InAdvance

D1.4 – Data Management Plan, first version

WP1 – Management and Coordination

Submission Date	28th June 2019
Responsible Partner	UVEG

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

This project has received funding from the European Union's Horizon 2020 research and1 innovation programme under grant agreement No 825750

Project Documentation	
Project Acronym	InAdvance
Project Full Title	Patient-centred pathways of early palliative care, supportive ecosystems and appraisal standard
Grant Agreement	825750
Call Identifier	H2020-SC1-2018-Single-Stage-RTD
Topic	SC1-BHC-23-2018
Funding Scheme	Research and Innovation action (RIA)
Project Duration	48 months
Coordinator	Universitat de Valencia
Website	www.inadvanceproject.eu

Deliverable Documentation	
Number	D1.4
Title	Data Management Plan, first version
Related WP	WP1 Management and Coordination
Lead Beneficiary	UVEG
Nature of Deliverable	ORDP: Open Research Data Pilot
Dissemination Level	Public
Submission Due Date	28/06/2019
Actual Submission Date	30/06/2019

Quality Control	
Author(s)	
Ascensión Doñate	UVEG
Estrella Durá	UVEG
Jorge Garcés	UVEG
Contributor(s)	
Lucas Espuig	UVEG
Sofia Reppou	AUTH
Lucia Aparici	UPV-Sabien
Vania Dimitrova	UNIVLEEDS
Mª Eugenia Gas	HULAFE
Editor(s)	
Ascensión Doñate	UVEG
Reviewed by	
Javier Plaza (DPO)	UVEG
Lucas Espuig	UVEG
Siok Swan	ERASMUS MC
Vicent Blanes	UPV-BDSLab
Juan M. García-Gómez	UPV-BDSLab
Jane Moreno	UVEG
Approved by	
Javier Plaza (DPO)	UVEG

This project has received funding from the European Union's Horizon 2020 research and2 innovation programme under grant agreement No 825750

Revision History			
Version	**Date**	**Comment**	**Author(s)**
V1	23/04/2019	Written first version	Ascensión Doñate
V1.1	24/04/2019	Contributions	Lucia Aparici
V1.2	03/05/2019	Contributions	Sofia Reppou
V1.3	06/05/2019	Revision first version	Javier Plaza
V1.4	08/05/2019	Contributions	Lucas Espuig
V1.5	08/05/2019	Contributions	Vania Dimitrova
V1.6	08/05/2019	Contributions	Mª Eugenia Gas
V2	10/05/2019	Second version	Ascensión Doñate
V2.1	16/05/2019	Revised version	Ascensión Doñate
V2.2	21/05/2019	Contributions	Estrella Durá
V3	28/05/2019	Revised version	Ascensión Doñate
V3.1	03/06/2019	Revision and contributions	Mª Eugenia Gas Lucía Aparici Adriano Fernandes Sofia Reppou
V3.2	04/06/2019	Revision and contributions	Vania Dimitrova Frances Hines
V4	07/06/2019	Revised version	Ascensión Doñate
V4.1	14/06/2019	Revision	Siok Swan
V4.2	14/06/2019 16/06/2019	Revision and contributions Revision and contributions	Vicent Blanes Juan M. García-Gómez
V4.3	17/06/2019	Revision	Jane Moreno
V5	28/06/2019	Revised version	Ascensión Doñate

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

Table of Contents

List of Acronyms

DAC	Data Access Committee
DMP	Data Management Plan
DOI	Digital Object Identifier
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
EC	European Commission
EHR	Electronic Health Records
FAIR	Findable, Accessible, Interoperable and Re-usable
FNC	File Naming Convention
GA	General Assembly
GDPR	General Data Protection Regulation
IoT	Internet of Things
IPR	Intellectual Property Rights
ORD	Open Access to Research Data
PC	Palliative Care
PHR	Public Health Report
RA	Research Action
WHO	World Health Organization
WP	Work Package

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

0. Presentation of the deliverable

This is the first version of the Data Management Plan (DMP) for the InAdvance project, prepared for the 6-month point in the project. A DMP describes the data management life cycle outlining how the research data collected, processed or generated by the InAdvance project will be handled during and after the end of the study. Thus, this deliverable is aimed to inform the InAdvance partners about main guidelines and good practices to be applied to the data processed and generated at the project. This DMP defines the methodologies and standards to be applied at InAdvance in order to make data FAIR: findable, accessible, interoperable and re-usable.

The DMP has been drafted following the guidelines on FAIR Data Management in Horizon 2020[1] and the specific template for projects funded under the Societal Challenge 1[2] (Health, demographic change and wellbeing).

The DMP is intended to be a living document in which information will be made available on a more detailed and specific level through periodic updating as the implementation of the project progresses and when significant changes or milestones occur. Thus, the DMP will be updated at least in the context of the periodic reports of the project at months 18, 36 and 48.

[1] H2020 Programme (2016). Guidelines on FAIR Data Management in Horizon 2020.
[2] https://ec.europa.eu/research/participants/data/ref/h2020/other/gm/reporting/h2020-tpl-oa-data-mgt-plan-annotated_en.pdf

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

1. Introduction

1.1 InAdvance project

The focus of the InAdvance project is on health science as well as socio-demographic data. The purpose of this Data Management Plan (DPM) is to make these data Findable, Accessible, Interoperable and Re-usable (FAIR). By enabling access to the data and knowledge generated under the project, the lifetime and usefulness of the scientific research conducted will be greatly enhanced which will be of benefit to health scientists, service providers, policy makers as well as professionals from the educative field.

InAdvance entails 11 international and interdisciplinary partners working on the development of 5 research and development actions in the context of 5 clinical sites: Valencia (Spain), Amadora-Lisbon (Portugal), Thessaloniki (Greece), Leeds and Inverness (United Kingdom). The previously mentioned research actions (RA), framed in different Work Packages (WP) within the project, are described briefly below:

- **RA1 Development of risk stratification models.** The objective of this RA is to identify complex chronic patients in need of early implementation of palliative care (PC). Thus, the RA entails a retrospective study in order to develop four stratification models: a 1-year time mortality model, a regression model to estimate the time until exitus, a decline prediction model based on frailty indices and a model for predicting hospital resources consumptions. These models will be developed through machine/deep learning. This RA is framed under the WP2 of the project (*Early identification of complex patients in need of PC*).
- **RA2 PC needs assessment.** The objective of this RA is to develop ways to conduct equitable and prompt PC needs analysis among complex chronic patients. Thus, this RA will gather sound and key knowledge information to design adapted and individualized care pathways. For this purpose, several data collection methods will be used: qualitative analysis, automated text analysis of digital content and historic cohort data. This RA is framed under the WP3 of the project (*Intervention modelling through equitable multilevel needs assessment*).
- **RA3 Supportive ecosystem augmentation of PC patients.** The objective of this RA is to create supportive environments in order to augment and improve the engagement of the whole ecosystem of the patient in need of PC. The RA expects to reach this through multifaceted training programmes and digital health platforms aimed to promote empathy and resilience, improve mental health, emotional and social support. This RA is framed under the WP4 of the project (*Patient centred-supportive PC ecosystem augmentation*).
- **RA4 Clinical trials.** The core of the InAdvance project will be the implementation of designed interventions during RA2 and RA3 through five clinical studies through randomized clinical trials. These 18-month clinical trials will be assessed through a continuous and iterative assessment in terms of feasibility, effectiveness and cost-effectiveness

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

collecting both qualitative and quantitative data. On the basis of these repeated evaluations and analysis, necessary adjustments will be introduced in the interventions in order to obtain the most refined intervention versions at the end of assessment cycle. This RA is framed under the WP5 (*Clinical Trials*) and WP6 (*Formative and summative assessment of the trials*) of the project.

- **RA5 Appraisal Standard Dashboard.** The objective of this RA is to provide a general overview of the users and interventions aimed to support clinical decisions and to assess outcomes. This RA will be elaborated iteratively following co-creation techniques and will be based on process mining, longitudinal data analytics and the re-use and integration of data from hospitals and public health repositories. This RA is framed under the WP7 of the project (*Design of a consolidated standard for appraising PC interventions*).

Table 1 summarizes the main characteristics of each RA in terms of data processing.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

Table 1 Summary of each InAdvance RA in terms of data processing

	RA1	RA2	RA3	RA4	RA5
Type of data to be collected	Health data. Potentially, socio-demographic data and psychological data.	Socio-demographic data. Health data. Needs and preferences related to PC.	Socio-demographic data. Health data. User metadata.	Sociodemographic data. Health data. Preferences and values related to care interventions. Costs associated to patient's care.	
Size of data	Tens to hundreds of thousands of patients' registers (estimated).	--	1 TB	--	--
Source / origin of data	Existing data (patients' medical files).	Generated data at study sites. Data retrieved from UK sources.	Existing data (patients' medical files). Generated data at study sites.	Generated data at study sites.	EHR. PHR. IoT systems.
Collection methods	Access to Big Data Platform.	Qualitative interviews. Focus groups. Comments on videos.	Questionnaires. Semi-structured interviews. Biosensors.	Questionnaires. Semi-structured interviews.	Health centre databases. IoT systems.
Data process	Multivariate regressions. Machine/Deep Learning.	Qualitative and quantitative analysis.	Pre-processing. Extraction. Deep learning. Json files.	Pre-processing. Quantitative (i.e. mixed models) and qualitative analysis.	Process Mining.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

Type of data generated	Score that reflects the patient's decline at different prospective moments, mortality predictions. Resources consumption predictions.	Transcribed interviews.	Personal data. Demographic data. Biosensor data. Mobile capabilities.	Health outcomes (since a comprehensive approach). Economic outcomes. Acceptability of the interventions.	Process maps (Enhanced Time Parallel Automaton – TPAs).
Purpose / Utility of data generated	To dispose basic criteria for a referral to a PC intervention.	To understand patients and health care professionals perspectives.	To create deep learning algorithms for prediction. To test the support platform deployed. To collect needed features to upload on cloud and examine.	To co-design, optimize and refine interventions based on end-users' preferences and needs. To assess the impact of interventions among end-users. To assess the 'value for money' of interventions. To assess the execution of the interventions.	To discover changes in human behaviour. To analyse adherence to treatments. To discover and evaluate PC protocols.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

1.2 InAdvance data collection

Data collection under the InAdvance project will be as diverse as the topics covered and methodologies used at the five RAs. An overview of the data to be collected is presented at Table 1.
Some data to be collected at the different RAs are not completely specified yet, but main general guidelines on how they will be collected and processed are already defined, which are described below. Additionally, it is important to highlight that InAdvance consortium claims that data to be collected under the different RAs are adequate, relevant and limited to what is necessary for the research purpose fulfilling the Data Minimization Principle of the GDPR.

Thus as tasks and methodologies are clarified in the future months, InAdvance researchers will establish the norms to be used in data collection and data sets to be collected.

Two main types of data will be used during the InAdvance lifetime:

- Data purposely generated by project partners under different activities through several means, such as validated questionnaires, individual interviews, etc.
- Background data gathered by partners before the start of the project that is relevant in order to perform the project tasks and achieve its objectives (i.e. patients' diagnostics, drugs prescribed, laboratory tests, etc.) retrieved from electronic health records (EHR). These will be de-identified data sets (patient data).

Personal data will be subject to de-identification and site control, unless written informed consent for very clearly defined research activities is specifically obtained.

Data formats and data types produced under the project depend on how data will be gathered and the software used for their analysis. Clearly, several data types require different file types. As general guidelines, InAdvance recommends:

- avoiding the use of property software format when possible;
- in case of data files in different formats, converting the data to the most common data format.

1.3 InAdvance data origins

Origin of the data is also different depending of the RAs:

- The information used for retrospective analysis at the **RA1** comes from several clinical information systems involving clinical activity and healthcare areas. Information regarding imaging activity, laboratory tests, clinical trials activity, use of resources and economic management may be also relevant. These will be de-identified where they relate to patients.
- **RA2** will use data generated within the project as well as data retrieved from UK sources.

This project has received funding from the European Union's Horizon 2020 research and10 innovation programme under grant agreement No 825750

- **RA3** will use both existing and generated data; existing patients' anonymized information from EHR and generated from the pilot testing of the technological applications deployed among them.
- **RA4** will use mostly data generated under the project directly involving end-users at the clinical trials and also some existing data from EHR where patient data is anonymised.
- **RA5** will use heterogeneous data sources, such as EHR, public health reports (PHR) or Internet of Things (IoT) systems.

As stated, there will be some re-use of existing data in InAdvance. RA1 will use data from EHR, such as emergency care settings, outpatient, hospitalization, clinical reports, surgical unit, intensive care unit, hospital at home care and attended in order to develop identification algorithms. RA2 will use data retrieved from UK sources. RA3 will use some existing data from EHR too. RA4 will collect anonymised health data from patients involved at the clinical trials to adapt PC interventions as well as to complement the data collected in vivo during the 18-month trials. And RA5 will be based also on the data collected under previous RAs as well as from EHR for instance.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

2. FAIR Data

2.1 Making data findable

In order to ensure that research data can be used, shared and re-used it is crucial to take care that those data are accessible, understandable and usable. This requires clear data description, annotation, contextual information and documentation explaining how data were created or digitized, what data mean or what their content and structure are.

InAdvance will encourage establishing strong links between data collected and stored and the associated documentation, following recommendations such as the following:

- Including information within the data or document itself, e.g. in the document properties function of a file or the file header.
- Keeping a database of metadata with links to files.
- Storing a readme.txt file alongside the data which provides basic explanatory details. Txt. Files will be acceptable to explain very simple stuff like how to connect to a database, not the database explanation itself
- Recording relevant context in lab notebooks or associated papers and reports.
- Including link to websites or web pages which explain the context of the research.

2.1.1 Metadata

The use of common data and metadata standards are a key aspect for semantic and technological data operability. Metadata can describe the content, context and provenance of datasets in a standardized and structured manner, typically describing the purpose, origin, temporal characteristics, geographic location, authorship, access conditions and terms of use of a dataset. This provides structured searchable information that helps users to find existing data resources, judge whether a particular dataset is suitable for their research purpose and provides a bibliographic record for citing data. Standard dictionaries of metadata will be used for all data types produced by the data creators.

InAdvance metadata will include different types of metadata:

- **Descriptive metadata**: common fields such as title, author, abstract, keywords which help users to discover online sources through searching and browsing.
- **Structural metadata**: relationships between different sets of associated data in the context of databases, such as schema describing associations between tables.
- **Administrative metadata**: preservation, rights management, and technical metadata about formats.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

Metadata should be simple and avoid administrative burden. However, given the nature of clinical research data objects, metadata created under InAdvance must be characterized by (Canham & Ohmann, 2016[3]):

- Unambiguously **identification of the research study** that the data object is about (or generated from or used within). Some options available are:
  - Registry identifiers (IDs), such as those provided by such as ClinicalTrials.gov.
  - The Universal Trial Number from the World Health Organization (WHO).
  - Protocol title, long or 'scientific' title of the study.
  - Ethics identifier.
- Define the **research object** itself – for example, its type, authorship, contents, size and language. Wherever possible, InAdvance data objects will also have a Digital Object Identifier (DOI). Once the decision is taken to make a data object available for sharing, a DOI should be assigned, probably most easily by the original data-generating organisation.
- Describe **where the object can be found** and the access regime under which the data object is available.

InAdvance encourages using the metadata scheme for clinical research data objects proposed by Canham & Ohmann (2016) based on the existing DataCite scheme[4] (see Table 2).

Table 2 Proposed elements to be included in the metadata scheme for clinical research data objects (Canham & Ohmann, 2016)

Mandatory	Recommended	Optional
Source story title	Study identifier records Study topics	
DOI Object title	Version	Object other identifiers Object additional titles
Creators		Contributors
Creation year		Dates
Resource type general	Resource type Description Language Related identifiers	Subjects (of data object)
Publisher Access type Access details Access contact Resources	Other hosting institutions	Rights.

Metadata Standards provide specific data fields or elements to be used in describing data for a particular use. Standards are created by international or internationally recognized bodies; some are generic, while others are domain-

[3] Canham, S. & Ohmann, C. (2016). A metadata schema for data objects in clinical research. *Trials, 17*: 557.
[4] https://schema.datacite.org/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

specific. Generic ones tend to be easy to use and widely adopted, but often need to be expanded in order to cover more specific information. Domain-specific schemas have a much richer vocabulary and structure but tend to be highly specialized and only understandable by researchers in that area.

Under the InAdvance project, the following general purpose metadata standards are proposed to be used:

- **CERIF** (Common European Research Information Format)[5]. It is the standard that the EU recommends to its member states for recording information about research activity. The physical CERIF model is a relational database model available as SQL (Structured Query Language) scripts based on common ERM (Entity Relationship Model) constructs. The CERIF format offers a model to structure the research domain into relevant objects and their relationships. Moreover, with the Semantic Layer it provides a powerful means for the management of contextual semantics.
- **DCMI** (Dublin Core Metadata Initiative)[6]. Small set of vocabulary terms that can be used to describe digital resources (video, images, web pages, etc.), as well as physical resources such as books or CDs, and objects like artworks. It may be used for multiple purposes, from simple resource description to combine metadata vocabularies of different metadata standards, to provide interoperability for metadata vocabularies in the linked data cloud and Semantic Web implementations.
- **MODS** (Metadata Object Description Schema)[7]. An XML-based bibliographic description schema developed by the United States Library of Congress' Network Development and Standards Office.

Moreover, there are a wide range of metadata standards for the management of research data and materials generated under specific disciplines. Thus, due to the clinical and health care research topics of the InAdvance project, it is expected to use the following specific purpose metadata standard:

- **SDMX-HD** (Statistical Data and Metadata eXchange-Health Domain) is a WHO implementation of the SDMX standard to allow medical facilities to share and exchange medical indicators and metadata between medical organizations.

2.1.2 File naming

Structured data storage is essential for a proper and secure storage of data files and records. For any file-based storage this includes clear and unambiguous file naming, the use of proper versioning as well as clear and intuitive folder structure. In this sense, a file naming convention (FNC) helps the project files stay organized by making it easy to identify the file(s) that contain the information from its title and by grouping files that contain similar information close together. A good FNC

[5] https://www.eurocris.org/cerif/main-features-cerif
[6] http://dublincore.org/
[7] https://www.loc.gov/standards/mods/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

also helps the consortium members to better understand and navigate through the work performed by other partners.

InAdvance will encourage the use of a standard FNC per data type and a clear versioning strategy. Table 3 shows some basic rules for developing InAdvance FNC:

Table 3 Guidelines for Developing a FNC

- Find the **right balance of components**. Too few components create ambiguity; too many limit discovery & understanding.
- **Avoid extra long folder names** and complex hierarchical structures but use information - rich filenames instead.
- **Abbreviate the content** of elements whenever possible and use meaningful abbreviations. File names that contain too many characters can be unwieldy and cause problems in transferring files.
- **Use the underscore (_)** as element delimiter. Do not use spaces or other characters such as: ! # $ % &**.**
- **Use the hyphen (-)** to delimit words within an element or capitalize the first letter of each word within an element.
- **Document your decisions** including what components you will use (the "project name" for example), what are the appropriate entries, what acronyms mean, etc.
- **Files will be grouped together** based on the first few components so start the FNC with the more general components and move to the more specific ones later on. Dates should always be yyyy-mm-dd to organize files chronologically.
- **FNC will be shared** among all partners of the project in order all individuals who needs to use the FNC are aware of it and knows how to apply it. A file naming convention breaks down if not followed consistently.

Content based on [8] and [9].

Within InAdvance a set of specific guidelines for file naming for the different data types will be drawn up which will be adapted to the specific RAs. Once these guidelines are finalized and validated, they will be made available to all partners to follow. The project coordinator will encourage partners about the importance of diligently applying the FNC and that deviating or compromising in guidelines adherence could have consequences to the project. Therefore, FNC adherence will be monitored as much as possible and if this becomes problematic, it is important to be aware that semi-automated structured file naming and storage software exists that can be used to help enforce and monitor adherence to naming and storage guidelines.

[8] Purdue University, Libraries and School of Information Studies: http://guides.lib.purdue.edu/c.php?g=353013&p=2378293

[9] Santaguida, V. (2010). Folder and file naming convention – 10 rules for best practices. Available at: http://www.exadox.com/files/pdf/en/Folder-File-Naming-Convention-10Rules-Best-Practice.pdf

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

2.2 Making data openly accessible

In the framework of the H2020 programme, the European Commission (EC) has launched a flexible pilot for Open Access to Research Data (ORD pilot) aimed at improving and maximising access to and re-use of research data generated by H2020 projects. The ORD Pilot applies primarily to the data needed to validate the results presented in scientific publications. Other data can also be provided by the beneficiaries on a voluntary basis. These guidelines will be applied taking into consideration the following:

- the need to balance openness and protection of scientific information;
- commercialisation and IPR (Intellectual Property Rights);
- privacy concerns;
- security; and
- data management and preservation questions.

Thus, the EC recognizes that some research data cannot be made open and applies the principle of *'as open as possible, as closed as necessary'*. It is therefore possible to opt out of research data sharing at any stage - before or after the signature of the grant agreement - but reasons have to be given e.g. for IPR concerns, privacy/data protection concerns, national security concern, if it would run against the main objective of the project or for other legitimate reasons (see General Annex L of the 2017 Work Programme adopted at 25 July 2016).

It is important to highlight that clinical trials data sharing will take into consideration the ethics requirements at each of the corresponding sites (e.g. in some cases, sharing may be limited to a specified storage location and listed research team).

In this regard, a Data Access Committee (DAC) will be specifically created in order to make proposals and take decisions about InAdvance data release to external requestors.

2.2.1 Open Access

According to the ORD pilot guidelines of the H2020 programme, all data collected or produced within the InAdvance project will be open by default. Some data may be shared only under restriction – for instance, at consortium level –, but providing a motivation for non-open access will be a requirement.

As a general guideline, InAdvance will encourage the whole consortium to publish research papers at peer-reviewed open access journals. More specifically, at this stage of the project, the following considerations have been already made regarding the open access of the data generated under the different RAs:

- **RA1:** The results obtained during the development of risk stratification models will be published (according to the IPR protection strategy) in dedicated conferences and journals with JCR index in health research, medical informatics, machine/deep learning and palliative care.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

- **RA2 / RA3:** The results obtained during the development of PC needs analysis and supportive ecosystems implementations will be published (according to the IPR protection strategy) in dedicated journals and peer reviewed conference, complying with the appropriate data access strategies.

- **RA4:** Using existing data from original clinical trials is more efficient than collecting new and additional data. Sometimes it is possible to analyse outcomes that were not included in the primary analysis or it enables meta-analysis of individual data. Also, it is favoured because of the transparency of the investigations and avoids patients' risks related directly to the investigations. However, data sharing must be carried out following indications and guidelines that strive especially for the confidentiality of data. Data de-identification for RA1/RA4 data set will be developed under the following requirements[10] [11]:
  a) Definition of the initial data sharing pack: the initial data-sharing pack should include the study protocol, annotated data collection forms, the statistical analysis plan, final report and data dictionary. Any relevant amendments during the study must be noted.
  b) Define a unique identifier for each trial participant using random number generator methodology (i.e. N bits hash or universally unique identifier – UUID). The link between the new code number and the original unique identifier should be maintained by primary researchers in case of any queries relating to the de-identified dataset from secondary researchers. Secondary researchers will not have access to original unique identifier.
  c) Categorize data as direct identifiers, indirect identifiers and superfluous information depending on the risk that individuals can be identified. Direct identifiers can immediately identify a participant or pose a very real risk that individuals can be identified. Indirect identifiers are those that may present a risk if present in combination with others. Problem variables must either be removed or modified, so that the risk of identification is diminished.
  d) Quality of data de-identification though a 'motivated intruder' test. The motivated intruder test involves considering whether an 'intruder' would be able to achieve re-identification of a participant if motivated to search for it. At least, two intruders should be designed. Intruders will act as independent reviewers of the dataset. One of them should have background knowledge about the clinical trial and one without knowledge.
  e) Controlled access process: The requestor must prove that she/he is a bona fide researcher with a sensible question to answer, and appropriate data-sharing agreements must be signed.

[10] Keerie, C., Tuck, C., Lewis, S., Milne, G., Eldridge, S., Wright, N. & Lewis, S.C. (2018). Data Sharing in Clinical Trials–Practical guidance on anonymizing trial datasets. *Trials, 19*(1): 25.
[11] Smith, C. T., Hopkins, C., Sydes, M. R., Woolfall, K., Clarke, M., Murray, G., & Williamson, P. (2015). How should individual participant data (IPD) from publicly funded clinical trials be shared? *BMC medicine, 13*(1), 298.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

Clinical trials data sharing will take into consideration the ethics requirements at each of the corresponding study sites (e.g. in some cases, sharing may be limited to a specified storage location and listed research team). For instance, in the United Kingdom (and expectations of in all other partner territories), for all patient data, access will only be made available in anonymised form unless individual participant written informed consent is given for data to be transferred to Partners. In addition, any such system will be subject to the UK Information Governance Tool complete process and approval at each participating Health Board / Health Trust within the UK, through the Information Governance Office and the IG Director authorisation system.

As the project moves forward, more specifications and details will be established regarding the open access of data collected and generated under the five RAs.

- **RA5** will provide open access results, even if the input data are not open access (for example, restricted to only consortium access), since the appraisal dashboard will process the data and provide only the results of such process and not the initial data.

As the project moves forward, more specifications and details will be established regarding the open access of data collected and generated under the five RAs.

2.2.2. Storage

The storage of the data collected and generated under the RAs carried out at InAdvance will primarily be the responsibility of the joint data controllers. However, InAdvance will select a data repository where all partners will be encouraged to store research data generated.
At this stage of the project, some potential international data repositories have been identified in order to store InAdvance research data (see Table 4). This list of repositories will be refined and/or updated for the second release of the DMP.

Table 4 Potential data repositories for InAdvance research data

Data repository	Subjects	Content types	URL
Global Health Observatory Data Repository	Medicine Public Health, Health Services Research, Social Medicine Life Sciences Medicine	Scientific and statistical data formats Raw data Standard office documents Structured text Structured graphics	http://www.who.int/gho/database/en/
HealthData.gov	Life Sciences Medicine Medicine Public Health, Health Services	Raw data Scientific and statistical data formats	https://healthdata.gov/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

	Research, Social Medicine		
ClinicalCodes.org	Public Health, Health Services Research, Social Medicine Medicine Medicine Life Sciences	Standard office documents Raw data Plain text other	https://clinicalcodes.rss.mhs.man.ac.uk/
Centers for Disease Control and Prevention, Data & Statistics	Medicine Epidemiology, Medical Biometry, Medical Informatics Public Health, Health Services Research, Social Medicine Social Sciences Psychology Basic Biological and Medical Research Medicine Life Sciences Social and Behavioural Sciences Humanities and Social Sciences Biology	Standard office documents Images Structured graphics Audiovisual data Scientific and statistical data formats Raw data Archived data Plain text	http://www.cdc.gov/DataStatistics/
Agency for Healthcare Research and Quality, Data & Surveys	Medicine Public Health, Health Services Research, Social Medicine Social Sciences Economic and Social Policy Statistics and Econometrics Life Sciences Medicine Social and Behavioural Sciences Humanities and Social Sciences Economics	Standard office documents Structured graphics Scientific and statistical data formats Plain text Raw data Archived data	https://www.ahrq.gov/research/data/index.html
International Clinical Trials Registry Platform	Life Sciences Medicine	Raw data Databases	http://www.who.int/ictrp/en/

2.2.3 Data Access

Several health-related repositories are listed at Table 4 that can be used by any researcher of the project to store or obtain data. Additionally, general or national repositories such as Zenodo [12] or DIGITAL.CSIC [13] may be also consulted throughout the different activities of the project.
The documentation describing how to find, get and access data will be described and stored in the selected data repository or software documentation. Thus, no additional documentation will be provided by the InAdvance project.

For the cases where access to the data needs to be restricted this can be identified at the time the data is submitted to the data repository. Registration to the selected repository will be necessary as well as access via password in order to obtain InAdvance data. For restricted data there will be a contact point for requests for data access.

Formats of data to be produced by the InAdvance consortium at the RAs can be handled by commonly used commercial or open software. Thus, data will be accessible using the same community-standard software.

2.3 Making data interoperable

<<*Interoperability is the ability to access and process machine-readable data from multiple sources, sometimes automatically, without that data losing meaning or integrity*>>[14]. In order to achieve this goal, data requires:

- **Syntactic interoperability:** widespread adoption of standard data formats, and the implementation of application programming interfaces and connectors that allow data from multiple sources to be accessed and integrated;
- **Semantic interoperability:** data and information must be exchanged across systems without its context and meaning being lost; and
- **Search interoperability:** it enables a user to conduct queries across two or more collections of data.

Thus, interoperability enables improved data usability and use. For this reason, putting in practice interoperability will be a priority for InAdvance in order to have the following benefits (among others):

- reducing time, effort and expense spent on data collection;
- eliminating frustration and risks associated with finding inconsistent and incomplete data;
- making available sustainable, disaggregated data for effective decision-making.

Table 5 presents a set of interesting guiding principles for interoperability to be implemented through InAdvance.

[12] https://zenodo.org/
[13] https://digital.csic.es/?locale=en
[14] https://www.publishwhatyoufund.org/wp-content/uploads/2017/11/JUDS_Report_Web_061117.pdf

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

Table 5 Guiding principles for interoperability

- **Use and re-use existing standards:** no new standards should be created in areas they already exist unless absolutely necessary. When new standards are created, they must be compatible with existing standards.
- **Do not overlook metadata:** metadata is crucial for discoverability, accessibility and fostering trust and understanding of the context in which the data was produced.
- **Use common classifications wherever possible:** it is important to ensure that to the extent possible, the language used to define and classify that data is the same.
- **Publish data in machine-readable formats:** it enables a computer to access, identify and filter data in an automated way.
- **Ensure that standards are user driven:** for data to be usable, it has to be driven by the needs of users themselves. It is crucial to long-term success and impact.

Based on Open Data Watch[15].

2.3.1 File formats

An overview of possible data formats and types for sharing, re-use and preservation of data to be used under the InAdvance project is presented at Table 6.

Table 6 Overview of possible file types for sharing, re-use and preservation of data

Type of data	Recommended formats	Acceptable formats
Quantitative tabular data with extensive metadata *Variable labels, code labels, and defined missing values*	• SPSS portable format (.por) • Delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) • Structured text or mark-up file of metadata information, e.g. DDI XML file	• Proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/.accdb)
Quantitative tabular data with minimal metadata *Column headings, variable names*	• Comma-separated values (.csv) • Tab-delimited file (.tab) • Delimited text with SQL data definition statements	• Delimited text (.txt) with characters not present in data used as delimiters • Widely-used formats: MS Excel (.xls/.xlsx), MS Access

[15] Open Data Watch: http://opendatawatch.com/blog/what-are-the-principles-of-joined-up-data/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

		(.mdb/.accdb), dBase (.dbf), OpenDocument Spreadsheet (.ods)
Qualitative data *Textual data*	• Rich Text Format (.rtf) • Plain text, ASCII (.txt) • eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema	• Hypertext Mark-up Language (.html) • Widely-used formats: MS Word (.doc/.docx) • Some software-specific formats: NUD*IST, NVivo and ATLAS.ti
Digital image data	• TIFF version 6 uncompressed (.tif)	• JPEG (.jpeg, .jpg) but only if created in this format • TIFF (other versions) (.tif, .tiff) • Adobe Portable Document Format (PDF/A, PDF) (.pdf) • Standard applicable RAW image format (.raw) • Photoshop files (.psd)
Digital audio data	• Free Lossless Audio Codec (FLAC) (.flac)	• MPEG-1 Audio Layer 3 (.mp3) if original created in this format • Audio Interchange File Format (.aif) • Waveform Audio Format (.wav)
Digital video data	• MPEG-4 (.mp4). • OGG video (.ogv, .ogg) • Motion JPEG 2000 (.mj2)	• AVCHD video (.avchd)
Documentation and scripts	• Rich Text Format (.rtf) • PDF/UA, PDF/A or PDF (.pdf) • XHTML or HTML (.xhtml, .htm)	• Plain text (.txt) • Widely-used formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx) • XML marked-up text (.xml) according to

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

	• OpenDocument Text (.odt)	an appropriate DTD or schema, e.g. XHMTL 1.0

Content based on UK Data Service[16].

As stated at Table 1, InAdvance RAs entails a wide variety of data and, consequently, different file formats and software. Table 7 lists the file formats by type of data that will be used by the InAdvance consortium.

Table 7 File formats by type of data to be used in InAdvance

RA	Data type	Data file types
RA1	Quantitative data	Comma-separated values (.csv) Tab-delimited file (.tab) Excel (.xls/.xlsx) SPSS (.sav)
RA2	Qualitative data	MS Word (.doc/.docx) PDF (.pdf) NVivo and ATLAS.ti
	Audio record	MPEG-1 Audio Layer 3 (.mp3)
RA3	Quantitative data	Excel (.xls/.xlsx) SPSS (.sav)
	Qualitative data	MS Word (.doc/.docx) PDF (.pdf) NVivo and ATLAS.ti
RA4	Quantitative data	Excel (.xls/.xlsx) SPSS (.sav)
	Qualitative data	MS Word (.doc/.docx) PDF (.pdf) NVivo and ATLAS.ti
	Audio record	MPEG-1 Audio Layer 3 (.mp3)
RA5	Quantitative data	Excel (.xls/.xlsx) CSV (.csv) Delimited text(.txt)
	Qualitative data	Excel (.xls/.xlsx) CSV (.csv) Delimited text(.txt)

2.3.2 Vocabulary standards

In addition to selecting a metadata standard, whenever possible it is important using a controlled vocabulary, which provides a consistent way to describe data. Using a controlled vocabulary will aid in searching and finding data and will make research data more shareable among researchers in the same discipline.

In the field of health sciences and medicine, InAdvance will encourage the use of the following controlled vocabularies proposed by international initiatives:

[16] UK Data Service. Recommended formats. Available at: https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

- **ICD** (International Classification of Disease) [17]. International standard diagnostic classification for all general epidemiological, many health management purposes and clinical use.
- **MeSH** (Medical Subject Headings) [18]. Hierarchical organization of terminology for indexing and cataloging of biomedical information such as MEDLINE/PubMed and other National Libraries of Medicine databases.
- **SNOMED-CT** (Systematized Nomenclature of Medicine) [19]. Systematic organization of computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. It is considered the most comprehensive, multilingual clinical healthcare terminology in the world.
- **LOINC** (Logical Observation Identifiers Names, and Codes) [20]. Database and universal standard for identifying medical laboratory observations.
- **Rx Norm** [21]. US-specific terminology in medicine that contains all medications available on the US market.

In order to minimize ambiguity and ensure the adequate interpretation of individual data items, they will be described and mapped ensuring interoperability between the different clinical sites as well as between the different RAs.

Thus, InAdvance will create a Data Mapping Specification, as a data dictionary, that is the process of mapping data fields from a source file to their related target fields. This procedure can be complex and time-consuming. Several data mapping techniques will be taken in consideration under InAdvance project:

- Manual data mapping: hand-coding the mappings between the data source and target database.
- Semi-automated data mapping: identification of two data objects that are semantically related and then building mappings between them.
- Automated data mapping: complete code-free environment for data mapping tasks by using automated data mapping tools.

2.4 Making data re-usable

Data re-use involves using research data for a research activity or purpose other than that for which it was originally intended.

Before data generated under the InAdvance project is made openly accessible and, thus, re-usable, the consortium will de-identified all data removing any personal information (or other kind of information) that may allow the identification of a person. Joint controllers, with the support and advice of the DPO (Data Protection Officer) and Advisory Boards, will establish the conditions to make date re-usable.

[17] https://www.who.int/classifications/icd/en/
[18] https://www.nlm.nih.gov/mesh/meshhome.html
[19] http://www.snomed.org/
[20] https://loinc.org/
[21] https://rxnav.nlm.nih.gov/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

Clinical partners will keep de-identified mapping tables for data review in case relevant information for patients is discovered.

2.4.1 Long-term use of data

The final goal is to make data openly accessible for as long as possible. The decision about the long-term provision of data collected and processed under the InAdvance project will be taken as the data are stored.

After completing all the RAs the data collected during the InAdvance project will be openly accessible if the following requirements are met:

a) Data collection and processing are completed.
b) Data checking – in terms of quality control – is performed.
c) After the completion of consortium partners' exploitation plan both in terms of scientific publications and commercial purposes.

A possible period to make InAdvance data openly accessible may be 2 years after data collection. However, this time frame will be discussed and agreed by the project General Assembly (GA).
Discussions about licenses and terms of re-use will be also done by the project GA, with the support and/or supervision of the DPO and the Ethics Advisory Board, if necessary.

Finally, data produced during InAdvance will be re-usable for as long as the information they contain are relevant for further research purposes. Specific discussions at this regard will be done also by the project GA.

2.4.2 Data quality control

Controlling the quality of data collected under any research is a critical part of the data collection procedures. Data need to be high quality in order the project outcomes and decisions from them can be done on the basis of reliable and valid data.
Moreover, data quality is a very important aspect in order to use them for purposes and actions beyond the project ones.

Quality control of data is an integral part of all research and takes place at various stages: during data collection, data entry or digitisation as well as data checking.

During data collection, InAdvance researchers must ensure that the data recorded reflect the actual facts, responses, observation and actions under study. Table 8 includes some quality control measures to be taken into consideration during data collection at InAdvance actions.

Table 8 Quality control measures during data collection

• Calibration of instruments to check the precision, bias and/or scale of measurement. • Taking multiple measurements, observations or samples. • Checking the truth of the record with an expert.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

- Using standardized methods and protocols for capturing observations, alongside recording forms with clear instructions
- Computer-assisted interview software to: standardize interviews, verify response consistency, route and customize questions so that only appropriate questions are asked, confirm responses against previous answers where appropriate and detect inadmissible responses.

Content based on UK Data Service[22].

Some of the InAdvance RAs entail qualitative data collection methods, such as individual or group interviews. In this regard, InAdvance will take into consideration several recommendations for a correct collection and subsequent process of data[18]:

- The level of sound or picture quality needed.
- The budget available for equipment and related consumables.
- How quickly the technology being used will become redundant.
- Whether consent is in place to allow fullest use of recordings.
- How the data created will be used.
- Whether data or information not allowed by consent can be excluded from recording.
- Whether the equipment will be simple to operate in the field

It is important to take measures in order to create audio recordings of good quality, which increases their usefulness. Good quality sound recordings should prevent mis-transcription and reduces the chance of sections of an interview remaining transcribed due to poor sound quality.

When InAdvance data are entered in a database or spreadsheet (through transcription, digitalization or codification), quality will be ensured by standardised and consistent procedures for data entry with clear instructions. This may include:

- Setting up validation rules or input masks in data entry software.
- Using data entry screens.
- Using controlled vocabularies, code lists and choice lists to minimise manual data entry.
- Detailed labelling of variable and record names to avoid confusion.
- Designing a purpose-built database structure to organise data and data files.
- Accompanying notes and documentation about the data.

Finally, quality of data can be annualized during data checking; that is when data are edited, cleaned, verified, cross-checked and validated. Some recommendations to be taken into consideration by the InAdvance consortium during data checking are stated at Table 9.

[22] UK Data Service: Quality Assurance. Available at: https://www.ukdataservice.ac.uk/manage-data/format/quality.aspx

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

Table 9 Quality control measures during data checking

- Double-checking coding of observations or responses and out-of-range values.
- Checking data completeness.
- Adding variable and value labels where appropriate.
- Verifying random samples of the digital data against the original data.
- Double entry of data.
- Statistical analyses such as frequencies, means, ranges or clustering to detect errors and anomalous values.
- Correcting errors made during transcription.
- Peer review.

Content based on UK Data Service[18].

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

3. Allocation of Resources

3.1 Costs for making data FAIR

Costs related to open access to research data in the Horizon 2020 programme are eligible for reimbursement during the project duration under the conditions defined in the H2020 Grant Agreement.

Resources for long-term preservation of datasets will be ensured by its storage in the repositories preliminarily outlined at Table 3.

Costs for making scientific publications open access may vary from $150 and $5000, depending on the journal. Converted to Euros, the average author fees for some of the most relevant publishers are: €1,637 by Elsevier's journals, €2,212 in the case of PLoS, €2,112 by Wiley's OA journals, €1,771 by BioMed's, but still higher in the case of Springer and Taylor & Francis journals[23].

These costs will be covered by the InAdvance budget.

3.2 Responsibilities

Data management responsibilities under InAdvance project are based on the following figures:

- **Joint controllers.** In the project there are joint controllers as there are several partners that jointly determine the purposes and means of data processing (art. 26 of the GDPR [24]). Controllers shall implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with the European GDPR.
- **Data processors** process personal data only on behalf of the controller.
- **DPO** provides expert professional knowledge in data protection law and IT security. The duties of the DPO includes: working towards the compliance with all relevant data protection laws; monitoring specific processes, such as data protection impact assessments; increasing employee awareness for data protection and training them accordingly; and collaborating with the supervisory authorities[20]. InAdvance has already appointed a DPO on January 2019 which has been duly communicated to the EC through the deliverable 10.7 (*'Appointment of the DPO'*).

Also, a **DAC** will be established specifically to address issues related to data release to external requestors.

[23] https://openscience.com/how-much-do-top-publishers-charge-for-open-access/
[24] Regulation (EU) 2016/679 (General Data Protection Regulation). Available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32016R0679

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

4. Data Security

Data security refers to the process of protecting data from unauthorized access and data corruption throughout its lifecycle.

Physical security, network security and security of computer systems and files will be considered and encouraged among the InAdvance consortium in order to guarantee security of data as well as to prevent unauthorized access, changes to data, disclosure or destruction of data (see Table 10).

Table 10 Security measures

Physical data security: • Controlling access to buildings, rooms, cabinets where data, computers, media or hardcopy materials are held. • Logging the removal of, and access to, media or hardcopy material in store rooms. • Transporting sensitive data only under exceptional circumstances, even for repair purposes; for example, giving a failed hard drive containing sensitive data to a computer manufacturer may cause a breach of security.
Network security: • Not storing sensitive data such as those containing personal information on servers or computers connected to an external network, particularly servers that host internet services. • Firewall protection and security-related upgrades and patches to operating systems to avoid viruses and malicious code.
Security of computer systems and files: • Locking computer systems with a password. • Ensuring computer software is up-to-date. • Protecting servers by power surge protection systems through line-interactive uninterruptible power supply (UPS) systems. • Implementing password protection and controlled access to data files, for example 'no access', 'read only', 'read and write' or 'administrator-only' permission. • Controlling access to files, folders or entire hard drives encryption. • Not sending personal or confidential data via email or other file transfer means without first encrypting them. • Destroying data in a consistent manner when needed: deleting files and reformatting a hard drive will not prevent the possible recovery of data; consult our guidance on data disposal. • Imposing non-disclosure agreements for managers or users of confidential data.

Content based on UK Data Service[25].

Data processed under the InAdvance project will be handled with appropriate confidentiality and technical security, as required by law in the individual

[25] UK Data Service: Data security. Available at: https://www.ukdataservice.ac.uk/manage-data/store/security

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

countries and EU laws and recommendations. All clinical sites of the project have to follow a common data protection protocol as follow:

- Only authorised researchers within the involved organizations should be granted access to data;
- All data must be made to be anonymous;
- The personal data gathered will be stored in a safe and secure manner following globally accepted standard ISO/IEC20013;
- Participants must be treated with respect at all times and their anonymity protected;
- Personal paper-based details of participants must be kept in locked filing cabinets;
- Audio recording will be kept in locked cabinets;
- All information that could be used to identify the participant must be separated from the data permanently before analysis; and
- Participants have the right to prevent data processing that is likely to cause damage to themselves or anyone else;

Deliverable 10.8 (*'Description of security measures'*) specifically addresses security measures in order to avoid unauthorized access to the project data or to the equipment used for processing. In that deliverable, a set of general guidelines has been established in order to guarantee confidentiality, integrity and availability of data gathered in the framework of the project. Moreover, measures at partner level (those entities processing data at the clinical sites) are described.

Cloud-based storage – specifically Google Drive – will be used by the project consortium in order to store and share relevant documents at management, administrative and technical level, such as meeting agendas, draft and final versions of deliverables or dissemination activities. Nevertheless, InAdvance is aware that this type of cloud storages are not necessarily permanent or secure. For this reason, data collected from the different RAs that include high-risk information – such as files containing personal or sensitive information or that have a very high intellectual property or commercial value – will not be stored on them. As stated at section 2.2.2 of this deliverable, research data collected at RAs will be store at specific data repositories.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750

5. Ethical Aspects

InAdvance will be compliant with ethical guidelines as well as with the EU regulations regarding the protection of personal data. Thus, in order to adapt those guidelines and materialize in accordance with InAdvance objectives and characteristics, several deliverables are being or will be released facilitating that the whole consortium are aware and compliant with them. These deliverables are the following:

- **D1.7** Setting-up the Ethics Advisory Board (submitted at month 3).
- **D1.3** Ethical Guidelines (month 6).
- **D10.8** Ethic Requirement describing security measures to prevent unauthorized access to personal data (month 6).
- **D10.9** Ethic Requirement evaluating ethics risks related to data processing activities of the project.

Specifically, the deliverable 10.9 addresses risks associated with processing personal data. On the basis of the analysis included in it, a Data Protection Impact Assessment (DPIA) has been considered necessary to being conducted in the following months of the project. The DPIA will be a useful tool to support InAdvance to be compliant with the GDPR as well as to manage the risks to the rights and freedoms of participants involved in the project resulting from the processing of personal data by assessing them and determining the measures to address them. The DPIA will be submitted as an attachment of the updated version of this DPM in month 18.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825750