# D3.2 First report on proof of concept technical solutions for RWE data harmonisation and integration

## 116020 - ROADMAP

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

## WP3 – Identification, mapping and integration of RWE



| | |
|---|---|
| **Lead contributor** | Stephanie Vos (4 – UM) |
| | s.vos@maastrichtuniversity.nl |
| **Other contributors** | Pieter Jelle Visser, Olin Janssen (4 – UM) |
| | Christoph Jindra, Sarah Baumeister (1 – UOXF) |
| | Johan van der Lei, Jan Kors (3 – EMC) |
| | Anna Ponjoan (6 - IDIAP) |
| | Martién Kas (15 - Rijksuniversiteit Groningen) |
| | Antje Hottgenroth (17- Eli Lilly and Company Ltd) |



| | |
|---|---|
| **Due date** | 31/10/2017 |
| **Delivery date** | 25/01/2018 |
| **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 | 26/10/2017 | First Draft |
| V2.0 | 23/11/2017 | Updated draft after internal peer review: Valéry Risson (Novartis), Anna Zettergren (UGOT), Andrew Turner (UOXF) |
| V3.0 | 15/12/2017 | Draft Changes after Full consortium review: Mihaela Benea, Michele Potashman (BIOGEN). |
| V4.0 | 22/01/2018 | 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.
- **MTA.** Material Transfer Agreement

# Publishable Summary

WP3 facilitates the work of WP2, 4 and 5 by identifying and providing access to relevant data sources for answering the research questions defined within ROADMAP. WP3 has developed a preliminary workflow to achieve this. After approval by the WP leads, ROADMAP researchers are required to fill in a specifically developed scientific research question form. This will be send to WP3, which then triggers the search for relevant data in all ROADMAP partner platforms. If the data are already uploaded to one of the platforms, data access can be provided fast, conditional on the approval of the data owners. Researchers can also identify data that are not yet included in one of the partner platforms. In this case, the data owner will be approached and asked for participation in the study.

4 use cases have been identified so far and this report shows how WP3 has made use of the existing partner infrastructures in ROADMAP to identify relevant sources. Data for the validation of the preclinical model, developed by Novartis, were identified from within EMIF AD and DPUK. EMIF AD has additionally approached cohorts that were identified as potentially interesting by WP4. Data for the validation of the model of cognitive decline in people with AD were identified by EMIF EHR and DPUK so far, but additional cohorts will be identified in EMIF-AD and it is considered to use CT placebo data as well. The dementia diagnosis validation study makes use of the SIDIAP database, while a pilot study for the estimation of the costs of dementia identified six cohorts from EMIF AD. ROADMAP partner VUMC and RUG additionally work on the feasibility of using mobile phone applications for people with AD to collect information on social communication and social exploratory measures.

Main differences between EMIF EHR, EMIF AD and DPUK were identified in terms of data extraction, harmonization and analysis. While EMIF AD routinely harmonises the data before uploading them to TransSMART, DPUK does not employ a common data model but instead uploads the data in the original format provided by data owner. EMIF-EHR on the other hand uses Jerboa for the extraction of the data and code mapping is in the hand of the data custodian. In case of EMIF AD, analysis of the anonymised data can be done either on TranSMART or locally. EMIF EHR uses the remote research environment Octopus for the analysis. In case of DPUK, the analysis takes place via the remote research environment provided within the platform and data are not allowed to leave the servers. Combining data from the different platforms would thus require new contracts. How to best integrate CT placebo data in one of the platforms is under evaluation and will be reported in the Deliverable 3.4 (The Final Report on proof of concept technical solutions for RWE data harmonisation and integration), which is due later during the project.

# 1. Introduction

WP3 is responsible for identification, mapping, and integration of real world evidence (RWE) data. WP3 facilitates the access to data needed to answer the defined research questions within the ROADMAP consortium by WP2, 4, and 5.

WP3 makes use of established data search tools and data platforms that have been developed in IMI EMIF and DPUK projects. These tools help to facilitate rapid and secure data acquisition, access, integration, and analyses. Both in EMIF and DPUK, Data Catalogues have been developed that can be used to search for appropriate data. To securely store and analyze cohort data, EMIF makes use of the TranSMART data platform. All data is harmonized and uploaded on the TranSMART platform to allow large-scale data pooling if desired. Also, data export is allowed to analyze the data offline. For EHR data, EMIF makes use of Jerboa and Octopus. DPUK uses a virtual platform (VMware), which requires the researchers to analyse the data on the platform. Using data from EMIF or DPUK data platforms will require approval and new contracts with data owners.

A procedure for requesting data access has been developed separately for EHR data and cohort data. Researchers within ROADMAP fill out a scientific research question form that includes information on the study aim, inclusion criteria and requested variables (see Annex I). Researchers can also propose some cohorts or EHR data of interest. Research questions need to be in line with those defined within the consortium and approved by the respective WP leads. Next, the data request form is sent to WP3. After clarifying the specifics of the study proposal (if needed), the WP3 team identifies (additional) cohorts or EHR data of interest using the EMIF and DPUK Catalogues. Then, the WP3 team approaches data owners for participation in this proposed study. First ROADMAP partners are approached because partners only need to sign a MTA, which makes the process go faster. We will develop a MTA template for use in ROADMAP but if data owners prefer to use their own MTA template that is possible as well. If requested, next non-partner data owners will be approached for participation. This requires setting up and signing of contracts. WP 1 will support the contracting and WP8 will have a look at the data sharing requests and approval by data owners from an ethical perspective. Upon approval of participation and after agreeing and signing of the contracts or MTAs, the data owners prepare their database or subset of database for sharing. Data will then be uploaded on the secure data platforms, for EMIF after data harmonization. If the data is already available on the EMIF or DPUK platforms, data can be shared with the researchers immediately. The data access procedure is summarized in Figure 1.

Below we will describe the EMIF and DPUK data tools in more detail and explain the procedures of data access used in our first use cases.

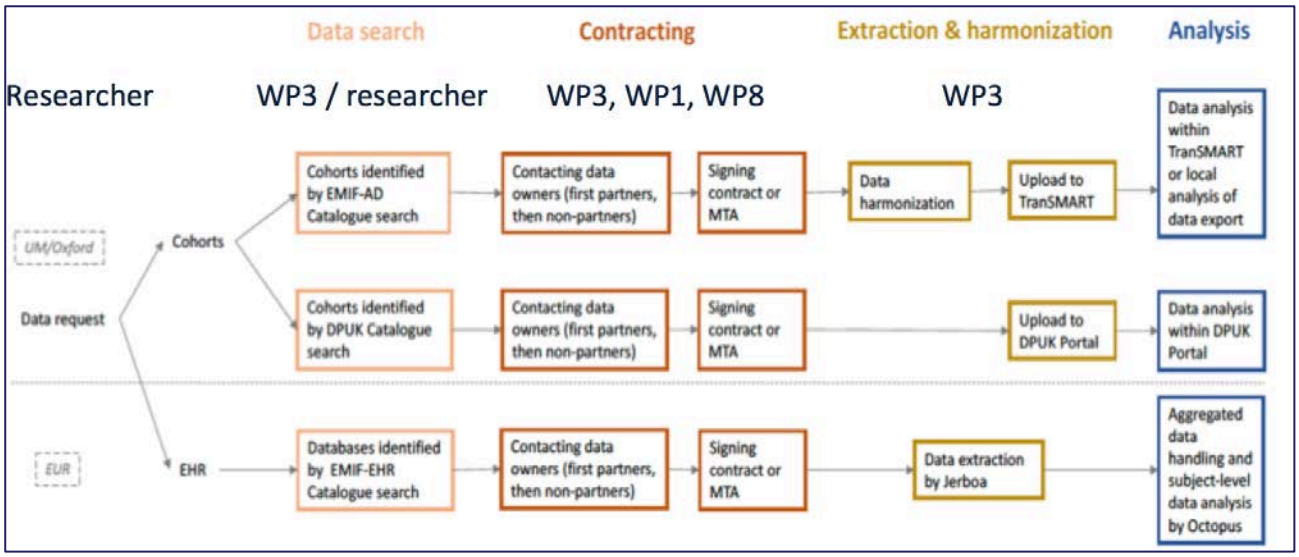


***Figure 1*** *General data request flow*

## 1.1. Introduction to the EMIF EHR platform

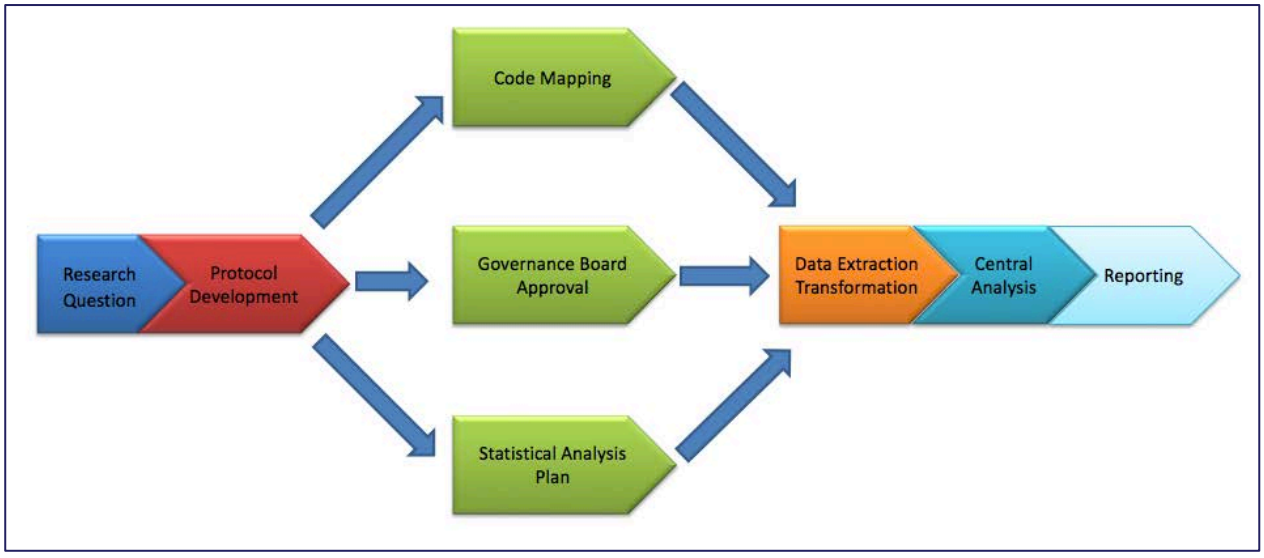


***Figure 2.*** *High-level workflow for performing complex studies in EMIF.*

EMIF EHR is one of the platforms in ROADMAP. The figure above (Figure 2) presents a high-level overview of the necessary steps to perform a complex collaborative study in the distributed network of data sources in EMIF EHR. The summary here focuses on the steps following code mapping, approval and statistical analysis plan. Jerboa and Octopus play a crucial role in these processes and are described below.


***Figure 3.*** *Jerboa Reloaded model for distributed data transformation*

Jerboa is the main tool for the data extraction and harmonisation (Figure 3). The Jerboa software is used in a so-called distributed network design, i.e. it runs de-identification and analysis locally at each data source site. Analytical datasets are produced that contain all relevant variables in an aggregated or patient level format. Jerboa runs a script that contains all parameters of a specific study design. This has the advantage that the local analyses are performed in a common way and are not subject to differences in implementation by local statisticians.

Jerboa additionally includes a Quality Control model, which is executed on the input files, and several models to perform necessary data transformation steps as described in the statistical analysis plan.

The data custodian will extract all the necessary data following the input files specifications provided in the statistical analysis plan. For the diagnosis (or clinical events), for example, this would include mapping to event labels based on the final mappings provided by the Code Mapping Step. The data transformation is then done locally by running Jerboa Reloaded on the common input files. Jerboa Reloaded then produces encrypted output files that can be uploaded by the data custodian to the Remote Research Environment.

The Octopus infrastructure (Figure 4), hosted at the Erasmus Medical Center, is used as a prototype for the private remote research environment (RRE). It allows for secure file transfer from and to the data custodians and can be used to collaborate on the analytical dataset generated by Jerboa Reloaded.

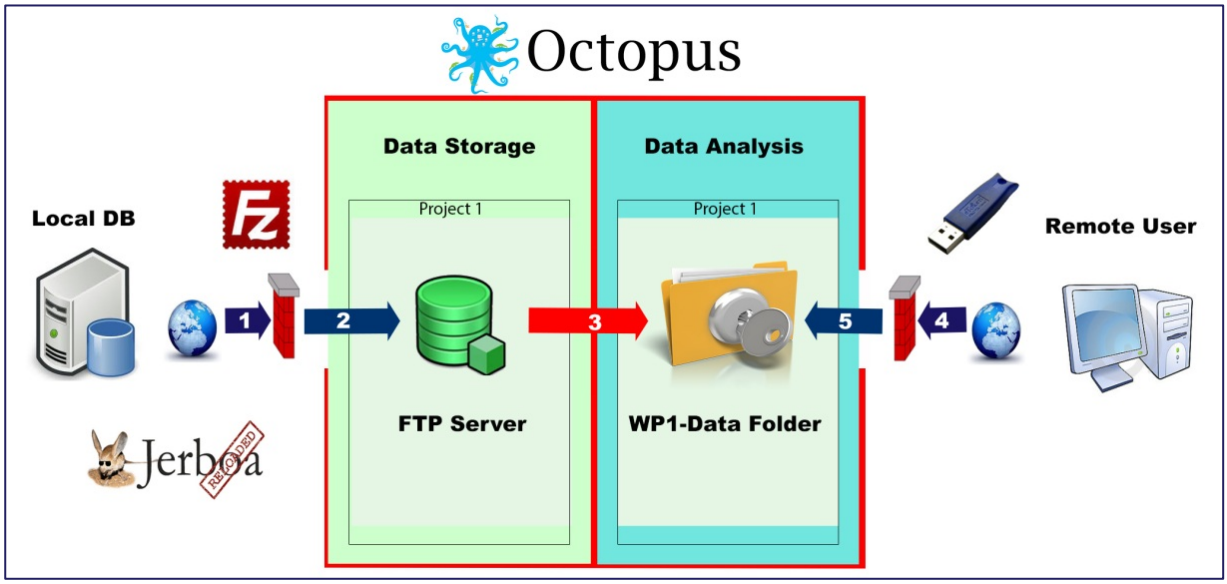


***Figure 4*** *The Octopus infrastructure.*

The infrastructure consists of an application server (Windows 2008 R2) that contains several analytical tools, word processing software, and utilities. It can host multiple research projects, each with its own secured area to share data and results. This facilitates the distribution of tasks, e.g., post-processing of Jerboa Reloaded output files. Procedures have been developed to ensure data protection and secure file transfer from and to the collaborating partners. On the server the researchers need to develop code in for example SAS, R or Stata, to produce the final tables from the analytical datasets of each participating database.

## 1.2. Introduction to the EMIF AD/TranSMART platform

### *1.2.1. EMIF-AD Catalogue*

The European Medical Information Framework (EMIF)-Alzheimer's Disease (AD) project developed an online catalogue aimed at supporting collaborative studies in the AD research field (https://emif-catalogue.eu). The EMIF-AD Catalogue contains meta-data of participating studies and can be used by researchers to browse and search for information on the different cohorts. Each cohort has its own 'fingerprint' which contains an overview of information on data access, study characteristics, in- and exclusion criteria, number of subjects, clinical information, dementia rating scales, subjective cognitive impairment, neuropsychiatric scales, quality of life, caregiver, cognitive screening tests, neuropsychological tests, physical examination, blood collection (including genetic analyses), CSF collection, urine collection, MRI, PET, CT scans, SPECT scans, electrophysiology and neuropathology. The Catalogue enables researchers to compare and explore different cohorts that could potentially participate in their research projects and currently includes 44 cohorts.

The Catalogue is currently being enriched with more information about cohorts related to AD that are also of interest to ROADMAP. 17 new cohorts have been approached to enter information about their cohort to the Catalogue and more cohorts will be approached the coming time. Also the content of the Catalogue has been slightly adjusted to be in line with the needs of ROADMAP. Questions on the availability of data of mortality and health resource utilization have been

incorporated in the fingerprints. Questions on mobile health data and additional physical information will soon be added. This information will be completed for new and existing cohorts in the Catalogue.

### *1.2.2. TranSMART*

TranSMART is a secured data platform on which data can be safely stored, managed and analysed. All cohort data stored on tranSMART is anonymised. Data uploaded to TranSMART will be harmonized according to the EMIF-AD common data model to enable pooling of different cohort data. TranSMART can store clinical data as well as high dimensional data such as genomics data. Within TranSMART, it is possible to compare cohorts, generate summary statistics and analyse data. Data files can be exported and analysed locally as well. However, control of the data remains with the data owner and access to the cohort data is restricted to users approved by the data owner based on research questions.

## 1.3. Introduction to the DPUK catalogue and platform

DPUK's portal provides data storage, management and a secure environment for the analysis of cohort data relevant to ageing and dementia. Once the portal's capabilities are fully developed, DPUK will host over 44 cohorts with over 2 million participants. DPUK does not employ a common data model. Instead the data are stored on DPUK servers in their original, cohort-specific, format alongside a DPUK converted csv-file. DPUK's system is designed to leave the ownership of the data with the cohort PIs and the application process makes sure that the cohorts have the necessary level of governance over their data to ensure that they are only used for research that is approved by them. This also implies that the data cannot leave DPUK.


***Figure 5.*** *Cohort comparison tool and cohort matrix*

Cohort search and application takes place on the DPUK website. Three main tools are provided to assist with the identification of relevant cohorts. A high-level overview over the available information is provided for each of the cohorts via a cohort matrix, which shows the categories of data that is collected, like lifestyle factors, cognition or brain imaging. The main source of information is the cohort comparison tool, which provides in-depth, specific information on ten types of data categories

 

alongside cohort descriptives (see Figure 5 Cohort comparison tool and cohort matrix for both tools). Cohort specific information is additionally available in form of primary publications or codebooks if available. Once the relevant cohorts are selected, researchers can apply for access to the data on DPUK's website. One of DPUK's aims is to make data access mechanisms as simple as possible. Part of this strategy is a streamlined application procedure. Only a minimal set of core information is requested for all applications, consisting of project title, public interest, data requested, scientific context, start and end dates, as well as key words. As cohorts might have different needs, an additional dynamic set of information is requested based on the selected cohorts. Applications are firstly screened by DPUK and then sent to the cohort contact persons for final approval.

Data access and analysis within DPUK is graphically shown in Figure 6 Data access and analysis in DPUK. Most of the cohort data are uploaded and stored on the UK Secure eResearch Platform (UKSeRP) shared infrastructure at Swansea University. However, some of the cohorts prefer to share their data on a study-to-study basis. In case data are not pre-anonymised by the cohort owner, they are anonymised using the Swansea anonymization services in conjunction with NHS Wales Informatics Service (additional restrictions are in place in case an analysis would lead to results based on five or less participants). The core infrastructure of DPUK in UKSeRP consists of 15 Intel 40 core 96GB servers with 720TB of fully backed up storage. Upon approval of the application and signing of the DPUK Data Access Agreement, data are released into the secure virtual desktop infrastructure and are accessible for the researcher via a two-stage authentication based on username/password and Yubikey token. Data can be accessed and analysed via a remote virtual desktop that uses VMware Horizon client in a secure analytical space within UKSeRP. The virtual desktop contains a variety of tools for data management and statistical analysis like SQL Server Management Studio, R, Eclipse and Stata. Additional software can be installed on request. For data that is derived from different sources on a modality basis (genetic, imaging) the Portal has integrated the DPUK Genomics Platform for search and analysis of genetic data that is provided in raw form (PLINK available for analysis for example); and an instance of XNAT for imaging, that allows search, upload and analysis of imaging data in a node and hub format (9 nodes across UK partner institutions and a central hub in Swansea). Both platforms allow for the derivation of summary data that can then be added to, or analysed alongside, any other cohort data within the VDI. Data linkage for DPUK datasets can be done via a common ID model. The common ID model allows cross-cohort and cross-modality participant linkage. Additionally, linkage to routine data from sources such as the NHS is possible, if consent is given and identifiers are available to perform anonymised matching. All analyses are done within the secure analytic space and only manually approved result files for publishing will leave UKSeRP. Cohort data will never leave UKSeRP.

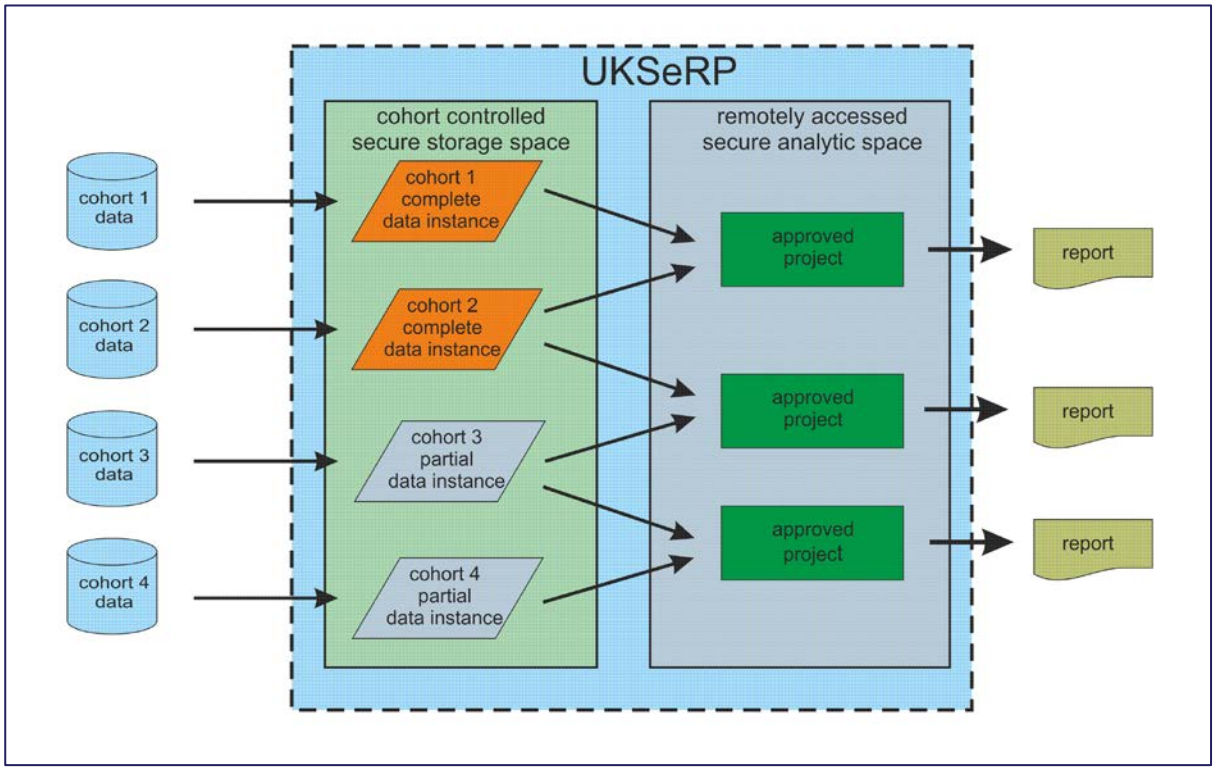


***Figure 6.*** *Data access and analysis in DPUK*

# 2. Use case 1 - Novartis WP4 Model

WP3 received a data sharing request from WP4 requesting data to be used for validation of the Novartis Alzheimer's disease (AD) prevention model, an existing disease progression model (insert link to data request here). After clearly defining the inclusion criteria, required variables and optional variables, we searched for suitable cohort data in the EMIF and DPUK Catalogues. Determining which cohorts are suitable for each data request depends on the model to be validated. For this specific request, the Novartis pre-symptomatic model validation, the process of defining the inclusion criteria, required variables and optional variables is discussed in more detail below (Section 2.2, subheadings 'Data request' and 'Identification of cohorts'). Next, we approached data owners for participation in the proposed research project. Currently two cohorts are interested to share data and are preparing the database for sharing. Below we will describe these procedures and used tools more in detail.

## 2.1. Data Sources

Data sources to validate the Novartis AD prevention model were identified by searching the AD cohort catalogues: The EMIF-AD and DPUK catalogues.

### *2.1.1. EMIF-AD Catalogue*

For the proposed study, we identified 8 cohorts that would be of interest:

- Athens, Greece: Hellenic Longitudinal Investigation of Aging and Diet (HELIAD)
- Kuopio, Finland and Stockholm, Sweden: Cardiovascular Risk Factors, Aging and Dementia (CAIDE)
- Gothenburg, Sweden: Prospect Population Study of Women (PPSW)
- Gothenburg, Sweden: Population Study (H70)
- Bonn, Germany: German Study on Ageing, Cognition and Dementia in Primary Care Patients (AgeCoDe)
- Cambridge, UK: Medical Research Council Cognitive Function and Ageing Study (MRC-CFAS)
- Duisburg, Germany: Heinz Nixdorf Recall Study (RECALL-HNR)
- Brescia, Italy: Alzheimer's Disease Repository Without Borders (ARWIBO)

### *2.1.2. DPUK Catalogue*

Using DPUK's cohort comparison tool and cohort specific-information, 6 cohorts were identified as potentially useful for the validation of the model. Focus during the selection process was on the non-optional variables in the data request. The 6 cohorts are

- EPIC Norfolk
- Whitehall II

- Lothian Birth Cohort 1936 (LBC 1936)
- Caerphilly Prospective Study (CAPS)
- English Longitudinal Study of Ageing (ELSA)
- MRC National Survey of Health and Development (MRC NSHD).

## 2.2. Methods, Tools and Processes

**Data request**

The WP4 data request for validation of the Novartis AD prevention model (Annex II) contained a description of the model, and the aims, data requirements and inclusion criteria of the proposed study. First, some ambiguities were clarified and, in consultation with the WP4 team, the inclusion criteria were adjusted. The initial criteria included solely cohorts with abnormal biomarker data on over 1000 cognitively normal individuals. Since this was not feasible, abnormal biomarker data was no longer required for inclusion and also smaller sample size were allowed for inclusion. Figure 7 provides a graphical depiction of the data access flow in Use Case 1.

**Identification of cohorts**

The WP4 team identified four cohorts they considered especially relevant for their research question: the Amsterdam, Rotterdam, Memento and BioFINDER cohorts. These cohorts were not identified in our initial Catalogue search as the Amsterdam and Memento cohorts did not fulfil the initial inclusion criteria while the Rotterdam and BioFINDER cohorts have not been fingerprinted yet. We proceeded to contact these data owners. Considering that 3 of these cohorts are not (yet) partners in the ROADMAP project, the process of contacting these cohorts and requesting access to their data requires extra time. An initial contact with these four data owners has been established, but the contact is progressing slowly and we are not sure whether we will be able to include their data for this use case shortly. The Amsterdam and MEMENTO cohorts are already included in the EMIF-AD Catalogue, while the fingerprinting of the Rotterdam and Biofinder study is still ongoing.

We then proceeded to search the EMIF-AD and DPUK Catalogues for additional cohorts that could be of interest for the proposed study. The Catalogue searches are described in more detail below. Two ROADMAP partner cohorts that met the inclusion criteria were contacted first. We contacted them asking them to check whether their cohorts indeed have most of the variables available, and if so, whether they would be willing to cooperate and share their data.

**Data sharing approval**

Approval for data sharing has been realized for four cohorts so far: the Memento, Amsterdam, H70, and PPSW cohorts. The Biofinder cohort has expressed interest to participate in the longer term.

The Memento cohort just became partner and approved to share data in Roadmap so contracts need to be set up for data sharing. Access to a subset of the Amsterdam data is approved but if more data is desired additional contracts need to be signed. The Biofinder cohort has indicated to have insufficient resources to participate in ROADMAP by sharing data at the moment. They may provide data at a later stage of the project. The Rotterdam study will first complete the fingerprinting before data sharing will be discussed.

From our additional search in the EMIF Catalogue two cohorts already agreed to share data: The H70 and PPSW studies. They are preparing the datasets and we expect to be able to include their data shortly. At a later point in time alternative, non-partner cohorts can be contacted.

**Data sharing**

After approval of data use of identified cohorts in the EMIF and DPUK Catalogues and signing of MTA/contracts, data will be shared with researchers via the available data platforms. If data is already available on the EMIF or DPUK data platforms this can go rather fast. If data is not yet available, it will be uploaded on the secure data platforms, for EMIF after harmonisation on TranSMART, and for DPUK unharmonised on VMware. Once cohort data is uploaded on the data platform(s), the request process for future studies with the cohort data will require less time and effort since the data is then already uploaded on the platform. For each new research question, the data owners of the appropriate cohorts included in TranSMART or VMware will be asked whether they are interested in sharing their data for that particular research question. Only in case additional variables will be requested, additional harmonisation and data upload will be required.

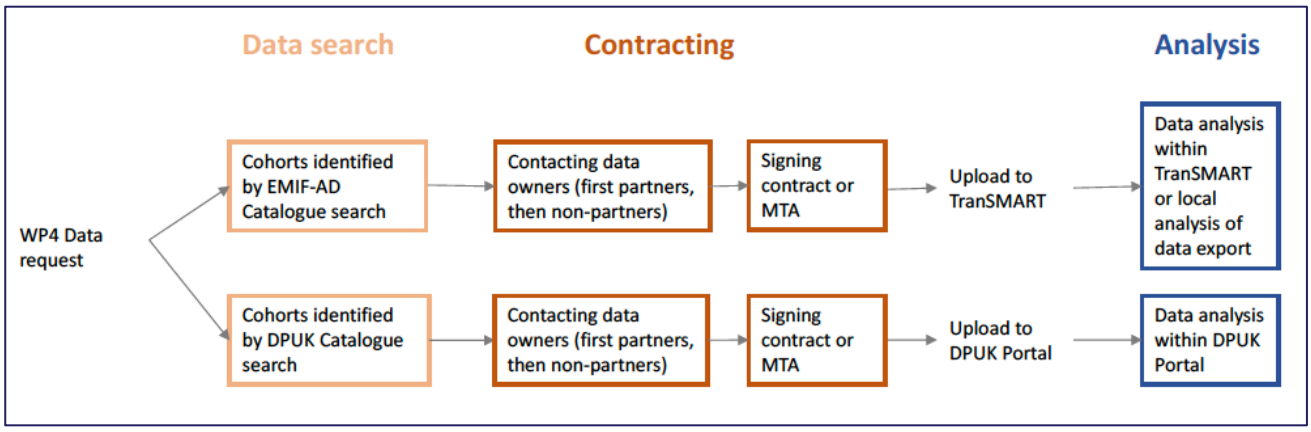


***Figure 7*** *Data flow for Use Case 1*

In the case of DPUK, the chronological order of past and immediate future steps for the validation of the time-to-diagnosis model are as follows:

1. Data request received
2. Data search within DPUK based on data request
3. Application for data access from DPUK

## 2.3. Assessment of Availability and Suitability of Data Sources

### *2.3.1. EMIF-AD Catalogue*

We consulted the EMIF-AD Catalogue for availability of data sources meeting the Prevention model validation request. We used the following search criteria: (cognitively normal subjects OR subjects with SCI) AND (follow-up performed AND information on dementia at follow-up) AND (age AND gender AND education) AND APOE ε4 tested AND (dementia screening test OR neuropsychological testing OR cognitive screening test). The search returned 8 cohorts, each

containing data on over 1000 cognitively normal subjects or subjects with SCI. See Table 1 for an overview of these cohorts and the availability of the variables requested.

Since Gothenburg is a partner of the ROADMAP project, we proceeded to contact these cohorts and informed whether the PPSW and H70 studies indeed had most of the requested variables available and whether they would be willing to share their data for this research question.

***Table 1*** *EMIF-AD Catalogue cohorts and availability of variables*

| Requested variables | HELIAD | CAIDE | PPSW | H70 | AgeCoDe | MRC-CFAS | RECALL-HNR | ARWIBO |
|---|---|---|---|---|---|---|---|---|
| | *Athens* | *Finland* | *Gothenburg* | | *Bonn* | *Cambridge* | *Duisburg* | *Brescia* |
| N | 1100 | 1270 | 1141 | > 1000 | 1338 | 18000 | 1460 | 1500 |
| Age, gender, education | X | X | X | X | X | X | X | X |
| Cognitive outcome | X | X | X | X | X | X | X | X |
| Neuropsychological outcome | X | X | X | X | X | X | X | X |
| APOE ε4 status | X | X | X | X | X | X | X | X |
| Follow-up diagnosis | X | X | X | X | X | X | X | X |
| Comorbidities | X | X | X | X | X | X | X | X |
| Medication use | X | X | X | X | X | X | X | X |
| *Optional* | | | | | | | | |
| Family history of dementia | X | X | X | X | X | | | |
| Amyloid-beta in serum/CSF | | | X | X | | | | X (subgroup) |
| Tau in serum/CSF | | | X | X | | | | X (subgroup) |
| FDG-PET | | | | | | | | X (subgroup) |
| PET amyloid beta imaging | | | | | | | | |
| Alcohol intake | X | X | X | X | X | X | X | X |
| Smoking | X | X | X | X | X | X | X | X |

### *2.3.2. DPUK Catalogue*

***Table 2*** *Overview DPUK Catalogue and availability of variables*

| | **EPIC Norfolk** | **Whitehall II** | **CAPS** | **ELSA** | **MRC NSHD** | **LBC1936** |
|---|---|---|---|---|---|---|
| N | 25639 | 10308 | 2512 | 12100 | 5362 | 1091 |
| Age, gender, education | X | X | X | X | X | X |
| Cognitive tests/neuropsychological outcome[1)] | X | X | X | X | X | X |
| APOE ε4 status | X | X | X | X | X | X |
| Follow-up diagnosis[2)] | X | X | X | X | TBC | TBC |
| Comorbidities | X | X | X | X | X | X |
| Medication use | X | X | X | X | X | X |

Notes: 1) Each cohort has some measures of cognition. However, the cohorts vary regarding of the amount of information that is available, the type of information that is collected and how often the information is collected. 2) Data linkage can be used to access potential diagnoses in case of EPIC Norfolk and Whitehall II. CAPS has clinical assessments, while ELSA only has self-reported diagnosis. For MRC NSHD and LBC1936, the availability of clinical diagnosis has yet to be confirmed.

The three main reasons for immediate exclusion of a DPUK cohort from further consideration were lack of biomedical information, focus on a different type of dementia and failure to fulfil the sample size restriction set out in the data request. However, if the sample size restriction would be softened, which could be interesting given that the ultimate aim is to see how different data can be used simultaneously to increase statistical power, the Aberdeen Birth Cohorts from 1921 and 1936

 

(L. J. Whalley et al. 2011) and other studies could be of interest as well. For nearly all studies, cognitive function is not necessarily assessed using a complete scale but instead items from different scales are used. Only specialised studies can afford to include long, specialised item batteries. For MRC NSHD, the overlap between waves regarding the items for cognitive function is unclear. Generally, concluding that a certain study has a measure of cognitive function that is richer than MMSE seems difficult. Three measures of cognitive function before diagnosis are not or not yet available for all studies. From the available information, it is not entirely clear whether or not LBC1936 and MRC NSHD differentiate between types of dementia. ELSA has only self-reported diagnosis. Availability of comorbidities varies by cohort but is comprehensive for those that allow data linkage with electronic health records. ELSA data are available on DPUK servers, while the others are still in the process of sharing their data. See Table 2 for an overview of the selected cohorts and availability of requested variables.

# 3. Use case 2 - SIDIAP Dementia Diagnosis Validation Study

The ROADMAP project aims to set new standards for the collation and evaluation of real world evidence on Alzheimer disease. Real world data can be provided by EHR databases, which offer research opportunities with many advantages, such as capability to show the reality of dementia in primary care practice, large sample size, representativeness, and relatively low economical cost.

One constraint of EHR databases from primary care services is the accuracy of diagnoses. General practitioners play a pivotal role in the recognition of dementia, in gatekeeping primary health care services, the accuracy of dementia diagnoses registered in the electronical health records is crucial as general practitioners play a pivotal role in the recognition of dementia. The diagnosis of dementia is not only a key point for the patient, their family, and caregivers, but it might also affect the quality of research studies. Thus, accuracy of the diagnostic records is imperative in studies based on electronic medical records.

The use case 2 in WP3 aims to assess the accuracy of dementia diagnosis in electronic health records. We will examine SIDIAP, one of the primary care databases previously listed in EMIF-EHR catalogue in task 3.1. SIDIAP contains structured records from about 5.8 million people attended in primary care centers in Catalonia (Spain). The validation will be based on algorithms that compare the diagnoses of dementia registered in SIDIAP with data from an electronic survey administrated to general practitioners, and with data related to anti-dementia drugs from an external source - the Catalan Advisory Board for Treatment of Alzheimer Disease. Through the assessment of dementia diagnosis accuracy, this validation contributes to develop task 3.4, which aims to evaluate availability and suitability of data from real world data sources. Therefore, the ROADMAP project provides a unique opportunity to address the necessity of assessment of not only accessibility but also suitability of data sources across the disease stages of Alzheimer disease.

## 3.1. Data Sources

The SIDIAP Database (www.sidiap.org) contains routine records of consultations from nearly 275 primary care practices from the National Health Service of Catalonia. SIDIAP includes anonymised longitudinal medical records related to demographics, symptoms, diagnoses, prescriptions, and socio-economic deprivation from about 5.8 million people (>80% of the Catalan population) [Med Clin (Barc) 2012;138(14):617–21]. The quality of these data for research purposes has been previously evaluated for certain diseases, such as cancer [Qual Prim Care 2012; 20(2):135-45], cardiovascular diseases [Rev Española Cardiol 2012; 65(1):29–37], and rheumatoid arthritis [Clin Rheumatol 2016; 35(3):751-7], but not for dementia. Thus, accuracy of dementia diagnoses in SIDIAP needs to be assessed prior to use the records for research purposes.

## 3.2. Methods, Tools and Processes

A survey will be conducted to request additional information on dementia diagnosis from general practitioners, one of the most robust methods of validation. The questionnaire will be administered to the general practitioners that integrate the Agency of Clinical Research Management in Primary

Care (AGICAP), from the Primary Care Research Institute IDIAP Jordi Gol. The AGICAP encompasses about 200 accredited general practitioners from 70 Catalan primary care centers (80% of the SIDIAP population). These general practitioners are trained and experienced in recruitment of patients in clinical trials, and in reviewing diagnoses recorded in the electronic medical history. The AGICAP offers fast recruiting and paperwork (including payments), data quality, and overall agility in the process. Therefore, the AGICAP network facilitates communication and coordination between SIDIAP and primary care professionals.

ROADMAP researchers will email an invitation to participate in this study to the 200 general practitioners from AGICAP. Amongst those who accept to participate in this study, 24 general practitioners will be randomly selected. Each general practitioner will review the electronic medical history of a maximum of 10 randomly selected patients following the questionnaire. The selected patients will be aged 18 years or older and will have a record of one of the following dementia ICD-10 (International Classification of Diseases 10$^{th}$ revision) codes: Alzheimer's disease (G30 and subtypes, F00 and subtypes), vascular dementia (F01 and subtypes), unspecified dementia (F03 and subtypes), other types of dementia (F02.0-F02.4, F02.8, G31.0, G31.8, G31.0).We have defined a broad inclusion criterion considering patients with a registered diagnosis of any kind of dementia subtype because some degree of misclassification between dementia subtypes is plausible in electronic medical records. A broad inclusion criterion will ensure identification of cases of Alzheimer's disease misclassified as any other dementia subtype.

A sample size of 240 cases – that is, 24 general practitioners and 10 patients per practitioner - will suffice to estimate a positive predicted value of 85% - similar to that seen in diagnoses recorded in clinical histories in England [Br J Clin Pharmacol 2010; 69(1):4-14] with a non-response rate of 20%, and a precision of 5%.

Patients who have been prescribed anti-dementia drugs will be defined directly as cases, without evaluation through the questionnaire, because such cases have been previously confirmed by the Catalan Advisory Board for treatment of dementia. In Catalonia, the prescription of anti-dementia drugs needs approval from an Advisory Board, a group of experts who evaluate all patients with dementia who may need pharmacological treatment.

The questionnaire was designed based on one main question about the basis of the dementia diagnosis with 5 possible answers. Depending on the chosen answer subsequent questions are laid out (figure 8). The recorded information included diagnosis from a hospital specialist; cognitive, functional, and behavioral tests; subtype and severity of dementia; and accomplishment of the DSM-IV or ICD-10 diagnosis criteria.

**Question 1 - Which is the current base of the diagnosis?**

**Answer 1.1- The diagnosis was made by a hospital specialist**

- Please indicate the subtype of dementia
- Please indicate the severity of dementia

**Answer 1.2- The diagnosis is based on cognitive and functional tests**

- Please indicate the tests and the scores
- Please indicate the subtype of dementia
- Please indicate the severity of dementia
- Are the DSM-IV or
- ICD-10 diagnosis criteria compiled? (Yes/no)

**Answer 1.3- The diagnosis is based on a clinical impression.**

- Symptoms observed by the GP, care giver or patient relative.
- Please indicate the subtype of dementia
- Please indicate the severity of dementia
- Are the DSM-IV or ICD-10 diagnosis criteria compiled? (Yes/no)

**Answer 1.4- The diagnosis is inconsistent (evolved or incorrect diagnosis)**

**Answer 1.5- Uncertain base due to lack of information (institutionalized; diagnosis was done by another GP…)**

***Figure 8.*** *Survey to be administered to general practitioners to evaluate dementia diagnosis records in SIDIAP.*

## 3.3. Assessment of Availability and Suitability of Data Sources

The WP3 of ROADMAP will contribute to a 'filled in' data cube, offering a view on which data sources are able to detect specific outcomes based on high quality data. Therefore, the ROADMAP project offers an opportunity to evaluate the quality of data in electronic medical record databases for research purposes.

In order to assess the suitability of dementia diagnoses, we will contrast the records in the SIDIAP database with the questionnaire applied by general practitioners and with the criteria of the Catalan Advisory Board for treatment of dementia.

Patients will be defined as true positives (patients with a dementia diagnosis recorded in SIDIAP who really have the disease) if they have (figure 9):

1. A prescription of anti-dementia drugs (ATC Index: N06DA, N06DX). In Catalonia the prescription of anti-dementia drugs needs approval from an Advisory Board. General practitioners must request to this Advisory Board evaluation of all patients with dementia who may need pharmacological treatment. The Advisory Board is composed of a group of experts who review the diagnosis of dementia and decide the treatment for each patient.
2. A record of dementia based on a hospital specialist's judgment (e.g. neurologist or psychiatrist).

3. A record of dementia based on the results of cognitive and/or functional tests and the fulfilment of the DSM-IV or ICD10 diagnostic criteria.
4. A record of dementia based on the clinical impression of the general practitioner and the fulfilment of the DSM-IV or ICD10 diagnostic criteria.

We will consider the diagnosis of dementia as false positive (patients with a dementia diagnosis registered in SIDIAP who really do not have the disease) if the record of dementia was based on (figure 2):

1. the results of cognitive and/or functional tests without fulfilment of the DSM-IV or ICD10 diagnostic criteria.
2. the clinical impression of the general practitioners without fulfilment of the DSM-IV or ICD10 diagnostic criteria.
3. Inconsistent evidence (the diagnosis has evolved or is incorrect).
4. Insufficient or inadequate information.

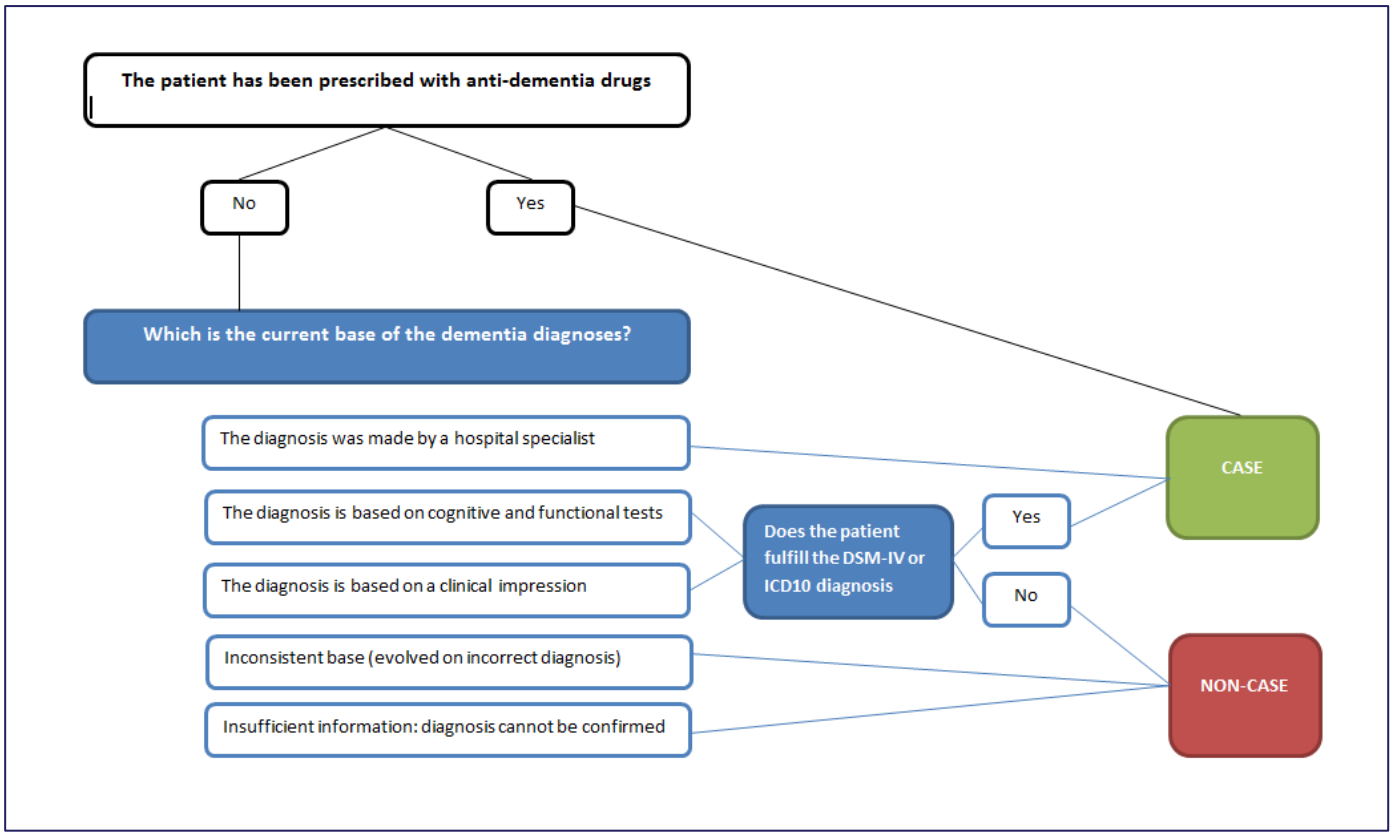


***Figure 9.*** *Algorithm to define true and false cases of dementia.*

We will calculate the positive predictive value (PPV) of dementia diagnosis, which indicates the probability that a person with a record in SIDIAP suffers from this disease. Higher values of PPV indicate better accuracy of the dementia records in SIDIAP.

# 4. Use case 3 – WP4 Ron Handels Model Validation Study

Another model that is going to be evaluated is the natural disease progression model of cognition among people with AD, developed by Handels et al. (2013). The validation exercise focuses on the prediction of MMSE scores in incidence cases of AD in a population of people 75 years and older. The original model was developed using data from the Kungsholmen project, a population-based cohort following all registered inhabitants of the Kungsholmen district in Stockholm, Sweden. Clinical assessments of the 1082 cognitively healthy people took place three times, with three years between the waves. Global cognitive function was assessed using MMSE scores and a potential dementia diagnosis was carried out by physicians based on clinical examination and cognitive tests using DRM-III-R/NINCDS-ADRDA criteria. 323 cases of AD were identified within the 9 years of follow-up and onset of AD was assumed to have taken place in the middle of the follow-up interval. The final equation for the development of MMSE scores over time includes time since diagnosis and age. The equation for the validation is as follows:

$$\mathrm{MMSE} = 26.87 - 3.26 * \mathrm{Time} - 0.35 * (\mathrm{Age} - 75) + 0.10 * \mathrm{Time} * (\mathrm{Age} - 75),$$

where Time is years after being diagnosed with AD. For the validation, we will look at incidence cases of AD in people 75 years or older. The data are required to have AD diagnosis, age, age at diagnosis and at least one MMSE measurement after diagnosis.

## 4.1. Data Sources

So far data from EMIF-EHR were used for the validation and potential cohorts in DPUK were identified for future analyses.

Source of data from EMIF-EHR is a longitudinal observational database of electronic patient records of Dutch general practitioners (GPs), the Integrated Primary Care Information (IPCI) database. The setting of the data is primary care. About 485 Dutch GP participate. IPCI covers roughly 2.4 million subjects. The full medical record is available, including free text. For most practices, the communication with other care providers is available (referrals, etc.).

Another potential source of data from EMIF-EHR is The Information System for the Development of Research in Primary Care (SIDIAP). SIDIAP is a Catalan primary care database with continuous data collection since 2006 on a total of almost 7.5 million individuals, of which 5.5 million are currently active. Electronic medical records related to dementia or outcomes listed by the WP2 are available.

Dementias Platform UK (DPUK) was searched for data that match the requirements for the validation of the model. Relevant data were identified using DPUK's cohort comparison tool and searching for relevant information in cohort-specific publications. Based on the requirements for the MMSE prediction model and the information on the cohorts available up to date, 3 cohorts were identified as potentially valuable for the validation exercise and to pilot data access procedures for ROADMAP within DPUK. The three cohorts are

- Brains for Dementia Research Initiative (BDR)

- Whitehall II
- Lothian Birth Cohort 1936 (LBC1936.

## 4.2. Methods, Tools and Processes

The validation of the MMSE model is nearly completed in EMIF EHR using IPCI data. A TRIPOD development statement was completed in a first step based on information from the original publication. Subsequently, a TRIPOD validation form was completed to facilitate transparent reporting of the model validation study. There will be one TRIPOD statement for each dataset that is being used for the validation. IPCI data have been transformed for Jerboa based on the relevant SAP (see appendix). Jerboa was executed on the data from IPCI, quality control outputs were produced and the anonymised and encrypted data were uploaded to Octopus for the final analysis.

The validation process using SIDIAP data includes several stages. Firstly, the TRIPOD checklist for the prediction model validation was adapted to the characteristics of SIDIAP and its framework. Thus, a specific TRIPOD validation form for SIDIAP data was created. Secondly, we are currently in the data management stage, in order to transfer SIDIAP data to Jerboa platform. Then the statistical analysis plan (SAP) using SIDIAP data will be run in Jerboa. The final statistical analysis using data from several sources –not only SIDIAP- will be conducted using Octopus platform.

As the cohorts in DPUK are not permitted to leave the servers in Swansea, the EMIF EHR approach cannot be translated one-to-one to DPUK and changes to the procedure are necessary. Particularly, Jerboa cannot send data to Octopus for analyses, despite the anonymisation process implemented in Jerboa. It was decided to still run Jerboa to produce aggregated quality control outputs and ensure that sample restrictions are applied consistently across cohorts. For this reason, Jerboa will be installed locally on DPUK servers but in the case of DPUK, no data will be sent to Octopus. DPUK is in the process of assessing the implications of Jerboa's capabilities to send data.

DPUK does not employ a common data model. This implies that data harmonisation and coding will be completed individually for each cohort on DPUK servers. The data will be transformed into the same format as EMIF's ICIP data for the use in Jerboa using statistical software available on DPUK's virtual research environment. Subsequently, Jerboa will produce quality outputs and apply sample restrictions. The main analyses will then be run locally using the same scripts as the ones used by Octopus. The main difference is that the scripts will be run in the specific statistical software locally instead of in the private remote research environment Octopus provides.

The chronological order of the steps in the process of the validation of the MMSE model with DPUK data is then as follows:

1. Tripod statement received with information on validation exercise
2. Data search within DPUK based on Tripod statement and relevant primary publication
3. Tripod statement for DPUK data completed
4. Submission of application for access to data from DPUK
5. Jerboa will be installed on DPUK servers upon approval from cohort owners
6. Data access and formatting according to Jerboas requirements

7. Execution of Jerboa on DPUK data
8. Statistical scripts run on DPUK data within the UK Secure eResearch Platform

Model performance will be evaluated using linear regression between observed and predicted values as well as using median absolute deviation between predicted and observed values.

## 4.3. Assessment of Availability and Suitability of Data Sources

Among the cohorts listed on DPUK, the three main reasons for immediate exclusion were lack of full MMSE scores; the cohort does not reach the age restriction; it focuses on a different type of dementia, like dementia due to Parkinson's disease. For one, there is no follow-up after the diagnosis. Thus, despite seemingly little requirements of the model, a number of cohorts cannot contribute to the validation of the model. Assessment of the suitability of all cohorts was complicated by a substantial variation in the quality of documentation between cohorts. Lack of availability of codebooks and data dictionaries means that outcome categories can often not be assessed and the timing of variables remains unclear. As the cohorts are not necessarily focused on dementia or Alzheimer's disease, it is unclear whether LBC1936 and Whitehall II (as well as other studies) follow-up on people who are diagnosed with a major neurocognitive disorder. Additionally, due to the age restriction in combination with the focus on incidence cases, it is difficult to assess in advance how many data points the final selection of cohorts will be able to add to this specific kind validation exercise. Two of the three identified studies are population based cohorts, while BDR is a registry that is more selective in terms of its participants compared to the original study. All three cohorts will be made available online on DPUK in the near future. However, only BDR is imminent for upload to the platform at the time of writing.

# 5. Use case 4 – Estimation costs of dementia – pilot study

Alzheimer's disease (AD) can now be diagnosed in non-demented subjects by the assessment of amyloid pathology in cerebrospinal fluid or by PET scanning. Treatment for AD is probably most effective in non-demented individuals because neuronal damage is still limited. However, clinical decline is limited as well, which makes it difficult in trials to detect cost-effectiveness of clinical effects within a reasonable follow-up. This is in particular the case in the preclinical stage of AD, when amyloid pathology is present but cognition is unimpaired. Aim of the present use case within WP4 was:

- To estimate the duration of preclinical, mild cognitive impairment (MCI), mild dementia, and moderate dementia stage of AD
- To estimate healthcare and societal costs at each stage
- To model how treatment with a disease modifying drug in non-demented subjects affects costs across the total disease duration

## 5.1. Data Sources

We identified prospective studies in which data were collected on non-demented subjects with known amyloid status or studies with subjects with a clinical diagnosis of AD-type dementia. It was required to have data on mortality as well.

Subjects were selected from four cohorts:

- Memory clinic based Amsterdam dementia cohort (amyloid positive subjects with subjective memory decline (SMD), MCI, dementia)
- Memory clinic based European multicenter study Descripa (amyloid positive subjects with SMD, MCI)
- Memory clinic based European multicenter study ICTUS (subjects with clinical diagnosis of AD-dementia)
- Research cohort ADNI (amyloid positive subjects with normal cognition, SMD, MCI, dementia)

We are now in the process to update the dataset with subjects from:

- Research cohort AIBL (amyloid positive subjects with normal cognition, SMD, MCI, dementia)
- Population-based Gothenburg study (amyloid positive subjects with normal cognition, SMD, MCI).

## 5.2. Methods, Tools and Processes

We harmonized a minimal dataset (age, education, gender, MMSE score, amyloid status, clinical diagnosis, death, APOE genotype) in a common data format from EMIF-AD at Maastricht University and VU University medical center. Healthcare costs at each clinical stage were taken from published studies. We modeled disease progression by multistate modeling, taken mortality into

account (Figure 10). We estimated conversion rate from one stage to the next stage and estimated the average duration for individual in each stage. We modeled trial effect on costs using a number of scenario's assuming an effect of treatment on reduction of conversion rate in subjects with preclinical or prodromal AD.

## 5.3. Assessment of Availability and Suitability of Data Sources

From the 4 cohorts listed above access to data was obtained with 6 months. Data were successfully pooled and analysed. We noted differences in disease duration between memory-clinic-based settings and research settings and age groups (Table 3).

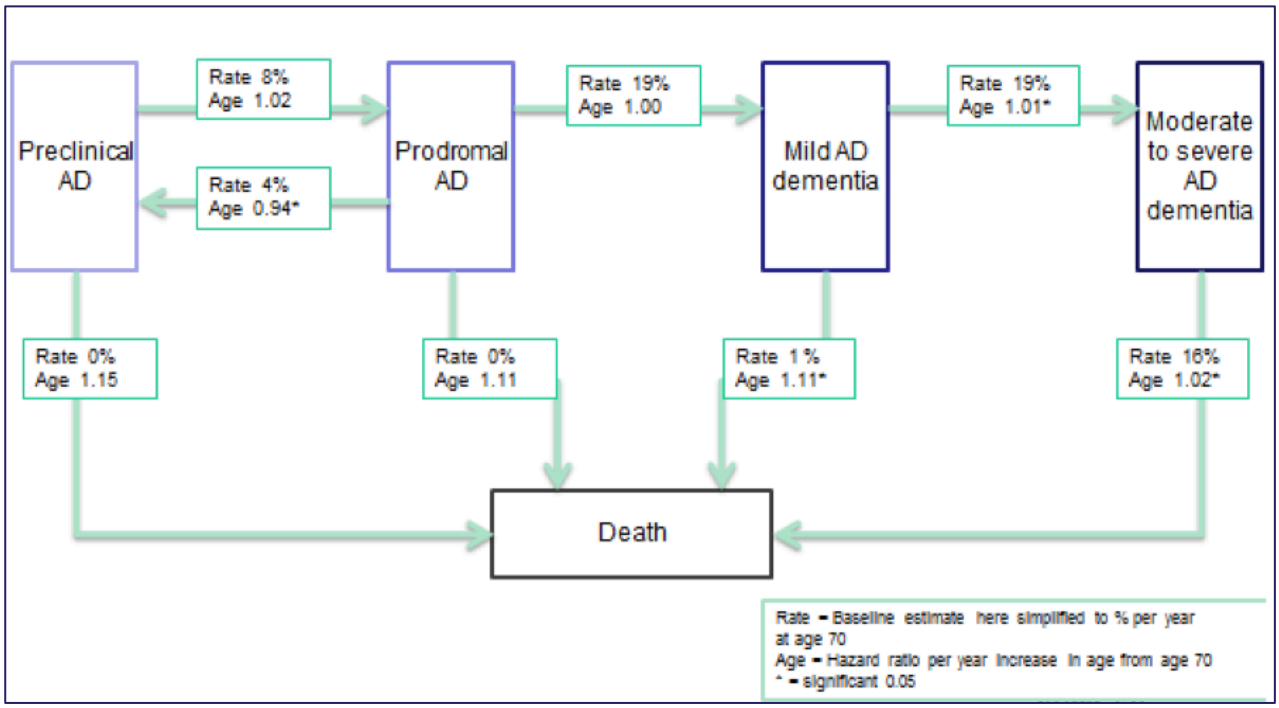


***Figure 10.*** *Multistate model with translation probabilities*

***Table 3.*** *Duration of preclinical, prodromal, mild dementia and moderate-severe dementia stages as a function of setting and age at diagnosis in participants with preclinical AD at baseline*

| Age at baseline | **Age 60** | **Age 70** | **Age 80** |
|---|---|---|---|
| **Time in preclinical AD** | **13.2 (10.8-15.1)** | **10.2 (8.7-12.1)** | **7.8 (5.7-10.3)** |
| Population-based | 14.9 (12.3-17) | 10.7 (9.1-12.4) | 7.5 (5.7-10.1) |
| Memory clinic | 5.1 (2.9-6.9) | 3.5 (2.5-5) | 2.2 (1.3-3.7) |
| **Time in prodromal AD** | **4.6 (3.9-5)** | **4.2 (3.5-4.8)** | **3.6 (2.3-4.6)** |
| Population-based | 4.5 (3.7-5) | 4.2 (3.3-4.9) | 3.8 (2.4-4.7) |
| Memory clinic | 5 (3.1-5.6) | 4.6 (3.4-5.1) | 3.7 (2-4.8) |
| **Time in mild dementia** | **3.7 (3.2-4)** | **3 (2.4-3.3)** | **2.1 (1.4-2.7)** |
| Population-based | 3.5 (3-3.8) | 3 (2.3-3.4) | 2.2 (1.5-2.8) |
| Memory clinic | 4.5 (2.7-4.8) | 3.6 (2.6-4) | 2.5 (1.4-3.1) |
| **Time in moderate to severe dementia** | **3.9 (3.2-4.7)** | **3.1 (2.4-4)** | **2.2 (1.4-3)** |
| Population-based | 3.8 (3-4.5) | 3.1 (2.3-3.9) | 2.3 (1.4-3.1) |
| Memory clinic | 4.9 (2.9-5.7) | 3.9 (2.8-4.5) | 2.7 (1.5-3.5) |

# 6. Short Report - Using Smart Phone App in Dementia patients Pilot

As part of WP3, VUMC and RUG are implementing a feasibility study in which a smartphone application (developed by RUG) will be installed on the smartphone of AD patients (recruited by VUMC). This pilot study is meant to identify feasibility of using passive behavioural monitoring as a way of real word assessment of social communication and (social) exploratory measures in these patients. Social withdrawal is a major burden in this patient group, and objective longitudinal measures to assess this behaviour are highly needed, both in view of monitoring and for potential intervention strategies. As part of this pilot study, VUMC has received approval for the protocol from their ethical committee.

As a collaborative effort between RUG and VUMC, we intend to publish on these results from the study. We want also integrate our findings (e.g., on feasibility) as part of the Task 3.3. deliverable report on the utility of digital technology for ROADMAP.

# 7. Conclusion and next steps

This first report on proof of concept technical solutions for RWE data harmonisation and integration shows that solutions for the harmonization and integration of data are available in order to support Alzheimer Disease research in the EU. Main capabilities are currently focusing on cohort data and EHR data, which are either harmonized and available in the same location (EMIF AD/TranSMART), kept in original format but co-located in a secure environment (DPUK) or kept at original locations and ad-hoc extracted into a common data format including aggregation and analysis steps for each research project (EMIF EHR/Jerboa). New capabilities to integrate continuous patient-generated data and validation of diagnosis in EMR records are explored as well. The usage of CT placebo data is considered for disease model validation and data harmonization and integration for this data type is starting.

During the coming months the ongoing Use Cases as described above and further research studies will provide deeper evidence on best practices for the various ways of integrating and analysing the data. Here we foresee that there will be no one-model-fits-all approach finally, based on the fact that various data sources require different approaches due to:

- variation in privacy and ethical requirements
- level of data integration allowed by data custodians
- variations in data structure and original data standards of source data
- geographical location of the data.

We have to look into a way to integrate knowledge and data access pathways at another layer, representing the central data cube, which guides and leads researchers to the right data at the right location with the appropriate methodology for their research question.

Overall, it is of benefit to continue and expand close collaboration with disease-related projects (e.g. IMI EMIF, IMI EPAD etc.) or overarching IMI projects (IMI BD4BO) to explore deeper synergies, shared learning and prepare for sustainability of capabilities and tools which were developed during the project phases.

# ANNEXES

# Annex I. TEMPLATE for ROADMAP Scientific Questions

## TEMPLATE - ROADMAP: Scientific Questions

**BACKGROUND**

Please describe your Scientific Question for ROADMAP by completing 1 form per Scientific Question and send it to WP3 leads Antje Hottgenroth, hottgenroth_antje@lilly.com and Pieter Jelle Visser, pj.visser@maastrichtuniversity.nl. Please also cc Alba Jené, ajene@synapse-managers.com, Stephanie Vos, s.vos@maastrichtuniversity.nl and your own WP leads. The WP3 team will evaluate your proposal and provide support in contacting data-owners to provide data needed to solve your specific Scientific Question.

**RESEARCH PLAN *add short title/acronym here***

| **1. Project title** |
|---|
| *Please add a descriptive project title here, from which the goal/topic of the Scientific Question (SQ) should be clear.* |
| **2. Aims and objectives** |
| *Please add a short description of the general aim & hypothesis to be assessed for this SQ and explain why it is important to address this SQ, and what is the benefit of doing this in the framework of ROADMAP.* |
| **3. Study design & methods** |
| *Please add a short overview of the number of subjects needed, inclusion/exclusion criteria to be applied, measures of interest, techniquees and tools to be used, overall approach – including any risks and alternative approaches for assessing this SQ.* |
| **4. Outcomes** |
| *Please add a comment on the expected outcomes for this SQ and how this links to the work done in the different WPs of ROADMAP (and any relevant milestones/deliverables).* |
| **5. Timelines** |
| *Please add a general overview of timelines, including clear (interim) deliverables.* |

| **6. Cohorts of interest** |
|---|
| *Please indicate if you would like the WP3 team to search appropriate cohorts for your SQ and/or add any suggestions on which cohorts to be approached based on your own search.*<br>• **Cohorts of DoW:**<br>• **Other cohorts:** |
| **7. Budget needed for analyses** |
| *Please give a ball-park estimation of the budget needed for the proposed study. Please indicate the source of that budget (i.e allocated budget in ROADMAP, proposed budget from ROADMAP, other available budget etc.)* |
| ***8. Public interest*** |
| *Please describe shortly what the public interest is of your research question.* |

## APPLICANT INFORMATION

| **9. Principal investigator** |
|---|
| *Please provide contact details for the PI for this SQ – This PI will also take responsibility for the monitoring of the work done for this SQ and reporting the results to the ROADMAP leadership team* |
| **10. Key team members** |
| *Please provide an overview of the key team members to be involved in solving this SQ.* |

# Annex II: WP4 data request for validation of Novartis AD prevention model

## BACKGROUND

The Novartis AD prevention model (unpublished) is a longitudinal model which explores the natural course of pre-symptomatic changes and was developed to optimize the design of a clinical trial in healthy individuals at risk for developing AD dementia. The model aims to describe the progression of the Alzheimer's Prevention Initiative Composite Cognitive test (APCC) score in relation to time to first diagnosis of mild cognitive impairment (MCI) or dementia due to AD in elderly, cognitively healthy individuals.

- **Aims**

- To test, refine and validate the AD prevention model for APCC or another relevant cognitive test score capable to capture cognitive decline in cognitively normal individuals *before* they are diagnosed with AD dementia in other data bases, and
- To explore the potential of the AD prevention model to identify patients at risk to develop AD dementia.

Inclusion criteria

Subjects of any age and gender are of interest. Included subjects should be cognitively normal at baseline and eventually develop MCI and/or AD-type dementia or remain cognitively normal. Subjects with subjective cognitive impairment (SCI) can be included as well.

- **Requested variables**

- Demographics
  - Age
  - Gender
  - Education
  - *Optional*: family history
- At least one cognitive or neuropsychological outcome that is richer than Mini-Mental State Examination (MMSE) score
- Biomarkers
  - APOE *ε4* status
  - *Optional*: amyloid-beta in serum and/or cerebrospinal fluid (CSF)
  - *Optional*: tau in serum and/or CSF
  - *Optional*: FDG-PET
  - *Optional*: PET amyloid-imaging
- Clinical variables:
  - Clinical diagnosis of MCI and/or AD-type dementia at follow-up

- Comorbidities
- Medication use
- *Optional*: alcohol intake
- *Optional*: smoking behavior

The cognitive, biomarker and clinical variables should comprise baseline measurements and at least two follow-up measurements per patient on average.

- **Key study team members**

Helene Karcher (h.karcher@analytica-laser.com)

Noemi Hummel (n.hummel@analytica.laser.com)

Billy Amzal (B.Amzal@analytica-laser.com)

 