The ADLIFE Project has received funding from the European Union under the Horizon 2020 Programme, grant reference number 875209.

Horizon 2020
European Union Funding
for Research & Innovation

D1.1 Data Management Plan

Deliverable No.	D1.1	**Due Date**	31/08/2020
Description	This document is the Data Management Plan of the ADLIFE project. It presents the intended data processing and flows, and specifies the intentions of the project to make open research data available at the end of the project.		
Type	Report	**Dissemination Level**	CO
Work Package No.	WP1	**Work Package Title**	Co-ordination and management
Version	0.4	**Status**	Draft

Authors

Name and surname	Partner name	e-mail
Dipak Kalra	i~HD	dipak.kalra@i-hd.eu
Ana Ortega Gil	Kronigrune	aortega@kronikgune.org

History

Date	Version	Change
16/07/2020	0.1	Initial content for all sections
28/07/2020	0.21	Updated Figure 1 and revised text on data processing and flows
28/07/2020	0.22	Tidied version for partner review
05/08/2020	0.3	With changes and comments from Omar and Gökçe, incorporated by Dipak
08/08/2020	0.4	Incorporating feedback from Lisa, Rachelle, Anne and email discussions with several other partners

Key data

Keywords	Data management, data protection, open research data,
Lead Editor	Dipak Kalra
Internal Reviewer(s)	Omar Khan, Ana Ortega Gil, Lola Verdoy Berastegi, Gökçe B. Laleci Ertürkmen, Anne Dichmann Sorknæs, Lisa McCann, Rachelle Kaye

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Abstract

This document is the Data Management Plan of the ADLIFE project. It has been deliberately scheduled early in the project lifetime to explain how the project intends to acquire and collect health data via its seven pilot sites in order to help design the various components of the ADLIFE solution, and then to establish evidence of its benefit to patients and health systems through a large-scale pilot and clinical trial.

The first part of this deliverable presents the data processing and flows that will largely take place within each of the seven pilot sites, to identify patients who match the eligibility criteria for the ADLIFE study, and then how these data will be processed including pseudonymisation and anonymisation steps, before being made available for wider consortium use. This wider use includes the design and development of technical components, the training and validation of artificial intelligence algorithms and, later in the project, the generation of evidence of health outcomes and health economic impact from use of the ADLIFE solutions.

That chapter is a data-oriented summary of the clinical research protocol, which is itself separately documented in detail and will be submitted as a forthcoming project deliverable after approval by the ethics committees of the seven pilot sites. There are other project deliverables that will specify the information governance and data protection policies, information security policies and measures to be adopted. These will be primarily within three WP11 deliverables that were added in response to the EC Ethics Review. These items are only briefly mentioned in this report.

The second half of this document focuses on the formal Data Management Plan template published by Horizon 2020. This mostly confirms the intention of the project to make available some open research data at the end of the project, and how it intends to comply with the FAIR principles. Some of the responses to the template questions are provisional, because certain decisions about exactly what data is permitted to share with others, how the data will be made available and how it will be documented with suitable metadata will be determined later in the project.

There is a closing short section on other types of knowledge asset that will be developed in the project, most of which we hope to make available as open source or open access.

This Data Management Plan will be maintained as a living document throughout the project. If it is considered appropriate, a final version of this Data Management Plan will be published as an updated deliverable, in the final year of the project, to provide definitive answers to the template questions.

Table of contents

1 The ADLIFE project from a data management perspective

ADLIFE is a Horizon 2020 funded project developing innovative digital health solutions to support the healthcare planning and care delivery for patients with advanced (severe) long term conditions or multiple conditions (multimorbidity, with chronic obstructive pulmonary disease and/or heart failure). ADLIFE's solutions will include integrated health information for the multi-professional care team, personalised care planning with patient inclusion, a patient self-management and empowerment platform, and an AI-driven Early Warning System monitoring for acute health deterioration. ADLIFE aims to demonstrate positive patient and clinician experience of using the solutions, improved health outcomes including better quality of life and fewer hospitalisations, and a reduced burden of disease on the family and the health system.

The **ambition** of ADLIFE is to:

- demonstrate that the ADLIFE personalised care model can be deployed and replicated on a large scale in different environments and be trusted with regard to data access, protection and sharing;
- achieve quantified gains in health status, preventing unnecessary suffering (by qualitative analyses), slowing down clinical and functional deterioration (through functional assessment) and improving Patient Reported Outcomes (PROMs);
- obtain improvements in efficiency by making a better use of resources and increasing the coordination among care stakeholders;
- protect functionality and enhance autonomy, empowering patients to participate in decisions making on their own health and adapting to their changing conditions and context.

ADLIFE's technology innovations will be deployed, used and evaluated in seven healthcare environments in Spain, UK, Sweden, Germany, Poland, Denmark, Israel. The aim of these large-scale pilots is to demonstrate the effectiveness of the ADLIFE intervention when deployed in clinical real conditions. The project will generate evidence of benefits for the patients, the families and carers, the professionals and the health system.

These seven sites will each enrol patients into a multi-site quasi-experimental design (non-randomized, non-concurrent, controlled trial) and mixed method analysis, to enable the health outcome and economic impact of ADLIFE to be assessed. The sites will therefore be the primary project sources of patient level data for the trial. Their data will also be used for the technical developments such as the implementation and validation of the digital tools, including the training and validation of the AI components.

A robust approach to data protection and data management will be adopted prior to any contact with patients or their health data. This approach is summarised in this report and will be expanded in subsequent WP11 deliverables.

If it is agreed and approved that some individual level or aggregated health and care data can be made available as open data after the project, policies regarding its data management, storage, protection, access and ownership will be developed and included as an updated Data Management Plan before the end of the project.

2 ADLIFE Data Summary

This chapter summarises the categories of patient level data that will be processed through the project and the ways in which it will be processed.

Figure 1 presents a high-level diagram of the overall study design, the details of which are documented in the research protocol (D11.1).

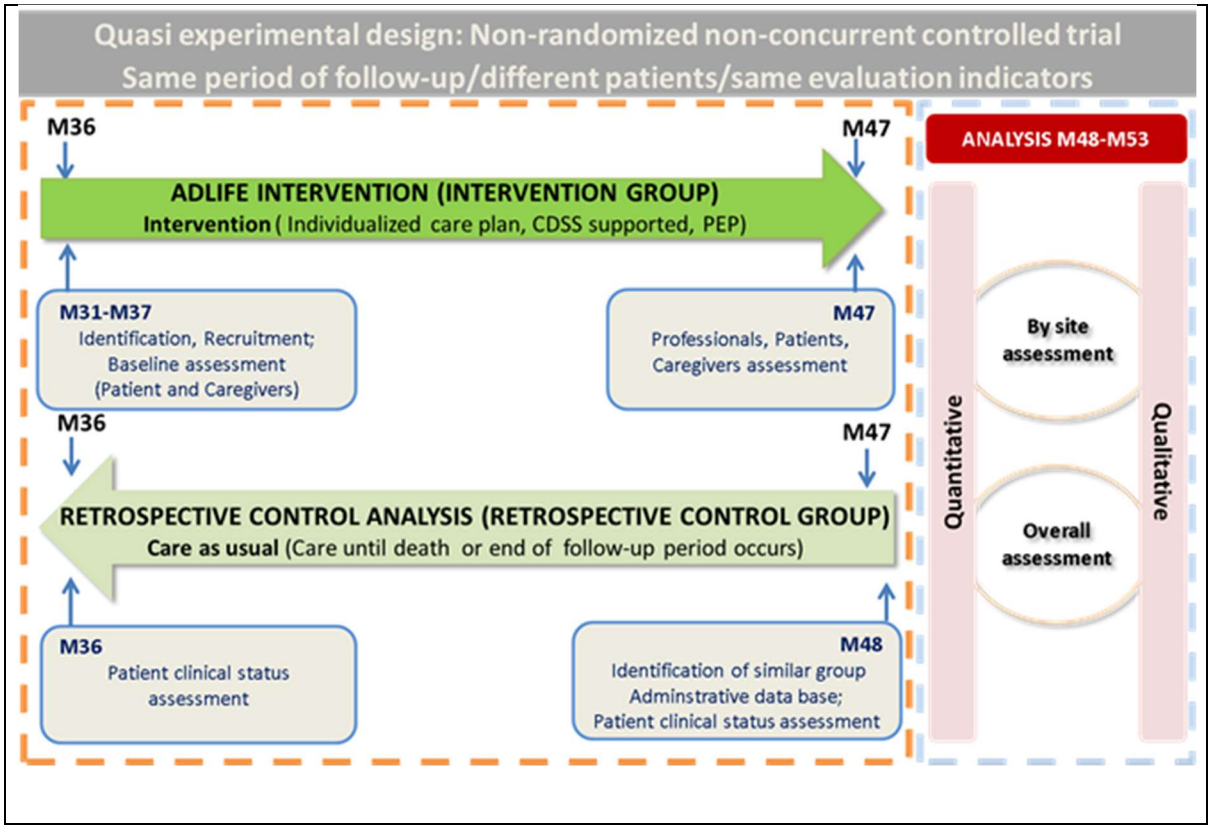

Figure 1: ADLIFE study design

The empirical methodology of the pilot study involves the allocation of eligible patients to intervention and control groups at seven sites. Informed consent will be obtained from the intervention group for the use of their health data, as they will also receive care that is supported by ADLIFE innovations. No consent will be obtained from the control group patients, and so their data may only be used anonymously, with ethics committee approval and subject to approved safeguards. Data from potentially eligible patients (prior to the allocation of patients to study arms) will be used for some system development activities, as training data. Mock data that is unrelated to real patients will also be used for some developments. The details of this empirical methodology are given in the ADLIFE research protocol (Deliverable 11.1, to be submitted in Month 12 of the project).

The information architecture of the ADLIFE study design involves the following categories of patient level data collection, processing and communication:

- Mock/synthetic healthcare data for software development and testing.
- Training healthcare data obtained without consent, if permitted by the site, for defining and refining the prediction models behind the clinical decision support systems (artificial intelligence algorithms). However, the developed algorithms will only be used in the care of patients who have given consent.
- Control healthcare data without consent and intervention healthcare and patient/health care professional reported data with consent, routinely collected for software usage and project evaluation purposes.

Figure 2 presents the planned data flows, focusing on the pilot sites which will do most of the patient-level data processing, and all of the personal data processing.

For readability, arrows are not shown on the diagram, but the data flows are vertical from top to bottom on the diagram. The vertical colour coded zones represent the data collections by category. The numbered steps on the edges of the diagram are elaborated in the text below, by number. The data tenants are represented by colour coded boxes to reflect the sovereignty of data in every step. The seven pilot sites are:

- Basque Country (Osakidetza), Spain
- NHS Lanarkshire, United Kingdom
- Region Jämtland Härjedalen, Sweden
- Werra-Meißner Kreis, Germany
- FALKIEWICZ Hospital (Lower Silesia), Poland
- Odense University Hospital, South Denmark
- Assuta Ashdod Hospital and Maccabi Healthcare Services Southern Reg, Israel

These are all consortium partners or contractual third parties. Each of them will act as the primary legal entity and data controller for the processing of personal data originating from their site on behalf of ADLIFE. In that data controller role, they will be responsible for all of the data objects and data flows shown in blue on Figure 2.

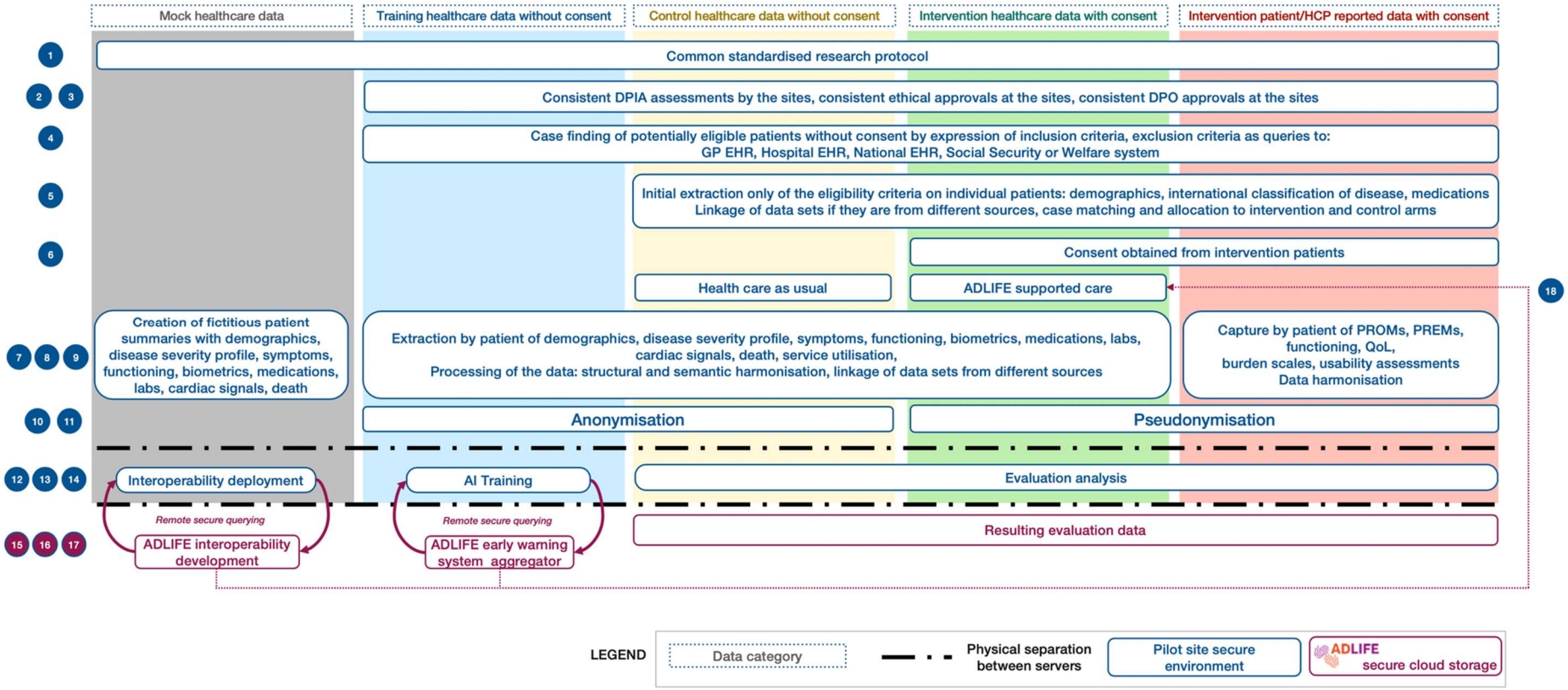

Figure 2: The envisaged data flows and processing within the ADLIFE pilot sites

Abbreviations used: DPIA = Data Protection Impact Assessment, DPO = Data Protection Officer, GP = General Practitioner, EHR = Electronic Health Record, PROM = Patient Reported Outcome Measure, PREM = Patient Reported Experience Measure, AI = Artificial Intelligence

2.1 The envisaged data flows and processing within the ADLIFE pilot sites

Please note that this is a data-oriented summary. The full details of the project's scientific methodology will be given in the research protocol (Deliverable 11.1, to be submitted in Month 12 of the project). The numbered points below relate to the numbers shown on the left and right edges of Figure 2.

1. The seven partners, alongside other partners in the consortium, are developing the research protocol that specifies the details of how the trial will be conducted. This includes the eligibility criteria, the recruitment methodologies, the data that will be needed for the study, how it will be processed and communicated - and in which forms - with other partners, and how it will be processed at the site and by other partners. Specifically, sharing of data with other partners is being documented using 'Data Sharing Storyboards' with the common theme that no identifiable personal data leaves the healthcare organisation/data controller. The protocol includes a brief summary of the data protection measures to be adopted, which are elaborated further in this data management plan and will be presented in detail in D11.1.

2, 3. Each of the sites will conduct a Data Protection Impact Assessment (DPIA) as required by the GDPR. They will be guided in the conduct of this through a generic DPIA template and support from partner Kronikgune, the co-ordinator and WP11 leader. Each site has to conduct a DPIA because it is the legal entity overseeing the processing of its data. The project itself is not a legal entity and cannot be a data controller. The partners are also collaborating on their ethics committee applications, so that these can also be consistent with each other. Kronikgune recommends presenting two separate applications, namely Figure 2 point **2.** for the training process of the predictive model on healthcare data without consent and Figure 2 point **3.** for the evaluation phase of the project on the control and intervention data. Sites will be also guided in this stage through generic application templates and support from Kronikgune.

Finally, and importantly, each site will consult and adhere to its data protection and information security policies and guidance. They might also require obtaining the approval of the local data protection officer to the research protocol, the intended data flows across the consortium and the data protection measures to be adopted.

4. The research protocol defines a set of inclusion and exclusion criteria for participant recruitment. All of the seven sites will run the same search on their electronic health record system in order to identify equivalent candidate recruitment pools of patients. Where necessary, the extraction query may be extended to other healthcare repositories such as local GPs, a national EHR system or other social security or welfare databases. However, it is likely that the case-finding query can be run on a limited number of repositories because the expression of inclusion criteria will only comprise a small number of data items. Despite the source, the query will need to be run by personnel authorised to access the electronic health record system, and the results of this extraction will need to be restricted to authorised personnel.

All eligible patients from this case-finding query will be utilised as a source of training data, extracted in Step 8, for the artificial intelligence algorithms. This dataset, which will be robustly anonymised, will also be used for guiding the development and initial testing of some of the other ADLIFE components: the data repository, the care pathway system and the patient empowerment platform (Step 7). Wherever possible, development work carried out on the technical partner's/secure cloud infrastructure will use synthetic data whereas testing at the sites will use anonymised or pseudonymised data. This will limit or avoid actual patient data (anonymised or not) from leaving each site.

The case-finding resulting data cannot be anonymised at this stage because there will be a need to identify patients who will be allocated to the intervention arm (Step 5). Only authorised personnel will have access to the identifiable information and only for the purposes of determining eligibility and allocation to a study arm.

5. It should be noted that this initial extraction of patient level data from the eligible candidates will need to be run prior to obtaining any informed consent. It will therefore need to be run by personnel authorised to access the electronic health record system, and the results of this extraction will need to be restricted to authorised personnel who will identify patients to be allocated to the intervention arm and will be directly contacted for recruitment. The remaining eligible patients may be allocated to the control arm whose (retrospective) electronic health record data will also be extracted. The methodology for case matching of patients to these study arms is defined in the research protocol.

6. Intervention arm patients will be contacted by telephone, email, post or face-to-face encounters, and briefed according to ethically approved Patient Information Sheets. Written informed consent will be sought from all of the patients allocated to the intervention arm. (Patients who do not consent will be excluded from the study). The patient information sheet and consent form template, and GDPR Transparency Notice are part of the ethics committee application which will have been approved prior to use. They will all be expressed in a locally relevant language. Informal caregivers may also be nominated by intervention arm patients to be included (not shown on Figure 2).

As a result of the above processes, each of the seven sites will have identified three sets of patients who all meet the inclusion and exclusion criteria: an intervention arm, a control arm and a (potentially overlapping) set of patients as a source of anonymised training data.

7. It is not necessary for the mock healthcare data to faithfully contain actual clinical data. Sites may therefore utilise the case-finding dataset as a basis for synthesising mock data which will predominantly be used for the technical development. The mock data needs to reflect the structure and semantics of the real data but need not be directly based on real patient data. Synthetic data that draws on the profile of real patient populations but does not actually include patient data might be used for some component development and testing (e.g. for the semantic interoperability suite).

8. A detailed dataset has been defined in the research protocol, but additional data items may be identified through interacting with the technical partners in the project and with the evaluation partners. It is expected that at least some data items would be extracted for the following categories of health and health care information, as indicated in Figure 2:

- demographics (especially age, gender, ethnicity)
- disease and its severity profile
- current symptoms
- functioning such as mobility
- biometrics such as blood pressure, pulse rate and rhythm, body weight
- medications
- laboratory test results, radiology investigation results, cardiac signals such as ECG
- information about if and when a patient has died and the cause of death
- healthcare service utilisation such as hospital admissions, clinic visits.

This information will be required retrospectively on the control arm patients and for extracting a training dataset from the source systems (but retained at each site). The same data would be extracted retrospectively and on a regular prospective basis for patients in the intervention arm. These data have been determined as being relevant and necessary for the conduct of the ADLIFE project and study, to comply with the data minimisation principle of the GDPR.

The ADLIFE study, in this context, means the development, testing, deployment and patient care use of components such as clinical decision support and MDT care planning (for intervention arm patients).

9. On patients in the intervention arm, there will be additional patient reported data collected through a combination of the patient empowerment platform and occasional questionnaires. This will include PROMs, PREMs, functioning, Quality of Life scores, measures of the burden of care. Both patients and healthcare professionals will complete standardised instruments: usability assessments and satisfaction questionnaires relating to the use of the ADLIFE system and its perceived benefits. Caregivers will also complete the burden of care, wellbeing and satisfaction questionnaires and interviews.

10. In order to make healthcare data accessible for research within the consortium, it needs to be protected. Intervention arm patients have given consent and have a GDPR legal basis to be used as personal data. Anonymization is necessary for control arm patients and for the training data since these will not have a GDPR legal basis. The anonymisation method will utilise the ARX[1] or OnFHIR Anonymisation[2] tools, according to each site's preference, which can both be configured to apply a number of statistical anonymisation methods including k-anonymity, blurring (or generalisation) and additionally to implement small cell size suppression and rounding of dates. Date of birth will be rounded to year of birth. Healthcare encounter dates will be converted into a date offset from an index date that is patient specific. Anonymization rules will be collated in a document by the partner Kronikgune, the co-ordinator and WP11 leader.

11. In order to link the retrospective healthcare data for the purposes of study evaluation, the prospectively extracted healthcare data and the patient reported data for each patient, these intervention arm datasets need to be pseudonymised by staff within the partner's healthcare organisation (data managers) who have authority to access personal health data. (For healthcare delivery to control and intervention arm patients, staff will continue to use identifiable data, see step **18**.)

Only the demographic traits required for the research will be extracted from the data sources, no other identifiers. The pseudonymisation will be undertaken using the ARX or OnFHIR Anonymisation tools, according to each site's preference. The mechanisms for linkage, such as linkage tables, will be kept securely. Pseudonymised rules will be collated in a document by the partner Kronikgune, the co-ordinator and WP11 leader.

The research protocol, supplemented by the wording and agreement of each consented patient, will determine what happens to already-collected data in situations where a patient withdraws from the study or dies.

12, 13, 14. The end result of all of this data processing, under the data controllership of each of the seven healthcare partners, will be 6 categories of anonymised/pseudonymised data:

- Mock healthcare data
- Training healthcare data
- Control arm healthcare data
- Intervention arm healthcare data
- Intervention arm patient reported data

[1] https://arx.deidentifier.org/
[2] https://onfhir.io/technology.html

- Intervention arm healthcare professional reported data

Data for the conduct of the trial will only be available to the site from which the data originated. Data for training will also be kept at each site (with no linkage between the sites and no centralised project repository). Partner EVERIS will employ a federated learning model approach. This will allow them to train the algorithms on data that remains local to each site, only aggregating the model parameters centrally. The data itself will not be held centrally. For study evaluation, the relevant individual-level data only needs to be made available to the evaluation partner, with no other consortium members needing access to even the anonymised data unless they have identified a legal basis and need to do so. (Aggregated analysis results will be shared and used by the consortium for scientific publications and dissemination.)

The preferred storage and usage for research within the consortium of the above six datasets will be in physical servers that are different from the ones where all the above processes took place but within the secure environment of each pilot site, under the terms of the ethical approval. These datasets will be transferred to these servers for partners to use for:

- the design, implementation and validation of the software components within the ADLIFE solution (Step 12)
- the training and validation of the AI used within the early warning system (Step 13)
- the population of ADLIFE components for the provision of healthcare, by MDT members, patients and caregivers (Step 14)
- the clinical evaluations of the impact of the ADLIFE solution on health outcomes and burden of care (Step 17)
- the health economic implications of the impact of the ADLIFE solution on cost of care (Step 17)

Much of the statistical analysis required for the research will be undertaken directly within these seven site-governed secure environments, using analysis and machine learning software that ADLIFE developer teams will also install within this secure environment at each site, to ensure no patient level data leaves the site, personal or anonymised.

15. The iterative communication process between site systems (Step 12) and ADLIFE developer teams at the development phase will allow the developer teams to access mock data for developmental and testing purposes in a preproduction environment. The ADLIFE secure cloud storage will be the repository of the ADLIFE components and allow the access of pilot sites and developers only for on-site testing and then for deployment purposes. The testing with real, pseudonymised data will be conducted at the site, by authorised personnel on a staging environment prior to going live.

16. The iterative communication process between site systems (Step 13) and ADLIFE developer teams will allow the recurring training on the anonymized healthcare data in the site secure environment. The generation of the risk prediction models which are the core of ADLIFE early warning systems will take place in the ADLIFE secure cloud storage, whereby the models are trained at the site, and only model parameters are aggregated within the ADLIFE secure cloud.

17. The evaluation team will proceed with the credible and feasible analysis on control and intervention data within the pilot secure environment. The results of such analysis, which will not comprise patient level data, including the clinical evaluations of the impact of the ADLIFE solution on health outcomes and burden of care and the health economic implications of the impact of the ADLIFE solution on cost of care will be uploaded to the ADLIFE secure cloud storage. They will then be used by consortium partners scientifically. Some aggregate data sets may also be made available as open research data.

Kronikgune will host the ADLIFE cloud storage. The access and processing of data from this ADLIFE centralised cloud platform will be audited.

18. The above numbered steps have described the way in which data will be processed in order to inform the development of the ADLIFE toolkit components, and to conduct the trial of their use. For patients in the control arm, healthcare delivery will continue as normal throughout the study period, and routinely collected health data will continue to be accumulated, which will be part of the eventually extracted control study data. Care for the intervention arm patients will be supported by the ADLIFE components. Healthcare staff will access identifiable patient data, as usual.

2.2 ADLIFE information governance instruments

In order to safeguard the data flows and data processing that has been described in Figure 1, the following instruments are being developed. They will be included in their first versions in forthcoming deliverables and will be formally adopted across the consortium before patient level data is accessed and before any patients are contacted for recruitment.

- The research protocol
- In each applicable language:
  - Patient and caregiver information sheets
  - Informed consent forms for patients and for caregivers
  - GDPR Transparency Notice
- Data Protection Impact Assessment template and guidance for completion by each pilot site
- Code of conduct for all consortium partners when dealing with personal data and anonymised data
- Data access and data sharing agreements, if needed by the sites
- Generic ADLIFE information security policy, for customisation at each partner site to additionally align with local policies
- Specific guidance on pseudonymisation and the minimum safeguards for this
- Anonymisation methodology and instructions for the use of the anonymisation tools
- Specification and instructions for configuration and use of the ADLIFE-operated secure cloud storage facilities, including the audit log and the monitoring of this. This will include instructions on appropriate use of tools, such as issue tracking, to avoid leaks of any personal data.
- The SPS component itself will ensure only authorised access to the ADLIFE solution (housing real data during the study) and will audit use of the system and access to patient data.

The project coordinator, Kronikgune, will store a copy of each of these instruments, for the record and for possible inspection by the European Commission if required. Kronikgune will additionally store copies of all ethics committee submissions and approvals, and any constraints or requirements imposed by the committees.

3 ADLIFE open research data and open access

The data sets and knowledge assets that will be created through ADLIFE and have potential for wider reuse or open access are:

1. One or more anonymised population health datasets of health and care information on patients with advanced long-term conditions and/or multimorbidity (chronic obstructive pulmonary disease and/or heart failure)
2. Aggregated health outcomes and health economic data that has been used within the study and its published results, and might be further reused by others
3. Clinical guidelines and decision support services that have been specifically tailored for advanced conditions and for combined use in cases of multimorbidity, expressed in a human readable form and as clinical decision support services
4. AI algorithms suitable for inclusion within an Early Warning System for patients with advanced long-term conditions
5. Academic publications, conference presentations and posters, and other dissemination materials that showcase the methodology, results and solutions of the project

3.1 Anonymised population health datasets

The project cannot commit at this stage to make anonymised population health data sets available as open access data. This is desirable but will be subject to ethical approval and organisational approval at the pilot sites, possibly also additional verification of the robustness of the anonymisation. The challenge we will face is that rich clinical data sets are very difficult to anonymise robustly. If the whole of the dataset cannot be made available, selected subsets of the data, which are not so rich, might be possible to release as open research data. This will be explored towards the end of the project when the dataset can be comprehensively assessed.

The responses given in the next section are based on the assumption that some data may be available for open data sharing. The project will therefore take care to ensure that relevant contextual metadata, in accordance with the FAIR principles, are captured and included within the data sets. Descriptive metadata to allow for discovery will be added later in the project. Data access arrangements, also in accordance with the FAIR principles, will also be specified. Data will be made open access via the EU Open Research Data Pilot (ORDP).

The Data Management Plan template specified by the European Commission has been completed for this category of data asset, in the next chapter.

3.2 Knowledge assets and publications

The numbered items 2 to 5 in the list above, covering aggregated data sets, clinical guidelines and decision support rules, algorithms and dissemination materials are discussed later in this report in a chapter on open access.

4 ADLIFE Data Management Plan template

This section is based on the Data Management Plan template provided by the European Commission for Horizon 2020 projects[3].

4.1 Data summary

Please see the previous section for a detailed description of the data processing activities to be undertaken, primarily at the pilot sites. The responses given in this section refer to the expected generation of anonymised population health data sets. These will have underpinned the main research results that will be published at the end of the project, and also have the potential to be reused by other researchers.

Aspect	Response/explanation
Purpose of the data collection/ generation and its relation to the objectives of the project	Health and care data collected from patients with advanced long-term conditions from seven healthcare sites across Europe and Israel, plus patient reported data on quality of life and health outcomes. The purpose of the data collection and processing is to conduct an evaluation of effects of a digital solution enabling cross sector care on health and healthcare utilisation for patients with multimorbid chronic conditions. The dataset will include patients who have utilized the ADLIFE project digital solutions, who will enable the health and economic evaluation of the solutions.
Types and formats of data generated or collected by the project	Database tables, the format of which will comply with the HL7 FHIR standard and will be documented in D3.1.
Any re-use existing data and how this will be done	The existing data will be electronic health record information on these patients, that had been collected as part of routine care in the years prior to the commencement of the project and its trial. Data will be extracted using the electronic solutions, either the ADLIFE solution for data included in the study, or the anonymisation tools outlined above. Data will be reused either for evaluation of the ADLIFE solution or to ensure the ADLIFE solution contains the full relevant clinical history of participants during the study.
The origin of such data	Hospital and general practice electronic health records, national electronic health records and Social Security or welfare system

[3] Full template available here for reference: https://ec.europa.eu/research/participants/data/ref/h2020/gm/reporting/h2020-tpl-oa-data-mgt-plan_en.docx

	records. The exact combination of which data sources will vary between the seven pilot sites.
Expected size of the data	To be confirmed by the end of 2022
Likely users of the data	Health and care professionals providing care to participating patients and/or involved in evaluating the ADLIFE digital solution. Public and population health researchers investigating the impact of long-term conditions and multi morbidity, health economists and health informatics researchers.

4.2 FAIR data

4.2.1 Making data findable, including provisions for metadata

Aspect	Response/explanation
Are the data produced and/or used in the project discoverable, identifiable and locatable by means of a standard identification mechanism	Once we have defined an anonymised dataset that can be made available as open research data we will liaise with the EC's OpenAire project (https://www.openaire.eu) and with the Fair4Health project (https://www.fair4health.eu) to make the data available.
What standard identification mechanism used (e.g. persistent and unique identifiers such as Digital Object Identifiers)	We will use DOI.
Is meta-data available through catalogue?	It will be, once the data that we will be sharing is finalised and the repository/portal that will be used has been decided.
Can meta-data be used for search?	Yes, it will be. Please see https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0040-1713684
Naming conventions used	We will align with the Horizon 2020 FAIR4Health on this point (one of our partners is in this project). We will adopt the naming conventions of HL7 FHIR. (https://wiki.hl7.org/FHIR_Guide_to_Designing_Resources#Naming_Rules_.26_Guidelines)
Clear versioning supported?	It will be.
Additional keyword search supported?	It will be.
What metadata will be created using which standards?	We will use HL7 FHIR Profiles as the metadata of the data to be shared. To be elaborated later in the project, on advice from FAIR4Health, and any other advice from the European Commission.

4.2.2 Making data openly accessible

Aspect	Response/explanation
Will data be made openly available as the default?	Once we have defined a dataset that we can robustly anonymise, we will make this available by default.
Which datasets will NOT be openly available and why?	We will not make data available if we believe, or we are advised, that it is not possible to robustly anonymise the data, because of distinctive patterns in the data due to some of the population profiles that are included.
How will the data & meta-data be made accessible (e.g. by deposition in stated repository)?	By deposition in the chosen repository. We expect to use the (partner) SRDC onFHIR repository as a means of sharing anonymised data and its metadata.
If known repository, what arrangements explored?	To be determined later in the project.
If project-specific access, then:	To be determined later in the project.
– Data Access Committee	A committee comprising some or all of the partners involved in the project will determine the policies for which data will be made open access.
– Any conditions for access (i.e. a machine-readable license)	The conditions that the committee will determine will include confirming which data sets (patient level and aggregate level, all robustly anonymised) are suitable for open data access and ensuring that the necessary approvals have been obtained from the originating sites. The committee will determine the time interval after the project when the data will be made available, how it will be discovered and accessed, the open access licence terms that will apply and how data access will be requested and granted. The committee will also make long-term sustainability decisions, including a sustainable business model for the data sets.
What methods or software tools will be needed to access the data?	To be determined later in the project. We expect to use the (partner) SRDC onFHIR repository as a means of sharing anonymised data, and guideline modelling tools as appropriate once the language for CIGs has been finalised.
– Documentation for software	This will be provided.
– Availability of software	The SRDC onFHIR repository, which is shared as Open Source on GitHub
Institution and researcher vetting process/procedures - describe	To be determined later in the project.

4.2.3 Making data interoperable

Aspect	Response/explanation
Are the data produced in the project interoperable	Interoperability standards will be adopted by design, including the use of HL7 FHIR, in order to harmonise the data coming from seven different pilot site. This is necessary for the project itself, for the data analytics work that will be undertaken, but also serves the benefit that any research data we make openly available later will also be standardized.
If not, explain why not	N/A
Data and metadata vocabularies, standards or methodologies used	To be determined later in the project. Within the project, for coded data, we will be adopting ICD10 and ATC for diagnoses and medications, and will be mapping site specific terminologies to these. It is expected that any patient level dataset shared would include these terminologies.
Standard vocabularies used	To be determined later in the project.
Mappings from uncommon or project-specific ontologies or vocabularies to more commonly used ontologies	We do not anticipate the need to adopt uncommon ontologies or vocabularies.

4.2.4 Increase data re-use (through clarifying licences)

Aspect	Response/explanation
Will data be available for onward data-sharing/re-use?	Yes
Approach to data licensing for onward use	To be determined later in the project, but it will aim to facilitate onward sharing.
Likely date for data availability for onward use	During 2024
Explain any restriction on date of availability	None is anticipated.
Possible restrictions on onward data-sharing	To be determined later in the project, but it will aim to facilitate onward sharing.
Data retention policy (including availability for data-sharing)	Will follow data retention policies as determined at sites. Evaluation/open access data will be retained according to EC guidance.
Description of data quality assurance processes	To be determined later in the project.

4.2.5 Allocation of resources

Aspect	Response/explanation
Estimated project costs for making data FAIR	This will be determined later. However, we do not anticipate substantial costs because we will be establishing the anonymized data repository according to standards and with suitable metadata as an intrinsic part of the project. We will use the Open Source OnFHIR Repository, which is supported by SRDC.
Data management responsibility across the project	The co-ordinator, Kronikgune, will take lead responsibility for this, but other technical partners will support.
Resources required for long term preservation (costs and potential value, who decides and how what data will be kept and for how long)	To be determined later in the project, under the responsibility of the co-ordinator, Kronikgune. This would include a sustainability business model for data sets that will be made available open access, covering costs of long-term storage, the maintenance of the data sets if necessary, and any human resources required to manage data sharing.

4.2.6 Data security

Aspect	Response/explanation
Data security measures used (including data recovery as well as secure storage and transfer of sensitive data)	As indicated earlier in this report, a number of information governance and data protection and information security instruments are being developed and will be included within forthcoming project deliverables. Additionally, a core component of the ADLIFE solution will be focussed on ensuring security and privacy of the solution (SPS). The onFHIR repository, integrated with SPS components, will enable authentication, authorization, anonymization and audit logging. As a minimum, no personal, identifiable data will leave study sites.
Where data will safely be stored (in certified repositories for long-term preservation and curation). Provide detail	To be determined later in the project, under the responsibility of the co-ordinator, Kronikgune.

4.2.7 Ethical aspects

Aspect	Response/explanation
Any ethical or legal issues that can have an impact on data sharing	There are no moral ethical issues. Processes for handling the withdrawal of patient from the study, for their death during the study period, including what should happen to their study data, are documented within the research protocol, D 11.1. There are potentially data protection issues which we will examine carefully before determining which data items and on which population profiles can be made available as open research data.
References to ethics deliverables and ethics chapter in the Description of the Action (DoA) – if relevant	Work package 11 deliverables D11.1, D11.2, D11.3
Questionnaires dealing with personal data	Some of the anonymised data will have been derived from questionnaires completed by patients, covering topics such as health outcomes, quality-of-life and burden of care.
How is informed consent for data sharing and long term preservation sought in such questionnaires?	It will not be a GDPR requirement to obtain informed consent for the scientific use of anonymised data. However, this data reuse will be described in the Transparency Notice accompanying the informed consent forms, and to be approved by the seven ethics committees representing the pilot site. These statements will cover all of the intended uses of the data post project, by project partners for future research or teaching, as well as external researchers.

5 Open access strategy for knowledge assets and publications

The section summarises the intentions of the consortium towards other assets that will be developed through the research in addition to research data sets.

5.1.1 Aggregated data sets

As part of conducting the research evaluations of the health outcomes impact and health economic impact of ADLIFE, the consortium will generate a number of aggregated health outcomes and health economic data tables. Some of the content of these will be derived from our research data, and other comparative data may have been derived from public sources such as the academic literature. Many of these data tables will be used to develop and publish our results, and may therefore be also held by publishers as ESCROW research data. However, we will also undertake to publish these data tables alongside the deliverables to which they relate, on relevant data repositories, with links to these on the project website, so that they are open and available without any data sharing or licensing restrictions. A specifically formulated data access committee, comprising some or all of the project partners, will determine the policies for this data access. These policies will include which data sets (patient level and aggregate level, all robustly anonymised) are suitable for open data access and ensuring that the necessary approvals have been obtained from the originating sites. The committee will determine the time interval after the project when the data will be made available, how it will be discovered and accessed, the open access licence terms that will apply and how data access will be requested and granted. The committee will also make long-term sustainability decisions, including a sustainable business model for the data sets. A sustainability business model will be developed for covering the costs of long-term discovery and storage, the maintenance of the data sets if necessary, and any human resources required to manage the data sharing arrangements.

5.1.2 Clinical guidelines

ADLIFE will use clinical guidelines that have been developed by professional societies or Health Technology Assessment organisations. We will need to adapt these in order that they focus on the necessary care pathways for patients with advanced conditions and for patients who have multi morbidity, for which multiple guidelines need to be used in parallel.

We would ideally like to make the human readable version of these adapted guidelines available publicly, to some extent in academic literature and also on our website. We would also ideally like the computable form of the guidelines, as clinical decision support service specifications, which will have utilised guideline representation standards, to be available as open source code, held within a mainstream open source repository. However, because the source materials we use may have copyright restrictions, we will need to negotiate this on a case-by-case basis with those original authors. We will liaise with the publishers of these single disease guidelines in order to engage them in the possibility of endorsing and publishing our multimorbidity guidelines. We cannot therefore confirm at this stage what guideline material we can publish.

5.1.3 AI algorithms

ADLIFE will develop novel AI algorithms for inclusion within an Early Warning System for patients with advanced long-term conditions. We have not yet determined the extent to which

these algorithms will be commercially exploited in the future, and what levels of detail could be made available as open source components. We will make this clear later in the project. However, the methodology surrounding development of these algorithms will be published as part of our dissemination.

5.1.4 Dissemination resources

The partners are committed to investing significant effort in creating academic publications, conference presentations and posters, and other dissemination materials that showcase the methodology, results and solutions of the project. We would ideally like all of this material to be open access, and we will prioritise open access journals. Many conferences request to host presentation slides, and sometimes video recordings of presentations, on their website, to which we will always agree. In situations where conferences do not make these materials available, we will seek permission to host them on our website ourselves, for public access. Publications made by ADLIFE partners will additionally need to comply with the project's publication policy.

The ADLIFE website will be used to disseminate project findings, by sharing links to publications, conferences, abstracts etc. (Access to the papers from our website will be dependent upon the copyright and sharing policies of the journals.) Academic partners are required to deposit author accepted copies of publications in their institutional repositories as per relevant open access policies. Our social media channels will also be used as dissemination channels, as appropriate, throughout the project