D3.6 Guidelines for combining RCT, cohort, and EHR-based data for RWE in AD

116020 - ROADMAP

Real world Outcomes across the AD spectrum for better care: Multi-modal data Access Platform

WP3 – WP Identification, mapping and integration of RWE

Lead contributor	Antje Hottgenroth (17 – Eli Lilly and Company Ltd.)
	hottgenroth_antje@lilly.com
Other contributors	Stephanie Vos (4 – UM)
	Olin Janssen (4 – UM)

Due date	31/10/2018
Delivery date	23/10/2018
Deliverable type	R
Dissemination level	PU

Description of Work	Version	Date
	V2.0	08/11/2017

Table of contents

Document History

Version	Date	Description
V1.0	15/08/2018	First Draft
V1.1	13/09/2018	Comments from WP leads + Use Case Owners
V2.0	28/09/2018	Final for Full Consoritum Review
V3.0	23/10/2018	Final Version

Definitions

- Partners of the ROADMAP Consortium are referred to herein according to the following codes:
  - **UOXF**. The Chancellor, Masters and Scholars of the University of Oxford (United Kingdom) – **Coordinator**
  - **NICE**. National Institute for Health and Care Excellence (United Kingdom)
  - **EMC**. Erasmus University Rotterdam (Netherlands)
  - **UM**. Universiteit Maastricht (Netherlands)
  - **SYNAPSE**. Synapse Research Management Partners (Spain)
  - **IDIAP JORDI GOL**. Fundació Institut Universitari per a la Recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (Spain)
  - **UCPH**. Københavns Universitet (Denmark)
  - **AE**. Alzheimer Europe (Luxembourg)
  - **UEDIN**. University of Edinburgh (United Kingdom)
  - **UGOT**. Goeteborgs Universitet (Sweden)
  - **AU**. Aarhus Universitet (Denmark)
  - **LSE**. London School of Economics and Political Science (United Kingdom)
  - **CBG/MEB**. Aagentschap College ter Beoordeling van Geneesmiddelen (Netherlands)
  - **IXICO**. IXICO Technologies Ltd (United Kingdom)
  - **RUG**. Rijksuniversiteit Groningen (Netherlands)
  - **Novartis**. Novartis Pharma AG (Switzerland) – **Project Leader**
  - **Eli Lilly**. Eli Lilly and Company Ltd (United Kingdom)
  - **BIOGEN**. Biogen Idec Limited (United Kingdom)
  - **ROCHE**. F. Hoffmann-La Roche Ltd (Switzerland)
  - **JPNV**. Janssen Pharmaceutica NV (Belgium)
  - **GE**. GE Healthcare Ltd (United Kingdom)
  - **AC Immune**. AC Immune SA (Switzerland)
  - **TAKEDA**. Takeda Development Centre Europe LTD (United Kingdom)
  - **HLU**. H. Lundbeck A/S (Denmark)
  - **LUMC.** Academisch Ziekenhuis Leiden – Leids Universitair Centrum (Netherlands)
  - **Memento**. CHU Bordeaux (France)
- **Grant Agreement.** The agreement signed between the beneficiaries and the IMI JU for the undertaking of the ROADMAP project (116020).
- **Project.** The sum of all activities carried out in the framework of the Grant Agreement.
- **Work plan.** Schedule of tasks, deliverables, efforts, dates and responsibilities corresponding to the work to be carried out, as specified in Annex I to the Grant Agreement.
- **Consortium.** The ROADMAP Consortium, comprising the above-mentioned legal entities.
- **Consortium Agreement.** Agreement concluded amongst ROADMAP participants for the implementation of the Grant Agreement. Such an agreement shall not affect the parties' obligations to the Community and/or to one another arising from the Grant Agreement.

1. Introduction

During the course of ROADMAP, we investigated through several Use Cases the availability, suitability and acceptability of data, methods and tools including the point of view of different stakeholders in order to use multi-modal data access platforms and create an overview of Alzheimer's disease (AD)-relevant data which is represented by the AD Data Cube.

A major aim of the ROADMAP project phase 1 was to increase understanding of the progression of AD across the disease spectrum, from preclinical stages to severe AD dementia. By integrating and combining different datasets and sources, understanding of the AD disease spectrum and its complex mechanisms can be improved. Based on our learnings, Figure 1 provides an overview of the availability of data in different data sources in relation to the AD spectrum. This overview was created based on the data that is most commonly available in the different data sources (i.e. population cohorts, research cohorts, clinical cohorts, electronic health records (EHR) clinical data, EHR claims data and Clinical Trial data). As can be seen in Figure 1, none of the data sources has information available on all the different outcomes or cover the full AD spectrum in detail. This points to the need of combining information from different data sources to eventually develop a model of the AD progression across the full disease spectrum.

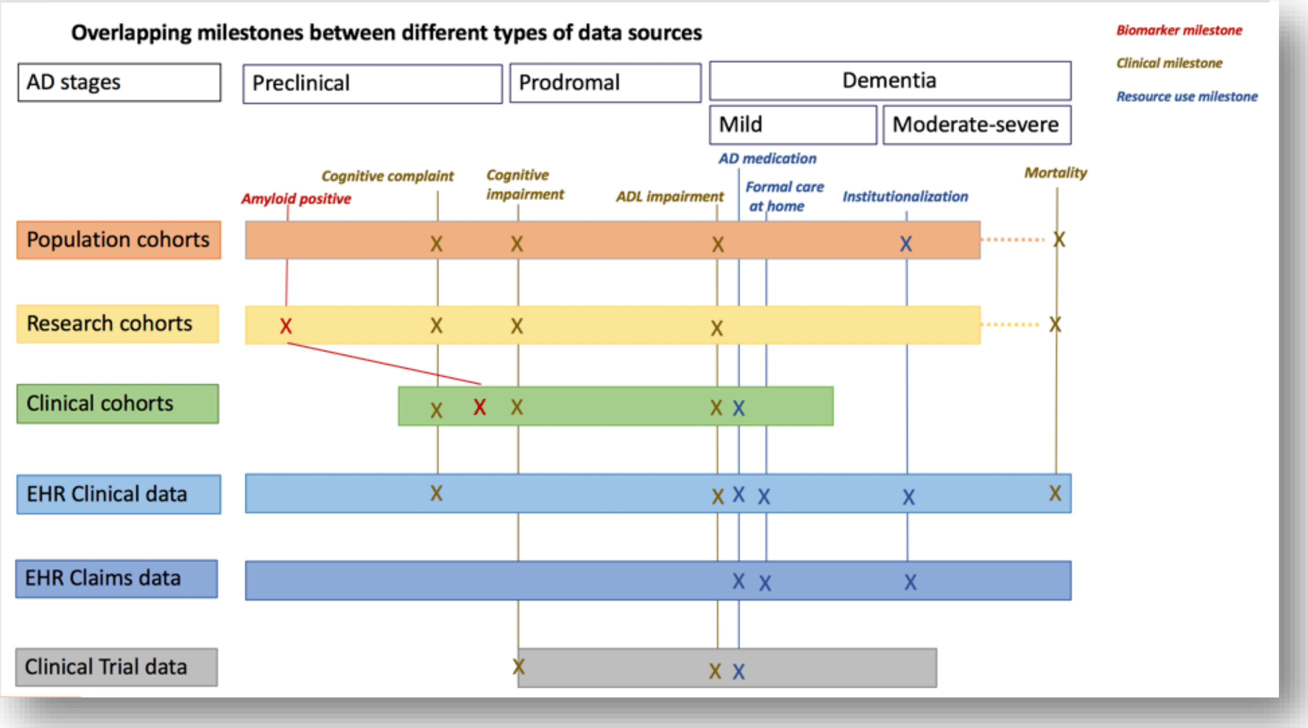

Figure 1 Overview of the availability of key events in different data sources across the AD spectrum. Each X marks the earliest possible assessment of this milestone in the different data sources.

The guidelines in this document should provide a summary of the project outcomes and should guide the development of further data capabilities for Alzheimer's Disease Research across the full spectrum of the disease.

Combining heterogeneous data types such as research cohort data, large population cohorts, national registries and controlled clinical trial data across multiple countries has to take into account several aspects. These are not only technical, ethical and legal/privacy considerations, but also the influence of heterogeneous health care systems, data documentation practices and physician behaviours, when dealing with cognitively impaired or demented patients through the course of the progression of the disease, has to be taken into account.

These aspects were explored with Use Cases of Disease and Economic Modeling as well as in Data Validation Studies and novel approaches using digital technology (further details are outlined in Deliverable 3.3) .

The guidance will provide insights into practices combining data sources originating from one country and practices for combining data sources across countries and will add considerations around the importance of knowledge management and meta-data quality as a starting point for access to data fit-for-purpose.

Although there is no ideal solution for a data governance and data flow process, this guidance is pointing out, that a federated approach to access data would be most feasible, specifically, when multiple countries contribute data to a common platform or research project.

2. Data Availability and Suitability

2.1. Data Knowledge Management and Meta-data

2.1.1. Data Landscaping

The first step of investigating the availability of existing data for the scope of work is to scan the environment. This landscaping is predominantly done through:

- Literature Research
- Dedicated internet searches by key words
- Targeted searches by visiting internet sites or contacting medical associations and large research institutes, or by investigating with data brokers and data custodians
- Catalogues (databases of databases)

It is important to mention, that remote research such as literature research and internet searches might not reveal all data sources of relevance, since they are not detecting recent data sources, which did not publish yet or which might not be conscious of their potential benefit to secondary research and keep silent. Therefore, it is important to build or join academic, medical and scientific networks dedicated to the disease of interest and listen to and probe for data access opportunities. Since RWE research is currently receiving more attention in the health care community, data custodian's awareness of the benefit of their data is rising. That is leading to a more comprehensive landscape of data available for secondary research.

***Use Case:** ROADMAP used the knowledge already available by these kind of activities and assembled information from consortium data custodians, previous landscaping projects (public and by consortium members) and targeted contacts to data brokers and data custodians and summarized the gained insights in Deliverable 3.3 Update on potential data sources with RWE data in Europe.*

2.1.2. Data Catalogue – Database of Databases

In order to have a sustainable library of the results from the landscaping activities, a catalogue or database of databases should be developed to document the characteristics of the data (meta-data catalogue), or an existing database of databases could be used. The definition of the meta-data items captured in the catalogue is an important step to assure best and most relevant search and display results, when the catalogue will be interrogated to identify most suitable data for a dedicated research project. Here it is important to balance the number of meta-data items against the documentation and updating efforts. The focus should be on the relevant outcomes defined by all stakeholders for the disease and research scope. Different outcomes are often captured using different measurements across data sources. In order to combine meta-data from different data sources, a common ground needs to be identified. Final conclusions of the suitability of a data source for a specific project or study has to be determined in collaboration with the data custodians of the pre-selected data.

The catalogue should be maintained as a database, which could be queried and should have functionalities to compare datasets or build reports to facilitate visualizations and data overviews. The documentation and updates should be done by the data custodians to have an accurate picture of the current status, either by direct entry or by review before publishing. The catalogue should be accessible via a Browser to facilitate collaboration with data custodians and researchers. As argued above, the final decision on suitability of data for a specific study will have to be made based on discussions with the datasources identified thorugh the cataloque.

***Use Case:** ROADMAP used pre-existing database catalogues (EMIF AD, EMIF EHR: https://emif-catalogue.eu/ and DPUK: https://portal.dementiasplatform.uk/CohortDirectory) and developed a Data Source Questionnaire in order to meet the ROADMAP goals and to create the ROADMAP AD Data Cube. The EMIF AD catalogue meta-data set was updated to add outcomes identified by ROADMAP stakeholder groups. The EMIF EHR catalogue did not contain sufficient meta-data for ROADMAP purposes. The consortium EHR dataset custodians were asked to fill-in the Data Source Questionnaire instead to be able to populate the ROADMAP AD Data Cube.*

Both the EMIF AD and EHR as well as the DPUK catalogues are accessible via Browser and provide search functionalities. In order to have a unified view across the catalogues and including databases not represented in the two catalogues, the ROADMAP AD Data Cube serves as the project-specific representation of a data catalogue. During the project, intermediaries from WP3 searched the catalogues and provided the combined recommendations for suitable data sources to the project leads of the other work packages.

2.1. The ROADMAP AD Data Cube

The identification and evaluation of the data sources suitable for research of Alzheimer's Disease across the spectrum of the disease led to the filled ROADMAP AD Data Cube, which will eventually provide a three dimensional representation of availability of outcomes across the data sources and across the disease stages. The cube is an evolution from a flat or two-dimensional database about databases and is helping to facilitate the matching of data sources against research protocols. Furthermore, it facilitates the gap analysis of any missing real world data, which might be needed to answer a research question or revertedly facilitates the design of research protocols to consider only outcomes, which are available currently. Last but not least it will guide considerations to prospectively plan and implement the collection of data in the standard care setting to fill the data gaps, which cannot be substituted by already available RWD (see Deliverable 3.3).

2.2. Data suitability

When data sources are identified by reviewing their meta-data, they still have to undergo additional investigations before being deemed suitable for a specific research project. The following activities should be conducted in order to determine the suitability of the data and might need the direct input of the data custodians in most cases:

- Review of more granular meta-data, when available (e.g. when an outcome measure is present in the data, how often is that measure documented over for a patient (longitudinality and frequency of data etc.)
- Review of documentation patterns, which should include amount of missing data for study-relevant outcomes, mode of identification of the study population in the data (e.g. are patients with Mild Cognitive Impairment (MCI) identifiable in the dataset by coded diagnosis or else)
- Conduct feasibility analysis, when applicable (e.g. identify number of patients in the data source meeting the study population criteria or sufficient number of patients having the required follow-up period data etc.)

To decide whether a data source is suitable for your investigations, it is in the first place important to understand the primary purpose of each of the data sources; administrative, reimbursement, quality of care, drug evaluation, regulatory, epidemiology etc. The amount and depth of suitability investigations differ by type of data source and depends on the grade of similarity between the primary purpose and secondary purpose of the data. Clinical trial data from Alzheimer disease studies as well as scientific cohort data of Alzheimer's patients might need less pre-project evaluations than large disease-agnostic EHR databases. The later ones should be characterized and investigated in the context of the health care systems they are originating from. That should include, but is not limited to a coding system review to understand the accuracy of reporting patients with the disease of interest including the pre-clinical disease stages and progression patterns as well as treatment and care practices, reimbursement and screening measures, which could influence the documentation in the data source. Furthermore, EHR databases might require additional free text mining in order to collect all required information, which is not so often needed in case of cohort data, for example.

For some purposes, e.g. mapping between disease-specific and generic outcome measures, the joint availability of several variables in datasets needs to be ascertained (Deliverable 5.3 Mapping algorithms from real world outcomes to preference-based measures).

2.3. Data Availibility – Permission to Use

When data were identified, the sole existence does not determine their availibilty for secondary research per se. Data sources and the permission to use the data for secondary research are based on a research protocol provided by the researchers wishing to use the data and are governed by laws and regulations on European and country-level and depends on the level of consent of the data originators (e.g. patients) provided as well as on the level of anonymization.

Data protection laws are not necessarily harmonized across countries which might potentially create challenges while pooling datasets from different geographic locations. GDPR is supposed to remedy this challenge within the EU but this comes at the cost of more stringent data protection laws introduced across Europe.

In addition, data custodians are adding further permission levels in order to restrict the access for specific stakeholders and to assure quality research done with the data. The later is often accomplished by a study review and approval process before access to the data is granted. Therefore, it is important to understand in the first place, the general willingness of the data custodian to share the data for research and whether there are restrictions to the kind of third party requesting the data (e.g. academic researcher, industry researcher, regulatory researcher etc.) as well as any study review and approval process expectations of the data custodian.

Permission to Use Levels and Access Request Management have to be negotiated and mapped clearly, when planning secondary research or data repositories and data platforms. That should include a clear picture of permission levels for each data source planned to be used or hosted on a data platform. Potential access levels might be, but are not limited to:

- Approval needed from the data custodian study-by-study
- Data Custodian approval for a study bundle
- Approval/license granted for a limited period of time
- No approval by the data custiodian, but limited scope defined for the use of data through the platform or research consortium

Request access management should be facilitated by a data or search platform in order to guide researchers to the applicable contacts and needed documentation.

Use Case: *The experience in the MMSE Model Validation Study illustrates the importance of understanding both the influence of the health care system and the primary purpose of the data collection on the data that are available for research. Although initially data may seem suitable for a specific model validation, the results of a validation study may well show that the data at particular sites are to be treated with caution (details are outlined in Deliverable 3.4 Final report on proof of concept technical solutions for RWE data harmonisation and integration and D4.4 Results from pilot model validation exercises). Several datasets, which were pre-selected based on the general* *availability of the required data items were dismissed because, for example, lack of longitudinally follow-up, inability to identify the type of dementia (e.g. codes such as dementia not further specified), or lack of full MMSE scores.*

Datasets were closely evaluated and selected to assure that the (few) needed variables and outcomes are present in the datasets. Problems, however, may readily emerge including the criteria for the diagnosis is identifiying Alzheimer's Disease or matching the population (for example, age range) to the results obtained from the other datasets (Memory Clinic setting vs. population-based hospital EHR). Differences in data collection or setting will vary compared to each other as well as to the results from the original disease model study – with, for example, different prevalence of AD in different countries and settings. This example underlines the importance to take into account of the context of data collection.

3. Methodology and Tools

3.1. Data Locations and Data Repositories

When data sources should be shared and potentially combined with other data sources for research projects there are several options, how to achieve that. The feasibility of the options is ruled by data privacy requirements in most cases and need to be evaluated early in the planning of a research project or consortium endeavour to understand and implement the necessary infrastructure and technology to support the chosen option. Often it will turn out, that several options need to be implemented, when combining various types of data across countries.

- Data reside at the data custodian
  - Data are kept and used in the native format - a common analysis is achieved by adaption of the analysis plan to the structure and content of the native data
  - Data are harmonized into a common model/format (Distributed Data Network Model) – a common analysis is achieved by execution of the same analysis algorithms on the standardized data
- Data reside at the data custodian, but are made available for external analysis
  - Data Platform tools and algorithms are able to access the data at the data custodian and only results are returned to the platform and platform user (e.g. EMIF AD Switchbox)
  - Study-specific algorithms are developed within a platform tool and executed locally by the data custodian. The algorithm could be enabled to harmonize, extract, aggregate and analyse the data locally and export results and limited/aggregated data to the platform only to enable additional analysis (e.g. EMIF EHRJerboa tool).
- Data Repository – data are hosted in one location
  - Data are kept in their native format as separate entities and code books/data dictionaries are provided with the data – analysis of data has to be adapted to the specific dataset or ad-hoc harmonization/standardization takes place study-by-study (e.g. DPUK)

- Data are harmonized into a common data model/format – data are standardized/harmonized as a whole dataset, when they are provided to the platform and a common analysis algorithm could be run in parallel on the datasets
- Data are standardized and integrated into a single research database – several data sets are integrated into a single dataset and a single analysis algorithm will be run on the database (e.g. EMIF-AD tranSMART)

While the concept of integrated databases is common for combining Clinical Trial data, cohort data are often hosted in Data Repositories in their native format or harmonized, but still kept as separate entities. The possibility to combine several datasets after harmonisation is however of large interest as big data analyses can provide innovative insights. EHR databases most often reside with the data custodians (e.g. big national EHR datasets or national registries). In addition EHR database harmonisation might be possible for different databases within one country, while harmonization of EHR databases across countries is more challenging.

Although data repositories have advantages from the data access and analysis point of view, administrative effort to host the data should not be underestimated. There is a wide spectrum of anticipated workload and personnel needed, based on the repository concept. Hosting native format cohorts, which are closed might need minimal administration. Hosting harmonized data sources, which are ongoing and providing regular data updates do need much more attention and dedicated personnel to handle the harmonization and update processes.

3.2. Data Harmonization and Standardization

There are at least two major aspects, which have to be taken into account when planning to use data from separate datasets or data sources for a common purpose. First the scientific and second the technical interoperability of the datasets. Both aspects should be investigated, when data should be combined (either on the data level or later downstream of evidence generation stream). Often they cannot be separated and scientific considerations should also help to choose the best standardization system and during data mapping of the native data to a data standard.

In order to understand, whether data are suitable to be combined for a common research objective or purpose from a scientific point of view, it should be determined, whether the same or similar methods were used to assess an outcome of interest. When that is not the case, it should be investigated, whether there might be methods to transform or translate an outcome measure to make it equal or comparable to another one. Some examples of scientific interoperability considerations:

- Are the same outcome measures used (e.g. e.g. ICD-10 codes for diagnosis of disease)?
- Are the same definitions used for a data item (e.g. cognitive impairment based on a MMSE score <27 or as cognitive z-scores below -1.5 SD)?
- Is the same or comparable version of an outcome measure used (e.g. three or five level version of EQ5D ?
- Is the patient/individual population the same or how much overlap (e.g. mild and moderate AD only, patients >55 years)?

- Is the population of interest identifiable in the datasets in a similar and reliable manner (e.g. is there a disease code for MCI in the coding system used for the dataset or is this defined mainly on cognitive test performance)

For all identified differences there should be an evaluation of impact on planned research purposes, studies or analysis plans as well as documentation of methods and algorithms used to equalize or mitigate these. That should also include an assessement of the potential impact of the local health care system on the data collection and availability in case of the broad population-based databases and registries as outlined in chapter 2.2. There are challenges not only with respect to the difficulties in combining EHR and Cohort data, even on the level of harmonized analysis. Even combining EHR data from multiple countries can be difficult as some of them originate from GP/EMR systems and others have more general coverage (registry based data systems). Even within one country, it can be difficult to harmonise different registries due to challenges in obtaining access to the secured environments of multiple registries. Since the data often cannot leave their secured environment, procedures need to be set in place to securely transfer data from the different registries to one environment.

Major challenges that are anticipated with EHR data compared to cohort data are missing data, misclassification (disease, exposure etc.) and lack of important variables.

In order to avoid the need to mitigate differences caused by variations in the methods and tools how to assess an outcome of interest it would be benefical to be able to harmonize or standardize the data collection in the first place. Guidelines for standardized outcomes or tests might be developed across countries and might be established in research cohorts as well as in health care system guidelines. This might be most feasible for research cohorts or registries in the first place and there are examples of efforts from other diseases, which could be translated to Alzheimer's disease research platforms as well. One example is the EULAR organization combining efforts on improving care, guidelines and research for rheumatoid diseases (https://www.eular.org/eular_strategy_2018.cfm). Standardization and harmonization of data is one aspect of their activities (Examples: https://www.eular.org/harmonicss.cfm, https://www.eular.org/epidemiology_study_groups.cfm).

Another aspect is the technical interoperability of data. There is no single common data model across the various types of data, which is accepted or feasible across all stakeholders and data custodians. Clinical Trial data are standardized according to Clinical Data Interchange Standards Consortium (CDISC), when they are originated from industry research (https://www.cdisc.org/). The majority of cohort studies or EHR databases are not standardized in the first place, but need to be mapped to a standard, when harmonization of data is planned. A potential standard to be used is the OHDSI I OMOP common data model (https://www.ohdsi.org/data-standardization/). The standard can be adjusted to align with common use in the field and to incorporate additional variables. When choosing a standard it should also be taken into account, which methods and tools might already be available (open source or commercial), which support data management and analysis of the data in that standard. The OHDSI organization for example is providing open source tools for database exploration, standardized vocabulary browing, cohort definition, and population-level analysis as well as a repository of Multiple Java based client applications to provide support creation and handling of OMOP standardized data (https://www.ohdsi.org/analytic-tools/). It should be noted that common data models might not provide standards for all variables already, which could limit the usefulness of the standard system or would need additional efforts to create standards in addition ot the data standardization process itself.

Use Case: *ROADMAP did not build a project-specific data repository, but used data hosted in existing repositories from consortium partners and related IMI projects (DPUK and EMIF). Data residing with their data custodians in native format were also included in the Use Cases. Clinical Trial Placebo data were provided in analysis datasets format to an academic consortium partner for consortium research within a local data infrastructure.*

DPUK as well as EMIF AD/EHR combine the features of a database catalogue, data respository, data harmonization/standardization and analysis tools within their respective infrastructures, which are illustrated in more detail in Deliverable 3.4 Final report on proof of concept technical solutions for RWE data harmonisation and integration. During the course of the conduction of Use Case studies by WP4 and 5, these infrastructures were used to identify cohort data suitable for the studies through an overarching ROADMAP process, tapping into both catalogues. Subsequent harmonization of the data, when applicable, were done wihin the infrastructure native to the platform, but in addition alternative ways were explored to use EMIF EHR harmonization and analysis tools (Jerboa) with DPUK cohort data at DPUK as well as on-site at the Dementia Registry of Gerona (ReDeGi).

DPUK provides analysis tools on a remote access platform and data are kept as separate entities in their original form. Data cannot be downloaded and the analysis has to take place in situ of DPUK. The EMIF EHR Jerboa data extraction, harmonization and analysis tool was tested on DPUK-hosted cohorts for WP4 Use Cases in order to introduce a harmonization step prior to analysis. This software tool can be installed locally. The purpose of Jerboa is to send data to Octopus for the final analysis which implied initial checks whether it can be used in its original form on DPUK given the local governance structure. However, as DPUK's security mechanisms do not allow any data to be send from the servers, the tool can be used without further adjustment on DPUK.. This is an example, that technology has to be reviewed and adjustment needs should be planned for to fullfill local governace requirements (further details are outlined in Deliverable 3.4))

For the BESIDE project, multiple registries from the Netherlands were combined. Since use of registry data was restricted to a secured environment, all three registry datasets were imported into one secured environment and combined using a crypted identification number for each individual (further details are outlined in Deliverable 3.4).

In Spain, the Register of Dementias of Girona (ReDeGi) was linked to the SIDIAP primary care database to assess the accuracy of Alzheimer's disease diagnoses (further details are outlined in Deliverable 3.4).

3.3. Data Hosting and Analysis Tools

Methods and tools to access, combine and analyse data depend on the type of data access, which is planned as outlined in the previous chapter. Therefore, data access planning and the setup of the process to manage, analyse and report the data have to be done in an integrated manner. Ideally, the platform hosting the data is providing tools and methods for preparing the data for analysis as well as analysis infrastructure and tools as it is the case for the EMIF-AD TranSMART platform (EMIF AD and TranSMART).

When a Distributed Data Network Model is chosen as in case of the EMIF EHR platform (http://www.emif.eu/about/emif-platform), there should be a common place or platform for initial access and project/study planning including a support infrastructure to liaise data custodian and researchers, as well as methods and tools to manage and analyse the data. A concept should be developed at which stage of the data processing the data or analysis result merging should take place in order to develop or purchase the appropriate tool set.

Concepts for Distributed Data Network Access can be:

- Data harmonization at data collection level (e.g. registries or cohorts are already implemented having in mind later secondary use through a DDN community) – no need for later data mapping and ready for usage with analysis tools
- Data harmonization (whole dataset) – the data custodian or platform host maps their entire database to a common data model on a regular basis, which is the case for the EMIF-AD TranSMART platform. Here the platform is taking over the task to harmonize the received datasets such that data pooling of different sources is allowed for large-scale analyses.
- Data harmonization (pre-study) – data are extracted and brought into a common format based on the study-specific needs (data subset) and structured and aggregated to allow release from the data custodian to a DDN analysis platform, which is the case in EMIF EHR.
- Analysis harmonization – a single analysis plan is developed and data custodians develop analysis algorithms to perform the analysis on their native datasets. Often that is no longer considered a Distributed Data Network methodology, but could be an option, when the data custodian is part of the DNN community, which could include adhering to common quality standards, data collection practices and scientific research exchange.

Although it might be benefical from the perspective of the data platform hosting point of view to choose one of the above outlined scenarios, how to access, store and analyse data, it might be more feasible to plan for a mixed model, even in cases of planning a platform for a single data type (e.g. cohort data platform), but data from various geographies. To accommodate the need to access the data at the data custiodians location or to host the data on the platform could help to enhance the number of available datasets through that platform (e.g. EMID AD TranSMART accommodate remote access (Switchbox) and platform hosting).

Use Case: *ROADMAP used the infrastructure of EMIF AD, EMIF EHR and DPUK in order to perform the Use Case Studies. For that purpose, some use case tested the application of the EMIF EHR pre-study data harmonization and aggregation tool Jerboa to cohort data from the EMIF AD and the DPUK catalogue. The feasibility of using this technology was shown, but needed some adaptions to the tools code in order to adhere to the data privacy compliance needs of the single-country cohorts in DPUK. There were some advantages having a dedicated data management infrastructure at DPUK ,which could handle the Jerboa application and adaption. In other cases the data custodians were not resourced to manage the adaptions of the application to their data set and data infrastructure.*

4. Data, Methods and Tools Acceptability

Data, methods and tools outlined in the chapters 2-4 do have benefits and short-comings, which are considered acceptable or less preferred by stakeholders, who should take actions or make decisions based on the outcome of the research conducted with these. Therefore, it is crucial to include all stakeholders in the planning of research platforms or research consortia and bench-mark the data, methods and tools acceptance in relation to the technical feasible options.

Stakeholders acceptance considerations include:

- Data Custodians – did not consent the data originators sufficiently to allow secondary use, might not trust platform hosting processes (losing control over their data) and would like to keep the data in-house, would like to run analysis by their own processes and standards, would like to share aged data only, potentially designed their study for another purpose.
- Data Originators (patients) – are refusing consent or consent was not obtained for research on their data, are eager to have their data shared to let benefit others with the same disease, want to maintain control of for which purpose their data are being used. Often the patients would like to share the data, but need to be consented appropriately and would like to benefit directly or indirectly from the secondary research. The would like to be involved beyond sharing their data and being informed about the results and outcomes.
- Researchers – would like to have most recent data (based on newly available techniques or digital patient reported data), data should contain granularity to identify sub-populations, disease severity, disease progression and disease or effectiveness markers. Data should be of large sample size, low variability and not be too stringent with inclusion criteria to allow epidemiologic research. At the expense of using large-scale combined data sources, there might be loss of information, when data are harmonized or standardized prior to analysis.
- Regulators – need evidence, that the data provide validated evidence and are not biased by un-controlled factors and unknown confounders, when used outside of their primary purpose they were collected for.
- HTA bodies – want best available data of relevance for their jurisdiction (e.g. geography, patient population, types of outcomes, time horizon)

Use Cases: *There is a complex landscape of stakeholder interests, which in turn might need a flexible and tailored approach to consent processes, data access governance and return of results and insights. Among the stakeholder groups, numerous considerations are relevant to evaluating the acceptability of proposed methods for re-using data for research. Often, considerations are shared across these groups, with each providing a different perspective or sets of interests relevant to a particular considerations.*

Seven cross-cutting themes of ethical concerns are relevant to multi-stakeholder evaluation of data sharing tools. Via a systematic review of literature discussing the ethics of biobanking and medical data repositories, the following themes were identified: (1) informed consent; (2) autonomy and participation; (3) transparency; (4) ownership; (5) data provenance; (6) privacy; and (7) group harms and discrimination (Deliverable 8.1 Review of ELSI issues in RWE approach).

- ***Informed consent*** *is one of the most discussed practical issues in the literature, where debate largely focuses on the different models for obtaining consent from research participants. For example, single-study consent, where participants are asked to consent to participate in a study for one specific purpose, is generally considered inadequate for increasingly common data collection and sharing practices in which data is collected by biobanks and cohort studies* *in order to provide a resource to the scientific community. To address this limitation broad consent approaches have been adopted where participants consent to studies governing access to data and ensuring, for example, that secondary research is scientifically rigorous and in the interests of participants and society. Equally other alternative consent models exist including tiered consent and dynamic consent, which aim to provide more fine-grained consent mechanisms.*

- *Other practical ethical issues concern the **ownership and provenance** of data. Questions of ownership are relevant to determining, for example, the extent to which participants receive a share in any benefits derived from data about them, or the extent to which commercial exploitation of findings may be possible. Again there are many possible ways to structure these relationships between participants, studies, and secondary researchers; those different possibilities illustrate how other ethical values, such as transparency and respect for participants' autonomy (see below) are instantiated. Questions of provenance are relevant to the validity of scientific inferences, since if the context of data is lost then poor understanding of data can lead to faulty inferences and interpretations. However, this becomes an ethical issue when such loss of context creates or reinforces inequalities.*
- *The remaining themes concern **ethical values** that underwrite various data sharing practices: **autonomy, transparency, privacy, anonymisation, and group harms and discrimination**. These values inform the creation of practices that protect the interests of participants and create stronger relationships between participants and researchers. All these themes feed into the production of the ELSI framework by highlighting key points to consider for the responsible development of data integration tools and governance structures.*

Data Custodians

There are three main concerns held by data custodians for the governance of consortium to custodian interfaces (Deliverable 8.2 Initial report on requirements for an ELSI framework for a RWE approach in AD).

- ***Interoperability of consent** taken by different studies and in different jurisdictions. That is, assessing the compatibility of the different positions that studies may take on the scope of research that is permissible with their data, based on the scope of participants' original consent.*
- ***Coordination of data access requests**: For example, many commentators and documents from existing consortia have noted that there is often duplication of effort when requesting access from multiple studies. Indeed, many have called for more streamlined systems in which studies are able to benefit from and defer to approvals given elsewhere.*
- ***Efficiency and Transparency of data access:** There are a series of recommendations for efficient organisation at the consortium level to ensure that the use of data from multiple sources is both efficient and transparent. For instance, centralising information on procedures for data access, and systems for reviewing and tracking access requests.*

The complexity of these challenges owes to the international data sharing landscape being governed by a patchwork of laws, regulation, and bespoke governance structures, policies, guidelines and best practices. National laws differ and data sharing policies are variously instantiated by data providers *according to their local context. As a result, there are a diverse range of processes for ethical, technical and access approval that can create a barrier to efficient data sharing.*

A series of principles reflected in codes of conduct and related governance frameworks authored by biobanks, medical data repositories, and other international organisations that promote sharing of medical data for research are identied as well (Deliverable 8.2). These principles describe norms covering social, epistemic and operational aspects of data re-use, intended to shape the design, implementation and evaluation of specific methods for RWD integration for AD research. The principles are as follows:

1. ***Protection of participants:*** *The protection of participants is paramount in data re-use. Beyond compliance with rules imposed by data providers, this includes: (1) protecting rights and interests, such as privacy and confidentiality, respecting autonomy and dignity, as well as minimising harm; and (2) protecting the collective interests of groups and communities, as well as individual participants.*

2. ***Accountable governance:*** *The governance of data re-use at the consortium-level must be organised to ensure processes are open and efficient, for example, by being clearly documented and open; flexible and sensitive to different kinds of research; streamlined and proportionate to the real risks that face research data sharing; and are compliant with relevant legal obligations that apply to research and data protection. Oversight should be complete and transparent, for example, by being accountable to all stakeholders through appropriate representation (e.g. tracking and auditing mechanisms to evaluate data sharing activities, reporting to and involvement of participants, data providers, and funders).*

3. ***Scientific quality:*** *The core purpose of data sharing and re-use is to advance scientific knowledge. This should be promoted through (1) ensuring data quality, integrity and interoperability, throughout their lifecycle beyond data providers, for example by adopting and developing best practices such as complete metadata and documentation; and (2) ensuring newly derived data are discoverable by the scientific community, are a sustainable resource for the future.*

4. ***Engagement with society:*** *Data re-use activities must be outward-looking and reflective about their social role and purpose. In particular this means (1) engaging research participants, patients and citizens about the results, value and implications of data reuse; and (2) promoting trust, and democratic and responsible data re-use through commitment to principles (1)-(3).*

Data originators

Empirical studies of patient and carer attitudes towards sharing data were reviewed by Work Package 8 (Deliverable 8.2). These studies suggest broadly supportive attitudes among patients and the general public to the sharing of health data for both research and care. However, that support is almost always premised on more nuanced views about the measures that need to be in place to secure acceptability. In terms of who data is shared with, researchers, universities, healthcare professionals, and health services were often trusted to protect individual's interests and act for the public good. In contrast, there is scepticism about the trustworthiness of commercial organisations along these lines. In terms of how data is shared, proper governance is key and individuals naturally *expect mechanisms to protect confidentiality and privacy of data. With that said, greater control over data sharing was also found to be valuable to individuals.*

Indeed, the common expectation that individuals would be asked for explicit consent to data sharing out of respect, even in cases where it was not strictly necessary, is particularly noteworthy and speaks to the importance of engagement to maintain trustworthiness, legitimacy and 'social license'.

In terms of what data is shared, supportive attitudes were often premised on only certain uses of data being permitted, or only certain types of data being shared. For example, that there would be no commercial use, or that types of data viewed as particularly being sensitive would be more strictly control, such as data about genetics, and mental and sexual health. The range of permissible uses of data also feeds into the question of why data is shared, where again supportive attitudes often rest on the promise of some benefit to individuals themselves, to individuals similar to them (for example, those with the same condition), or to the general population.

The views of data originators were also explored in focus groups with the European Working Group of People with Dementia (EWGPWD) and their supporters (Deliverable 8.3 Brief on findings of ELSI focus groups for RWE approach to AD). The focus groups echoed many themes already present in the literature. Participants in the focus groups highlighted concerns about data sharing and re-use that focused partly on data quality issues but most notably on risks associated with the disclosure of personal information. Participants used discourses of privacy and anonymity to challenge situations where seemingly more information than necessary about them was available to health care professionals. The risk of personal information being disclosed by hacking was a key concern among many participants. Participants deployed the language of trust to deal with uncertainties about the protection of their data. They held the view that strong relationships with research and healthcare institutions underwrote their ability to participate in research without being paralysed by the risks and concerns they identified. Participants in both focus groups shared the desire to be, at a minimum, informed about the results of research, and ideally to have a stronger more engaged relationship with the research process: to them this meant participation beyond the point where they contribute data and greater recognition of their contribution. A key finding from the focus groups was the view that greater engagement about secondary research enabled participants to deal with uncertainties around the disclosure risks that are seen as facing the sharing of health data. That is, participants wanted to be better informed about the ways data about them is re-used.

5. Conclusion

There is a diverse landscape of Real World Data, which are available for Alzheimer's Disease research across Europe. There are starting points to integrate data in research platforms to facilitate secondary research, such as the DPUK and EMIF AD and EMIF EHR platforms as well as initiatives in other parts of the world (CAMD institute – Clinical Trial placebo data, or GAAIN). ROADMAP investigated the data landscape and provided an integrated view of the data considered relevant and acceptable by various stakeholders, their availability across the data sources and the suitability to use them in specific research types such as disease or economic modelling, including model validation. Novel and upcoming data types (e.g. digital patient reported data, continuous monitoring data) were included in the evaluation and are reported elsewhere in more depths (Deliverables 3.4 Final report on proof of concept technical solutions for RWE data harmonisation and integration + 3.5 Guidelines for use of smart devices as a measure of RWE).

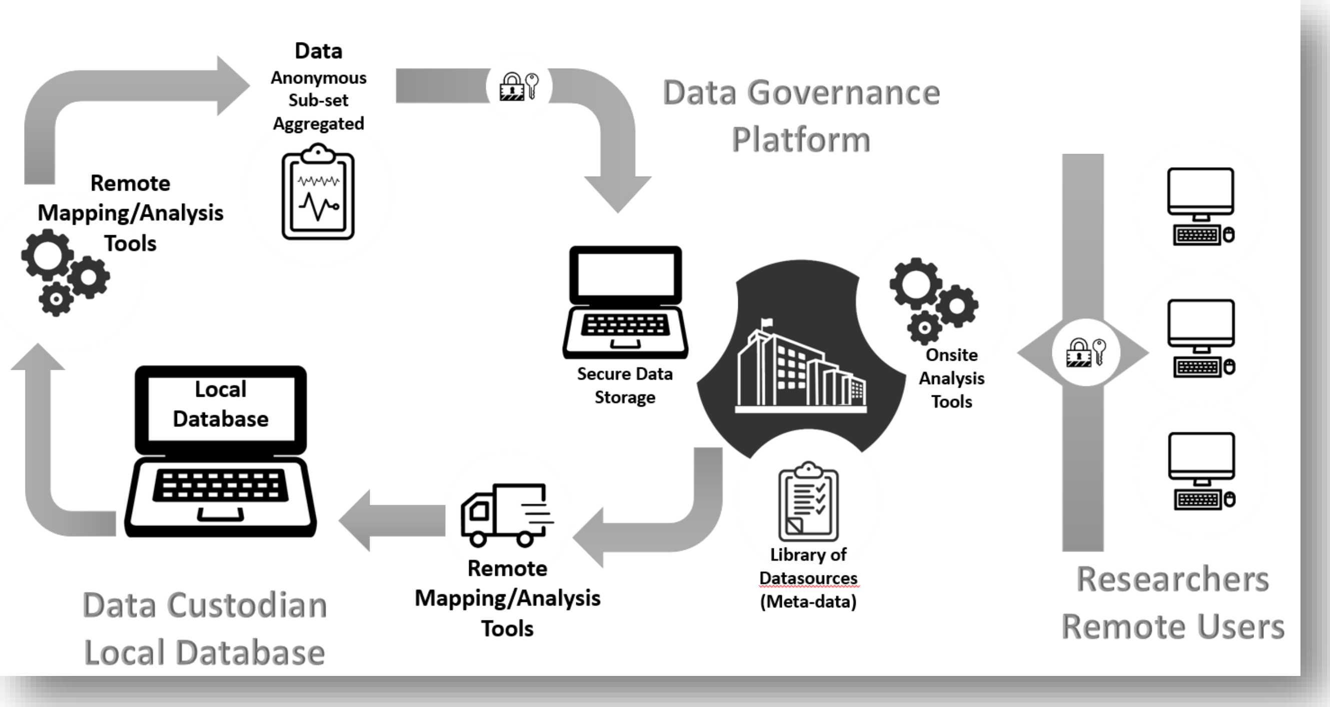

Figure 2. Distributed Data Network (DDN) governance approach. The Data Governance Platform provides the infrastructure to enable data searches through a library of databases, provides a secure storage location for hosting pre-processed data and an secured analysis environment. Data stay with their data custo+dians and analysis and/or data mapping tools are provided to the local infrastructure in order to extract data or aggregates in a data privacy compliant manner. Researchers are able to access the anonymized or aggregated data through secure access measures on the platform, but will never have direct access to the original data source, which stays with the data custiodian.

The consortium was able to carry out several Use Cases to provide an evaluation of suitability of data sources, methods and tools. These are the major conclusions with respect to the Real World Data:

- Data repositories including a database catalogue and analysis tools and methods are important to facilitate research projects. Once data are available on a data platform future research requests can be handled faster.
- Data Access and Permission processes have to be carefully planned and should not be underestimated, even when data are hosted on a platform.
- A federated approach to access data is most suitable for combining data from various countries or data types (see Figure 2).
- Distributed Data Network approaches need an evaluation of dedicated personnel at the data custodian in order to be able to implement tools and methods and perform to data management and analysis activities at the site as needed for the DDN projects.
- Highly suitable data sources are requested by third party researchers (academic + industry) at high frequency, and this is leading to a competition of resources at the site for in-house and third party research or consortium-based research.
- Different data sources capture different types of information and to various degrees of richness or coverage. In general, EHR data are covering a very broad spectrum of the general population and document actual care delivered to patients, but with lack of details, which often are needed for a specific population or disease of interest. Cohort data typically are well structured and detailed, but only cover a small part of the effected population. Therefore evidence generation should consider several types of data sources to combine insights, with in general more detailed information being available in cohort databases compared to EHR databases.
- Stakeholder-reported priority outcomes are often somewhat different from the outcomes captured in research studies or available in EHR databases.They are often of a more qualitative nature (e.g. losing the sense of who you are, judgement and insight, emotional issues) and therefore provide a challenge to map them to a direct measure or objective data capture. Future studies might take this into account when deciding on the assessment procedures of different outcomes by focusing on the dissimilarities between stakeholder-reported priority outcomes and outcomes captured in previous cohort/EHR studies or available in EHR data. For the later it might need national efforts to establish the reporting of these outcomes in the routine care of patients.
- To facilitate secondary use of data, collection of priority tests for each domain should be encouraged. Key tests should be identified for each different domain beforehand, as well as assessment procedures for each outcome and structures/standardization of the datasets. This will allow combining of different data sources.
- Identification of data gaps that are relevant for the various stakeholders is a crucial part of a data landscape evaluation, which should inform plans to setup stand-alone or supplementary data collection platforms to collect missing evidence in the future.
- Due to the primary purpose of each of the data sources (e.g. administrative, reimbursement, quality of care, drug evaluation, regulatory, epidemiology) some data items are not available to support secondary research most efficiently. There might be an effort to go back to the

primary sources to help them think ahead to make the data fit for secondary research as well, when feasible.

- There is a certain amount of diversity of tests or data items representing a specific outcome, which imposes a challenge to combine data from different data sources.

6. Recommendations

We would like to conclude these guidelines with some recommendatios in order to improve the efficiency and relevance of RWE research. Focus areas should be:

- Enhance global collaboration to create a common technical, legal and ethical model for RWE research and evidence generation
- Continue and enhance local/country as well as across-countries efforts to harmonize database/cohort protocols to improve interoperability of data at the level of the data generation and capture, when feasible
- Development and validation of outcomes of relevance as defined by all stakeholders in order to inform the development and harmonization of data collection to create the RWE foundation, which will be accepted and relevant to inform the Health Care community about the disease and intervention pathways.
- Integrate data custodians and patients into RWE research not only as contributors of data, but as research partners and implement measures to inform these stakeholders about the research results and immediate benefits.
- Broad population-based EHRs could be stimulated to include additional information for a disease of interest by more general means or by specific collaborartions to setup data collection add-ons for a specific research programms.