONSIBILITIES	
Partner Owner / Data Collector	**AUTH, MICROSOFT**
Partner in charge of the data analysis	**AUTH**
Partner in charge of the data storage	**AUTH, MICROSOFT**

WPs and Tasks
The data are going to be collected within the activities of WP3, WP4 and WP6, to mainly serve the research efforts of T3.2, T4.3, T6.1, T6.2, T6.3 and T6.4.

DATA SET REFERENCE NAME	**DS3.5-VoiceEnhancementAnalysis**
DATA SET DESCRIPTION	
Generic description	
Speech database for the design and implementation of the voice enhancement algorithms within the assistive interventions software. The database will contain speech data collected from phone calls recorded from the i-PROGNOSIS smartphone dialler application. The data set includes several annotations about the speech data and the underlying speakers.	
Origin of data	
The dataset will be collected by recording conversations using the i-PROGNOSIS smartphone dialler application during SData and Intervention Data collection phase.	
Nature and scale of data	
The goal of this dataset is to enable the development and implementation of voice enhancement algorithms for speech from PD patients. The dataset will contain the raw speech signals recorded by the i-PROGNOSIS smartphone dialler application during the SData and Intervention data collection phase. Annotations including date and time of the recordings along with speaker information (e.g. ID, gender, PD scale) will provide the necessary information and will facilitate the development and evaluation of the speech enhancement algorithms. The dataset will contain raw speech data (e.g., 44.1. 16 kHz sampling rate, wav unencoded) and text files for the annotations. Data Format: Raw audio (wav, raw) audio waveforms; XML or plain text files for annotations/metadata.	

The dataset will be on the order of 0.5 MB to 5 MB per minute of speech depending on the encoding and sampling rate.
To whom the dataset could be useful
The dataset will be used within the project for the development and evaluation of automatic signal processing speech enhancement algorithm for speech from PD patients. This data will be used in WP4 and WP6, supporting the assistive intervention software and for the overall assessment of the i-PROGNOSIS system. This dataset may also be useful in the future for other researchers who want to explore PD patients' speech and develop speech enhancement algorithms.
Related scientific publication(s)
This database will build the foundation of our research and development of algorithms for the automatic enhancement of the speech from PD patients within the assistive intervention software. We plan to propose our findings on ICASSP, Interspeech or other voice and biomedical-related conferences.
Indicative existing similar data sets (including possibilities for integration and reuse)
To the best of our knowledge no speech database for the development and evaluation of speech enhancement algorithms for the speech of PD patients is available.
STANDARDS AND METADATA
The dataset will be accompanied with detailed documentation of its contents. Indicative metadata include: (a) description of the experimental setup and procedure that led to the generation of the dataset, and (b) documentation of the variables recorded in the dataset.
DATA SHARING
Access type
Due to ethical reasons, only the data captured by a subset of the patients during the initial phases by normal healthy control subjects could become publicly available, while the rest of them will be private to serve the i-PROGNOSIS R&D objectives. The inclusion of a (normal healthy control) subject's data in the public part of this dataset will be done on the basis of appropriate informed consent to data publication.
Access Procedures
For the portions of the dataset that will be made publicly available, a respective web page will provide a description of the dataset, links to the data management portal and a playback possibility in case the playback manager approach is followed. The private part of this dataset will be stored at a specifically designated private space of FRAUNHOFER, in dedicated hard disk drives, on which only members of the FRAUNHOFER research team will have access.
Embargo periods (if any)

The applicable datasets will be publicly available two years after the end of the project to allow the consortium prepare and submit the scientific publications.
Technical mechanisms for dissemination
For the public part of the dataset, a link will be provided from i-PROGNOSIS site to the Zenodo service. The link will be provided in all relevant i-PROGNOSIS publications. A technical publication describing the dataset and acquisition procedure will be published.
Necessary S/W and other tools for enabling re-use
The dataset will be designed to allow easy reuse and access with commonly available tools (e.g., Matlab, Python, VLC, GVIM) and software libraries (e.g., Tensorflow, FFMPEG, HDF5 C++ API, C++ STL), because the data will be stored primarily in common file formats and generic data containers.
Repository where data will be stored
The public data will be hosted within the Zenodo service that will serve the needs of the Data Management Portal of the i-PROGNOSIS project.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The public part of the dataset will be preserved on the Data Management Portal of the i-PROGNOSIS project. The private part of the dataset will be preserved by FRAUNHOFER at least until the end of the project.
Approximated end volume of data
The dataset should be expected to consume up to 10 GB depending on the encoding and the length and quantity of the speech signals (e.g., single channel audio waveform 16 bit, 44.1 kHz is of size 5.3 MB/min, single channel mp3 192 Kbps is of size 1.44 MB/min)
Indicative associated costs for data archiving and preservation
There are no costs associated with data preservation in institutional servers or the Zenodo service.
Indicative plan for covering the above costs
There are no relevant costs.
PARTNERS ACTIVITIES AND RESPONSIBILITIES

Partner Owner / Data Collector	**TUD, KI**
Partner in charge of the data analysis	**FRAUNHOFER**
Partner in charge of the data storage	**FRAUNHOFER**
WPs and Tasks	

The data is going to be collected within activities of WP4 and WP6, to mainly serve the research efforts of T4.3, T6.3 and T6.4.

4.3.4 Requirements Data

DATA SET REFERENCE NAME	**DS4.1-FocusGroupsDataset**
DATA SET DESCRIPTION	
Generic description	
This dataset aims at providing a non-exhaustive list of end-user requirements, as these have been elicited through a series of focus groups and personal interviews with experts (both patient groups and clinicians) in PD.	
Origin of data	
KCL and relevant partners gathered pertinent data and established focus groups data set, so as to then design a final paper-based user requirement questionnaire that was used as an acquisition means in four focus groups. This dataset contains information from all stakeholders (developers and clinicians) involved in the development of the i-PROGNOSIS system and its potential users (healthcare professionals, and patients). The main components of the focus groups questionnaire included: a) demographics, e.g. gender, age, health problem related to PD, b) technology adoption, c) i-PROGNOSIS App design-oriented questions, d) i-PROGNOSIS App delivery through existing smart phones, e) usability aspects of the specialized gadgets, such as smartwatch (Microsoft Band was provided as an example), smart belt, Mandometer and a ECG-enabled TV remote control, f) general questions about the PGS interventions and finally, g) some specialized questions about specific interventions, such as the sleep and gait interventions. For a more detailed description on the questions and the results produced from the analyses, one is referred to D2.1 - *First version of user requirements analysis*.	
Nature and scale of data	
The focus groups were conducted among health professionals such as nurses, therapists, clinicians and researchers as well as patients and carers. The subjects of the focus groups were about the design of the i-PROGNOSIS project and expectations of it, the design of the application and special gadgets that will be used in i-PROGNOSIS (smart belt, smart watch, etc.). Finally, the questions focus on the i-PROGNOSIS interventions. The participants were carefully recruited for the face-to-face focus groups. The dataset contains the output of the systematic analysis Data format: The dataset will be made available as a single Microsoft Excel file (.xlsx) by merging the CSV files of all languages. All questions and answers will be translated in English.	
To whom the dataset could be useful	
The dataset and the respective outcomes and findings will be used to shape the user requirements and the technical specifications of the i-PROGNOSIS system.	

In addition, this kind of datasets could be exploited by researchers that investigate to build new technologies of patients with PD relating to self-care, as well as, developers, system architects and user experts that are interested to develop a similar to i-PROGNOSIS ICT-based system for health management.
Related scientific publication(s)
The dataset will accompany the research results regarding the prospective users' attitude towards ICT-based self-care solutions and the i-RPOGNOSIS system against PD. Research results and findings is planned to be published initially in one of the indicative conferences provided below: - International Conference on e-Health (Conference) - Human Computer Interaction International (Conference)
Indicative existing similar data sets (including possibilities for integration and reuse)
There are no similar datasets with open access. Results of relevant end-user focus groups target to participatory design of symptomatic domains to be measured as well as interventions (through serious games) design, i.e.: Serrano, J. A., Larsen, F., Isaacs, T., Matthews, H., Duffen, J., Riggare, S., ... & Graessner, H. (2015). Participatory design in parkinson's research with focus on the symptomatic domains to be measured. Journal of Parkinson's disease, 5(1), 187-196. McNaney, R., Balaam, M., Holden, A., Schofield, G., Jackson, D., Webster, M., ... & Olivier, P. (2015, April). Designing for and with People with Parkinson's: A Focus on Exergaming. In Proceedings of the 33rd annual ACM conference on Human Factors in Computing Systems (pp. 501-510). ACM.
STANDARDS AND METADATA
For this dataset, the metadata correspond to the participant profile as well as the settings of the focus group, and mainly include: a) occupational and expertise of the participants (nurse, therapist, clinician, researcher, patient, carers), b) the country where the focus group conducted, c) number of the participants, and d) health status relating to PD.
DATA SHARING
Access type
Personal identification as stated in the consent forms will be kept separate from any research and health-related data, which will be pseudonymised (only initials and date of birth will be kept). All paper based data will be stored in an access restricted, locked building with access for the research team. Electronic data will be saved on a secured NHS server and the i-PROGNOSIS data management portal.
Access Procedures
Personal and clinical data will be accessible by technical, medical and scientific personnel being involved in the i-PROGNOSIS project and not by any third-party users. When applicable, the publicly available dataset will be open for third-party stakeholders to download based on the procedure described in Section 3.3.2.1.

Embargo periods (if any)
The dataset is planned to be publicly available two (2) years after the end of the project to allow i-PROGNOSIS investigators to prepare and submit the respective scientific publications regarding the final version of user requirements.
Technical mechanisms for dissemination
The dataset will be available through the Zenodo web service. Links redirecting to the portal and the available dataset will be provided through a dedicated page of the i-PROGNOSIS project website (www.i-prognosis.eu).
Necessary S/W and other tools for enabling re-use
The dataset will be kept in a numeric manner and therefore designed to allow easy reuse with commonly available tools.
Repository where data will be stored
The data will be hosted within the Zenodo service that will serve the needs of the Data Management Portal of the i-PROGNOSIS project and part of them in a secured NHS server.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The dataset will be preserved as long as there are regular downloads. After that it would be made accessible by request and preserved by AUTH at least until the end of the project. Data set will be stored as per HRA regulations and requirements.
Approximated end volume of data
The approximate size of the dataset is a couple of tens of MB.
Indicative associated costs for data archiving and preservation
The downloaded CSV files and the cumulative Microsoft Excel file will be stored in an AUTH server, both not imposing any additional costs.
Indicative plan for covering the above costs
There are no relevant costs.
PARTNERS ACTIVITIES AND RESPONSIBILITIES

Partner Owner / Data Collector	**TUD, KCL, AUTH**
Partner in charge of the data analysis	**All**
Partner in charge of the data storage	**AUTH**

WPs and Tasks
The data are going to be collected within the activities of WP2, to mainly serve the research efforts of T2.1 and T2.4.

DATA SET REFERENCE NAME	DS4.2-WebSurveyDataset
DATA SET DESCRIPTION	
Generic description	
The dataset contains the questions and the qualitative answers to the questions of the i-PROGNOSIS Web-based questionnaire (Web survey), conducted within the context of the identification of user requirements and system specifications, from ~2000 anonymous survey participants.	
Origin of data	
The dataset is populated by the answers of participants to the questions of the i-PROGNOSIS Web survey. The Web survey consists of three sections corresponding to three groups of survey participants, i.e., healthy adults, PD patients, and experts in PD (including physicians and carers). Each participant answers the questions of her/his corresponding section based on the group s/he belongs to. Each section is further divided in three parts, i.e. demographics, PD detection and interventions. The section corresponding to the group of healthy adults does not include the interventions part. The participant is redirected to the corresponding section based on a question allowing her/him to denote the group s/he belongs to. If no option is selected, the participant is redirected to the survey section corresponding to healthy adults. There are six versions of the Web survey corresponding to six languages, i.e., English, Greek, French, German, Portuguese and Swedish.	
Nature and scale of data	
The Web survey is conducted electronically and it is available online. The questionnaires were designed using the Google Forms online software. The answers to the questions are qualitative and are stored in Google Sheets files (one per language) that are updated with new entries (an entry corresponds to a participant in the survey) automatically. The current dataset consists of six Google Sheets files with a total of ~2000 entries, but it is updated frequently as the Web survey is planned to be online until the updated user requirements will be produced (month 28 of the project). The dataset (Google Sheets files) can be downloaded manually as CSV files. Data format: The dataset will be made available as a single Microsoft Excel file (.xlsx) by merging the CSV files of all languages. All questions and answers will be translated in English.	
To whom the dataset could be useful	
The dataset and the respective results will be used to shape the identified user requirements and the technical specifications of the i-PROGNOSIS system. The dataset could be of interest to researchers that investigate the relationship of similar groups of users with new technology relating to self-care, as well as, developers, system architects and user experts that are interested to develop a similar to i-PROGNOSIS ICT-based system for health management. Adequate use of these data presupposes at least basic knowledge in data analysis methodology and experience in the use of statistical software packages.	

Related scientific publication(s)
S. Hadjidimitriou et al., "Active and healthy ageing for Parkinson's disease patients' support: A user's perspective within the i-PROGNOSIS framework," 2016 1st International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW), Vila Real, 2016, pp. 1-8. doi: 10.1109/TISHW.2016.7847785
Indicative existing similar data sets (including possibilities for integration and reuse)
There are no similar datasets with open access. Results of similar surveys are available through publications and may be used for shaping the i-PROGNOSIS user requirements, i.e.: Zhao, Y., Heida, T., van Wegen, E. E., Bloem, B. R., & van Wezel, R. J. (2015). E-health support in people with Parkinson's disease with smart glasses: a survey of user requirements and expectations in the Netherlands. Journal of Parkinson's disease, 5(2), 369-378.
STANDARDS AND METADATA
For this dataset, the metadata correspond to the participant-provided information via the first part of each web survey section, i.e., the demographics part, and mainly include: a) the age of the participant, b) the country of residence, c) their occupational category, d) health status relating to PD, e) expertise (in case of health care professionals), and f) familiarity with new technology and new means of communication. Thus, no standard is followed.
DATA SHARING
Access type
The dataset will be publicly available (see Section 3.3.2.1).
Access Procedures
The publicly available dataset will be open for third-party stakeholders to download based on the procedure described in Section 3.3.2.1.
Embargo periods (if any)
The dataset is planned to be publicly available after month 28 of the project to allow i-PROGNOSIS investigators to prepare and submit the respective scientific publications regarding the final version of user requirements.
Technical mechanisms for dissemination
The dataset will be available through the i-PROGNOSIS data management portal (see Section 3.2). Links redirecting to the portal and the available dataset will be provided through a dedicated page of the i-PROGNOSIS project website (www.i-prognosis.eu). A technical description providing information on how the web-survey was structured and conducted will accompany the dataset.
Necessary S/W and other tools for enabling re-use

As the compressed dataset will consist of a Microsoft Excel file (.xlsx), only a (de-) compression software and the Microsoft Excel software (or other compatible with this type of file software) will be required for data parsing and re-use.
Repository where data will be stored
The dataset will be stored in AUTH secure servers. External repositories will also be considered as alternatives, to increase visibility and dissemination efforts.
ARCHIVING AND PRESERVATION (including storage and backup)
Data preservation period
The dataset will be preserved online for as long as there are regular downloads and at least for one year after the end of the project. After that it would be made accessible by request.
Approximated end volume of data
The approximate size of the dataset is ~ 3 MB based on the size of each entry (~ 1.2 KB) and the expected final number of entries (participants in the survey) (~ 2500 to 3000 participants - entries).
Indicative associated costs for data archiving and preservation
The original Google Sheets files comprising the dataset are stored in the Google drive account of the i-PROGNOSIS project that is free of charge. The downloaded CSV files and the cumulative Microsoft Excel file are stored in an AUTH server, both not imposing any additional costs.
Indicative plan for covering the above costs
There are no relevant costs.
PARTNERS ACTIVITIES AND RESPONSIBILITIES

Partner Owner / Data Collector	**AUTH**
Partner in charge of the data analysis	**AUTH**
Partner in charge of the data storage	**AUTH**

WPs and Tasks
The data are going to be collected within the activities of WP2, to mainly serve the research efforts of T2.1 and T2.4.

APPENDIX I – DATASET DESCRIPTION TEMPLATE

The table template used to describe each dataset along with clarifications on each field:

DATA SET REFERENCE NAME	**DS#.#-<DatasetName>**
DATA SET DESCRIPTION	
Generic description	
<Provide a summary of data to be collected>	
Origin of data	
<How will the dataset be produced, e.g. mobile phone app logs or Microsoft Band and a sampling rate of xx Hz etc.>	
Nature and scale of data	
<Description of file format, e.g. TXT files, estimated size, number of participants>	
To whom the dataset could be useful	
<What research purposes will the dataset facilitate, e.g. The dataset will be valuable for benchmarking algorithms for activity analysis etc.>	
Related scientific publication(s)	
<Either existing ones in case we get open data from external to the project repositories, or future publications that we intend to make, e.g. description of indicative research area like neuroscience or name a few scientific conferences or journals that will be targeted>	
Indicative existing similar data sets (including possibilities for integration and reuse)	
<List other already existing datasets or repositories>	
STANDARDS AND METADATA	
<Reference to existing suitable standards of the discipline. Format, e.g. EDF format for biosignals. Metadata should contain information such as: (a) description of the experimental setup and procedure that led to the generation of the dataset, (b) documentation of the variables recorded in the dataset and (c) annotated experiment state of the monitored person per time interval. If these do not exist, provide an outline on how and what metadata will be created.>	
DATA SHARING	
Access type	
<E.g., open - can be publicly shared, protected - can be shared but the participants have to provide their consent, or confidential - cannot be shared outside the project>	
Access Procedures	
<Explain how will you handle private (if any) and publicly available datasets, e.g., depending on the access type account for each dataset a relevant access procedure should be defined. Access procedure should contain information about the developed area within the portal that will let third-party users download files>	

Embargo periods (if any)	
<When will the data be published and access provided>	
Technical mechanisms for dissemination	
<Accompany datasets with a technical description of the dataset and the way data were captured>	
Necessary S/W and other tools for enabling re-use	
<Code/libraries/open source software to read and process data so as to allow for reproducible research>	
Repository where data will be stored	
<i-PROGNOSIS portal dedicated download section, institutional portals or standard repository for the discipline>	
ARCHIVING AND PRESERVATION (including storage and backup)	
Data preservation period	
<For how long should the data be preserved? We should take into account any national regulations as well>	
Approximated end volume of data	
<E.g., ### GB raw signal>	
Indicative associated costs for data archiving and preservation	
<Hard disk drives, servers>	
Indicative plan for covering the above costs	
<Within project, whole consortium, local partner, centrally>	
PARTNERS ACTIVITIES AND RESPONSIBILITIES	
Partner Owner / Data Collector	*<Partner Acronym>*
Partner in charge of the data analysis	*<Partner Acronym>*
Partner in charge of the data storage	*<Partner Acronym>*
WPs and Tasks	
<During which stage of the project will data be captured?>	