D4.4 Results from pilot model validation exercises

116020 - ROADMAP

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

WP4 – Disease Modelling and Simulation

Lead contributor	Jan Kors (3 – EMC)
	j.kors@erasmusmc.nl
Other contributors	Luyuan Qi (16 – Novartis)
	Ron Handels (4 – UM)
	Vincent Bouteloup (26 – Memento)
	Josep Garre (6 – IDIAP JORDI GOL)
	Anna Ponjoan (6 – IDIAP JORDI GOL)
	Nazib Seidu (10 – UGOT)

Due date	30/04/2018
Delivery date	18/02/2018
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	15/10/2018	First draft
V1.1	16/11/2018	Integration of comments of WP4 members
V1.2	5/12/2018	Integration of second round of comments
V1.3	18/12/2018	Final version after Consortium review

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**. Göteborgs 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)
  - **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)
- **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.

Abbreviations

AD	Alzheimer's Disease
APCC	Alzheimer's Prevention Initiative Composite Cognitive test score
MCI	Mild Cognitive Impairment
MMSE	Mini-Mental State Examination
NACC	National Alzheimer's Coordinating Center
RRE	Remote Research Environment
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
TTE	Time to event

Publishable Summary

This deliverable reports on the external validation of three disease progression models for AD dementia: Handels' Mini-Mental State Examination (MMSE) model, Novartis' preclinical model, and Eli Lilly's institutionalization model. For each of these models, data sources for external validation were sought. A validation pipeline was developed to standardize the analyses across multiple data sources, reduce the workload of database custodians, and increase transparency and repeatability. The pipeline consists of the following steps: documentation of the model development and validation steps, specification of a statistical analysis plan, extraction and standardization of the data, and generation of the validation results.

For the MMSE model, we applied the validation pipeline to a variety of data sources, including electronic health record databases, population-based cohorts, and memory-clinic data sources. We also applied the pipeline to the data that were originally used to develop the model. Results indicate poor to moderate prediction of MMSE scores for individual cases, both for the validation sets and the development set.

The preclinical model consisted of three submodels: a time to event model that predicts time to first diagnosis of MCI and AD dementia, and two models that predict progression of the Alzheimer's Prevention Initiative Composite Cognitive (APCC) score. These models were validated on two external datasets. Since APCC was not available in the external data sets, APCC proxies were constructed. Validation results indicate satisfactory performance for the APCC models, while the TTE model tended to overestimate the overall survival probability.

For the institutionalization model, it was difficult to find external data sets that contained all variables required by the model. The one data source that was selected for external validation contained most of the variables but required conversion of the scales for functional ability. Validation results show a large overestimation of the predicted times for the patients who were institutionalized during follow-up.

We conclude that the validation pipeline was successfully applied to a large number of data sources, and provides a viable approach to generate validation results in a standardized and reproducible way while minimizing the workload of database custodians. The validation results indicate that the predictions of the MMSE model for individual patients are poor to moderate. Although the preclinical model showed satisfactory performance in predicting individual APCC time courses, the TTE submodel tended to overestimate the overall survival probability. The institutionalization model largely overestimates the time to institutionalization during follow-up. These results may partly be explained by differences in setting and patient characteristics between the model development set and the validation sets, but also highlight the importance of comparing external with internal validation results. Finding suitable data sets for external validation of models that are more complicated than the MMSE model can be extremely hard, mostly because there is a large variety of variables across data sets and the variables used in the model only partly match those available in the data sources.

1. Introduction

In the past decades, various disease progression models related to Alzheimer's Disease (AD) dementia have been proposed in the literature. Disease progression models play a crucial role in both the assessment of any therapeutic intervention in the disease process and understanding the (economic) impact of these interventions, and may inform patient recruitment for randomized clinical trials. In Deliverable 4.1, "Catalogue of RWE relevant AD models and simplistic disease stage framework", we performed a literature review of disease progression models and identified a total of 62 different models described in 43 studies. Very few of these models have been externally validated, i.e., their performance has not been assessed on other data sets than the ones that were used to develop the models. Thus, it is not clear how well these progression models generalize to other settings than the ones in which they were developed. Validation on external data sets could clarify this issue and is the subject of this deliverable.

External validation of disease progression models is not an easy task. Finding data sets that can be used for external validation may be problematic. The variables that are contained in the validation sets must match the input and outcome variables required by a given model (possibly after transformation if a variable cannot be matched directly). Also the patient population on which the model is validated has to be in line with the population that was used to develop the model. If multiple data sets are available for validation, one has to ensure that the analyses on the different data sets are standardized as much as possible. Finally, a validation exercise should allow for privacy and governance issues that may prohibit sharing of validation data sets and require data processing and generation of validation results to be done locally, at the site where the validation data reside.

As external validation of all disease progression models that were culled from the literature review is clearly out of scope, we limited ourselves to three different models: Handels' Mini-Mental State Examination (MMSE) model (Handels et al., 2013), Novartis' preclinical model (Caputo et al., 2017), and Eli Lilly's institutionalization model (Belger et al., 2018). The criteria and considerations that led to the selection of these three models have extensively been described in Deliverable 4.3, "Selection of appropriate disease models for validation". Briefly, the criteria included data availability (model variable requirements should not be so stringent that no other data source would be able to provide the right data; at least one of the selected models should have "low" data requirements, i.e., many data sources should be able to provide the data required for model validation), detailed understanding of the model (preferably the original developer of the model should be involved in the validation team), and specific interest of partners participating in ROADMAP.

Here, we will first describe a generic validation pipeline that was established to externally validate disease progression models and to standardize the generation of validation results across multiple external data sources. We will then, for each of the three selected progression models, describe the model in more detail, give an overview of the data sets that were selected for external validation, and present and discuss the validation results that were obtained. Finally, we will conclude with a general discussion on the issues and challenges in external validation of disease progression models.

2. Validation pipeline

A generic validation pipeline was developed to standardize the processing of multiple external data sets for a given disease progression model. The pipeline consists of five steps, where all but the first step are done for each external data source:

1. Fill in a TRIPOD development checklist

For each external data set:

2. Fill in a TRIPOD validation checklist
3. Specify data processing and analysis details in a statistical analysis plan (SAP)
4. Perform data extraction and transformation
5. Generate the model validation results

In the following, the different steps in the pipeline will be described in more detail.

Step 1 is meant to get a good insight into the purpose of the model and how it was developed. For this, we require that a TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) development checklist is filled in. TRIPOD checklists have been developed as tools to stimulate transparent reporting of prediction model studies (Collins et al., 2015; Moons et al., 2015). The checklists include a total of 22 items. The checklists come into two flavors: a development checklist that contains items relevant for model development, and a validation checklist that contains items relevant for model validation. Part of the items in both checklists overlap. Annex I provides a list of all items and whether they are part of the development or validation checklist, or both.

The remaining four steps in the validation pipeline are executed for each data set that is used to validate the model.

Step 2 involves filling in the TRIPOD validation checklist. Part of the items in this checklist will not change between different external data sets (such as assessment of validation results), but items that are specific for the external data set (such as data acquisition and size) need to be specified per data set.

Step 3 requires that the statistical analysis plan (SAP) is updated with information about the external data set. In particular, a general description of the data set has to be added, together with a detailed description of mapping of variables (if any) and how diagnosis was established. The SAP also specifies the study cohort, the variables that need to be collected and in what format, data transformations and data processing, and all output and validation results that will be generated.

In step 4 the data that are needed for the model validation, are extracted from the external data set and transformed as specified in the SAP. For this we use a software tool called Jerboa (Figure 1). Jerboa takes as its input three simple, standardized files with data about patients, events, and measurements, as specified in the Jerboa data preparation and processing manual and in the SAP. The output of Jerboa consists of anonymized analytical data sets that can be used for further

processing. The input files have to be supplied by the database custodian, but because of their simplicity, the effort in creating these files is minimized.

It should be noted that Jerboa is typically installed and executed locally, i.e., there is no need to transfer the data outside the local environment. Jerboa is under full control of the database custodian. Its output can be viewed and approved before it is allowed to be used for further processing. Jerboa runs on a Java platform, which practically means that it can be executed on any computer system. The tool was developed at the Erasmus MC, Rotterdam, and has been used in many multinational observational data studies (Coloma et al., 2011; Trifiro et al., 2014).

Figure 1. Local data extraction and transformation using the Jerboa tool.

Finally, in step 5 validation results are generated by an R script based on the analytical datasets produced by Jerboa, and made available in a secure remote research environment (RRE). As RRE we used the Octopus system (Figure 2) developed by Erasmus MC (Trifiro et al., 2014). There are two options here: either the R script can be executed locally and only the validation results are uploaded to the RRE, or the anonymized analytical datasets (after encryption) are first uploaded to the RRE and then the R script is run on the RRE.

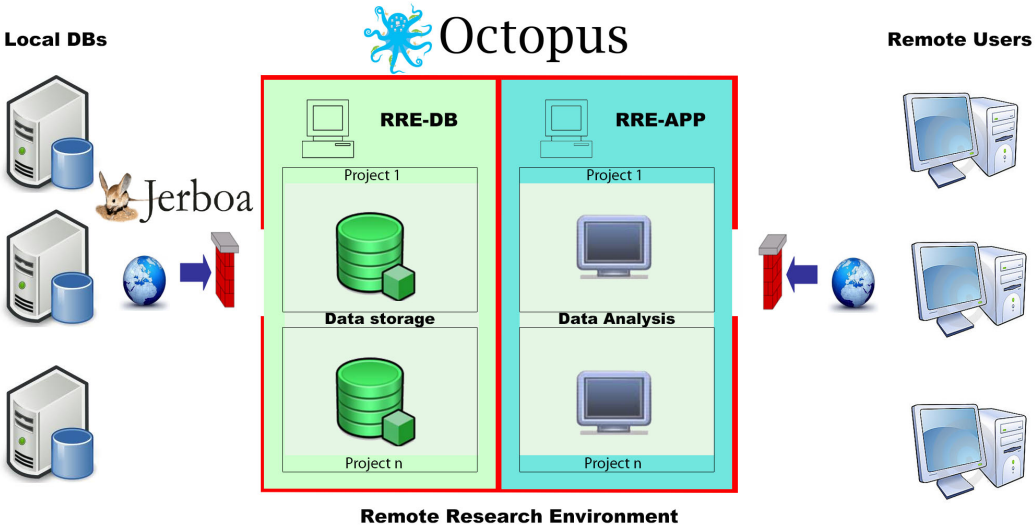


Figure 2. Octopus remote research environment.

3. Validation of Handels' MMSE model

3.1. Model description

Handels et al. (2013) developed a model to predict the natural progression of cognition as assessed by the MMSE score in incident cases of AD dementia in a population of people aged 75 years and older. The 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 1,082 cognitively healthy subjects at baseline took place at 3-, 6-, and 9-years follow-ups. A potential dementia diagnosis was carried out by physicians based on clinical examination and cognitive tests using DRM-III-R/NINCDS-ADRDA criteria. A total of 323 incident cases of AD dementia were identified during follow-up. The onset of AD was assumed to have taken place in the middle of the follow-up interval.

The prediction equation for MMSE is as follows:

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

where Time is years after being diagnosed with AD dementia, and Age is patient age at the time of the measurement that is to be predicted.

For the validation data sets, subjects were eligible if they were diagnosed as incident AD dementia cases, were aged 75 years or older at diagnosis, and had at least one MMSE measurement after date of diagnosis. The two predictors, Age and Time, were derived from the date of birth, date of AD dementia onset, and date of MMSE measurement. The onset of AD was assumed to have taken place in the middle of the follow-up interval preceding the visit at which dementia was diagnosed. Since follow-up intervals lasted an average of 3 years, this was operationalized by adding a time correction of 1.5 years to the Time variable.

3.2. Data sources

External validation of the MMSE model was performed on a variety of data sources. The selection of data sources has been documented in Deliverable 3.4, "Final report on proof of concept technical solutions for RWE data harmonisation and integration". Briefly, selection was based on information from the EMIF-AD and EMIF-EHR Catalogues, as well as additional information about data sources from ROADMAP consortium partners. The selected data sources can be divided in electronic health record databases, population-based cohorts, and memory clinic registries. They include the following:

- GOTHENBURG H70 Studies & Prospective Population Study of Women (PPSW) (Johansson et al., 2010; Arnoldussen et al., 2018). On-going studies with complex cohort-sequential design (longitudinal and cross-sectional cohorts) of 70-year-olds that started in 1971 (H70 Studies) and women of different ages (PPSW) that started in 1968. Participants were selected from the Revenue Office Register based on certain birth dates, without screening. They participate in longitudinal follow-ups at regular intervals until cohort extinction. Some cohorts have been enlarged with new individuals at ages of 85 years and older. Diagnoses are based on DSM-III-R.

- IPCI (Integrated Primary Care Information) (Vlug et al., 1999). The IPCI database is a Dutch primary care database with continuous data collection since 1995 on a total of 2.8 million individuals, of whom 1.8 million are currently active. The average follow-up for individuals is 3 years. The full medical record is available, including free text. IPCI uses the ICPC coding system.
- SIDIAP (Information System for the Development of Research in Primary Care) (García-Gil Mdel et al., 2011). SIDIAP is a Catalan primary care database with continuous data collection since 2006 on a total of almost 7.5 million individuals, of whom 5.5 million are currently active. The average follow-up for individuals is 7.4 years. SIDIAP uses the ICD-10 coding system.
- SIDIGI, a linkage between SIDIAP (García-Gil Mdel et al., 2011) and the Register of Dementias of Girona (ReDeGi) (Garre-Olmo et al., 2009). ReDeGi is an epidemiological surveillance device that provides information about the clinical and demographic characteristics of all new cases of dementia in Girona province (0.7 million inhabitants). This linkage encompasses all patients recorded in ReDeGi who had an EHR in SIDIAP (about 5,000 persons) and provides a unique data source that combines longitudinal real-world data with high-quality dementia records obtained from specialists.
- Copenhagen. The Copenhagen database includes approximately 2.75 million patients who have been admitted to hospitals in the capital region and the region of Zealand, Denmark, between 2006-2016. Data include all diagnoses coded in ICD-10, tests, procedures, drugs administered during the hospital stay, results from laboratory and biochemical tests, and free text. Patients with AD dementia were distinguished in those who received none, one, or more than one dementia drug. The drugs considered were donepezil, galantamine, rivastigmine, and memantine.
- EDAR (acronym of “Beta amyloid oligomers in the early diagnosis of AD and as marker for treatment response”) (Barnett et al., 2010). EDAR is a longitudinal observational cohort of individuals recruited from memory clinics across Europe. Individuals were recruited between 2008 and 2010 and follow-up assessments were performed within three years after baseline.
- Girona (Garre-Olmo et al., 2010). The GIRONA clinical cohort is a two-year prospective cohort study of patients recruited in a memory clinic located in the Santa Caterina hospital in Girona (Catalonia). A total of 905 individuals were recruited between 1998 and 2011 and the follow-up assessments were performed every 6 months. Among measures of disease progression, cognitive functions were assessed with the Cambridge Cognitive Examination and the MMSE by trained neuropsychologists.
- ICTUS (Impact of Cholinergic Treatment Use) (Canevelli et al., 2016). The ICTUS study is a prospective multicenter cohort study aimed at evaluating the clinical course, treatment outcome, and the socioeconomic impact of AD in Europe. It involved 29 participating centers from 12 European countries. After baseline assessment (from 2003 to 2005), participants were followed up to 2 years with midterm reevaluations every 6 months.
- MEMENTO (Dufouil et al., 2017). The MEMENTO cohort is a clinic-based study of patients presenting with a large variety of cognitive symptoms and subjective cognitive complaints, who are followed over a 5-year period. From April 2011 to June 2014, 2323 patients were enrolled in

28 centers of the French national network of university-based memory clinics (Centres de Mémoires de Ressources et de Recherche).

We wanted to compare model performance on the external validation sets with model performance on the Kungsholmen data that was used to develop the model, but the original publication (Handels et al., 2013) did not report those results. We therefore applied for access to the Kungholmen data, with the kind help of Ron Handels, and were granted access for this study.

3.3. Validation results

The validation pipeline was executed for all selected external data sources. Validation checklists were filled in (see Annex II for the development checklists and Annex III for an example of the validation checklists) and the SAP was constructed. Based on the SAP, the Jerboa tool was tuned for this study, a Jerboa installation and user manual was written (Annex IV), and an R script that generates the validation results was developed. Tuning of Jerboa was done by the Jerboa development team at Erasmus MC and took about half a week of work. Writing of the SAP and Jerboa installation and user manual took about one and a half week, and developing the R script another week.

For each data source, Jerboa was used to extract and transform the data. The resulting analytical data sets were processed by the R script. The output of the R script consisted of summary characteristics of the data, plots of the observed MMSE values as a function of time relative to index date (i.e., the date of AD dementia onset), plots of the observed versus predicted MMSE values, plots of the difference between predicted and observed MMSE values as a function of time since diagnosis, and model performance measures. Performance measures consisted of the slope, intercept, and R-squared value of a linear regression between observed and predicted MMSE values, the median absolute deviation of the difference between predicted and observed values (MAD = median($|diff_i$ – median($diff_i$)|)), and the median and interquartile range of the (absolute) difference between predicted and observed values. We chose to compute several performance measures as they capture different aspects of model performance and a single “best” measure does not exist.

The data summary characteristics of the development set (Kungsholmen) and the different validation sets are given in Table 1. The number of patients and number of MMSE measurements varied greatly across the different data sources, with the largest numbers occurring in the SIDIAP, Copenhagen, and IPCI EHR databases. Only the Gothenburg validation set consisted of population-based cohorts, like the Kungsholmen development set, but was smaller in size. The Gothenburg and SIDIGI data sets had the longest follow-up times, comparable to the Kungsholmen data. Follow-up times for IPCI and EDAR were very short.

Table 1: Characteristics of data sources that were used for development and validation of the MMSE model.

Data source	Type[1]	Patients (N)	Age, median (yr)	Female (%)	Measurements (N)	Follow-up[2] (yr)
Kungsholmen	*PC*	*344*	*85*	*83*	*499*	*1.5 (1.5, 4.5)*
Gothenburg	PC	118	82	94	131	3.5 (2.0, 5.5)
IPCI	EHR	2,014	84	68	2,829	0.1 (0.0, 0.8)
SIDIAP	EHR	11,181	83	69	14,466	1.2 (0.2, 3.0)
SIDIGI	EHR + MC	365	81	69	448	2.1 (1.1, 3.8)
Copenhagen	EHR	1,496[3]	84	69	3,107	0.4 (0.1, 1.3)
		2,054[4]	83	64	5,008	0.6 (0.1, 1.7)
		269[5]	82	60	849	1.4 (0.4, 2.6)
EDAR	MC	23	80	39	32	0.2 (0.2, 0.8)
Girona	MC	375	80	68	1,665	0.6 (0.0, 1.5)
ICTUS	MC	803	80	67	2,995	1.0 (0.4, 1.6)
MEMENTO	MC	115	82	52	399	0.5 (0.0, 1.5)

[1]EHR = electronic health record, MC = memory clinic, PC = population-based cohort; [2]Values indicate median (interquartile range); [3]Patients who did not receive any dementia drug; [4]Patients who received one dementia drug; [5]Patients who received more than one dementia drug.

Figure 3A shows the observed MMSE measurements as a function of time relative to the index date (i.e., the date of onset of AD dementia) for the Kungsholmen development set. As expected, the measured MMSE values show a clear declining trend after the index date, but it should also be noted that the variability in the MMSE measurements is extremely large. The predicted MMSE values show a similar decreasing trend (Figure 3B), but with much smaller variability of the estimates as compared to the observed MMSE values. Therefore, the predicted MMSE values for individual patients poorly match the observed values, as shown in Figure 3C. The poor prediction is also reflected in the large variability of the differences between the observed and predicted MMSE values as a function of time after index date (Figure 3D). However, Figure 3D also shows that the median differences over time are (almost) zero, indicating that the median of the MMSE predictions matches well with the median of the observed values.

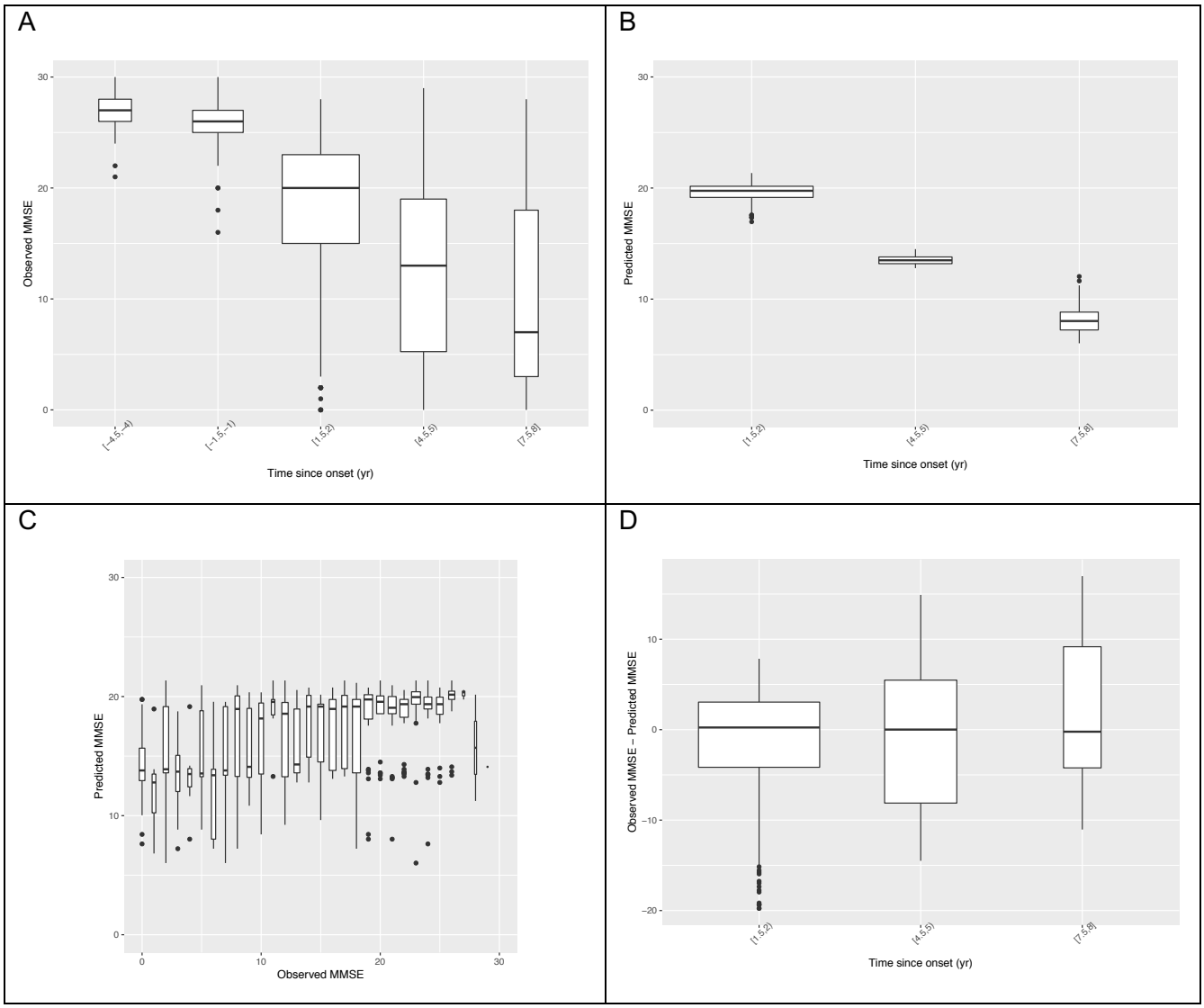


Figure 3. Validation results of the MMSE model for the Kungsholmen development set. (A) Observed MMSE as a function of time relative to the index date; (B) Predicted MMSE as a function of time relative to the index date; (C) Observed versus predicted MMSE; (D) Difference between observed and predicted MMSE as a function of time relative to the index date.

The measures that quantify model performance are shown in Table 2. For the Kungsholmen development set, the slope is only 0.21 with a moderate R-squared of 0.173. The MAD and median absolute difference between observed and predicted MMSE values are larger than 4, reflecting the limited accuracy of the model to predict MMSE for individual patients. Only the median of the overall differences between observed and predicted values is close to 0.

Table 2: Performance of the MMSE model for the development set and different external validation sets.

Data source	Linear regression			MAD[1]	Median difference (IQR)	Median absolute difference (IQR)
	Slope	Intercept	R-squared			
Kungsholmen	*0.21*	*14.00*	*0.173*	*4.01*	*0.24 (-4.75, 3.56)*	*4.16 (1.99, 7.35)*
Gothenburg	0.36	8.81	0.333	4.73	2.89 (-2.37, 7.13)	5.24 (2.84, 8.61)
IPCI	0.04	21.41	0.007	3.82	1.86 (-5.86, 1.85)	3.80 (1.86, 6.66)
SIDIAP	0.18	17.04	0.114	6.21	-6.17 (-12.48, -0.06)	7.41 (3.38, 12.85)
SIDIGI	0.19	15.20	0.108	5.64	-4.27 (-9.94, 1.35)	6.22 (3.10, 10.66)
Copenhagen	0.06[2]	20.17	0.011	4.40	-0.58 (-5.65, 3.38)	4.35 (2.11, 7.61)
	0.06[3]	19.84	0.010	4.16	-0.83 (-5.00, 3.31)	4.18 (1.96, 7.03)
	0.15[4]	17.19	0.050	3.82	-0.94 (-4.94, 2.77)	3.81 (1.83, 6.86)
EDAR	0.01	23.36	0.001	3.92	-3.05 (-6.34, 1.63)	4.13 (2.14, 6.34)
Girona	0.10	20.79	0.022	2.93	-4.34 (-7.30, -1.45)	4.55 (2.26, 7.32)
ICTUS	0.09	20.26	0.034	3.55	-2.67 (-6.25, 0.86)	3.87 (1.74, 6.66)
MEMENTO	0.14	19.11	0.059	3.31	-0.34 (-3.83, 2.90)	3.24 (1.58, 5.03)

[1]MAD = Median absolute deviation; [2]For patients who did not receive a dementia drug; [3]For patients who received one dementia drug; [4]For patients who received more than one dementia drug.

Table 2 also shows the model performance for the external validation sets. Overall, model performance is poor to moderate. In the following, we will illustrate and discuss the validation results for a number of data sources in more detail.

The population-based Gothenburg cohort obtained an R-squared value of 0.333, higher than for the development set. The Gothenburg cohort has a long follow-up. The observed and predicted MMSE scores as a function of time (Figure 4 and Figure 5) show a similar declining pattern as for the Kungsholmen data. Although the relatively high R-squared indicates that the model can explain part of the variability in the MMSE measurements and the model predictions clearly correlate with the observed values (Figure 6), the accuracy of the predictions is generally poor, with large MAD and median absolute difference. The median of the differences between observed and predicted MMSE was 2.89, indicating that the model overall underestimates MMSE for this data set. Figure 7 shows that this occurs mostly in the first five years after the index date.

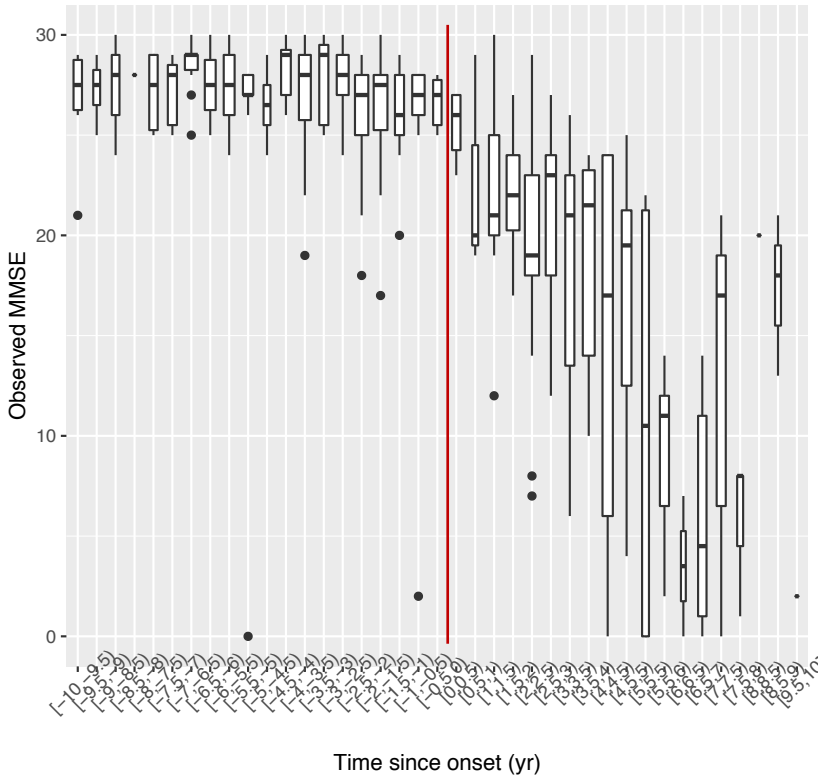


Figure 4. Observed MMSE as a function of time relative to the index date for the Gothenburg validation set. The red line marks the index date.

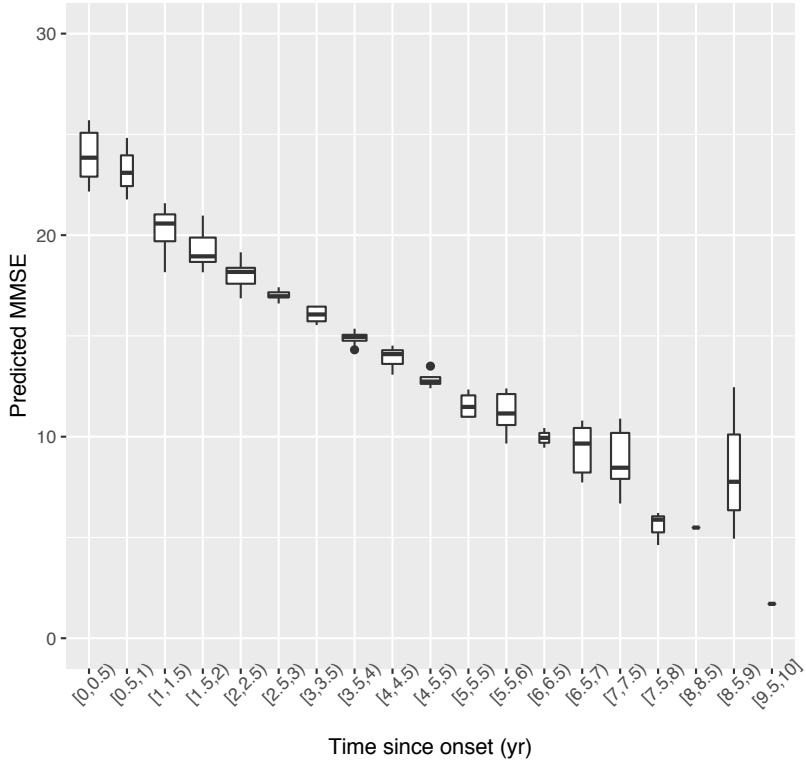


Figure 5. Predicted MMSE as a function of time relative to the index date for the Gothenburg validation set.

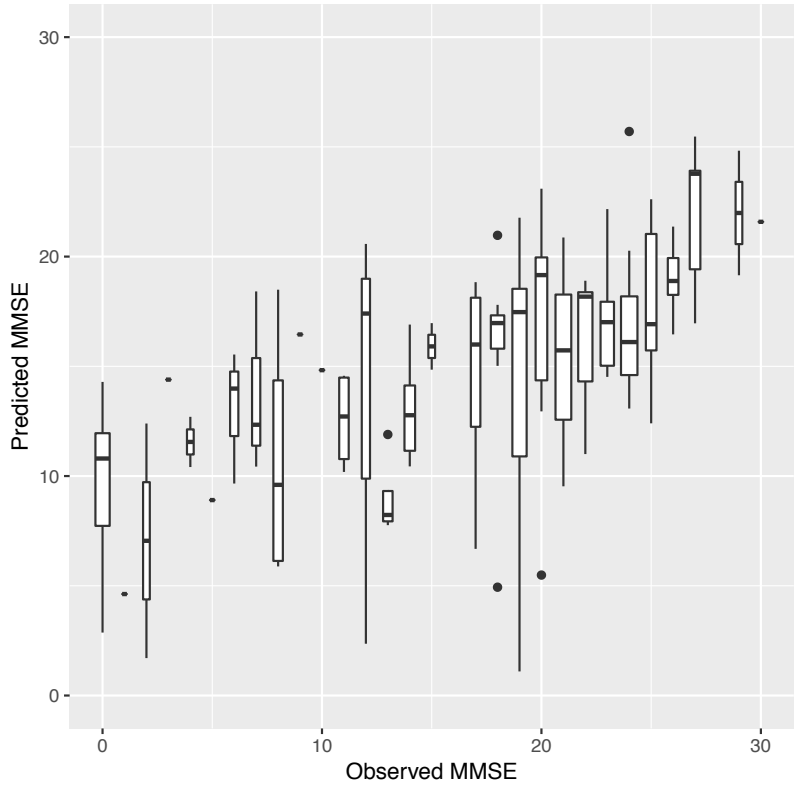


Figure 6. Observed versus predicted MMSE for the Gothenburg validation set.

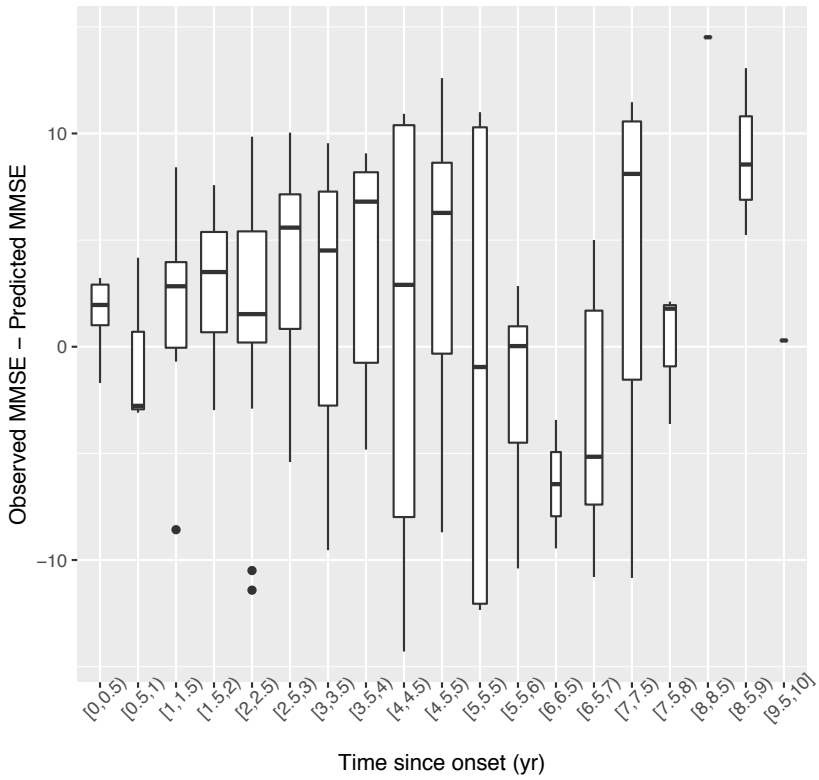


Figure 7. Difference between observed and predicted MMSE as a function of time relative to the index date for the Gothenburg validation set.

On the IPCI data set, the model performs poorly, with an almost flat slope and very low R-squared value. Inspection of the observed MMSE scores reveals that the measurements start to decline about a year before the index date, but instead of further declining essentially remain stable (Figure 8). This may be explained by the fact that IPCI is a primary care database, and general practitioners (GPs) in the Netherlands do not routinely collect MMSE measurements. In fact, the follow-up time in IPCI is very short, suggesting that most patients in the cohort only have an MMSE assessment at or shortly after the index date. There will generally be no need for GPs to measure MMSE again in patients diagnosed with dementia as the outcome will not affect the course of dementia, but reassessment may be more likely for patients who do not show the cognitive decline that usually comes with progressing stages of dementia. Figure 8 also illustrates the huge variability of MMSE scores within each of the time intervals.

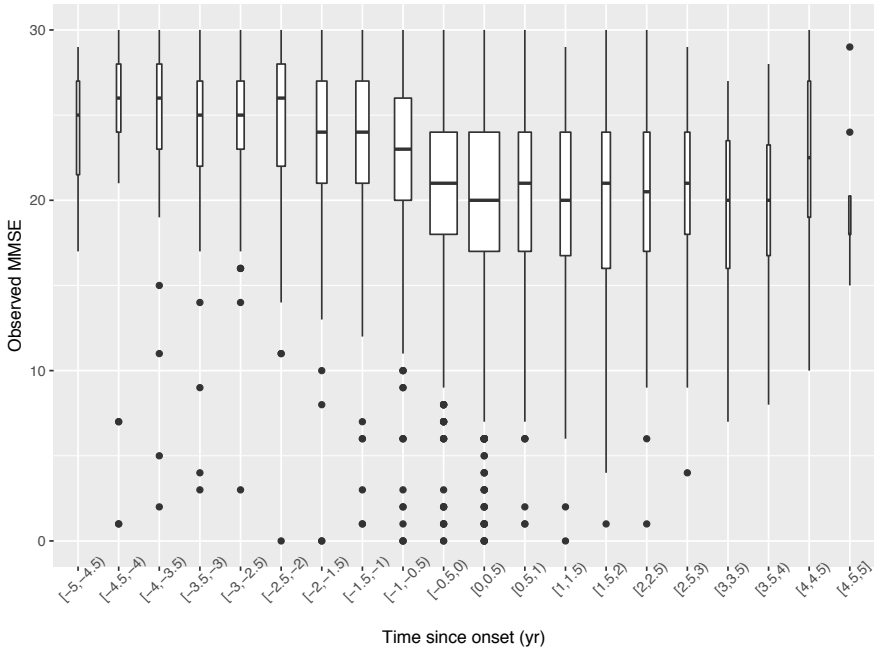


Figure 8. Observed MMSE as a function of time relative to the index date for the IPCI validation set.

The SIDIAP data set is the largest validation set in this study. Figure 9 shows that the observed MMSE measurements start to decrease about one and a half year before the index date (marked by the red line) and show a very large variability. The median of the observed MMSE values around the index date is rather low, at around 19. The predicted MMSE values also show a declining trend, but the median estimates are consistently higher than the observed medians (Figure 10). The model therefore substantially underestimates the MMSE scores (Figure 11), which is also reflected in the high MAD and median absolute difference scores for the SIDIAP data. Figure 11 also suggests that recalibration of the model by changing the constant term in the prediction equation could considerably improve the model performance for this data set.

A potential explanation for the underestimation in SIDIAP is that it may be difficult to determine an accurate index date. SIDIAP is a primary care database, and the diagnosis of AD dementia is made by the GP, who is not a specialist. The diagnosis may have been recorded in the EHR well after the onset of dementia. In an attempt to provide more accurate estimates of the index date, part of the SIDIAP data was linked to registry information in the Register of Dementias of Girona, which contains high-quality dementia records obtained from specialists. The observed MMSE scores for the resulting SIDIGI data set are shown in Figure 11. No substantial differences between the MMSE patterns in SIDIAP and SIDIGI were observed, with similar median MMSE values of about 19 at index date in both data sets. The model performance parameters for SIDIGI in terms of MAD and median absolute difference appear slightly better than for SIDIAP (Table 2), but overall indicate poor model performance.

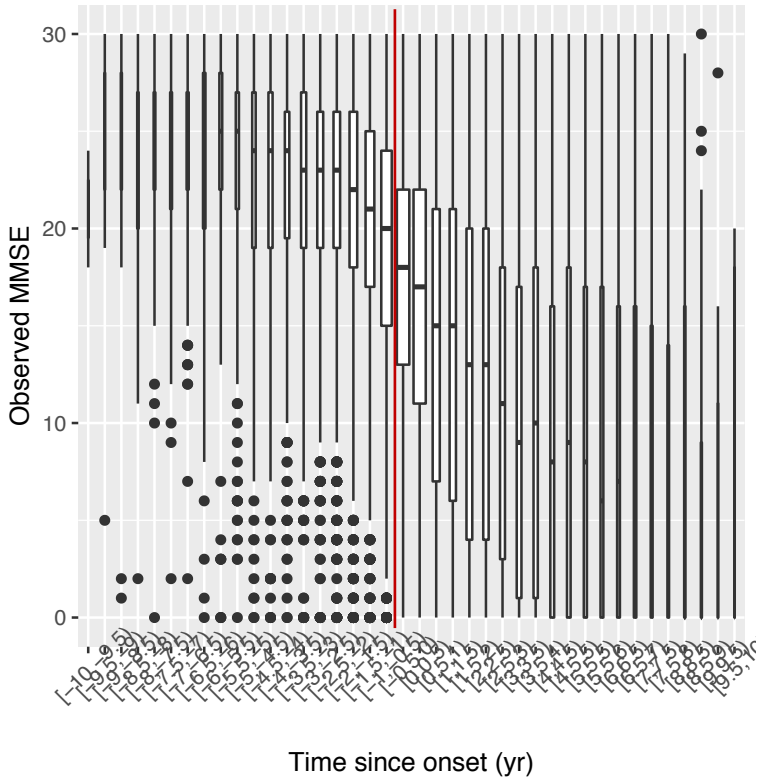


Figure 9. Observed MMSE as a function of time relative to the index date for the SIDIAP validation set. The red line marks the index date.

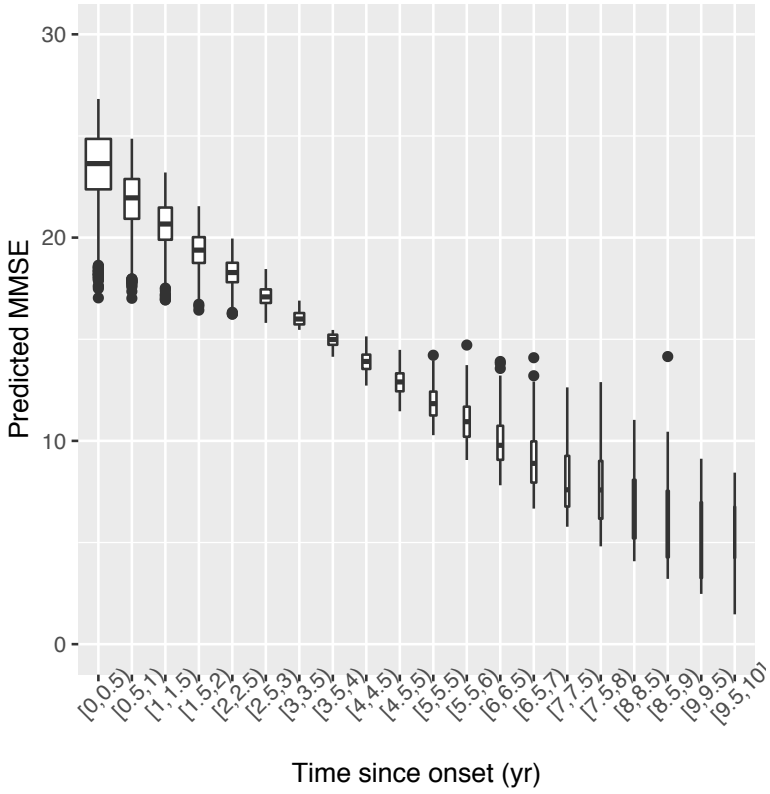


Figure 10. Predicted MMSE as a function of time relative to the index date for the SIDIAP validation set.

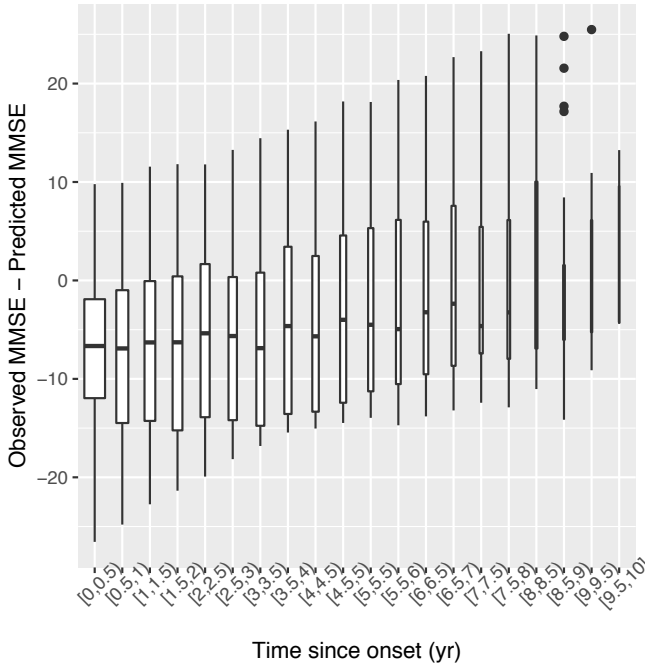


Figure 11. Difference between observed and predicted MMSE as a function of time relative to the index date for the SIDIGI validation set.

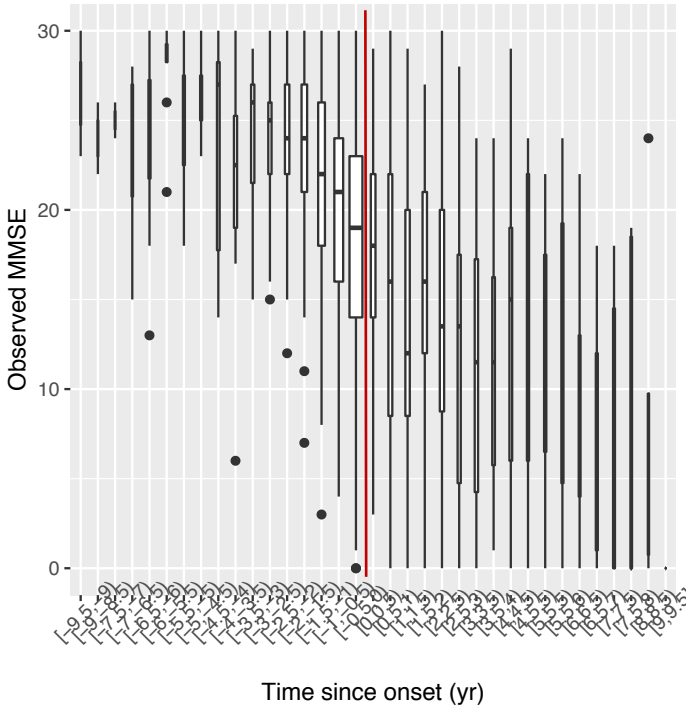


Figure 12. Observed MMSE as a function of time relative to the index date for the SIDIGI validation set. The red line marks the index date.

The Copenhagen data allowed us to investigate the effect of dementia drug use on model performance. Model performance was assessed for three patients groups: those who did not use any dementia drug, those who used one type of drug, and those who used more than one type of drug. Model performance seems to be slightly better for patients who receive more than one dementia drug (see Table 2), but overall validation results are still poor.

The lowest MAD and median absolute difference were obtained for MEMENTO, a memory-clinic based cohort study. The observed and predicted MMSE scores are shown in Figure 13 and Figure 14. Although the variability of the observed measurements within the time intervals is still considerable, it is smaller than the variability in the validation sets that were discussed above. As illustrated in Figure 15, the median predictions match the median observed measurements reasonably well, in particular for the first two and a half years after the index date.

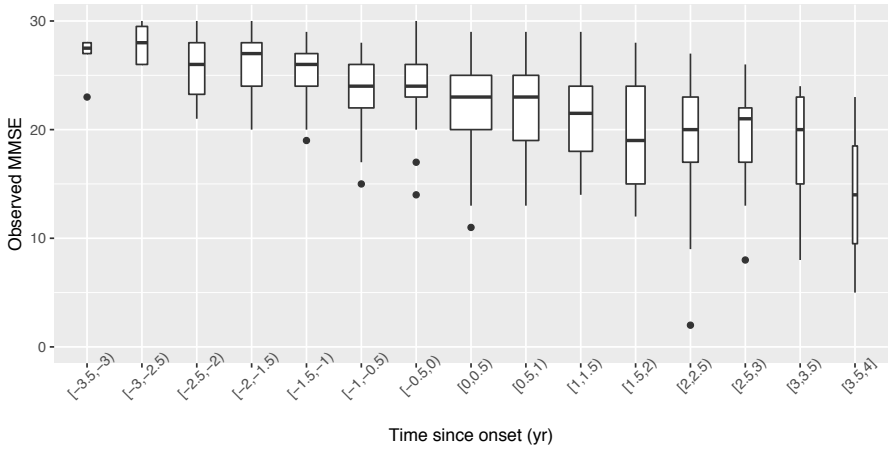


Figure 13. Observed MMSE as a function of time relative to the index date for the MEMENTO validation set.

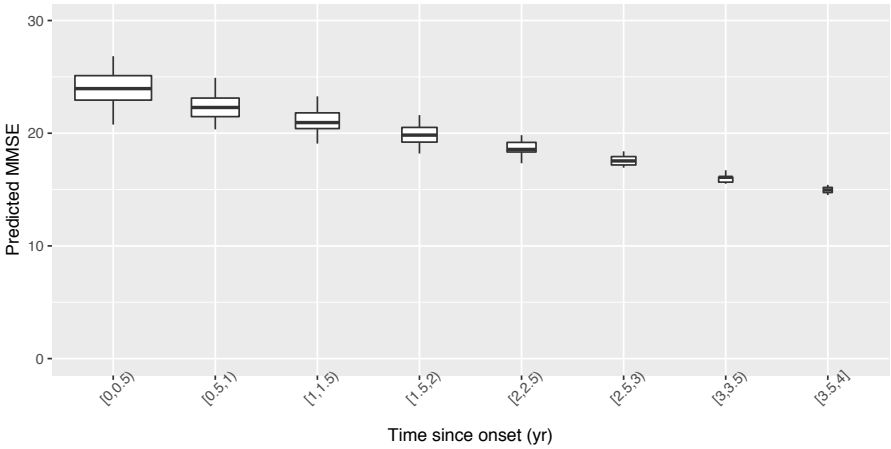


Figure 14. Predicted MMSE as a function of time relative to the index date for the MEMENTO validation set.

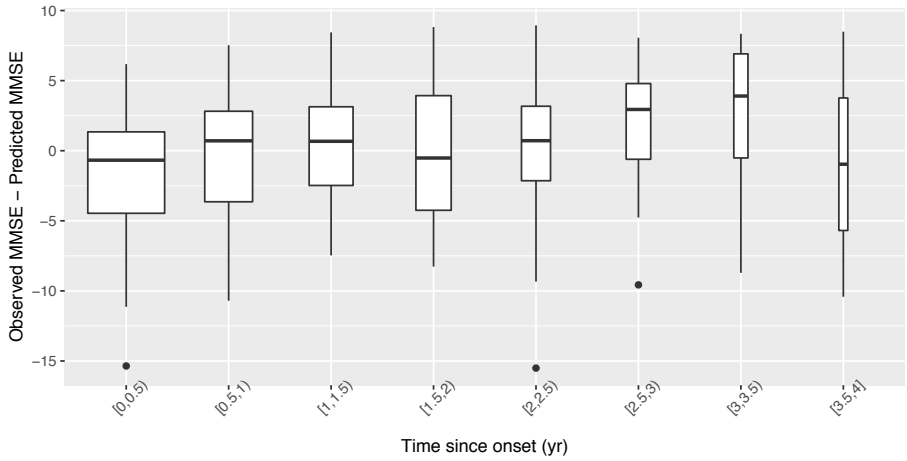


Figure 15. Difference between observed and predicted MMSE as a function of time relative to the index date for the MEMENTO validation set.

The plots for all the validation sets are provided in Annex V.

3.4. Discussion and conclusions

We validated the MMSE model on a variety of data sets. The validation results indicate poor to moderate performance of the model in predicting individual MMSE scores for AD dementia patients. Since the original publication about the model did not report evaluation results, we also assessed model performance on the Kungsholmen data that was used to develop the model. Interestingly, model performance with respect to individual MMSE predictions on the Kungsholmen data set was also moderate, and comparable with the validation results of some of the external validation sets. The median of the MMSE predictions in the Kungsholmen data matched the median observed MMSE values very well. This may not come as a surprise since the model was trained on the Kungsholmen data and the model estimation procedure will have minimized the overall differences between predicted and observed measurements.

One may speculate about the causes for the rather poor performance of the model in predicting individual MMSE scores. Probably an important reason is the huge variability in the observed MMSE measurements at a given point in time after the index date (illustrated by many of the plots in the preceding section), in combination with the relative simplicity of the model: it contains only two predictors, time since onset of dementia and age at the time of the MMSE prediction, which can only account for the variability in the data to a limited extent.

It should be noted that the model was originally developed mainly to predict MMSE progression at the population level (Handels et al., 2013) to serve a health-economic simulation study. The random intercept and slope type mixed model allows to describe and simulate the variation between individuals but does not allow to predict individual values without using each individually fitted model. An alternative model that would for example include the baseline MMSE score as a predictor may be expected to yield more accurate individual MMSE predictions over time.

There are a number of other factors that may also affect model performance. First, as we have seen in the IPCI data, selection bias may play a role. In a primary care setting, MMSE measurements may more likely be available for subjects whose cognitive function remains relatively stable.

Second, the criteria to establish AD dementia varied across the validation sets. Some applied DSM/NINCDS criteria, as were used in the Kungsholmen data, but for other validation sets dementia diagnoses were based on different coding systems, such as ICD-10, with less granularity and possibly less accuracy.

Third, the time of onset of AD dementia is an important parameter in the model, but can be very difficult to establish precisely. This is already true for the Kungsholmen data set, where the index date was set half-way the three-year follow-up interval preceding the visit to the study center when the dementia diagnosis was established. For some of the other cohorts, the interval between visits was shorter, which should reduce the uncertainty in establishing the index date. For EHR-type validation sets, there may also be considerable uncertainty in determining the precise onset of dementia as setting a diagnosis may also depend on considerations not applicable in a population-based cohort study (e.g., general practitioners following a wait-and-see scenario, people not willing to go through a diagnostic procedure).

A fourth factor that may affect model performance, is the length of follow-up. For the Kungsholmen data, the median follow-up was 1.5 year (interquartile range 1.5 to 4.5 years), but for most validation sets follow-up was shorter. If follow-up becomes very short, as for the IPCI and EDAR validation sets, the model is essentially used to predict MMSE at or very close to the time of diagnosis.

Fifth, the data sets vary greatly in size, and in particular for the smaller data sets the performance estimates may be less reliable.

Sixth, patient characteristics vary across the different data sets. Table 2 shows that both age at index date and gender distribution of the validation sets may be quite different from the Kungsholmen development data. Age is included in the MMSE model. Gender was considered during model development but did not prove to be a significant predictor, implying that for those aged 75 and older gender does not matter (Handels et al., 2013). However, other variables such as comorbidities are not accounted for and may affect model performance in an unpredictable way.

Finally, the data are likely to contain measurement errors that will affect model performance. We have already mentioned the difficulty of establishing a precise dementia onset. Also, some of the MMSE scores may have been inaccurate, possibly related to the varying experience of heterogeneous examiners (physicians, nurses, psychologists) in using and scoring the MMSE. The plots of the observed MMSE measurements for most of the data sets show outliers with improbable high or low MMSE values, which may be due to data entry errors. Sometimes the extraction of MMSE scores from the data source was challenging. For example, in the Copenhagen database all MMSE scores had to extracted from the clinical notes by text mining and were not always captured right, producing erroneous MMSE values. Although some scores were manually checked (all scores below 4 and score differences of more than 10 within a period of 60 days) and corrected if necessary, checking of all scores was infeasible within the available time frame.

4. Validation of Novartis' preclinical model

4.1. Model description

Shifting the focus of clinical trials that test disease-modifying interventions against Alzheimer's disease (AD) from the dementia stages of the disease to preclinical stages might increase the likelihood of success for these trials (Berk and Sabbagh, 2013). Although various models describing cognitive decline in later stages of AD existed (Chua, 2015), a model describing cognitive function in the preclinical phase of the disease and predicting time to first diagnosis of mild cognitive impairment (MCI) or dementia due to AD was lacking. Hence, there was an urgent need of such a model to, e.g., inform the design of trials targeting patients at risk to develop dementia due to AD. Under this context, Novartis developed the preclinical model to support the optimization of the clinical trial design at the preclinical stages (Caputo et al., 2017; Lestini et al., 2018).

The preclinical model consisted of three models which were fitted independently on natural history cohorts. The first model was a time to event (TTE) model describing time to first diagnosis of MCI and dementia due to AD using a Weibull parametric survival model. Then, two mixed-effects models were developed to describe the progression of the Alzheimer's Prevention Initiative Composite Cognitive (APCC) (Langbaum et al., 2014) for two subpopulations: the "progressors", i.e., patients with first diagnosis of either MCI or dementia due to AD within eight years from baseline, and "non-progressors", i.e., patients who were either not diagnosed or only diagnosed after eight years. Both APCC progression models followed a general power model structure, i.e., score (time) = intercept + slope*time$^{\gamma}$. The model covariates were chosen based on the clinical relevance, goodness of model fit, and statistical tests (Table 3). The preclinical model was developed on three cohorts (ROS, MAP, and MARS) from the Rush Alzheimer's disease center (Rush) (Bennett et al., 2005) and the National Alzheimer's Coordinating Center (NACC) (Viswanathan et al., 2015). The TTE model was developed on data of 2,159 subjects from Rush and 8,535 subjects from NACC who were cognitively normal at baseline and were diagnosed with MCI or dementia due to AD during follow-up. To develop the two APCC models, data of 2,336 subjects (732 progressors, 1,604 non/late progressors) from Rush who were cognitively normal at baseline and had no other diagnoses than MCI or dementia due to AD during follow-up, were used.

Table 3: Covariates and their positions in the preclinical model after systematic covariate selection consisting of backward elimination with AIC as the criterion for model selection

Model		Baseline APCC	Education	APOε4 status	Gender	Age
TTE		×	×	×		× (at BL*)
APCC "progressor"	Intercept	×	×		×	× (at event)
	Slope	×	×	×		
APCC "non-progressors"	Intercept	×	×		×	× (at BL)
	Slope	×	×	×	×	× (at BL)

*BL = baseline, the first APCC measurement when subjects entered the dataset

4.2. Data sources

In collaboration with WP3, two longitudinal datasets from two European electronic healthcare data sources were accessed for the external validation study, which aimed at testing and validating the performance of the preclinical model in predicting real-world AD progression.

The first dataset was from two ongoing Gothenburg prospective cohorts including participants sampled from the Swedish population register based on birth data (Prospective Population Study of Woman (PPSW) (Johansson et al., 2010) and Gerontological and Geriatric Population Studies (H70) (Arnoldussen et al., 2018). Diagnoses are based on DSM-III-R. From the received Gothenburg dataset, subjects who were non-demented at inclusion (= year 2000) and eventually developed AD-type dementia (probable AD, possible AD, AD plus vascular dementia, vascular dementia plus AD, mixed plus AD) or remained non-demented during follow-up, were selected for the external validation. A total of 617 subjects (complete-cases, with at least 2 APCC proxy values) were selected for the validation of the APCC models (558 for the APCC non-progressor model and 59 for the APCC progressor model), while 718 subjects (complete-cases) were selected for the validation of the TTE model.

The second dataset was from the 4C study (Clinical Course of Cognition and Comorbidity in Mild Cognitive Impairment and Dementia Study) (Liao et al., 2016). From the accessed 4C dataset, subjects with subjective cognitive impairment (SCI) at inclusion, who eventually developed MCI (Petersen criteria) or AD dementia (probable AD and possible AD, NINCDS-ADRDA criteria for AD diagnosis) or AD dementia (probable AD and possible AD) or did not progress during follow-up, were selected for the external validation. Data of 22 subjects (complete-cases, with at least 2 APCC proxy values) were used for validating the APCC progressor model, while data of 121 subjects (complete-cases) were used for validating the TTE model. Since the follow-up of the 4C study was less than eight years, it was not possible to identify non/late progressors from the validation dataset. Thus, the validation of the APCC non/late progressor model was not performed on this dataset.

4.3. Validation process and results

The overall comparison between the development dataset and the validation datasets initiated the external validation. As shown in Table 4, the validation datasets differed from the development dataset in four main aspects. Since APCC was not available in both validation datasets, APCC proxies were constructed before performing the predictions. APOε4 status was missing in the 4C dataset, thus the mean APOε4 value in the development dataset was used for the validation of the APCC progressor model on the 4C dataset, and a sensitivity analysis was applied for the validation of the TTE model by assuming that all subjects were APOε4 non-carriers, heterozygotes carriers or homozygotes. Because the diagnosis of MCI was not available in the Gothenburg dataset, the event in the TTE model was defined as the diagnosis of dementia due to AD.

Table 4: Major differences between the development dataset and the validation datasets

	Development dataset	Validation dataset	
		Gothenburg dataset	4C dataset
Diagnosis status at baseline	Cognitively normal	Non-demented	Subjective cognitive impairment (SCI)
APCC	Yes	No	No
Missing key covariate	-	-	APOε4
Event definition for the TTE model	First diagnosis of MCI or dementia due to AD	First diagnosis of dementia due to AD (no MCI)	First diagnosis of MCI or dementia due to AD

The APCC proxies were constructed following three steps. First, the test items from the original APCC were replaced by different test items from the same cognitive domains that were available in the validation datasets (Table 5). Then, the ranges of test item replacements were rescaled to be the same as those in the original APCC, and the APCC proxy was calculated. Finally, the range of the obtained overall APCC proxy was rescaled to be the same as the original APCC (i.e., 0-100).

The time courses of the original APCC scores and the constructed APCC proxies are compared in Figure 15. The APCC proxy obtained from the Gothenburg dataset showed a similar pattern as the original APCC. The cognitive capacity as characterized by APCC score started to decline a couple of years before the diagnosis for progressors (15A-15B), while the time courses of APCC and APCC proxy stayed quite flat over time for non/late progressors (15D-15E). However, the APCC proxy constructed for progressors in the 4C dataset did not show a clear decline in the preclinical stages (15C), possibly due to data size, short follow-up, or diagnosis status at inclusion.

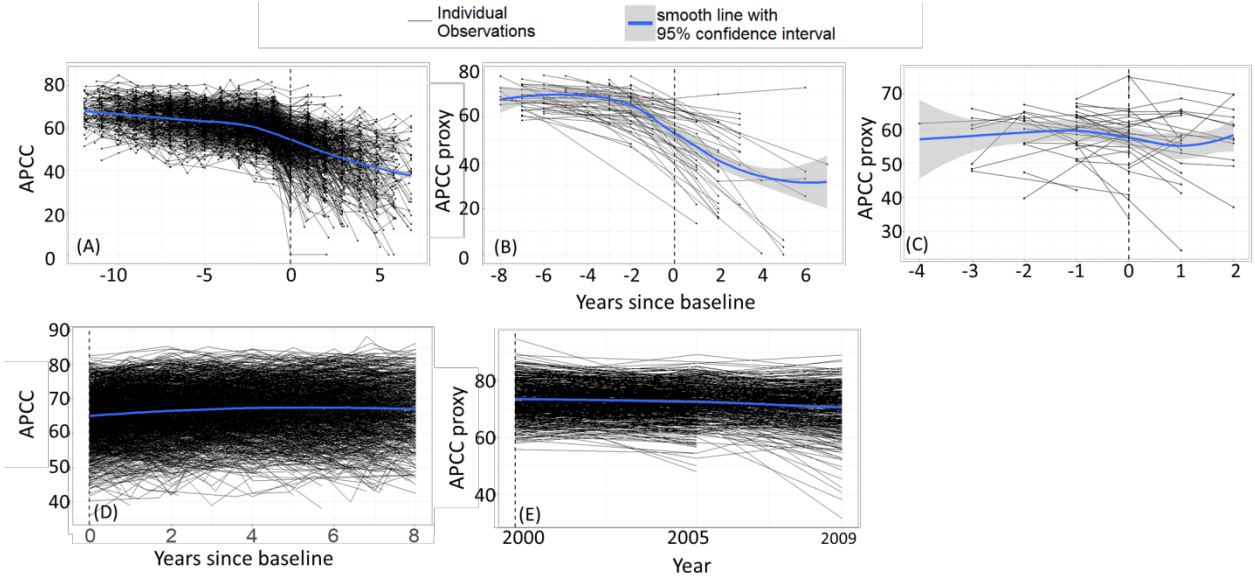


Figure 16. The time courses of APCC and APCC proxy. (A) original APCC on the development data for progressors; (B) APCC proxy on the Gothenburg dataset for progressors; (C) APCC proxy on the 4C dataset for

progressors; (D) original APCC on the development data for non/late progressors; (E) APCC proxy on the Gothenburg dataset for non/late progressors. The black line is the observation per subject. The dot on the black line shows the value of each assessment. The blue line is the smooth curve of the observations. The shaded area around the blue line is the 95% confidence interval.

Table 5: Mapping of available cognitive tests in the validation datasets for constructing APCC proxy

Original APCC (developed in Rush)			Comparable cognitive tests in validation dataset and the ranges	
Components of cognitive tests	**Domain**	**Ranges**	**Gothenburg dataset**	**4C dataset**
Word List recall	Delayed Memory	0-10	Item recall; 0-12	Word list recall; 0-15
Logical Memory Story A recall	Delayed Memory	0-25	-	-
Symbol Digit Modalities Test	Attention	0-110	-	Letter Digit Substitution Test; 0-125
Judgment of Line Orientation	Visual/Spatial ability	0-15	Visuoconstructional apraxia; 0-6	MMSE Copy Figure; 0-1
Ravens Progressive Matrices – subscale	Attention / Executive functioning	0-9	Word fluency(animals)	Word fluency (animals)
MMSE Orientation to Place	Visual/Spatial ability / Orientation	0-5	MMSE Orientation to Place; 0-5	MMSE Orientation to Place; 0-5
MMSE Orientation to Time	Visual/Spatial ability / Orientation	0-5	MMSE Orientation to Time; 0-5	MMSE Orientation to Time; 0-5

APCC Total Score = 1.36*WordListRecall + 0.528*LogMemDelRecall + 0.26*SymbDigModal + 0.68*JudgLineOr + 1.39*ProgrMatrSub + 2.14*MMSEPlace + 2.24*MMSETime

Before using the formula above to construct the APCC proxy, the ranges of the test item replacements should be rescaled to be the same as the ranges of the original test items in Rush. The range of the obtained APCC proxy should be rescaled to be the same as the range of the original APCC (range 0-100).

After construction of the APCC proxy, the distributions of the covariates required by the preclinical model (i.e., APCC (proxy), age, years of education, gender, APOε4 status) were compared between the development dataset and the validation datasets. As shown in Figure 17, the Gothenburg dataset had a higher median baseline APCC proxy, lower median baseline age, and lower median

years of education than the development datasets. The 4C dataset had a comparable median baseline APCC proxy, lower median baseline age, and lower median years of education compared to the development datasets. The differences in the number of years of education between the datasets might be partially explained by the different educational systems of the countries from which the data were sourced.

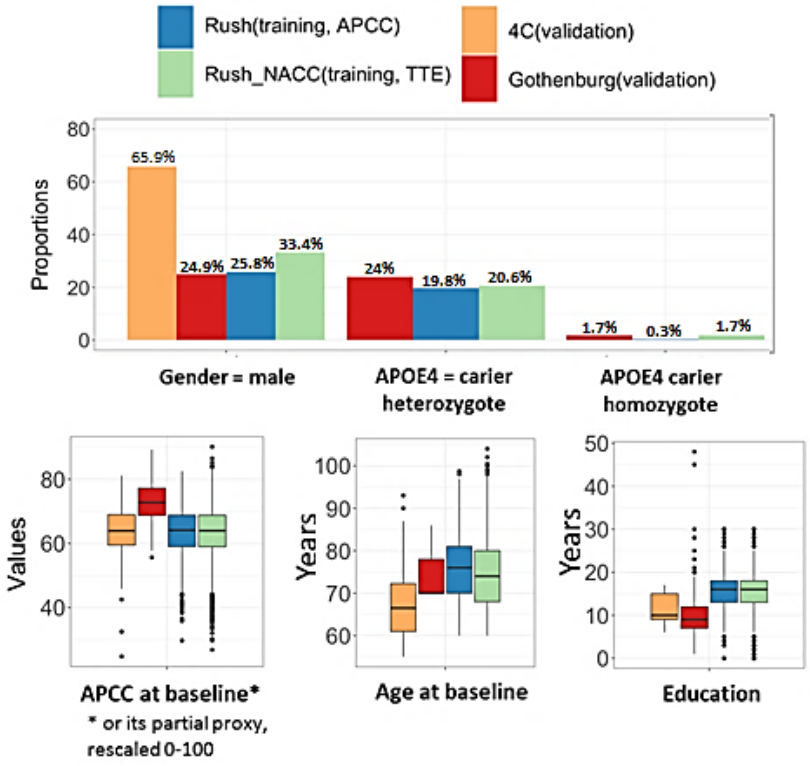


Figure 17. Comparison of the distributions of the model covariates between development dataset and validation datasets (box plot: median, 25% and 75% quantiles, maximum, minimum, and potential outliers).

The preclinical model was then externally evaluated by comparing the predictions with observations, and the performance of the model was assessed via visual checks and a goodness-of-fit measure (i.e., root mean square error). Figure 18illustrates the comparison between the observed and predicted APCC proxies. The points cluster around the diagonal line for both APCC models (i.e., progressors and non/late progressors) in both the Gothenburg and 4C datasets. The root mean square error was less than 10% for the APCC predictions. Clear systematic bias was not observed in the predictions for progressors when comparing the accuracy of the predictions between type of dementia and time since diagnosis (Figure 18A, B and D)). However, the residual progression along the nature time was not well captured for non/late progressors on the Gothenburg dataset (Figure 18C).

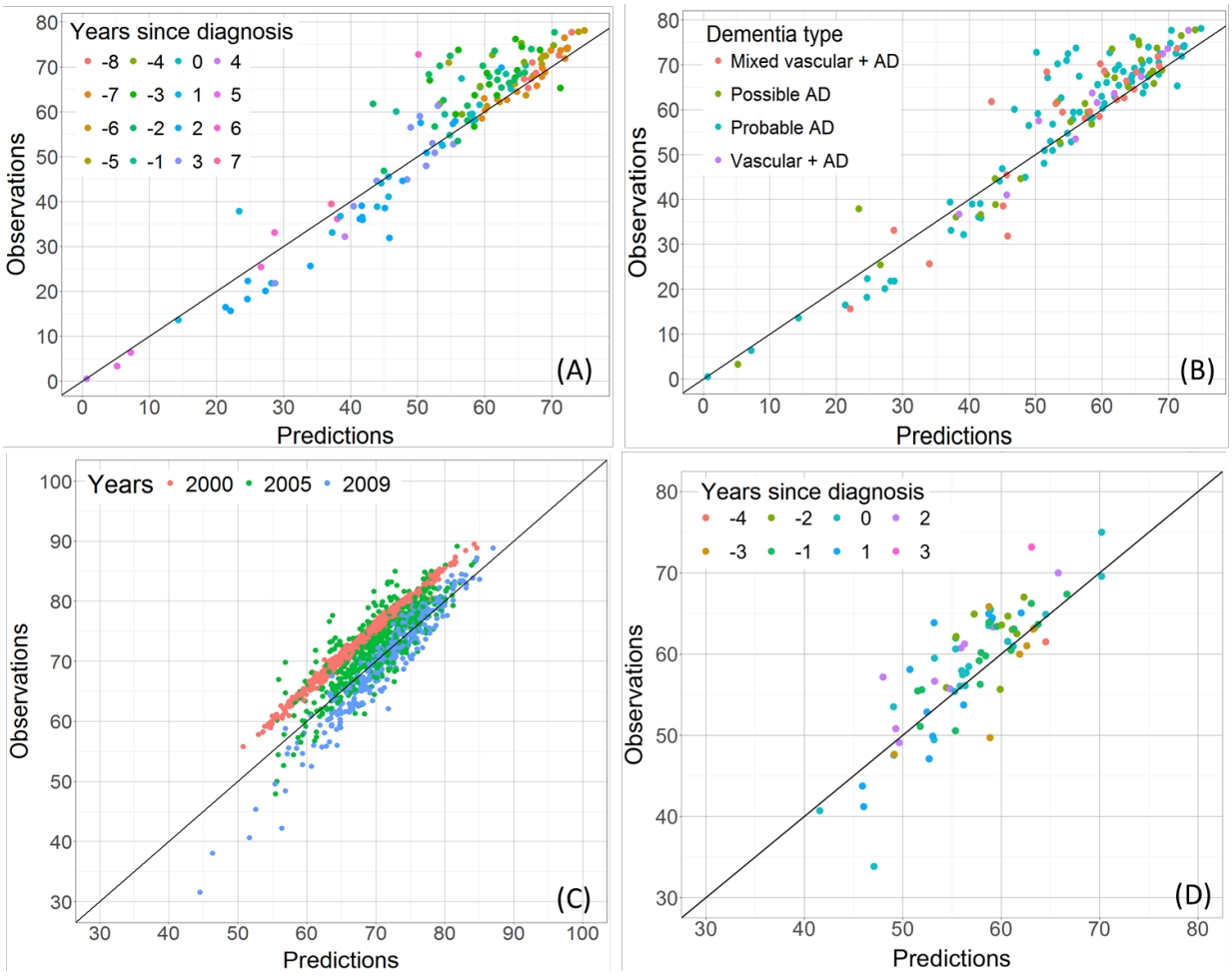


Figure 18. Comparison of the predicted APCC proxy and the observed ones: (A-B) predictions vs. observations on the Gothenburg dataset for progressors; (C) predictions vs. observations on the Gothenburg dataset for non/late-progressors; (D) predictions vs. observations on the Gothenburg dataset for non/late-progressors; (D) predictions vs. observations on the 4C dataset for progressors.

Error! Reference source not found. shows a comparison between the observed Kaplan-Meier survival curve and the survival curve predicted from the preclinical model. Note that the survival curve predicted by the model falls in the 95% confidence interval of the Kaplan-Meier estimation in the development dataset (Figure 19A, blue lines). The survival curves predicted by the model were far from the Kaplan-Meier curves on both external validation datasets (Gothenburg and 4C). This could be due to the differences (e.g., diagnosis status at baseline) between the development dataset and the validation datasets as shown in Table 4, and the large drop out in the 4C dataset (especially between the second year and the third year after inclusion).

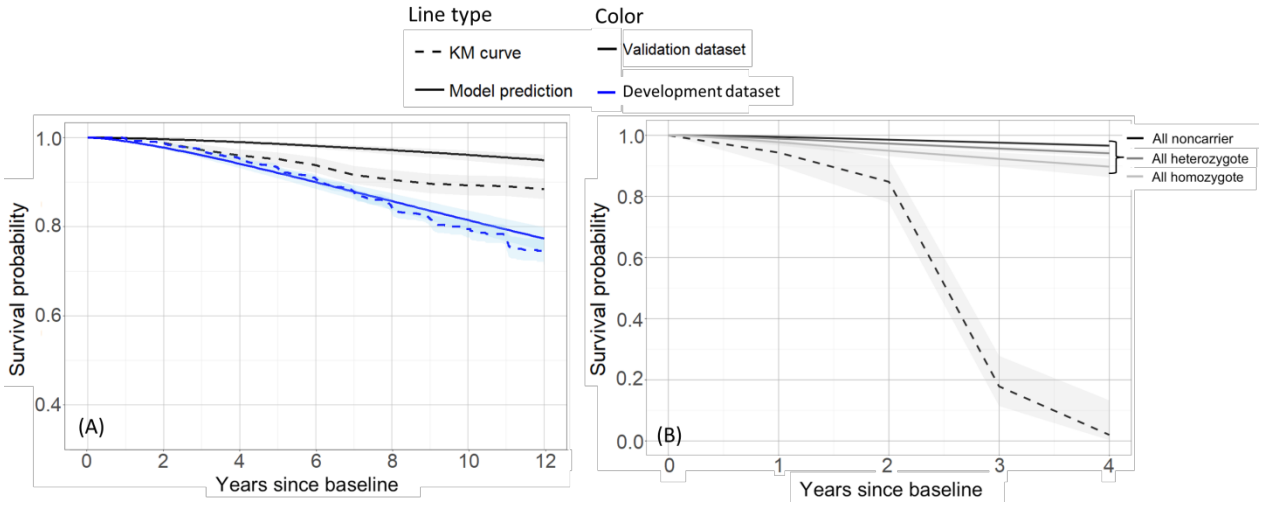


Figure 19. Comparison of the predicted survival curve and the observed Kaplan-Meier survival curve: (A) Gothenburg dataset and development dataset; (B) 4C dataset. The solid lines are predicted survival curves from the preclinical model. The dashed lines are observed Kaplan-Meier survival curves. The shaded area shows the 95% confidence interval. The blue color indicates the results on the development dataset, while the black/grey color indicates the results on the validation datasets.

4.4. Discussion and conclusions

The preclinical model had its specific prerequisites regarding the population, diagnosis, cognitive tests, and follow-up. Despite the differences between the development dataset and the current validation datasets (Table 4), the two APCC models were validated on different external datasets (i.e. the Gothenburg dataset and the 4C dataset) Figure 18). The APCC models were flexible enough to be generalized to predict individual APCC (proxy) time courses on new datasets. Indeed, the nature of the APCC models (mixed-effect models) compensated the impacts from the different patient characteristics to some extent. Nevertheless, the APCC model for progressors needs to be further validated on larger datasets to confirm the current results. Indeed, the current data size (59 progressors in the Gothenburg dataset, and 22 progressors in the 4C dataset) was small and may have limited statistical power.

The current results suggest that the TTE model part of the preclinical model tended to overestimate the overall survival probability on both external validation datasets, implying that this TTE model might have limited abilities in identifying patients at high risk to develop AD dementia symptoms on real-world datasets. However, because of the differences in population, diagnosis, cognitive tests and follow-up (censoring) between the development dataset and current validation datasets, the external validation of the TTE model should be further explored. For example, when merely the diagnosis of probable AD was selected instead of all AD-related diagnoses (i.e., probable AD, possible AD, AD plus vascular dementia, vascular dementia plus AD, mixed plus AD) for the external validation of the TTE model on the Gothenburg dataset, the obtained Kaplan-Meier curve (black dashed line) in Figure 19A was closer to (still under) the survival curve predicted by the model (black solid line). In terms of the 4C dataset, we observed a sharp decline in the Kaplan-Meier curve between the second year and the third year (Figure 19B). Therefore, an exploration of the nature of the censored data or a drop-out model would be needed to understand whether censoring occurred at random or whether censoring was conditional on any factor or covariate, thus

helping to refine the population selection for the external validation. In addition, a larger and relevant dataset is required for a more appropriate external validation of the TTE model. Ideally, the validation dataset should have a comparable population as the development dataset (especially the diagnosis status at inclusion), a larger number of subjects with a diagnosis during follow-up (e.g,. around 30% as in the development dataset), and relatively few dropouts.

The cognitive composite score (i.e., APCC) adopted by the preclinical model was expected to track preclinical cognitive decline in individuals who subsequently progressed to the clinical stages of late-onset AD. Since APCC was not available in the external datasets, a well-constructed APCC proxy became critical for the external validation of the preclinical model. We note that since not all the test items required by the original APCC were available in the validation datasets, the current APCC proxies derived from this simple and pragmatic rescaling approach (Table 4) might not adequately have captured the preclinical cognitive decline compared to the original APCC. Further efforts should be put into the construction of the APCC proxy, and a dataset with richer cognitive tests (e.g., the Memento dataset (Dufouil et al., 2017)) would facilitate this work.

Although the APCC models built under the mixed-effects model framework displayed a satisfactory performance during the external validation on the two datasets, a potential systematic bias was observed in the APCC model for non/late progressors (Figure 18C). The predicted APCC proxy values were smaller than the observed ones in the year 2000 which was used as the baseline. This might be due to the fact that the baseline APCC proxy was a covariate in the model. A time term might be further introduced into the APCC model structure to better capture the residual progression. The current TTE model only adopted the APCC at baseline to predict the risk of developing AD symptoms. In order to refine the risk prediction, a joint model (Hickey et al., 2018) that fits the longitudinal APCC progression and time to diagnosis at the same time could be developed, thus leveraging the APCC decline in the risk predictions.

The current results of the external validation are encouraging. In order to better describe the disease progression of AD, and identify subjects at high risk of developing AD symptoms from pre-clinical stages, the preclinical model should be further explored with respect to the refinement of the model structure (e.g., joint model development), the inclusion of additional relevant covariates, the refinement of the APCC proxy, the refinement of population and diagnosis selection, and especially perusing appropriate data sources for model development and validation. The journey of this external validation study also emphasizes the need for longitudinal real-world datasets of patients from the preclinical stage onwards.

5. Validation of Eli Lilly's institutionalization model

5.1. Model description

This model predicts the time to institutionalizatio for patients with AD dementia. The model was developed as part of a study that examined the costs of caring for community-dwelling AD patients in relation to the time to institutionalization (Belger et al., 2018), and builds on earlier work of Green et al. (Green et al., 2011). Data for the development of the model were taken from the GERAS study, a prospective observational study of costs associated with care of community-dwelling caregivers in three European countries (France, Germany, UK) (Wimo et al., 2013). GERAS enrolled community-dwelling patients aged at least 55 years, meeting the NINCDS/ADRDA criteria for probable AD, with an MMSE score equal to or less than 26. Patients were stratified by disease severity at baseline: mild AD dementia (MMSE 21-26), moderate AD dementia (MMSE 15-20), and severe AD dementia (MMSE <15). Data were collected at baseline and during routine care visits at 6, 12 and 18 months in all three countries, and at 24, 30 and 36 months in France and Germany. Patient cognitive function was assessed using the MMSE. Functional ability was determined using the Alzheimer's Disease Cooperative Study Activities of Daily Living inventory (ADCS-ADL) (Galasko et al., 2005), with a score range of 0-78 (higher scores indicate better functioning). Separate subscores were derived for the basic ADL (BADL, score range 0-22) and instrumental ADL (IADL, score range 0-56). Behavioral and psychological symptoms were assessed using the 12-item version of the Neuropsychiatric Inventory (NPI) (score range 0-144), where a higher score indicates more severe problems.

Patients characteristics considered for inclusion in the model were age, gender, years of education, time since diagnosis of AD, comorbidities, and baseline scores for MMSE, BADL, IADL, total ADL, and NPI. Caregiver characteristics considered were age, gender, relationship with the patient (spouse yes/no), and caregiver working for pay (yes/no). The following patient and caregiver factors were independently associated with time to institutionalization and were selected for inclusion in the final model: MMSE, BADL, IADL, NPI, and caregiver relationship (spouse yes/no), all measured at baseline. The prediction equation is:

$$\text{Time to institutionalization} = \exp(7.600 + 0.034 * \text{MMSE} + 0.025 * \text{IADL} - 0.044 * \text{BADL} - 0.015 * \text{NPI} - 0.640 * \text{NoSpousalCaregiver})$$

where Time to institutionalization is the predicted number of days till the patient is institutionalized, and NoSpousalCaregiver is 1 if there is no spousal caregiver, otherwise 0.

No external validation of the prediction model has been performed.

5.2. Data sources

As documented in Deliverable 3.4, "Final report on proof of concept technical solutions for RWE data harmonisation and integration", data sources for external validation of the time-to-institutionalization model were sought in the EMIF and DPUK Catalogues. The challenge was to find sources that contained informal caregiver information and measurements for the three specific cognitive, functional, and behavioral scales (MMSE, ADCS-ADL, NPI) that had been used for the model development. Searches in the Catalogues did not result in the identification of cohorts that

fulfilled these requirements. Also literature searches for cohorts that are not included in the Catalogues, did not reveal suitable data sources. Hence, studies were sought that used the same cognitive and behavorial scales and caregiver information as the development cohort, and a functional scale that was similar to ADCS-ADL. Two such studies were identified: the ICTUS study (Canevelli et al., 2016) and the 4C Dementia Study (Liao et al., 2016). Since access to the 4C data set was granted relatively late in the ROADMAP project, in the following the focus for external validation of the institutionalization model is on use of the ICTUS data.

The ICTUS study is a prospective multicenter cohort study aimed at evaluating the clinical course, treatment outcome, and the socioeconomic impact of AD in Europe. It involved 29 participating centers from 12 European countries. Inclusion criteria were: (1) diagnosis of probable AD according to NINCDS-ADRDA criteria; (2) MMSE score in the range of 10-26; (3) living in the community with an informal caregiver, and (4) absence of known conditions reducing the patient's life expectancy. After baseline assessment (from 2003 to 2005), participants were followed up for 3 years with midterm reevaluations every 6 months.

Variables in the ICTUS data set include MMSE and the 12-item version of NPI, like in the original development set. Functional ability in ICTUS was assessed by the Katz ADL (score range 0-6) (Katz et al., 1963) and the Lawton IADL scales (score range 0-8) (Lawton and Brody, 1969). We mapped the Katz index to the basic ADCS-ADL and Lawton's IADL to the instrumental ADCS-ADL scale by a simple linear transformation:

TargetScore = (maxTargetScore / maxSourceScore) * SourceScore,

where maxTargetScore is 22 and 56 for BADL and IADL, respectively, and maxSourceScore is 6 and 8 for the Katz index and Lawton's IADL, respectively.

Note that the items in the Katz index can largely be matched with BADL items (Rósza et al., 2009). For the IADL scales, mapping of the items is not straightforward, but we assumed that despite differences at the item level, similar percentage score on both scales indicated overall similar functional ability.

The ICTUS data also contains information about the primary caregiver, including caregiver status (husband, wife, child, friend, other). Information about date of institutionalization and death is also available.

5.3. Validation results

A total of 1,375 subjects with AD dementia were recruited in the ICTUS study. Of these, 132 were excluded because they did not have follow-up data, were not living at their own home at baseline, or were younger than 55 years. Of the remaining 1,243 patients, 421 were excluded due to missing data, mostly for IADL (n = 402). No data were missing for MMSE, and few for BADL (n = 2) and NPI (n = 27). The study population therefore consisted of 822 patients.

Descriptive characteristics of the ICTUS data at the baseline assessment are shown in Table 6. These can be compared with the baseline characteristics of the GERAS data that were used for model development (Table 7, information taken from (Berger et al., 2018)).

Table 6. Baseline characteristics in the overall study population and by AD dementia severity in the ICTUS data.

Characteristic	Overall	Mild AD	Moderate AD	Severe AD
Patient, n	822	428	312	82
Age, mean (SD)	76.8 (7.6)	76.3 (7.5)	77.2 (7.7)	77.7 (7.2)
Gender, % female	87.3	82.5	92.6	92.7
MMSE, mean (SD)	20.3 (4.0)	23.5 (1.7)	17.7 (1.7)	13.1 (1.1)
BADL, mean (SD)	19.8 (3.4)	20.6 (2.6)	19.1 (3.8)	17.6 (4.6)
IADL, mean (SD)	34.0 (15.9)	39.7 (14.1)	29.3 (15.4)	22.1 (14.0)
NPI, mean (SD)	13.8 (13.8)	12.1 (12.7)	15.2 (14.1)	17.3 (16.6)
Caregiver, % spouse	38.3	42.8	33.0	35.4

Table 7. Baseline characteristics in the overall study population and by AD dementia severity in the GERAS data.

Characteristic	Overall	Mild AD	Moderate AD	Severe AD
Patient, n	1,495	566	472	457
Age, mean (SD)	77.6 (7.7)	77.3 (6.9)	77.8 (8.0)	77.6 (8.1)
Gender, % female	54.8	47.9	57.0	61.1
MMSE, mean (SD)	17.4 (6.3)	23.3 (1.6)	17.9 (1.7)	9.5 (4.3)
BADL, mean (SD)	17.3 (5.2)	19.8 (3.1)	18.3 (3.8)	13.2 (6.0)
IADL, mean (SD)	29.1 (15.2)	38.5 (11.8)	29.9 (12.5)	16.6 (12.3)
NPI, mean (SD)	15.1 (15.3)	10.2 (10.7)	14.3 (12.6)	22.0 (19.4)
Caregiver, % spouse	65.9	70.6	63.1	62.9

The ICTUS data and the GERAS data differ in their distribution of patients over the AD severity groups, with ICTUS having relatively few patients in the severe dementia group. The percentage of women in the ICTUS study population is considerably larger than in GERAS, while the percentage of spousal caregivers is substantially lower. The mean values of the cognitive, functional and behavorial scales in the mild and moderate AD severity groups are largely comparable for both data sets, but values for the severe AD group on average suggest greater AD severity of the patients in this group in GERAS than in ICTUS.

Of the 822 patients, 117 (14.2%) were institutionalized during follow-up. This percentage is considerably lower than 20.5% institutionalization reported in the GERAS study (Berger et al., 2018). According to baseline severity, the number of patients institutionalized was 49 (11.4%), 54 (17.3%), and 14 (17.1%) for the mild AD, moderate AD, and severe AD dementia groups, respectively. For the patients who were institutionalized, a scatter plot of the predicted time to institutionalization versus the observed time to institutionalization is given in Figure 20. Clearly, the

prediction equation greatly overestimates the time to institutionalization, with largest overestimations for the mild dementia patients.

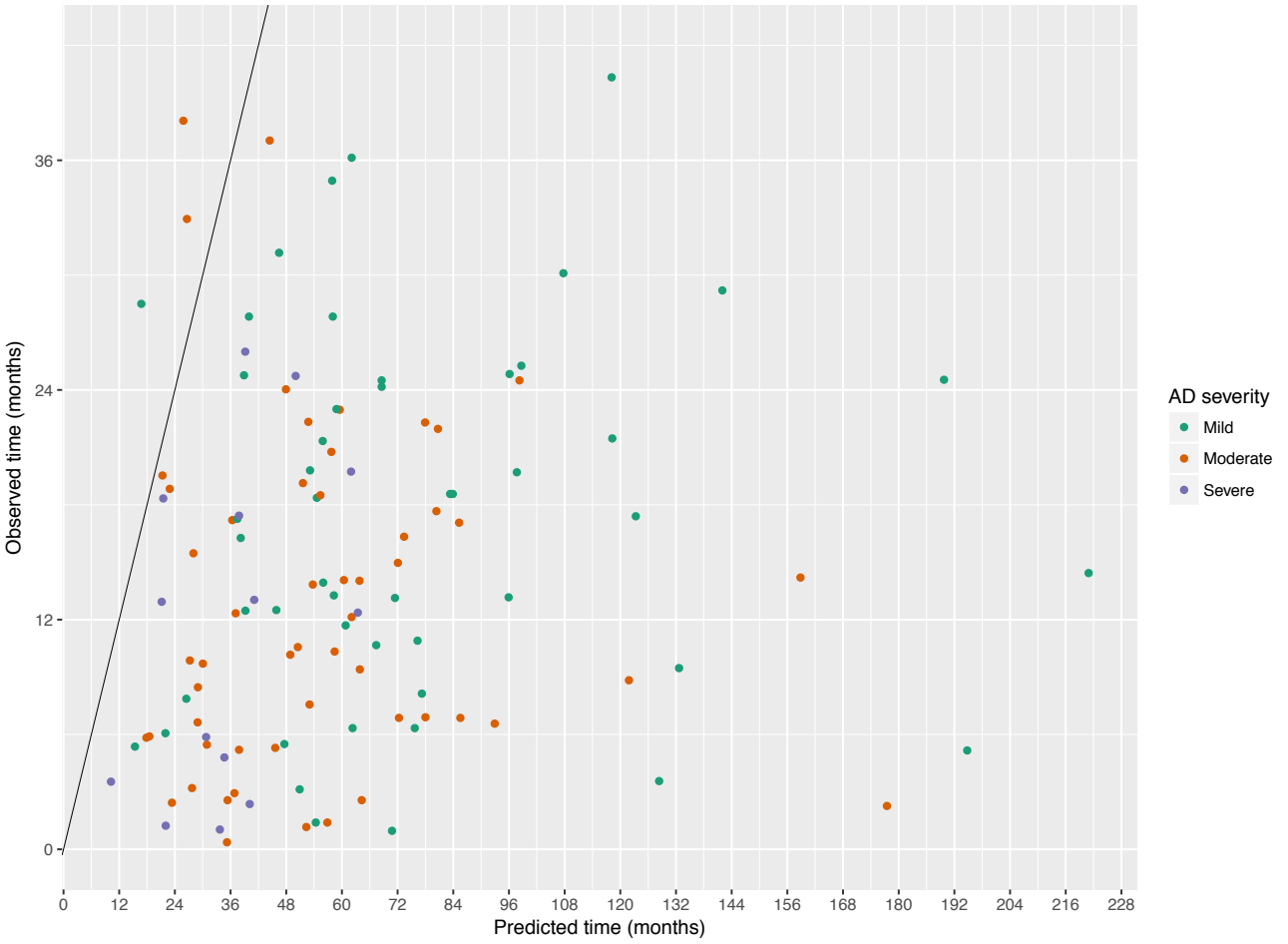


Figure 20. Predicted time to institutionalization versus observed time to institutionalization in the ICTUS data set.

For patients who were not institutionalized during follow-up (n = 705), we computed the predicted time to institutionalization relative to the date of last follow-up (Figure 21) or, if they died during follow-up, relative to the date of death (Figure 22). Figure 21 shows that for the patients not institutionalized till last follow-up (n = 644), the predicted date of institutionalization lies after the date of the last follow-up assessment for the far majority of patients. Only 32 patients were predicted to be institutionalized before last follow-up (3 mild, 19 moderate, and 10 severe AD dementia patients).

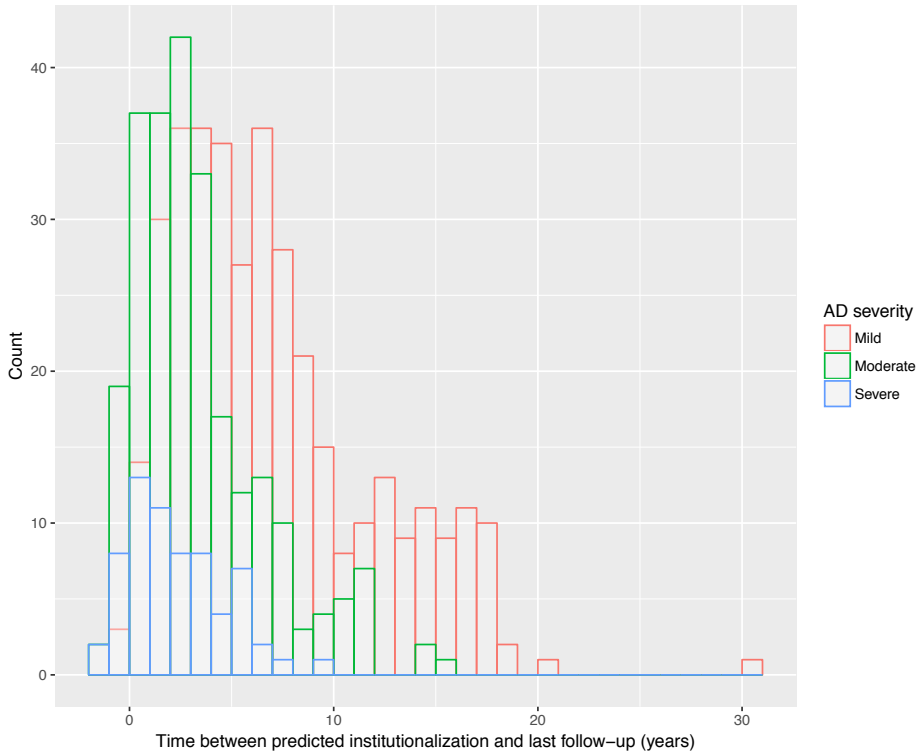


Figure 21. Time between predicted date of institutionalization and date of last follow-up.

Figure 22 shows that also for patients who died during follow-up without being institutionalized (n = 58), the predicted date of institutionalization lies later than the date of death for most of the patients. Five patients were predicted to be institutionalized before they died, one patient with mild dementia at baseline and four with moderate dementia.

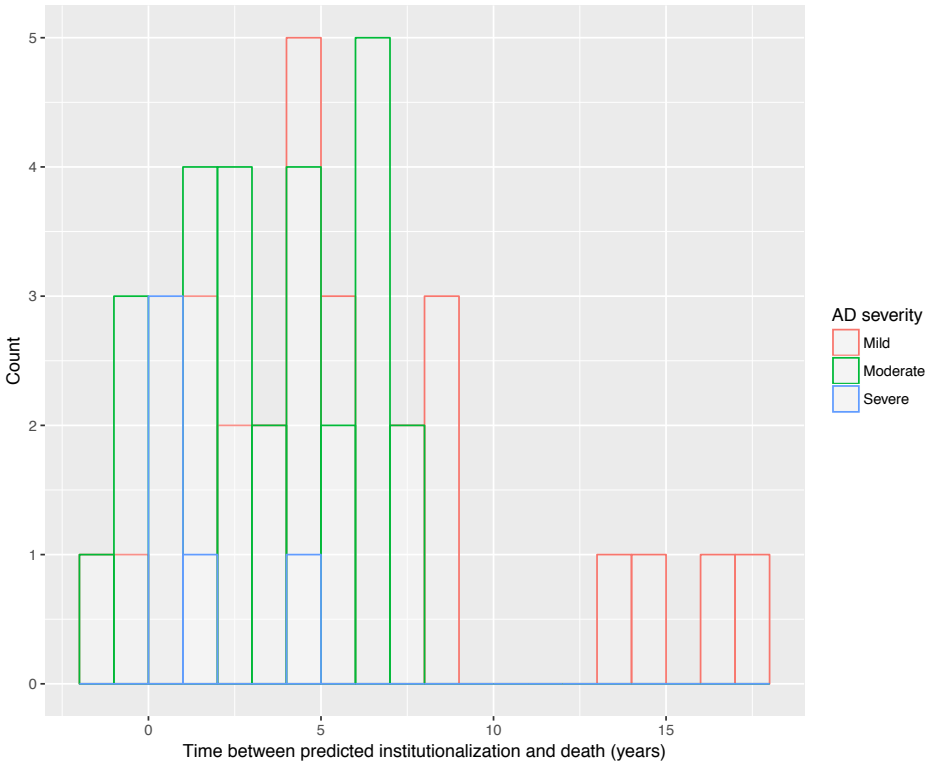


Figure 22. Time between predicted date of institutionalization and date of death.

5.4. Discussion and conclusions

We have evaluated the predictions of the time-to-institutionalization model on the ICTUS data set. Our results show a large overestimation of the predicted times for the patients who were institutionalized during follow-up. For the patients who were not institutionalized, either because they died during follow-up before being institutionalized or because they had not been institutionalized at the last follow-up round, the predicted date of institutionalization followed the event for most of the patients.

These results are difficult to compare with those from the study that used the GERAS data as a development set. The original paper does not compare predicted and observed time to institutionalization on an individual basis. We did not have access to the GERAS data to perform the same analyses as we did for the ICTUS data.

Comparison of the baseline characteristics of the ICTUS and GERAS study populations indicated that the mean differences for the cognitive, behavorial and functional scales were generally small for the mild and moderate AD severity groups. However, the mean scores for the severe dementia group suggested that the severity of this group in GERAS was generally worse than in ICTUS. Other differences between the two data sets include the percentage of women and the percentage of spousal caregivers, but it remains unclear to what extent these differences affect prediction accuracy. To get a better insight in this, it would be necessary to obtain the performance results on the GERAS data and compare those with the current results.

A possible source of error is the mapping procedure that we used to convert the Katz and Lawton functional scales available in ICTUS to the ADCS-ADL functional scales available in GERAS and required by the model. We applied a very simple, straightforward mapping based on the score ranges of the different scales. A proper validation of the mapping was not possible because we did not have a data set that contained both the source and target scales. Comparison of the baseline characteristics of the mapped functional scales in ICTUS with the functional scales in GERAS, shows comparable values for the mild and moderate dementia patients, but lower values for the severe dementia cases. This may be attributed to differences in the study populations, but a less accurate mapping for these more severe patients cannot be excluded.

6. General discussion and conclusion

We have done pilot exercises to validate three different disease progression models in the field of AD dementia. The validation results for each of these models have been described and discussed in the model-specific sections above. Here we want to conclude with a more general discussion on several issues and challenges in external model validation.

First, we had to address heterogeneity of external data sources and issues related to data access. For this purpose we developed a validation pipeline that offers a structured approach for external validation of a disease progression model. We made use of TRIPOD checklists to acquire basic information about the model development and implementation, input and output variables, the external data sources, and validation measures. A statistical analysis plan specified the details of the analysis. The Jerboa tool was used to transform and anonymize the data at the site where the data were stored. Jerboa took as its input a set of simple, standardized data files that could be generated locally by the database custodian with minimal effort. The output of Jerboa was further processed by an R script that implemented the model and generated the validation results, either locally or in a secured remote research environment. Thus, an important requirement for the use of many external data sets, viz. that patient level data should stay at the database site, could be met. Apart from addressing privacy and governance issues, the validation pipeline offers transparency by documenting all of the validation steps and allowing inspection of intermediate results. It also greatly simplifies the implementation of additional analyses, which amount to centrally changing the R script, locally rerunning it, and gathering the validation results in the remote research environment. A decentralized approach to implement changes would have been unwieldy, less transparent and more error-prone.

The full pipeline was tested with nine different data sources for the external validation of the MMSE model. We did not perform a formal evaluation of the effort to implement and tune the pipeline for the MMSE model, but estimate that adjustment of Jerboa and development of the R script took about two weeks of work. The generation of the Jerboa input files may vary across databases but typically took another week of work per database. Thus, we were able to quickly generate validation results for most data sources once data access had been granted. For a few data sources, generation of the Jerboa input files took longer. For example, the MMSE scores in the Copenhagen database had to be extracted from free text in the electronic health records, for which an automatic algorithm in combination with manual curation was used. The development and testing of this text-mining algorithm took time. It should also be noted that governance rules of many of the data sources that were used for external validation of the MMSE model, did not allow the data to leave the local environment. This made the ability to locally execute Jerboa and the R script a crucial feature to include these data sets in the validation exercise. For the other two models, we did not implement the full validation pipeline, mainly because these models were validated on only one or two external data sources of which the (anonymized) data was made available to the investigators who performed the validation exercise. Thus, there was no need to locally analyze the data and considering the small number of data sources and the effort to adjust the Jerboa tool for these models, it was decided to bypass the Jerboa processing step. If more data sources would become available for external validation of these models, use of Jerboa should be reconsidered.

Finding suitable validation sets and getting access to their data proved to be a second challenge in external model validation. An important reason for the difficulty to find suitable data sets is the large variety of variables across data sets. Although the preclinical model and the institutionalization model include only five input variables, it proved extremely difficult to find external data sources that contained these variables and also provided the required output variable. In fact, the data sources that were selected for the current validation exercise of these models required one or two variable conversions to meet the model requirements. An important consideration for selection of the MMSE model was its relative simplicity: it requires only two input variables (age and time since onset AD dementia) and one commonly available output variable, MMSE. We expected that many data sources could provide these variables, and indeed we were able to validate the MMSE model on nine external data sets. However, for more complex models involving variables that are less commonly available, it is likely that few data sources, if any, will meet model requirements.

The selection of data sources has currently been based on information from the EMIF and DPUK Data Catalogues, and by informal contacts between researchers and database owners, as described in Deliverable 3.4, “Final report on proof of concept technical solutions for RWE data harmonisation and integration”. Future searches for suitable data sources may be facilitated by use of the ROADMAP data cube, a resource that brings together information about variables, outcomes, and databases (see Deliverable D4.2, “Availability/suitability of data cube”).

Once potential suitable data sources were identified, data access had to be requested and granted. The time to complete this process varied greatly between data sources, from a few weeks to many months. For some data sources that were selected at an early stage in the ROADMAP project, it proved infeasible to arrange for data access within the project’s life span. We may conclude that finding suitable data sources for external validation and arranging for data access can be a lengthy process and generally takes much longer than generating the actual validation results.

Thirdly, there are issues related to the interpretation of the validation results. Our results indicate poor to moderate prediction performance on the validation sets for the MMSE model and the institutionalization model. For the preclinical models, good results were obtained for the APCC models, but not for the survival model. These results may partly be attributed to differences between the development set and the validation sets, which are many (as documented in chapters 3-5). However, a proper interpretation of the validation results should also take into account the model performance on the development set. For instance, the performance of the MMSE model on the validation sets was moderate at best for individual predictions, but it turned out that the performance on the development set was comparable. This suggests that the moderate performance is intrinsic to the model and cannot be explained by differences between the best-performing validation sets and the development set, or put differently, the differences between these data sets do not appear to affect model performance. For the institutionalization model, we could not compare the validation results with the results on the development data. It is therefore difficult to determine whether the poor prediction results are caused by differences between the validation set and the development set, or are model intrinsic. Interpretation of the external validation results of disease progression models requires that the model performance on the development set is known. Ideally, an internal validation is performed during model development, e.g., by dividing the development set in a training set to develop the model and an independent test set to evaluate its performance, or by cross-validation techniques. Unfortunately, very few publications about disease progression models report such internal validation results.

7. References

Arnoldussen, I.A.C., et al. (2018). A 10-year follow-up of adiposity and dementia in Swedish adults aged 70 years and older. J. Alzheimers Dis. 63, 1325-35.

Barnett, J.H., Blackwell, A., Scheltens, P., et al. (2010). Cognitive function and cognitive change in dementia, mild cognitive impairment, and healthy aging: The EDAR study. Alzheimers Dement. 6, S127.

Belger, M., Haro, J.M., Reed, C., et al. (2018). Determinants of time to institutionalisation and related healthcare and societal costs in a community-based cohort of patients with Alzheimer's disease dementia. Eur. J. Health Econ. [Epub ahead of print]

Bennett, D.A., Schneider, J.A., Buchman, A.S. et al. (2005) The Rush Memory and Aging Project: study design and baseline characteristics of the study cohort. Neuroepidemiol. 4, 163-75.

Berk, C., and Sabbagh, M. (2013). Successes and failures for drugs in late-stage development for Alzheimer's disease. Drugs Aging. 30:783-92.

Canevelli, M., Kelaiditi, E., Del Campo, N., et al. (2016). Predicting the Rate of Cognitive Decline in Alzheimer Disease: Data From the ICTUS Study. Alzheimer Dis. Assoc. Disord. 30, 237-42.

Caputo, A., Racine, A., Paule, I., et al. (2017). A model for Alzheimer's disease in the prevention setting. In: Proc. ISPOR 20th Annual European Congress. 20, A756.

Chua, L. (2015). Modeling Alzheimer's disease progression on a pathological timeline. Thesis, Department of Pharmacy, National University of Singapore.

Collins, G.S., et al. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann. Intern. Med. 162, 55-63.

Coloma, P.M., Schuemie, M.J., Trifiro, G. (2011). Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol. Drug Saf. 20, 1-11.

Dufouil, C., et al. (2017). Cognitive and imaging markers in non-demented subjects attending a memory clinic: study design and baseline findings of the MEMENTO cohort. Alzheimers Res. Ther. 9, 67.

García-Gil Mdel, M., Hermosilla, E., Prieto-Alhambra, D., et al. (2011). Construction and validation of a scoring system for the selection of high-quality data in a Spanish population primary care database (SIDIAP). Inform Prim Care. 19, 135-45.

Garre-Olmo, J., Flaqué, M., Gich, J., et al. (2009). Registry of Dementia of Girona Study Group (ReDeGi Group). A clinical registry of dementia based on the principle of epidemiological surveillance. BMC Neurol. 9, 5.

Garre-Olmo, J., López-Pousa, S., Vilalta-Franch, J., et al. (2010). Grouping and trajectories of the neuropsychiatric symptoms in patients with Alzheimer's disease, part I: symptom clusters. J. Alzheimers Dis. 22, 1157-67.

Green, C., Shearer, J., Ritchie, C.W., et al. (2011). Model-based economic evaluation in Alzheimer's disease: a review of the methods available to model Alzheimer's disease progression. Value Health. 14, 621-30.

Handels, R.L.H., Xu, W., Rizzuto, D., et al. (2013). Natural progression model of cognition and physical functioning among people with mild cognitive impairment and alzheimer's disease. J. Alzheimers Dis. 37, 357-65.

Hickey, G.L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2018). Joint models of longitudinal and time-to-event data with more than one event time outcome: a review. Int. J. Biostat. 14.

Johansson, L., et al. (2010). Midlife psychological stress and risk of dementia: a 35-year longitudinal population study. Brain J. Neurol. 133, 2217–24.

Katz, S., Ford, A.B., Moskowitz, R.W., et al. (1963). Studies of illness in the aged. The index of ADL: a standardized measure of biological and psychosocial function. JAMA. 185, 914-9.

Langbaum, B.S., et al. (2014). An empirically derived composite cognitive test score with improved power to track and evaluate treatments for preclinical Alzheimer's disease. Alzheimers Dement. 666–74.

Lawton, M.P., and Brody, E.M. (1969). Assessment of older people: self-maintaining and instrumental activities of daily living. Gerontologist. 9, 179-86.

Lestini, G., et al. (2018). Two-steps modelling approach of time to event and cognitive decline to inform Alzheimer's disease prevention trials. Population Approach Group Europe (PAGE), Montreux, Switzerland.

Liao, W., et al. (2016). A profile of The Clinical Course of Cognition and Comorbidity in Mild Cognitive Impairment and Dementia Study (The 4C study): two complementary longitudinal, clinical cohorts in the Netherlands. BMC Neurol. 16.

Moons, K.G.M., et al. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and elaboration. Ann. Intern. Med. 162, W1-W73.

Rósza, S., Brandtmüller, A., Nagy, B., et al. (2009). The psychometric properties of ADCS – activities of daily living inventory and comparison of different ADL scores. HEDS Discussion Paper 09/05, University of Sheffield.

Trifiro, G., Coloma, P.M., Rijnbeek, P.R., et al. (2014). Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? J. Intern. Med. 275, 551-61.

Viswanathan, A., Macklin, E.A., Betensky, R., et al. (2015). The influence of vascular risk factors and stroke on cognition in late life: analysis of the NACC cohort. Alzheimer Dis. Assoc. Disord. 29, 287-93.

Vlug, A.E., van der Lei, J., Mosseveld, B.M., et al. (1999). Postmarketing surveillance based on electronic patient records: the IPCI project. Methods Inf. Med. 38, 339-44.

ANNEXES

ANNEX I. TRIPOD model development and validation checklist

Section/Topic			Checklist Item	Page
Title and abstract				
Title	1	D;V*	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	
	5b	D;V	Describe eligibility criteria for participants.	
	5c	D;V	Give details of treatments received, if relevant.	
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	
Sample size	8	D;V	Explain how the study size was arrived at.	
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	
	10c	V	For validation, describe how the predictions were calculated.	
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	
	15b	D	Explain how to the use the prediction model.	
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V

ANNEX II. TRIPOD development checklists for selected models

This Annex provides the TRIPOD development checklists for the three selected models: Handels' MMSE model, Novartis' preclinical model, and Eli Lilly's institutionalization model.

Table 1: TRIPOD development checklist for Handels' MMSE model.

Section/Topic		Checklist Item	Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	357
		Natural Progression Model of Cognition and Physical Functioning among People with Mild Cognitive Impairment and Alzheimer's Disease (Handels RL, Xu W, Rizzuto D, et al. J Alzheimers Dis. 2013;37:357-65)	
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	357
		Objective: We aimed to estimate AD-free survival time in people with mild cognitive impairment (MCI) and decline of cognitive and physical function in AD cases. Methods: Within the Kungsholmen project, 153 incident MCI and 323 incident AD cases (international criteria) were identified during 9 years of follow-up in a cognitively healthy cohort of elderly people aged ≥75 at baseline (n = 1,082). Global cognitive function was assessed with the Mini-Mental State Examination (MMSE), and daily life function was evaluated with the Katz index of activities of daily living (ADL) at each follow-up examination. Data were analyzed using parametric survival analysis and mixed effect models. Results: Median AD-free survival time of 153 participants with incident MCI was 3.5 years. Among 323 incident AD cases, the cognitive decline was 1.84 MMSE points per year, which was significantly associated with age. Physical functioning declined by 0.38 ADL points per year and was significantly associated with age, education, and MMSE, but not with gender. Conclusion: Elderly people with MCI may develop AD in approximately 3.5 years. Both cognitive and physical function may decline gradually after AD onset. The empirical models can be used to evaluate long-term disease progression of new interventions for AD. In the following, we focus on one of the models developed in this study, which estimates the changes of cognition (as assessed by the MMSE) in incident AD dementia cases.	
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	358
		Natural progression models in AD have been developed in several studies, mostly among clinical samples or prevalent AD dementia cases. However, disease modifying treatments are supposed to be effective in early (pre-dementia) AD, thus long-term data on the natural course are required to evaluate their effectiveness. Such target populations have not been reflected by previous studies, leaving an urgent need for population-based empirical models that describe the long-term natural progression of the dementia and pre-dementia phases of AD.	
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	358
		Objective: Estimate the changes of cognition (MMSE) in incident AD dementia cased from a population-based cohort. The study describes the development of the model.	
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	358
		Source of data is the Kungsholmen Project, a population-based cohort study on aging and dementia	
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	358
		The Kungsholmen project started in 1987. Data were collected at baseline and at 3-, 6-, and 9-year follow-ups.	
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	358

		General population	
	5b	Describe eligibility criteria for participants.	358
		All registered inhabitants of the Kungsholmen district of Stockholm, Sweden, who were aged ≥75 years in October 1987, had no dementia, MCI, or an MMSE < 20 at baseline, with incident AD-type dementia (either AD or mixed AD & vascular dementia) during follow-up. A diagnosis of dementia (including both questionable and definite diagnoses) was established by the examining physicians, based on a comprehensive clinical examination and cognitive tests according to the DSM-III-R criteria. The diagnostic criteria applied were equivalent to probable AD according to the criteria of the National Institute of Neurological and Communicative Dis- orders and Stroke-Alzheimer's Disease and Related Disorders Association, and according to those of the National Institute of Neurological Disorders and Stroke-Association Internationale pour la Recherche et l'Enseignement en Neurosciences.	
	5c	Give details of treatments received, if relevant.	
		Not reported	
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	358
		Outcome MMSE. Assessment at follow-ups.	
	6b	Report any actions to blind assessment of the outcome to be predicted.	
		Not reported	
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	359
		Predictors tested: Age, Gender, Education, and Time after being diagnosed with AD. The onset of AD was assumed to have taken place in the middle of each follow-up interval (each lasting an average of 3 years). This was operationalized by adding a time correction of 1.5 years. Only Age and Time after being diagnosed turned out to be significant predictors.	
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	
		Not reported	
Sample size	8	Explain how the study size was arrived at.	359
		No sample size calculations done, entire cohort used	
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
		The mixed model with random effects takes missing or censored data into account.	363
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.	
		Not specifically reported, no categorization or transformation	
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	359
		Mixed model with random effects. A stepwise procedure was used and predictors were included if the goodness-of-fit statistics −2 log likelihood change and Wald z of the predictor were significant. The following steps were used to determine the final MMSE prediction model: (1) include time, as years after being diagnosed with AD; (2) include a random intercept; (3) determine if time is non-linear by stepwise adding a higher-order polynomial of time; (4) include a random time factor; (5) include gender, age, and education and all 2-way interactions and remove interactions with highest p-values first until p < 0.05, followed by predictors. No internal validation was performed.	
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
		No measures used, model performance not assessed	
Risk groups	11	Provide details on how risk groups were created, if done.	
		Not done	
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	358, 359
		At baseline, 225 of the 1,810 participants were diagnosed with dementia and 110 participants refused the extensive evaluations. Of the remaining 1,475 dementia-free persons, 355 with MCI (130 with amnestic MCI (aMCI) and 225 with other cognitive impairment not demented (OCIND)) at baseline and 38 with very low global cognitive status in the absence of a dementia diagnosis (MMSE) <20) were excluded, leaving 1,082 cognitively healthy subjects at baseline. Out of those, 323 developed AD during 9-year follow-up.	
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	359

		Age at diagnosis 86.7 (4.1) yrs, 83% female, education 8.2 (2.9) yrs, MMSE at diagnosis 19.7 (5.0), Katz ADL at diagnosis 1.2 (0.7). No specific information on missing data.	
Model development	14a	Specify the number of participants and outcome events in each analysis.	360
		For the 323 participants who developed AD during follow-up, 313 MMSE scores were available at the moment of AD diagnosis, 109 at 3 years after diagnosis, and 28 at 6 years after diagnosis. Forty-nine percent of the participants died during follow-up.	
	14b	If done, report the unadjusted association between each candidate predictor and outcome.	362
		Regression parameters estimates (95% CI) of univariate mixed effects regression model to predict MMSE: Age -0.41 (-0.57 to -0.26), Time after being diagnosed -1.84 (-2.10 to -1.57), Gender -1.14 (-2.89 to 0.60), Education -0.05 (-0.29 to 0.19).	
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	360
		MMSE = 26.87 – 3.26 Time – 0.35 (Age – 75) + 0.10 Time (Age – 75), in which Time is years after being diagnosed with AD.	
	15b	Explain how to the use the prediction model.	
		Not reported, but straightforward	
Model performance	16	Report performance measures (with CIs) for the prediction model.	
		Model performance not assessed	
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	363
		The Kungsholmen project included persons aged 75 and older, which resulted in attrition due to death and refusal. However, this reflects reality, since most demented people are older than 75, and the mixed model with random effects and the survival analysis take missing or censored data into account. Nonetheless, generalization to a younger population should be done with caution, A second limitation is that the Kungsholmen project started in 1987, when the current cholinesterase inhibitors and memantine treatments that affect cognitive decline were not available. Thirdly, the empirical models were not adjusted for comorbidities, as this information was not available to the researchers. Furthermore, the 1.5 year correction might limit the precision of the time-to dementia conversion. The regression and survival models have not been validated by external datasets, or by predicting the progress of similar patients in current clinical practice. The data available at follow-up was limited, resulting in uncertain predictions. Finally, generalizability to other countries is limited because differences in life expectancy might lead to differences in average disease progression rates or the effect of age.	
Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	361-363
		A population-based study including 95 incident dementia participants [20, 21] found an average rate of cognitive decline of 1.71 MMSE points per 6 months, whereas we found a lower average rate of decline (1.84 / 2 = 0.92 points per 6 months). The difference could be explained by the inclusion of a higher proportion of moderately severe dementia participants in the Kungsholmen Project, who decline less quickly due to the floor effect of the MMSE. According to the multivariate model using average age, subjects decline by 1.2 MMSE points in the first 6 months after being diagnosed. Mendiondo et al. [22] and Mohs et al. [23] parameterized the annual rate of cognitive decline and found a U-shaped pattern with low decline rates in mild and severe dementia and a higher decline rate in between. We explored this model, but the results were not significant and could be attributed to the use of a population-based sample instead of a clinical sample, as the latter probably includes persons with a poorer prognosis because consulting a medical professional is probably initiated by the person's memory complaints. Han et al. [24] reviewed studies largely based on clinical samples of prevalent cases with an average of 2 years of follow-up, and found a mean annual rate of decline of 3.3 MMSE points per year. Our estimates are at the lower bound of their confidence interval. Besides the use of incident community participants, this difference could be explained by the long follow-up time, in which some participants reach the floor level of the MMSE.	
Implications	20	Discuss the potential clinical use of the model and implications for future research.	363
		The empirical models developed in this study (including the MMSE model) could be used to simulate the natural disease progression in a cohort and compare this with a scenario where a hypothetical future treatment is available. Such predictions can be integrated with evidence on health care resource usage and quality of life, and enable policy makers to address questions about the potential of new diagnostic or treatment interventions from a cost-effectiveness point of view. Such analyses could provide added value to randomized controlled trials which are limited in terms of follow-up time or the number of scenarios to compare.	
Other information			

Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	
		No supplemenary resources mentioned	
Funding	22	Give the source of funding and the role of the funders for the present study.	364
		Dutch Alzheimer's Society, Center for Translational Molecular Medicine	

Table 2: TRIPOD development checklist for Novartis' preclinical model.

Section/Topic		Checklist Item	Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
		Model for APCC time profile and time to first diagnosis of mild cognitive impairment or dementia due to Alzheimer's disease (AD) in elderly, cognitively normal individuals at risk to develop symptoms of AD	-[1]
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
		OBJECTIVES: Shifting the focus of clinical trials testing disease-modifying interventions against Alzheimer's disease from the dementia stages of the disease to pre-symptomatic stages may increase the likelihood of success for these trials. The aim of this research was to develop a model for the pre-symptomatic time course in the AD prevention setting to inform clinical trial design. METHODS: We developed a statistical model describing time to first diagnosis of mild cognitive impairment (MCI) and dementia diagnosis using a Weibull parametric survival model and the progression of the Alzheimer's Prevention Initiative Preclinical Composite (APCC, see Langbaum et al. 2014[2]), a measure for cognitive decline, using a non-linear mixed-effects model. We chose model covariates based on clinical relevance, goodness of model fit and statistical tests. We trained the model on cohorts from the Rush Alzheimer's disease center (Rush) (ROS, MAP and MARS) and the National Alzheimer's Coordinating Center (NACC), US databases including healthy as well as cognitively impaired and demented subjects. For the time-to-diagnosis model, we used N=2159 subjects from Rush and N=8535 subjects from NACC who were cognitively normal at baseline and were diagnosed with MCI or dementia due to AD during follow-up. For the APCC model, we used N=2336 subjects from Rush who were cognitively normal at baseline and had no other diagnoses than MCI or dementia due to AD during follow-up. RESULTS: We identified age, apolipoprotein E ε4 (APOε4) status, APCC at baseline and education level as important model covariates. Patient simulations showed a good fit between model predictions and observed values, for both time to first diagnosis and progression of APCC. Simulations also showed that an enrichment strategy focusing on elderly participants yielded a higher power for a given hazard ratio of the investigated interventions. CONCLUSIONS: The 2-step model linking APCC decline and time to MCI or AD diagnosis is the first AD disease progression model for pre-symptomatic stages of the disease. It can be used in the context of optimizing design of clinical trials in the prevention setting. Further refinements of the model, e.g. including biomarkers such as amyloid-beta and tau as covariates and covering other relevant endpoints, external validation of the model, and incorporation into a health economic model to evaluate interventions in the prevention setting, are objectives of future research.	
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	
		Shifting the focus of clinical trials testing disease-modifying interventions against AD from the dementia stages of the disease to pre-symptomatic stages may increase the likelihood of success for these trials. Various models describing cognitive decline in later stages of AD exist so far, but a model describing cognitive function in the pre-symptomatic phase of the disease and predicting time to first diagnosis of MCI or dementia is lacking. Hence, there is an urgent need of such a model to e.g. inform the design of trials targeting patients at risk to develop dementia.	
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	

[1] No publication available.

[2] http://www.sciencedirect.com/science/article/pii/S1552526014000636

		The aim of this study was to develop a model for the pre-symptomatic time course in the AD prevention setting. The study describes the development of the model.	
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	
		Source of data are the Rush and the NACC longitudinal cohorts. Rush: Cohort study cohort study of common chronic conditions of aging with emphasis on decline in cognitive and motor function and risk of AD. NACC: Prospective cohort study with participants from Alzheimer's Disease Centers (ADCs).	
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	
		Rush: Started in 1997, still ongoing. NACC: Started in 2005, still ongoing.	
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	
		Rush: Participants are older adults recruited from 37 retirement communities and subsidized senior housing facilities throughout Chicagoland and north-eastern Illinois. NACC: Participants are followed at 39 past and present U.S. ADCs (with or without dementia). Subjects may come from clinician referral, self-referral by patients or family members, active recruitment through community organizations, and volunteers who wish to contribute to research on various types of dementia. Most centers also enrol volunteers with normal cognition.	
	5b	Describe eligibility criteria for participants.	
		Rush: - older persons without known dementia - must agree to an assessment of risk factors, blood donation, and a detailed clinical evaluation each year NACC: - participant at a contributing ADC	
	5c	Give details of treatments received, if relevant.	
		Not reported	
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	
		Outcomes: - APCC, assessed continuously throughout the study (from Rush) - Diagnosis of MCI and dementia due to AD, assessed throughout the study (from Rush and NACC)	
	6b	Report any actions to blind assessment of the outcome to be predicted.	
		Not reported	
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	
		Tested predictors are APCC at baseline, age at baseline or at time of diagnosis, gender, APOε4 status and educational level (years of education).	
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	
		Not reported	
Sample size	8	Explain how the study size was arrived at.	
		No sample size calculations done, entire cohort used	
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
		APCC model: The mixed effects model takes missing data into account. Time-to-first-diagnosis model: The Weibull survival regression model takes censored data into account, but removes subjects with missing covariates.	
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.	
		Continuous predictors were log transformed and centered around their median.	
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	
		APCC model: Non-linear mixed effects model (power model). Time-to-first-diagnosis model: Weibull survival regression model. Model structures were chosen because of their flexibility to fit the data. Covariate were chosen based on investigating the predictive value of a set of candidate predictors in a systematic way. No internal validation performed.	

	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
		Model performance was assessed using diagnostic plots.	
Risk groups	11	Provide details on how risk groups were created, if done.	
		Not done.	
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	
		APCC model: We evaluated a total of N=2336 subjects from Rush who were cognitively normal at baseline, had at least two visits and had no other diagnoses than MCI or dementia due to AD during follow-up. Of those subjects, 732 were first diagnosed with MCI or dementia within eight years, and 1604 stayed cognitively normal within eight years of follow-up. Time-to-first diagnosis model: We evaluated a total of N=10694 subjects from Rush and NACC who were cognitively normal at baseline, had at least two visits and had no other diagnoses than MCI or dementia due to AD during follow-up. Of those subjects, 2870 were first diagnosed with MCI or dementia, and 859 were first diagnosed with dementia.	
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	
		APCC model: Mean APCC at baseline 61.0, mean education 16.1 years, 1.1% homozygote carriers of APOε4 and 23.6% heterozygote carriers (8.3% missing values) for subjects diagnosed with MCI or dementia. Mean APCC at baseline 64.9, mean education 16.0 years, no homozygote carriers of APOε4 and 18.0% heterozygote carriers for subjects staying cognitively normal. Time-to-first diagnosis model: Mean age at baseline was 74.4 years, mean APCC at baseline was 63.5 (16.4% missing values), 1.7% homozygote carriers of APOε4 and 20.6% heterozygote carriers (39.7% missing values).	
Model development	14a	Specify the number of participants and outcome events in each analysis.	
		APCC model: APCC was available for all subjects diagnosed with MCI or dementia at baseline, at four subsequent follow-up visits on average, and maximally at seventeen subsequent follow-up visits. APCC was available for all subjects staying cognitively normal at baseline, at three subsequent follow-up visits on average, and maximally at eight subsequent follow-up visits. Time-to-first diagnosis model: 2870 subjects were first diagnosed with MCI or dementia (7824 censored), 859 were diagnosed with dementia (9835 censored).	
	14b	If done, report the unadjusted association between each candidate predictor and outcome.	
		Not reported	
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	
		APCC model for converters, i.e. subjects diagnosed with MCI or dementia within eight years: mixed-effects power model with predictor APCC at baseline for the intercept and the slope, and predictors education and APOε4 carrier status for the slope. APCC model for non- or late-converters, i.e. subjects staying cognitively normal within eight years: linear mixed-effects model with predictors education and age at baseline for the intercept, and predictors APCC at baseline, APOε4 carrier status and age at baseline for the slope. Time-to-first-diagnosis of MCI or AD model: Weibull survival regression model with predictors age at baseline, APCC at baseline and APOε4 carrier status. For clinical trial simulations, the models were linked in the following way: first, time to first diagnosis of MCI or AD was simulated. Second, if a subject was diagnosed within 8 years, the APCC model for converters was applied to simulate APCC progression for that subject. If a subject was not diagnosed within 8 years, the APCC model for non-/late-converters was applied to simulate APCC progression for that subject. A further link between the two models exists via the time to event: The APCC models the time course using TTE minus 8 years as the baseline and not calendar time t=0.	
	15b	Explain how to the use the prediction model.	
		Straightforward	
Model performance	16	Report performance measures (with CIs) for the prediction model.	
		Model performance not assessed	
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	
		- APCC in the NACC database is just a proxy - Number of subjects in specific subgroups of interest is rather small. Example: APOE4	

		homozygote carriers - No biomarker data available, hence, no information on important prognostic factors - Model structure needs to be justified, i.e. compared with other model structures - Choice of model covariates needs to be justified, i.e. should be done systematically	
Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	
		The 2-step model linking APCC decline and time to MCI or AD diagnosis is the first AD disease progression model for pre-symptomatic stages of the disease. It can be used in the context of optimizing design of clinical trials in the prevention setting, although results have to be considered with care since a validation of the model is lacking. Some limitations of the model may be due to the fact that the model was originally not developed as a disease model with a broader and more general interpretation, but as a basis for trial simulations in a specific setting. Hence, the strategy of the model development and model fit was tailored to the requirements of the clinical trial setting. These limitations need to be investigated and modifications of the model may be explored to leverage the model to a broader application and interpretation.	
Implications	20	Discuss the potential clinical use of the model and implications for future research.	
		APCC starts to decline in cognitively normal individuals ~5 years before MCI/dementia diagnosis, therefore the model could also be used to predict time to MCI/dementia diagnosis in healthy individuals once APCC decline has started, i.e. ~2 years before.	
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	
		None	
Funding	22	Give the source of funding and the role of the funders for the present study.	
		The model was developed within Novartis.	

Table 3: TRIPOD development checklist for Eli Lilly's institutionalization model.

Section/Topic	Checklist Item		Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
		Healthcare and societal costs related to the time to institutionalisation in a community-based cohort of patients with Alzheimer's disease dementia Mark Belger[1], Josep Maria Haro[2], Catherine Reed[1], Michael Happich[1], Josep Maria Argimon[3], Giuseppe Bruno[4], Richard Dodel[5], Roy W. Jones[6], Bruno Vellas[7], Anders Wimo. Publication has been submitted to European Journal of Health Economics. The Modelling structure is also described in the following publications (PENTAG model) 1. Green, C., Shearer, J., Ritchie, C.W., Zajicek, J.P.: Model-based economic evaluation in Alzheimer's disease: a review of the methods available to model Alzheimer's disease progression. Value Health 14(5), 621–630 (2011). doi: 10.1016/j.jval.2010.12.008 2. Bond, M., Rogers, G., Peters, J., Anderson, R., Hoyle, M., Miners, A., Moxham, T., Davis, S., Thokala, P., Wailoo, A., Jeffreys, M., Hyde, C.: The effectiveness and cost-effectiveness of donepezil, galantamine, rivastigmine and memantine for the treatment of Alzheimer's disease (review of Technology Appraisal No. 111): a systematic review and economic model. Health Technol. Assess. 16(21), 1–470 (2012). doi: 10.3310/hta16210.	
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
		Objectives: To examine the costs of caring for community-dwelling patients with Alzheimer's disease (AD) dementia in relation to the time to institutionalisation. Methods: GERAS was a prospective, non-interventional cohort study in community-dwelling patients with AD dementia and their caregivers in three European countries. Using identified factors associated with time to institutionalisation, models were developed to estimate the time to institutionalisation for all patients. Estimates of monthly total societal costs, patient healthcare costs and total patient costs (healthcare and social care together) prior to institutionalisation were developed as a function of the	

		time to institutionalisation. Results: Of the 1495 patients assessed at baseline, 307 (20.5 %) were institutionalised over 36 months. Disease severity at baseline (based on Mini-Mental State Examination [MMSE] scores) was associated with risk of being institutionalised during follow-up ($p < 0.001$). Having a non-spousal informal caregiver was associated with a faster time to institutionalisation (944 fewer days versus having a spousal caregiver), as was each one-point worsening in baseline score of MMSE, instrumental activities of daily living and behavioural disturbance (67, 50 and 30 fewer days, respectively). Total societal costs, total patient costs and, to a lesser extent, patient healthcare-only costs were associated with time to institutionalisation. In the five years pre-institutionalisation, monthly total societal costs increased by more than £1000 (€1166 equivalent for 2010) from £1900 to £3160 and monthly total patient costs almost doubled from £770 to £1529. Conclusions: Total societal costs and total patient costs rise steeply as community-dwelling patients with AD dementia approach institutionalisation.	
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	
		The PENTAG model has been used for economic models to assess the cost effectiveness of ACHEi's. During these submissions NICE identified a number of weaknesses to the submitted model, these focused around the relevance of the data used to build the models. The recent work has focused on developing models using the GERAS study data for both time to Institutionalisation, time to death and costs and quality of life related to pre-institutionalisation time.	
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	
		The work is an update on the PENTAG model, using more recent data from The GERAS study. No external validation has been performed on the equations used within the model. The publication includes equations to predict the time to institutionalisation and equations for cost as a relationship to pre-institutionalisation. These are taken from the three-year follow-up data from the GERAS study. Additional models are available based on 60 month follow up data from GERAS, and including models on time to death, and the relationship of pre-Institutionalisation to to quality of life (EQ-5D)	
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	
		The data comes from the GERAS study (reference is: Wimo, A., Reed, C.C., Dodel, R., Belger, M., Jones, R.W., Happich, M., Argimon, J.M., Bruno, G., Novick, D., Vellas, B., Haro, J.M.: The GERAS Study: a prospective observational study of costs and resource use in community dwellers with Alzheimer's disease in three European countries – study design and baseline findings. J. Alzheimers Dis. 36(2), 385–399 (2013). doi: 10.3233/JAD-122392	
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	
		GERAS is an 18-month, multicentre, observational study designed to assess the direct and indirect country costs associated with AD for patients and their caregivers in France, Germany and the UK. Patients in France and Germany were being followed for a further 18 months. An addendum to the study collected information on Date of death and date of institutionalisation. Recent database lock on the 60-month follow up data is available. The study enrolled patients between October 1 2010 and September 31 2011. Patients and caregivers were evaluated at baseline and every six months	
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	
		Patients enrolled were in a community dwelling with a probable AD diagnosis according to the National Institute of Neurological and Communicative Disorders, and stroke and Alzheimer's disease and related disorders association (NINCDS-ADRDA) 94 sites were enrolled from three countries	
	5b	Describe eligibility criteria for participants.	
		Community dwelling Age ≥55 years; Probable AD (NINCDS-ADRDA)	

		An MMSE score of ≤26 Presented within the normal course of care Patients were excluded if they had a history, clinical signs or imaging of stroke or transient ischemic attack, patients with an history of Parkinson's disease prior to or at the start of AD onset; Probable Lewy-body disease. Patients were required to have a caregiver who was willing to participate in the study, and were defined as an informal carer who would normally take care of day to day activities (not for a health care professional)	
	5c	Give details of treatments received, if relevant.	
		Patients were on standard of care, there was no requirement for patients to be treated with any specific AD medication at study entry.(78% received ACHEi's; 21% were receiving Memantine at study enrolment)	
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	
		Time to Institutionalisation Total societal cost as a function of Pre-Institutionalisation Patient medical cost as a function of Pre-Institutionalisation Patient medical and social care cost as a function of Pre-Institutionalisation Models are also available for, but not in publication. Time to death Quality of as a function of Pre-Institutionalisation	
	6b	Report any actions to blind assessment of the outcome to be predicted.	
		This is an observational study, no blinding occurred.	
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	
		Two models were considered one including only patient characteristics, and a second model which included both patient and caregiver characteristics: All predictors measured at baseline: Patient characteristics considered: Age Gender Years of education Time since diagnosis of AD Number of comorbidities MMSE score Total ADCs-ADl Instrumental ADCS-ADL Basic ADCS-ADL NPI AD medication Caregiver factors considered Age Gender Relationship with patients (spouse yes/no) Caregiver working for pay Sensitivity analysis were considered which looked at interaction terms, and sub-domains of the ADL and the NPI Details of the scales used can be found in the Wilmo publication	
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	
		No blinding	
Sample size	8	Explain how the study size was arrived at.	
		Enrolment was over a 12 month period, with sample size based on country and MMSE severity group. Sites were selected within the three countries to aim for approximately equal numbers of patients in each MMSE severity group. Sample size was based on the precision obtained for estimating costs Further details are provided in the Wilmo publication	

Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
		Survival analysis was used for models predicting time to Institutionalisation, patients were censored at last visit or at time of discontinuation from the study No imputation was performed on missing baseline data as over 97% of baseline data available Missing Cost data was imputed based on the reason for missing cost data.The following rules were applied: 1. For institutionalised patients, mean monthly costs from the last visit were used for the period until institutionalisation and monthly costs for institutionalisation were used from institutionalisation up to 18 months for the UK and up to 36 months for France and Germany. For patients who died, last observation carried forward was used such that costs from the last known visit were extrapolated up to the date of death (no costs after death were computed). For patients with other reasons for discontinuation, the multiple imputation regression method [19] stratified by MMSE group and country was applied to missing costs. The list of factors used in the multiple imputation procedure was selected from those identified by Dodel et al. (Dodel, R., Belger, M., Reed, C., Wimo, A., Jones, R.W., Happich, M., Argimon, J.M., Bruno, G., Vellas, B., Haro, J.M.: Determinants of societal costs in Alzheimer's disease: GERAS study baseline results. Alzheimers Dement. 11(8), 933–945 (2015). doi: 10.1016/j.jalz.2015.02.005)	
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.	
		No transformations were conducted on the continuous variables. The caregiver relationship categorical variable was dichotomised into spouse (yes/no).	
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	
		Factors associated with time to institutionalisation were explored using Cox proportional hazards models of the 36-month data; time to institutionalisation was censored at the time of last follow-up or time to death for those subjects who did not report being institutionalised. One hundred different models using forward and backward selection were run, selecting 67 % of subjects at random for inclusion in the model, and the factors identified in each model summarised. Entry and exclusion of individual factors was based on a significance level of 0.05. Any factor found to be significant in over 75 % of the models was included in the parametric models used to predict time to institutionalisation. To allow for different assumptions around the distribution of the data, the parametric models considered exponential, log-logistic, Weibull, log-normal and gamma distributions. Model fit was assessed using AIC and BIC model fit statistics, and the best fitting model was selected for use in the model that estimated societal and patient costs as a function of time to institutionalisation. Models were fitted to estimate costs (y) as a function of time to institutionalisation (x). Separate models were developed for total societal costs, total patient costs (patient healthcare plus social care costs) and patient healthcare costs. For each patient, the predicted time to institutionalisation (Pred_Inst) was calculated from the parametric model. Then, for each 6-month visit, the patient's time to institutionalisation (Pre-Inst) was calculated as: Pre-Inst = Pred_Inst – visit. Each individual subject time point was treated as independent, had an associated cost and any missing cost visits used the imputation methods described earlier.	
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
		See above, Time to institutionalisation models were assessed by AIC and BIC, and then a visual inspection of the extrapolated curves.	
Risk groups	11	Provide details on how risk groups were created, if done.	
		No risk groups created	
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	
		1495 patients were enrolled into the study, 307 were institutionalised during the 36-month follow up, while 152 patients died before being institutionalised. 298 patients discontinued the study before end of follow up period.(18 months UK, and 36 months France and Germany)	
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	

<table>
<tr><td></td><td></td><td>1495 patients enrolled; 566 with Mild AD, 472 moderate and 457 with moderate severe/sever AD at baseline.
Mean (sd) age 77.6 (7.7) years
, 55% female; 72% married/cohabitating; 76% living in urban area; 96% living in own home; 10.4(3.2) years of education; 2.2 (2.2) years since AD diagnosis; baseline MMSE score 17.4 (6.3); ADLscore 46.5 (19.5); NPI_12 score 15.1 (15.3)</td><td></td></tr>
<tr><td rowspan="3">Model development</td><td>14a</td><td>Specify the number of participants and outcome events in each analysis.</td><td></td></tr>
<tr><td></td><td>1495 patients enrolled
307 institutionalised in first 36 months
152 died in first 36 months

Updated figures using the 60 month addendum data are available</td><td></td></tr>
<tr><td>14b</td><td>If done, report the unadjusted association between each candidate predictor and outcome.</td><td></td></tr>
<tr><td></td><td></td><td>Not available</td><td></td></tr>
<tr><td rowspan="3">Model specification</td><td>15a</td><td>Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).</td><td></td></tr>
<tr><td></td><td>Submitted publication uses just the 36-month data and is reporting the model including caregiver factors. The equations using just patient factors and the 60-month addendum data can be provided. The models using just patient factors and the 60 month addendum data may be more appropriate for use in external validation:

Models from submitted publication:

<table>
<tr><th>Variable</th><th>Regression coefficient[a]</th><th>Standard error</th><th>95 % confidence limits</th><th>Chi-square</th><th>p value</th><th>Change in time to institutionalisation, days (months)[b]</th></tr>
<tr><td>Intercept</td><td>7.600</td><td>0.209</td><td>7.190; 8.010</td><td>1321.79</td><td>< 0.0001</td><td>–</td></tr>
<tr><td>MMSE total score[c]</td><td>0.034</td><td>0.009</td><td>0.016; 0.052</td><td>13.75</td><td>0.0002</td><td>−67 (2.3)</td></tr>
<tr><td>Instrumental ADCS-ADL[c]</td><td>0.025</td><td>0.005</td><td>0.015; 0.036</td><td>23.44</td><td>< 0.0001</td><td>−50 (1.7)</td></tr>
<tr><td>Basic ADCS-ADL[c]</td><td>−0.044</td><td>0.014</td><td>−0.071; −0.017</td><td>10.13</td><td>0.0015</td><td>+89 (3.0)[d]</td></tr>
<tr><td>NPI-12 total score[c]</td><td>−0.015</td><td>0.003</td><td>−0.021; −0.010</td><td>27.35</td><td>< 0.0001</td><td>−30 (1.0)</td></tr>
<tr><td>Spousal caregiver, No (Ref = yes)</td><td>−0.640</td><td>0.095</td><td>−0.826; −0.453</td><td>45.18</td><td>< 0.0001</td><td>−944 (31.5)</td></tr>
<tr><td>Scale</td><td>1.206</td><td>0.054</td><td>1.105; 1.317</td><td>–</td><td>–</td><td>–</td></tr>
</table>
Analysis of maximum likelihood parameter estimates of patient and caregiver factors associated with time to institutionalisation from the log-normal model

Models showing the relationships of costs to pre-Institutionalisation (pre_Inst)

Q1: Total societal costs (£) = 3159.68 – (334.03 Pre-Inst) + (18.57 Pre-Inst^2) – (0.43 Pre-Inst^3)
Q2: Total patient costs (£) = 1528.96 – (208.53 Pre-Inst) + (12.73 Pre-Inst^2) – (0.28 Pre-Inst^3)
Q3: Patient healthcare costs (£) = 348.31 – (14.88 Pre-Inst) + (0.35 Pre-Inst^2)</td><td></td></tr>
<tr><td>15b</td><td>Explain how to the use the prediction model.</td><td></td></tr>
<tr><td>Model performance</td><td>16</td><td>Report performance measures (with CIs) for the prediction model.</td><td></td></tr>
<tr><td></td><td></td><td>Within the economic model alternative parametric models are run in the form of sensitivity analysis (These models are available for both the 36 and 60 month analysis)</td><td></td></tr>
<tr><td colspan="4">Discussion</td></tr>
<tr><td>Limitations</td><td>18</td><td>Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).</td><td></td></tr>
<tr><td></td><td></td><td>Patients with no formal caregiver were not eligible for the study
In the model with patient and caregiver factors, patient age was not selected. If caregiver factors were excluded then model uses:
Patient age, NPI, ADL and MMSE
Other factors not collected may influence the likelihood of institutionalisation are not considered, also reasons for institutionalisation may be country specific (UK model is available)
There is a possibility of selection bias due to the recruitment of the study participants mostly from memory clinics, which may limit the generalisability of the findings as the sample is not fully representative of all AD patients living in the community</td><td></td></tr>
</table>

Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	
Implications	20	Discuss the potential clinical use of the model and implications for future research.	
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	
		The economic model framework described in PENTAG, also has models for time to death, and QoL as a function of pre-Institutionalisation. Within the model there is also an equation looking at MMSE overtime The submitted publication described above is just focusing on the methods used to take a model for predicting time to Institutionalisation and relating that to costs. For the development of the economic models to update the PENTAG model we have the following information available that makes use of the 60 month follow up data: Time to Institutionalisation Time to death Costs as a function of pre-Institutionalisation Qol as a function of pre-Institutionalisation MMSE over time Models for UK only cohort have also been developed	
Funding	22	Give the source of funding and the role of the funders for the present study.	
		The GERAS study was sponsored by Eli Lilly, and analysis was conducted by Eli Lilly	

ANNEX III. TRIPOD validation checklist for IPCI

As an example of the TRIPOD validation checklists, we provide the checklist for the IPCI data set below. The topics marked in yellow are data-source specific, and were adjusted accordingly for the other validation sets. The topics that were not marked, remain the same for all validation sets.

Section/Topic		Checklist Item	Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
		Validation of a model to predict MMSE in incident AD dementia cases	
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
		TBA	
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	
		Validation of an existing MMSE disease progression model, described in: Handels RL, Xu W, Rizzuto D, et al. Natural progression model of cognition and physical functioning among people with mild cognitive impairment and alzheimer's disease. J Alzheimers Dis. 2013;37:357-65.	
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	
		Validation of an existing MMSE model	
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	
		Source of data is a longitudinal observational database of electronic patient records of Dutch general practitioners (GPs), the Integrated Primary Care Information (IPCI) database.	
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	
		TBA	
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	
		Primary care setting. 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.).	
	5b	Describe eligibility criteria for participants.	
		Participants were eligible if they were diagnosed as incident AD dementia and had at least one MMSE measurement after date of diagnosis and were ≥75 at diagnosis.	
	5c	Give details of treatments received, if relevant.	
		Not relevant	
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	
		Predicted outcome is MMSE (Mini-Mental State Examination). Date of MMSE assessment is available, no information on how assessment has been done.	
	6b	Report any actions to blind assessment of the outcome to be predicted.	
		Retrospective study, MMSE assessment did not involve information about the predictors	
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	
		Two predictors were used: Age (in year) and Time after being diagnosed with AD (in year). Age is derived from the date of birth, Time since AD is derived from the date of AD dementia diagnosis.	
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	
		Retrospective study, predictors assessed independently of other observer information	
Sample size	8	Explain how the study size was arrived at.	
		The study includes all incident cases of AD dementia in the IPCI database that had one or more MMSE measurements after diagnosis	
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	
		Complete-case analysis	

Statistical analysis methods	10c	For validation, describe how the predictions were calculated.	
		MMSE = 26.87 – 3.26 Time – 0.35 (Age – 75) + 0.10 Time (Age – 75), in which Time is years after being diagnosed with AD.	
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	
		Model performance assessed by linear regression and median absolute deviation (MAD) between predicted and observed MMSE measurements	
	10e	Describe any model updating (e.g., recalibration) arising from the validation, if done.	
		Not done	
Risk groups	11	Provide details on how risk groups were created, if done.	
		Not done	
Development vs. validation	12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	
		Differences in setting of original study (population-based cohort study), eligibility criteria (dementia dx based on DSM-III-R, NINCDS-ADRDA), and predictors (onset AD assumed in the middle of follow-up interval)	
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	
		TBA	
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	
		TBA	
	13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	
		TBA	
Model performance	16	Report performance measures (with CIs) for the prediction model.	
		TBA	
Model-updating	17	If done, report the results from any model updating (i.e., model specification, model performance).	
		NA	
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	
		TBA	
Interpretation	19a	For validation, discuss the results with reference to performance in the development data, and any other validation data.	
		Model performance on the development data has not been reported	
	19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	
		TBA	
Implications	20	Discuss the potential clinical use of the model and implications for future research.	
		TBA	
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	
		TBA	
Funding	22	Give the source of funding and the role of the funders for the present study.	
		ROADMAP	

ANNEX IV. Jerboa installation and user manual

Validation of Handels' Prediction Model

Data Preparation and Quality Control Run

1 Introduction

This document describes the creation of the input files for the validation study of the MMSE prediction model developed by Handels et al. (J Alzheimers Dis. 2013;37:357-65). The first run that will be done on these input files is a Primary Data Extraction and Quality Control run using Jerboa. Jerboa is used in a so-called distributed network in which each database is elaborated locally and analytical anonymized datasets can be shared (figure 1).

Jerboa runs on the JAVA platform on any modern computer.

Figure 1. Jerboa model for distributed computing on databases.

The output of the program can be viewed and approved by the data custodian locally. An encrypted copy of the output is also created that needs to be uploaded to the Remote Research Environment called Octopus at Erasmus MC for further analysis as shown in figure 2.

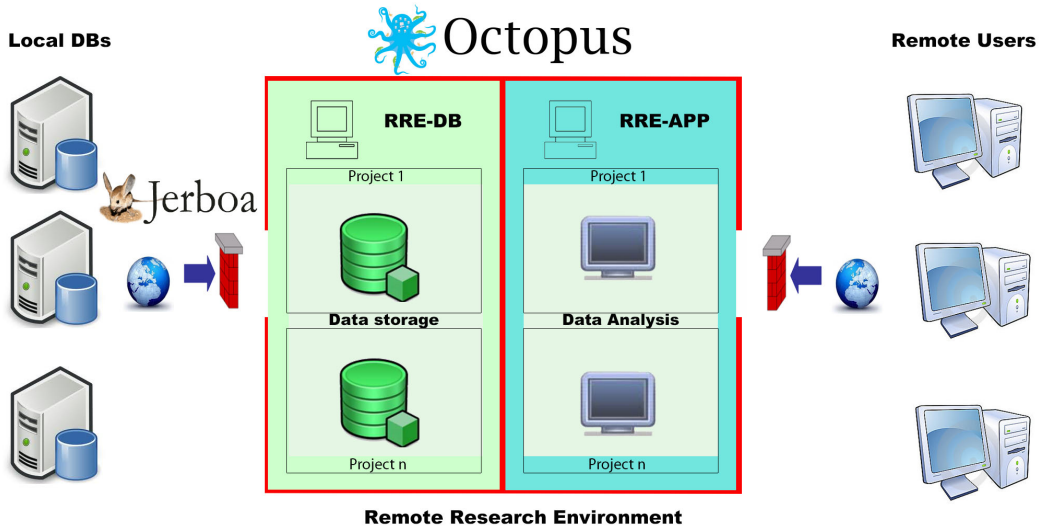


Figure 2. Octopus Remote Research Environment.

2 Jerboa data preparation

For the the validation study of Handels et al. we need three input files: patients.txt, measurements.txt, and events.txt. We do not use prescriptions.

The following format is used for the patient IDs and dates:

Patient IDs

A patient ID is an alphanumeric string of characters that uniquely identifies a patient. Patient IDs can be numbers (1, 2, 3, etc.) or combination of numbers and letters (a01, a02, b01, etc.). Maximum length of the Patient ID is 32 characters.

Important: No duplicate patient IDs are allowed.

Date formatting

All dates should be formatted as YYYYMMDD.

For example: the 28th of March, 2008, is formatted as:

20080328

File format

The input files should be in CSV (Comma-Separated Values) format. The first row should contain the column headers (the column header names provided below per file are mandatory). The order of the columns and rows is not important.

Note that missing values are not allowed except if specified! If a record in an input file contains a missing value, the entire record is considered inconsistent and placed in a list of erroneous records. Jerboa will automatically detect the patient file based on the header so the name of the file is irrelevant (we suggest to add a date to your input filenames, for example 2013-05-10-Patient.txt, to keep track of multiple versions).

Important: The input file is always checked for integrity before processing. If inconsistencies are found, an error log is produced for each input file and the user is asked to correct all errors before continue.

Patient.txt

The patient file has one record for each patient in your source population, containing the following variables:

PatientID	Patient Identifier
BirthDate	Date of birth
Gender	Gender of the patient
StartDate	Date from which the patient is eligible to start follow-up in the study. This is typically the date the patient is entered into the registration system (date of registration with insurance/region, date GP started to collaborate)
EndDate	Date that follow-up for this patient ends from a database perspective (e.g. end of registration with GP, insurance, moving out, death, last data draw down (whichever is earliest)

NOTE: include *all* patients of your source population. The cut-offs required by the validation study will be performed by Jerboa (e.g., implementing the age limit of 75 years of age of older).

Example of patients input file:

patientid,gender,birthdate,startdate,enddate
1,F,19590601,19950802,20050701
2,M,19830301,19960912,20060903

Gender
The gender of a patient can have one of the following values:

FEMALE F
MALE M

Events.txt

This input file contains information about the diagnostic events of the patients in *Patient.txt*.

PatientID	Patient identifier
Date	Date of the event
EventType	Type of event (see below)
Code	Optional diagnostic code or free text

No missing values for these variables are allowed, except for Code.

The following events need to be extracted (see the Statistical Analysis Plan for details on mapping):

EventType	**Description**
DementiaAD	Dementia, Alzheimer
DementiaVascular	Dementia, vascular
DementiaOther	Dementia, other causes Note: Create this event for cases where another type of dementia (that is, not Alzheimer or vascular) is known, e.g., frontotemporal, Morbus Pick, Lewy bodies, Kreutzfeld, posterior cortical atrophy. The specific type does not have to be specified.
DementiaNOS	Dementia, Not Otherwise Specified Note: Create this event for cases where further sub-classification by type of dementia is not possible.

Example of events input file:

patientid,date,eventtype,code
1,20040601,DementiaAD,,

NOTE: include *all* events for *all* patients of your source population. The cut-offs required by the validation study will be performed by Jerboa (e.g., implementing the age limit of 75 years of age of older).

NOTE: Different event types may be specified per patient (e.g., DementiaNOS and DementiaAD), as well as multiple occurrences of the same event type (e.g., DementiaAD diagnosed at more than one date). Jerboa determines the final event type and occurrence.

NOTE: If a patient has been diagnosed with a mixed dementia (e.g., AD and vascular dementia), the event types corresponding with the constituent types of dementia (i.e., DementiaAD and DementiaVascular) have to be specified with the same date of diagnosis.

Clinical definitions:

We distinguish three types of dementia in this study: Alzheimer's disease dementia (DementiaAD), vascular dementia (DementiaVascular), and dementia due to other causes (DementiaOther). In addition, the diagnosis Dementia Not Otherwise Specified (DementiaNOS) refers to patients with dementia where the type of dementia is not known.

Alzheimer's dementia is the most common cause of dementia (for the different types of dementia, we used definitions provided by https://www.alzheimers.org.uk). The word dementia describes a set of symptoms that can include memory loss and difficulties with thinking, problem-solving or language. Alzheimer's disease, named after the doctor who first described it (Alois Alzheimer), is a physical disease that affects the brain. During the course of the disease, proteins build up in the brain to form structures called 'plaques' and 'tangles'. This leads to the loss of connections between nerve cells, and eventually to the death of nerve cells and loss of brain tissue. People with Alzheimer's also have a shortage of some important chemicals in their brain. These chemical messengers help to transmit signals around the brain. When there is a shortage of them, the signals are not transmitted as effectively.

Vascular dementia is the second most common type of dementia. The word dementia describes a set of symptoms that can include memory loss and difficulties with thinking, problem-solving or language. In vascular dementia, these symptoms occur when the brain is damaged because of problems with the supply of blood to the brain.

Dementia due to other causes includes the following diagnoses:
Dementia with Lewy bodies. Dementia with Lewy bodies (DLB) is a type of dementia that shares symptoms with both Alzheimer's disease and Parkinson's disease. It may account for 10-15 per cent of all cases of dementia. Lewy bodies are named after the German doctor who first identified them. They are tiny deposits of a protein (alpha-synuclein) that appear in nerve cells in the brain. Researchers don't have a full understanding of why Lewy bodies appear, or exactly how they contribute to dementia. Lewy bodies are the cause of DLB and Parkinson's disease. They are two of several diseases caused by Lewy bodies that affect the brain and nervous system and get worse over time. These are sometimes called Lewy body disorders. The way someone is affected by DLB will depend partly on where the Lewy bodies are in the brain. People with a Lewy body disorder can have problems with movement and changes in mental abilities at the same time.
Frontotemporal dementia. Frontotemporal dementia is one of the less common types of dementia. The term covers a wide range of different conditions. It is sometimes called Pick's disease or frontal lobe dementia. The word 'frontotemporal' refers to the lobes of the brain that are damaged in this type of dementia. The frontal lobes of the brain, found behind the forehead, deal with behaviour, problem-solving, planning and the control of emotions. An area of usually the left frontal lobe also controls speech. The temporal lobes – on either side of the brain – have several roles. The left temporal lobe usually deals with the meaning of words and the names of objects. Frontotemporal dementia occurs when nerve cells in the frontal and/or temporal lobes of the brain die, and the pathways that connect the lobes change. Some of the chemical messengers that transmit signals between nerve cells are also lost. Over time, as more and more nerve cells die, the brain tissue in the frontal and temporal lobes shrinks. When the frontal and/or temporal lobes are damaged in this way, this causes the symptoms of FTD. These include changes in personality and behaviour, and difficulties with language. These symptoms are different from the memory loss often associated with more common types of dementia, such as Alzheimer's disease.
Creutzfeldt-Jacob disease. Creutzfeldt-Jakob disease (CJD) is caused by an abnormally shaped protein called a prion infecting the brain. Sporadic CJD, which normally affects people over 40, is the most common form of the disease. It is estimated that the disease affects about one out of every 1 million people each year. It is not known what triggers sporadic CJD, but it is not known to be inherited or otherwise transmitted from person to person. A more recently identified form of CJD, called new variant CJD, was caused by eating meat from cattle infected with bovine spongiform encephalopathy (BSE). This typically affected younger adults. In new variant CJD, there may be many years between a person being infected and the development of symptoms. In sporadic CJD, the disease usually progresses within a few months. Early symptoms include minor lapses of memory, mood changes and

loss of interest. Within weeks the person may complain of clumsiness and feeling muddled, become unsteady walking, and have slow or slurred speech. Symptoms progress to jerky movements, shakiness, stiffness of limbs, incontinence and loss of the ability to move or speak. By this stage the person is unlikely to be aware of their surroundings or disabilities. People affected by CJD usually die within six months of their early symptoms developing. In a small number of patients the disease may take longer to run its course.

Posterior cortical atrophy. Posterior cortical atrophy (PCA), also known as Benson's syndrome, is a rare degenerative condition in which damage occurs at the back (posterior region) of the brain. In the vast majority of people, the cause of PCA is Alzheimer's disease. The first symptoms of PCA tend to occur when people are in their mid-50s or early 60s. However, the first signs are often subtle and so it may be some time before a formal diagnosis is made. Initially, people with PCA tend to have a relatively well-preserved memory but experience problems with their vision, such as difficulty recognising faces and objects in pictures. They may also have problems with literacy and numeracy. These tasks are controlled by the back part of the brain, where the initial damage in PCA occurs. As damage in the brain spreads and the disease progresses, people develop the more typical symptoms of Alzheimer's disease, such as memory loss and confusion. There are no specific medications for the treatment of PCA but some people find medications for Alzheimer's disease helpful.

Mixed dementia. A given patient can have more than one type of dementia – a mixed dementia. For patients who have been diagnosed with mixed dementia, the data custodians are asked to create the different dementia event types that constitute the mixed dementia, with the same date of diagnosis.

Measurements.txt

This input file contains information about measurements, vital signs and laboratory values of the patients in *Patient.txt*.

PatientID	Patient identifier
MeasurementType	Type of the measurement (see table below)
Date	Date of the measurement
Value	Value of the measurement

No missing values for these variables are allowed.

MeasurementType	**Description**	**Possible values**
MMSE	Mini-Mental State Examination	0-30

Example of measurements input file:

patientid,date,measurementtype,value
1001,20150202,MMSE,16

NOTE: Please include *all* measurements of *all* patients of your source population, but only if the possible values of the MMSE measurements range from 0 to 30. Exclude MMSE measurements if the maximum attainable value is less than 30, e.g., when not all areas of cognitive function that are part of the MMSE were tested.

How to run Jerboa?

Prerequisites

Jerboa requires the latest Java version in order to run. You can download it from here: http://www.java.com/en/download/manual.jsp/

Following this link, choose the appropriate Java version for your operating system (e.g., Windows, Mac OS X, Linux) and its type (e.g., 32 bits or 64 bits). To find out what type of operating system you are running, do the following:

- On Windows : right click on My Computer → Properties → see System Type
- On Linux: open a terminal (Ctrl + Alt + T) and type "getconf LONG_BIT"

Instructions on how to install Java can be found here as well, but if you need help please let us know. Possibly, you need the help of your local technical staff with administrator's rights to install new software on your machine.

Downloading the latest version of Jerboa and the script from Octopus

The database owners need to have access to Octopus to be able to download and upload files. If you do not have access yet, please ask for an application form by sending an email to rre@erasmusmc.nl.

The latest version of Jerboa and the script for the current run can always be found in Octopus using FileZilla. Instructions on how to use FileZilla can be found in the documentation and video sent to all Octopus users.

When you login using FileZilla you will see a folder named Jerboa-<Project Name>. In this folder you can find a zip file with Jerboa, the script (.jsf), and documentation.
Download and unzip the zip file into a folder and copy the script file into the folder containing your input files.

Running Jerboa

Double-click on the JerboaReloaded.jar file to start Jerboa. After accepting the license, you will see the screen in Figure 1.

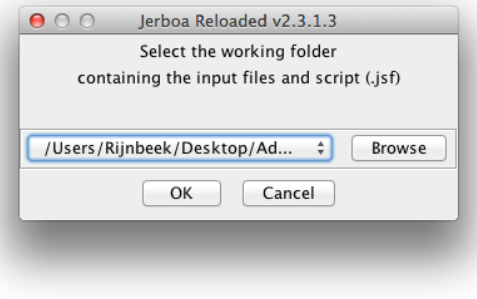


Figure 1. Opening a working folder

1. You can choose your working folder containing the input data file by clicking the browse button. In the first run of the Jerboa software, the folder where the JerboaReloaded.jar file is located is selected as default. If this folder corresponds to the location of your input file(s), just press OK.

Previously used workspaces are remembered and available to open by clicking the dropdown list at the left of the browse button.

Important: Make sure that the provided script file (e.g., script.jsf) is in your chosen working folder.

2. Once a working folder is selected press OK to continue. The screen in Figure 2 will appear. As long as the patient file contains all mandatory columns, it will automatically be loaded and recognized, as shown in the Patients file panel on the upper side of the screen. If no patient file is found this will be indicated. **Note** that multiple patient files in the same folder are not allowed for this run.

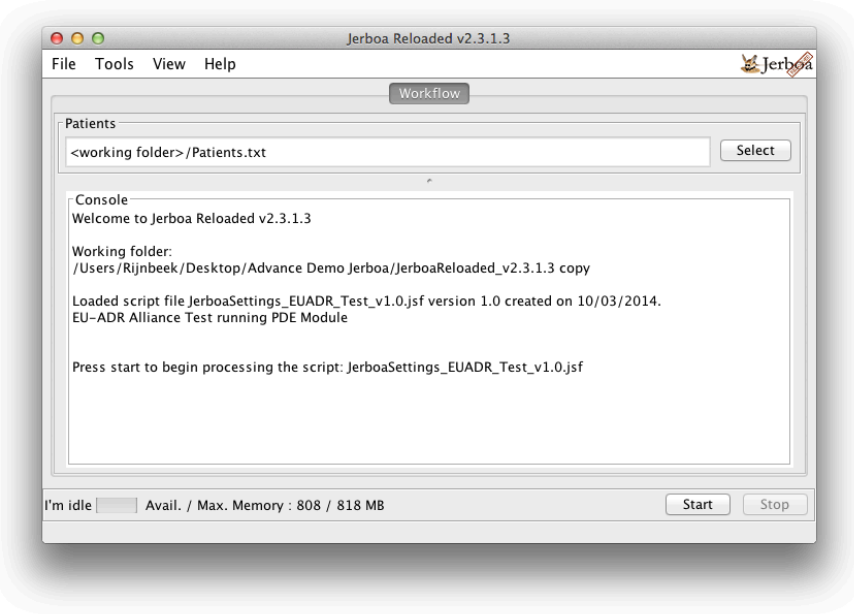


Figure 2. The application has successfully loaded the patients file

3. Now click the start button and select your database. If your database is not listed, you can add it using the add button.

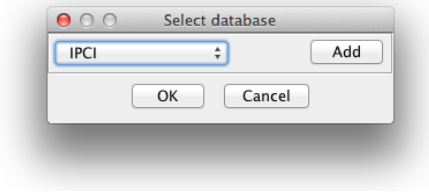


Figure 3. Selecting your database

4. The application will check the input files and will report any errors found.

4.a If errors are found in the input file(s), the user is informed as shown in Figure 4.

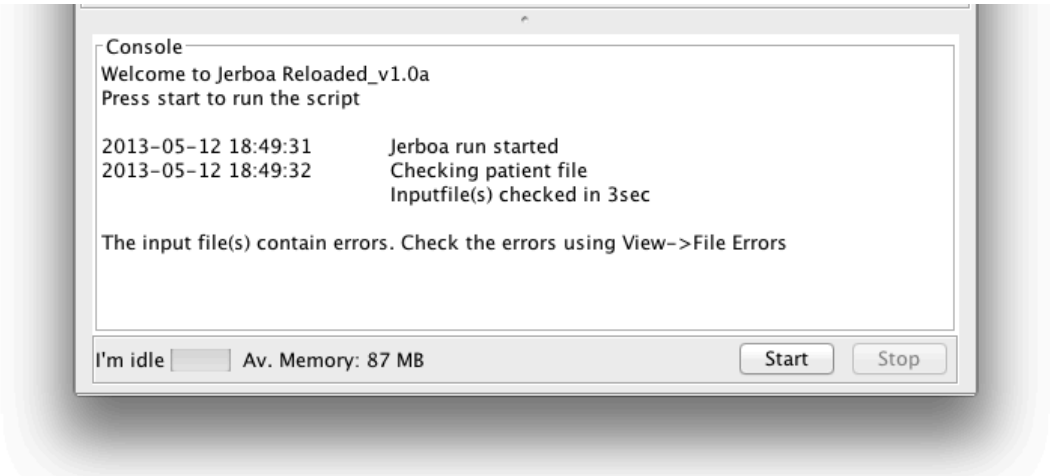


Figure 4. Errors in the input file(s)

The user can check the errors by clicking the « View » menu and selecting « File Errors ». The screen shown in Figure 5 will appear showing on the left side the error message and on the right side the actual content of the record in the input file. Alternatively, an error log file is generated for each input file. These files can be found in the « logs » folder of the current Jerboa run.

Important: In the working folder, a folder called « jerboa » is created. This folder contains all the files generated during each run of the Jerboa software. For each run, an individual folder is created inside the « jerboa » folder. The folder name is formed by the date of the run and the run number. This will allow you to keep a log of previous runs. Figure 6 shows an example of the folder structure created after a run.

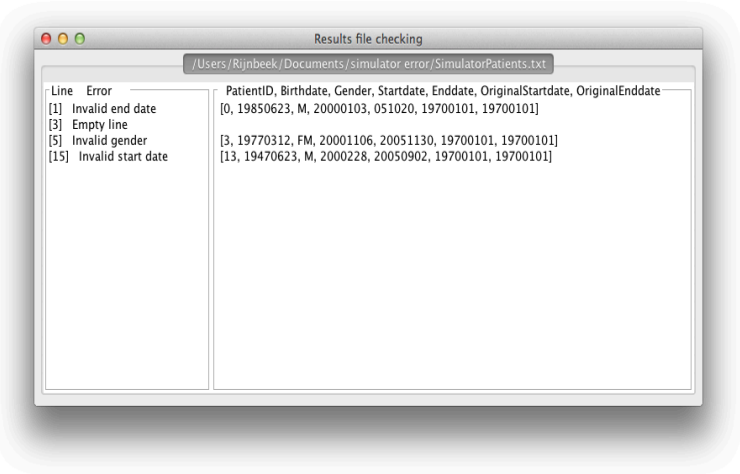


Figure 5. Results file checking

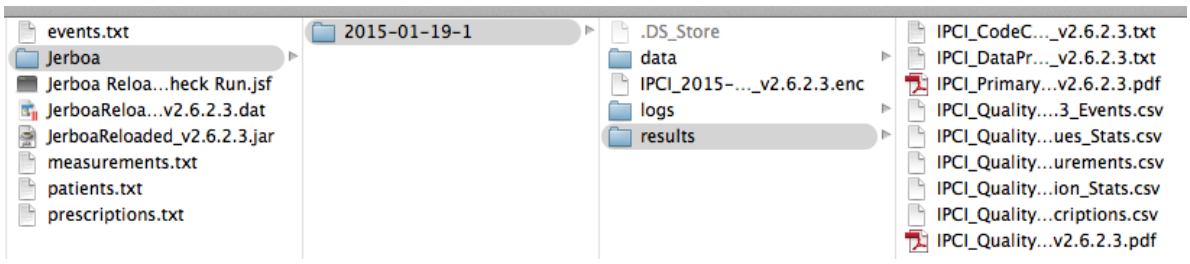


Figure 6: Folder structure created during Jerboa runs

4.b. If no errors are found in the input file(s) the application will proceed. An indication of the time left to finish the current step is given in the progress bar on the bottom of the screen.

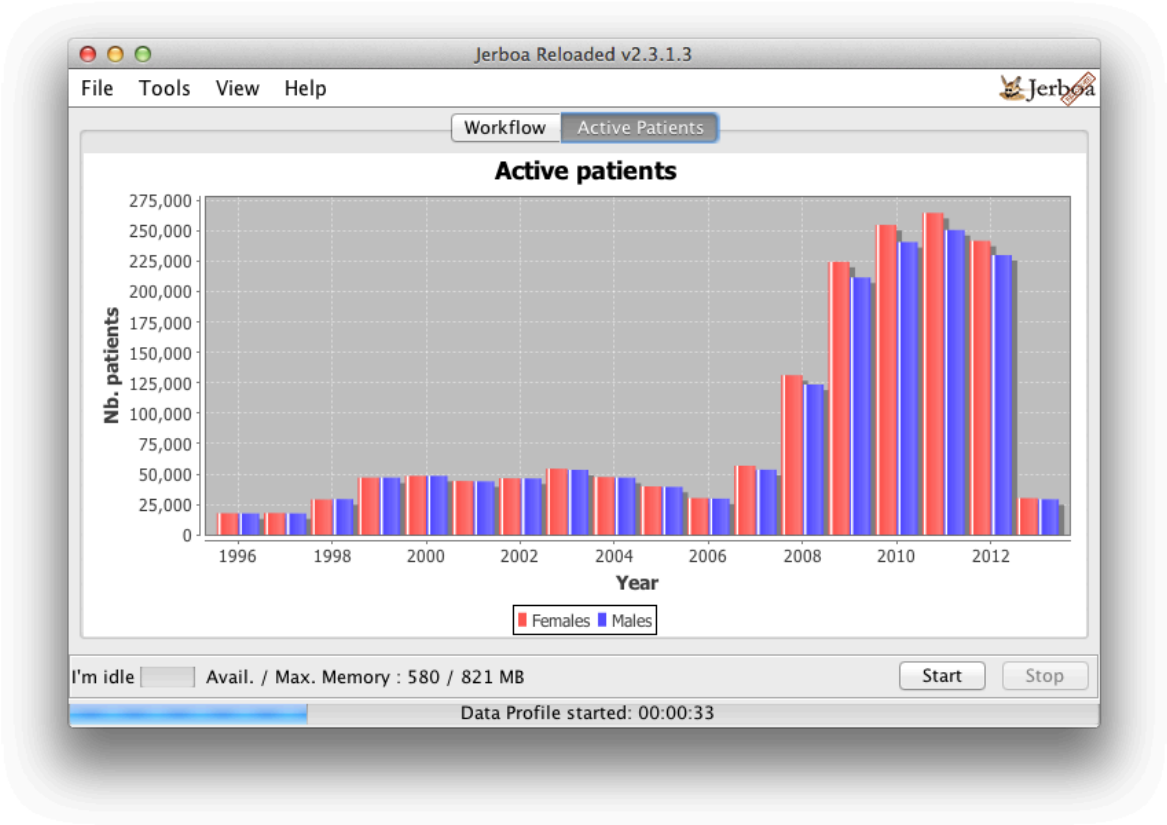


Figure 7. Input data checking was successful and processing the data following the script

During the run feedback is given in the form of a graph showing the active male and female patients in your database per year. For each newly generated graph a tab is created on the top of the window.

Primary Data Extraction Module (PDE)

The PDE extracts some basic information about your patient input file. For example, the number of active patients, births, start dates, etc.

In Figure 8 some examples are shown.

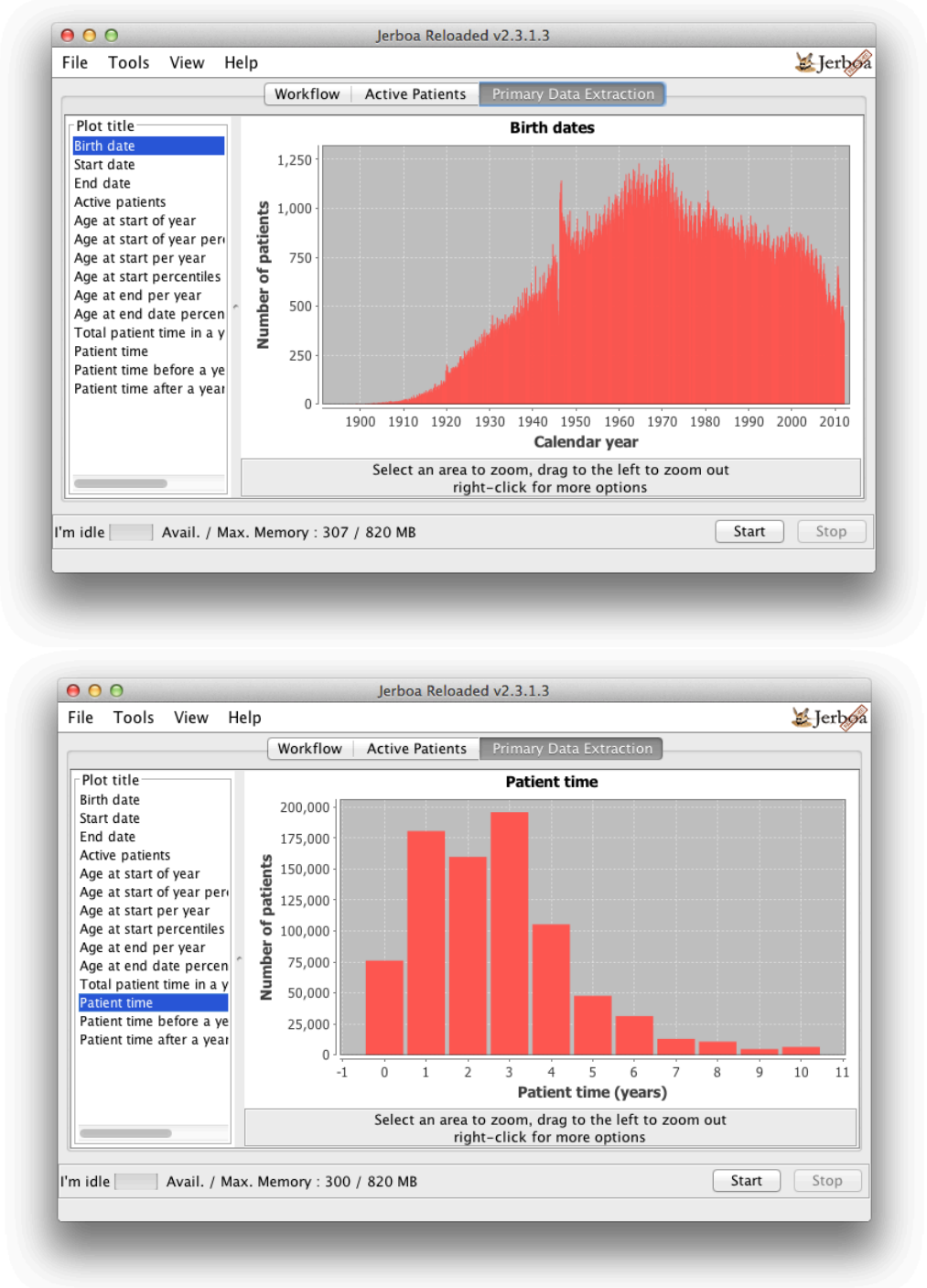


Figure 8. Examples on simulated data

Some of the graphs are generated for males and females seperately (use Next and Previous buttons). It is possible to zoom in by drawing a zoom window in the Graph. If you drag to the left the graph will zoom out to its original view. Right-click for more zooming options like zooming only one axis or print the graph. In the result folder a pdf is created containing all the plots.

Quality Control Module (QCM)

The QCM creates an overview of the events and their codes, and measurements in your input files. This allows you to double check that you are not missing items or see unexpected data. Please have a good look at these graphs!

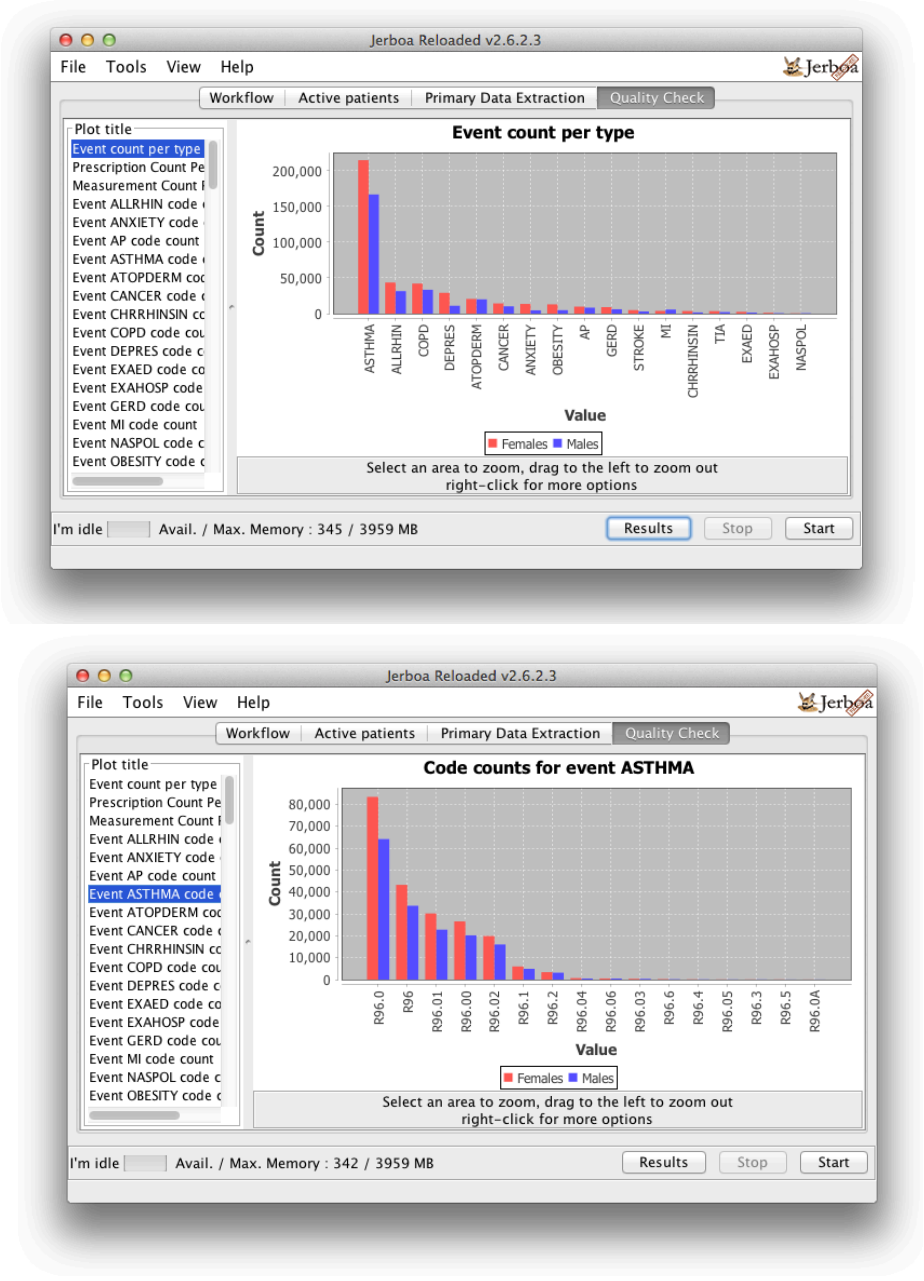


Figure 9. Example of Quality Control Module Graphs

On the left (see figure 9) you can select many parameters that might be of interest to you. This module also generates a pdf file in the result folder and a number of txt files that will be encrypted in the .enc file (see below) for sharing. The PI of the study will double check these files as quality control after uploading to our server.

5. In the final step Jerboa will produce an .enc file. This is an encrypted file containing the output files. The file is to be found in the folder of the current run (e.g., MyFolder/Data/jerboa/2013-05-09-03/). The location is also shown in the console (or click on Results). This file should be sent to EMC following the procedure described below.

Sending data to EMC with the use of Octopus

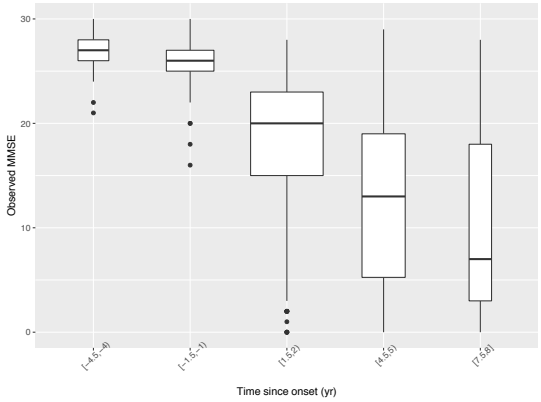
The .enc file should be uploaded to Octopus using the FileZilla procedure as described in the Octopus instructions. Please create a folder with the name of the project in your upload folder and upload the Jerboa output there.

Send an email to rre@erasmusmc.nl with subject "[RRE FTP] ROADMAP Val1 upload from database <database name>".

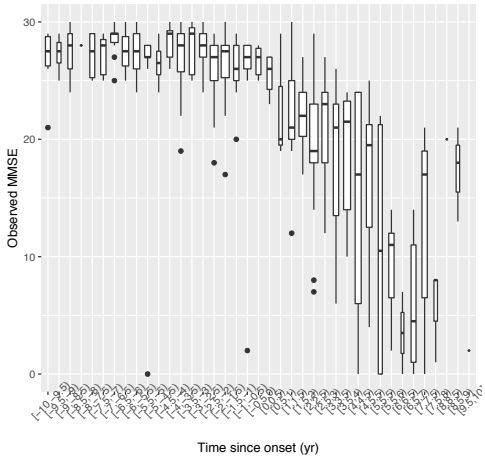
For any questions regarding Jerboa or Octopus, please use the same email address.

ANNEX V. Plots of observed and predicted MMSE values for all data sources

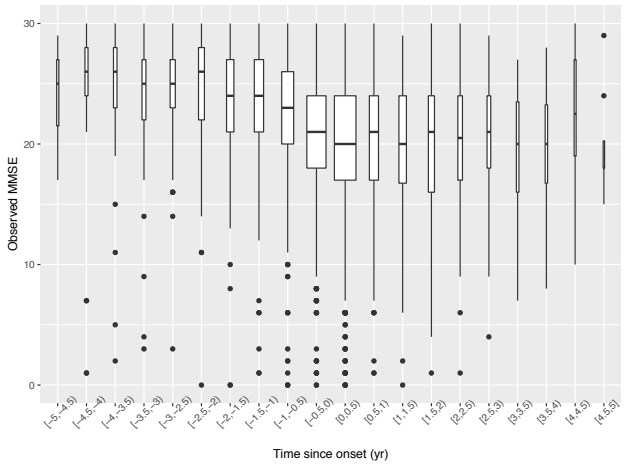
Observed MMSE as a function of time relative to the index date for the development set (Kungsholmen) and 11 validation sets.



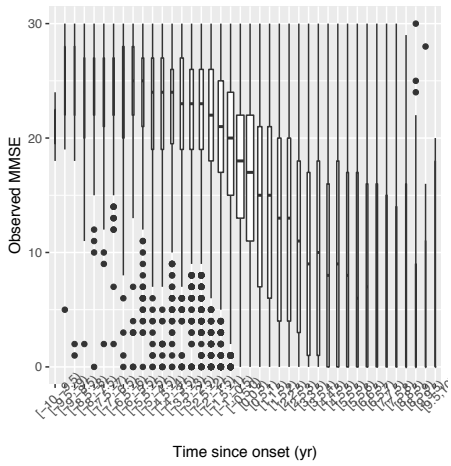
Kungsholmen



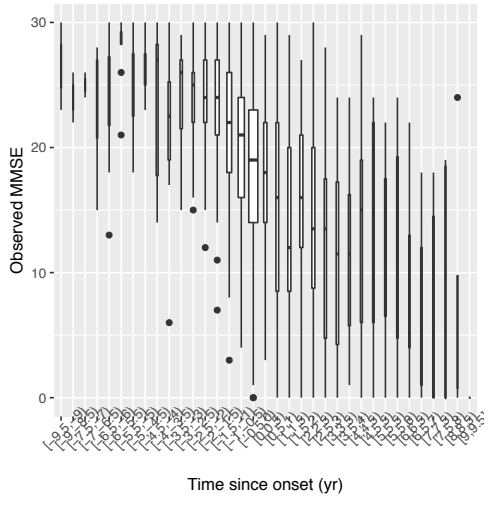
Gothenburg



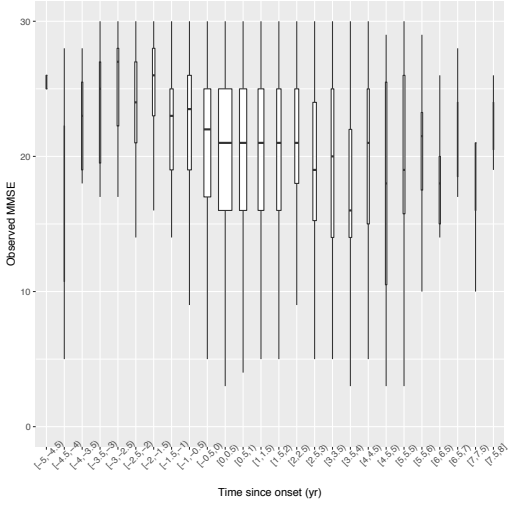
IPCI



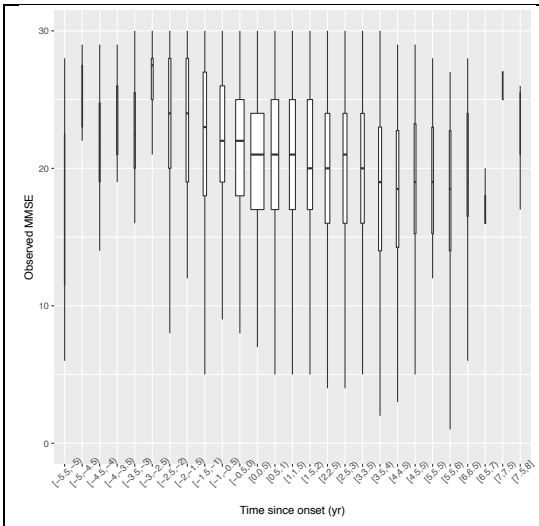
SIDIAP



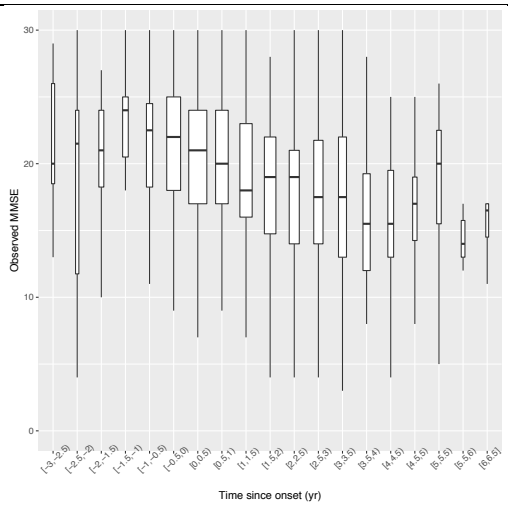
SIDIGI



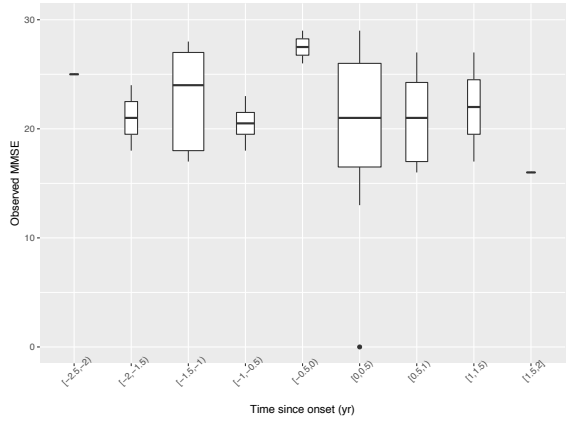
Copenhagen (0 drug)



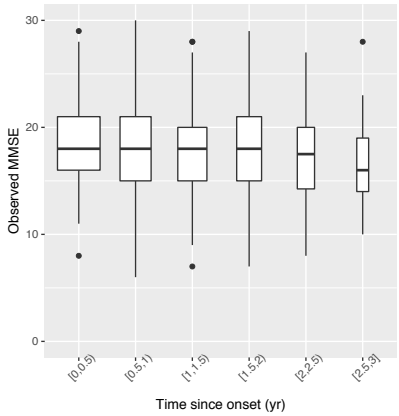
Copenhagen (1 drug)



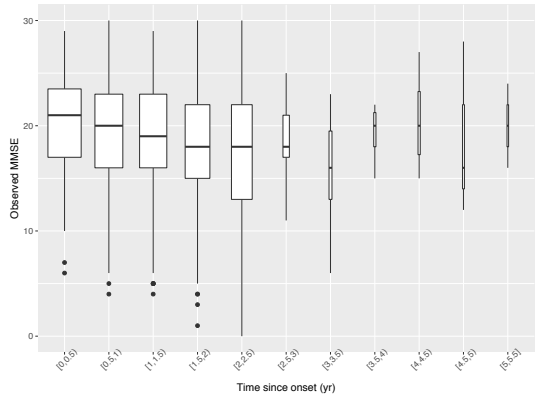
Copenhagen (>1 drug)



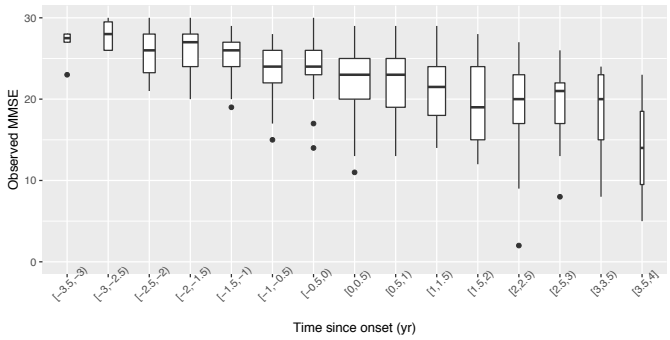
EDAR



Girona

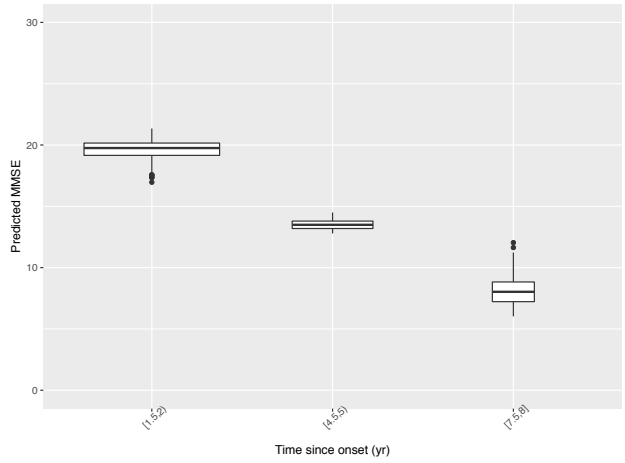


ICTUS

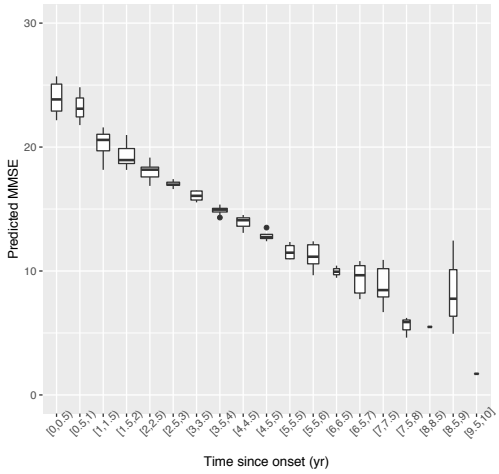


MEMENTO

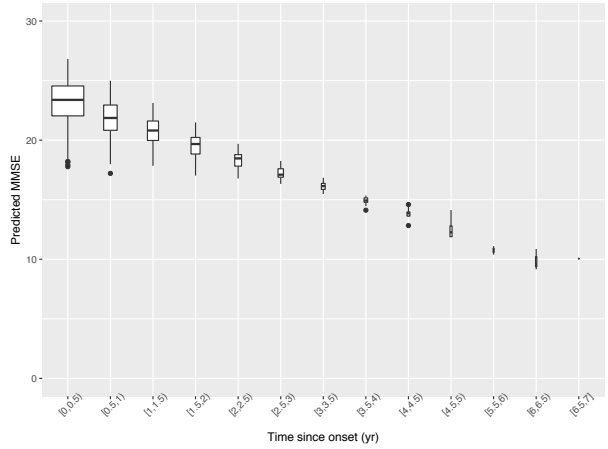
Predicted MMSE as a function of time relative to the index date for the development set (Kungsholmen) and 11 validation sets.



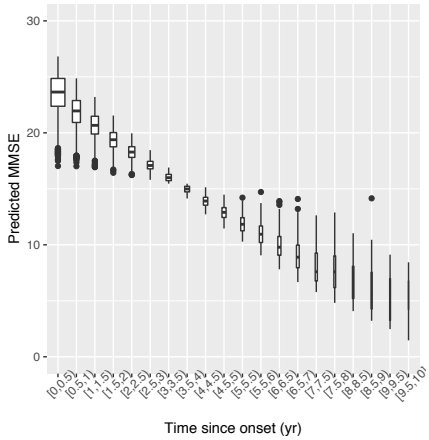
Kungsholmen



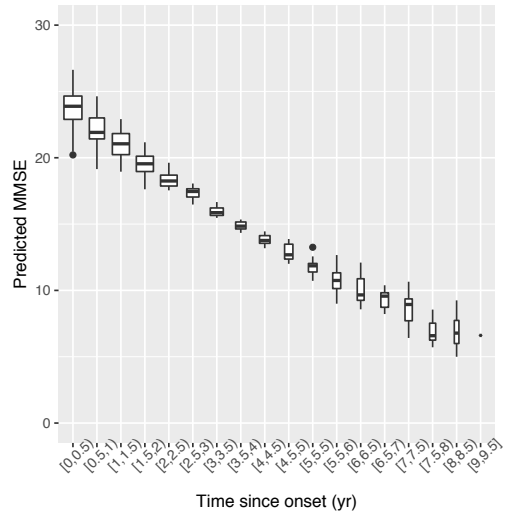
Gothenburg



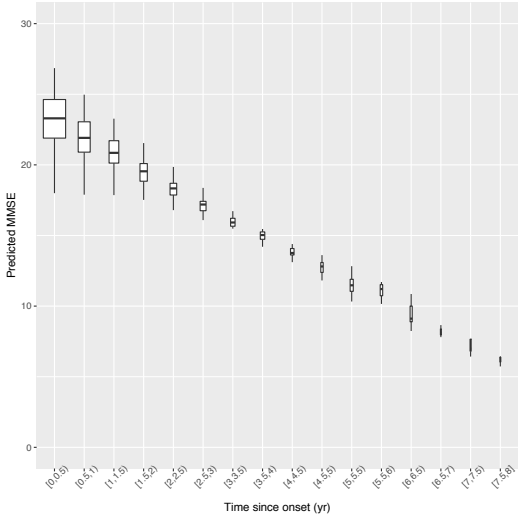
IPCI



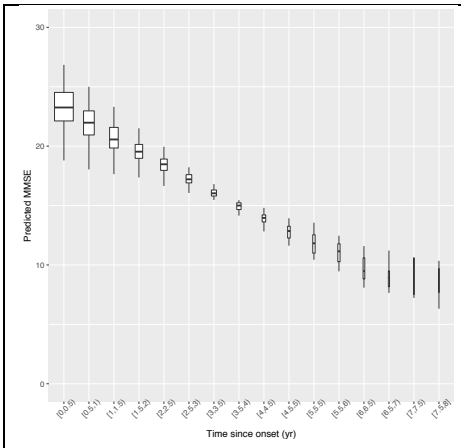
SIDIAP



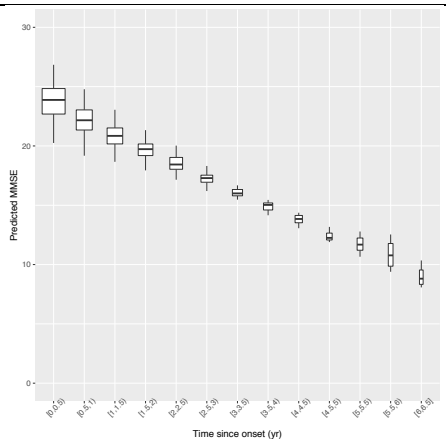
SIDIGI



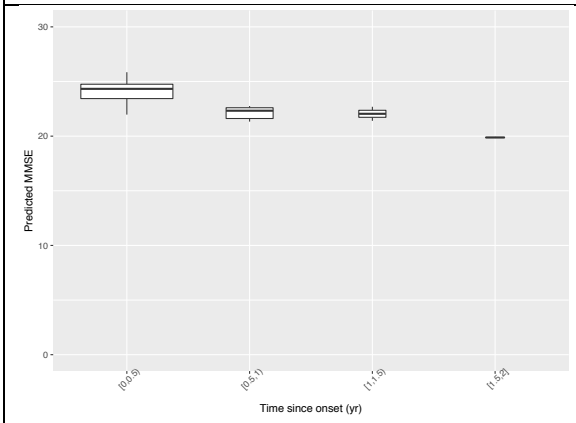
Copenhagen (0 drug)



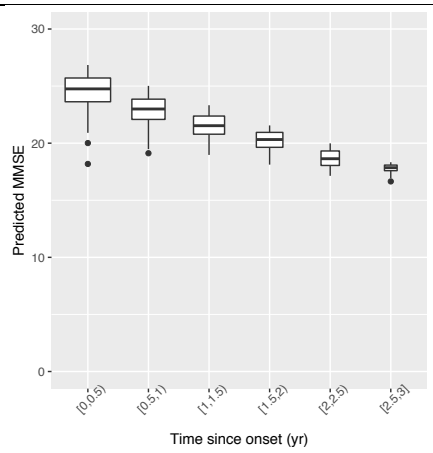
Copenhagen (1 drug)



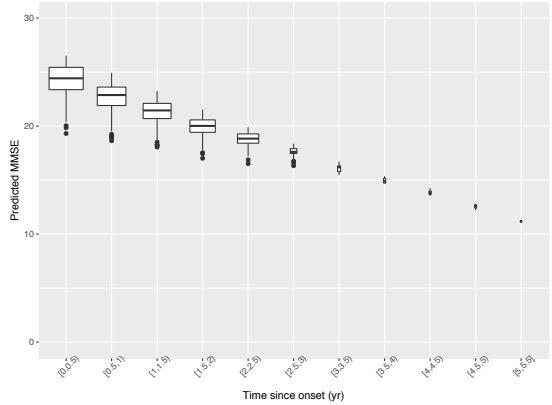
Copenhagen (>1 drug)



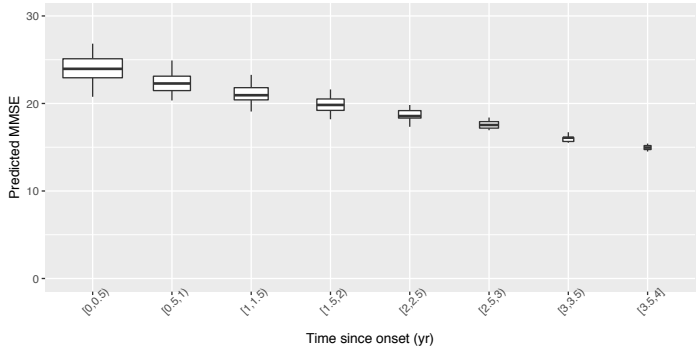
EDAR



Girona

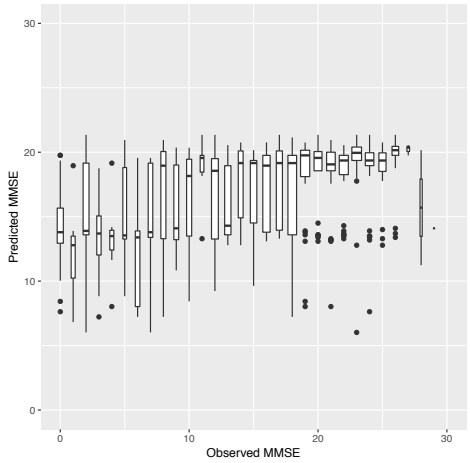


ICTUS

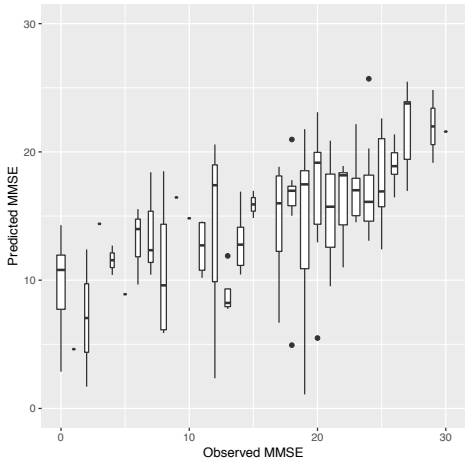


MEMENTO

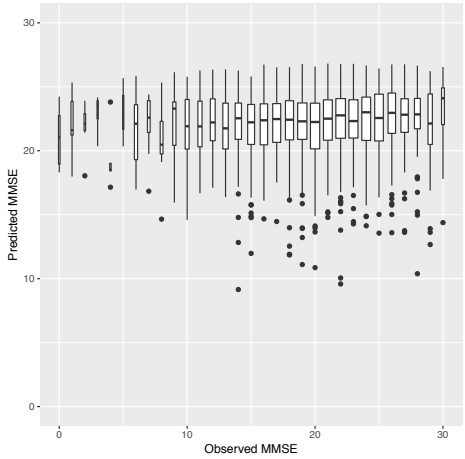
Observed versus predicted MMSE for the development set (Kungsholmen) and 11 validation sets.



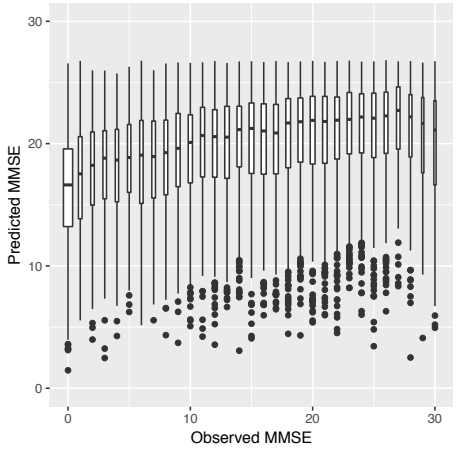
Kungsholmen



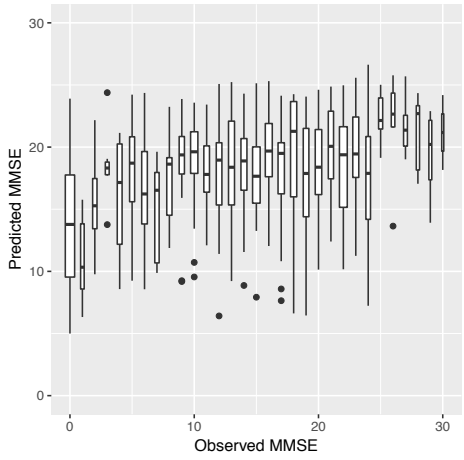
Gothenburg



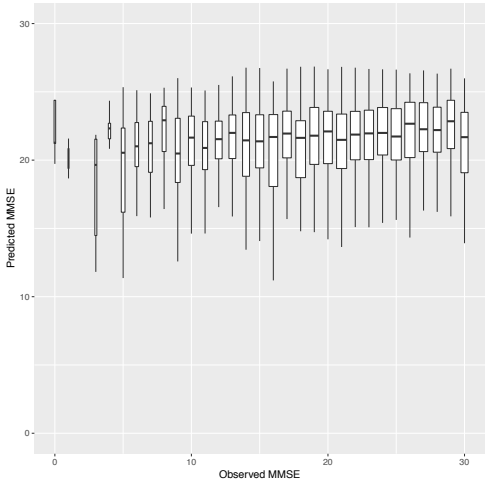
IPCI



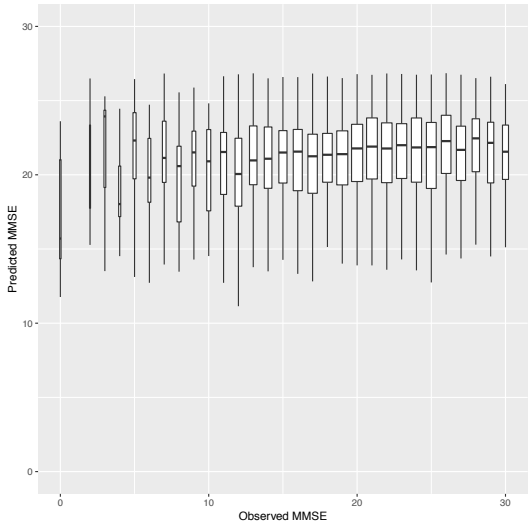
SIDIAP



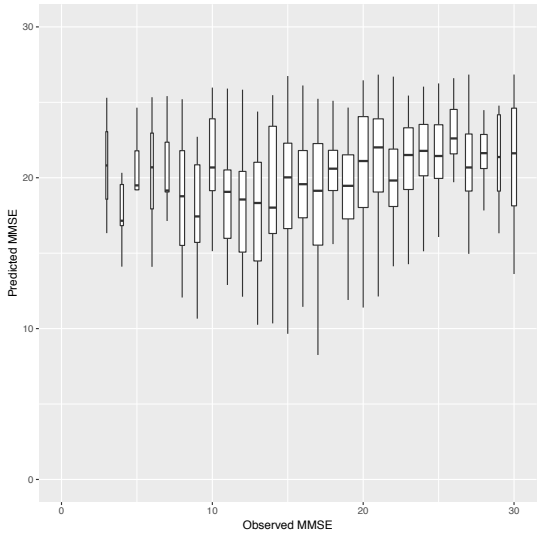
SIDIGI



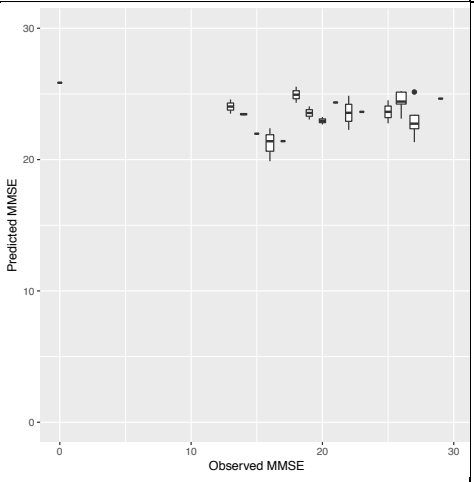
Copenhagen (0 drug)



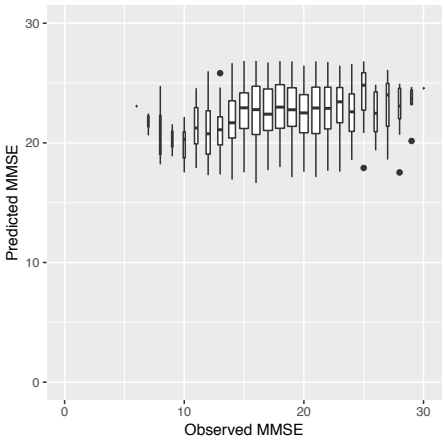
Copenhagen (1 drug)



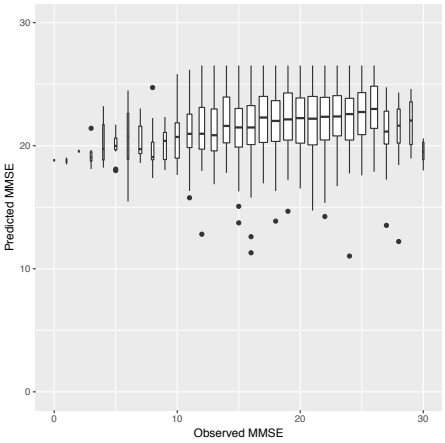
Copenhagen (>1 drug)



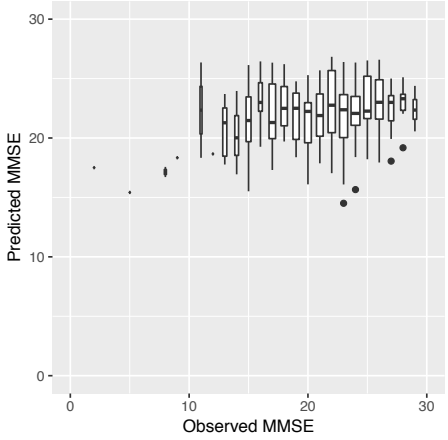
EDAR



Girona

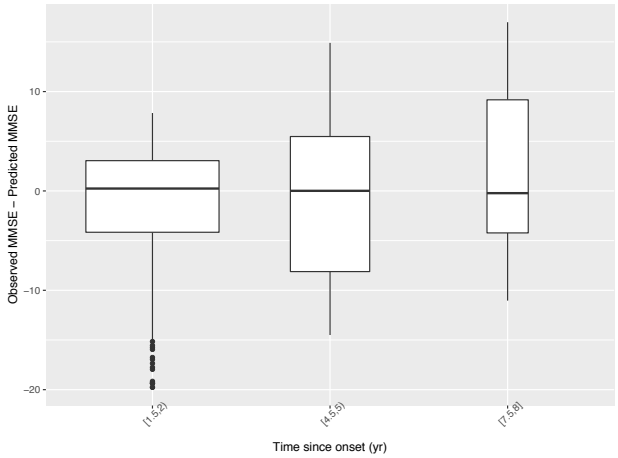


ICTUS

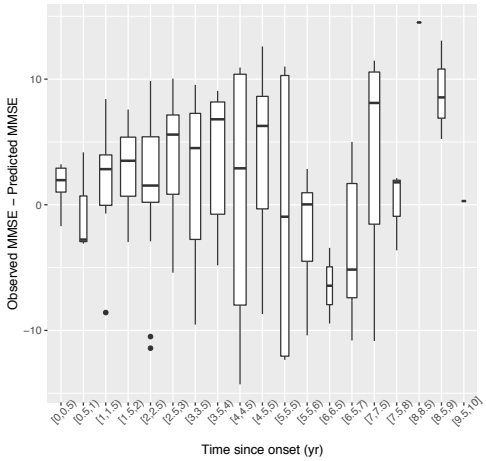


MEMENTO

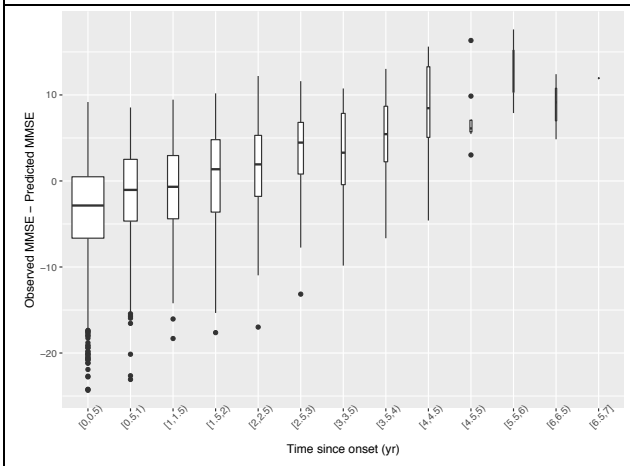
Difference between observed and predicted MMSE as a function of time relative to the index date for the development set (Kungsholmen) and 11 validation sets.



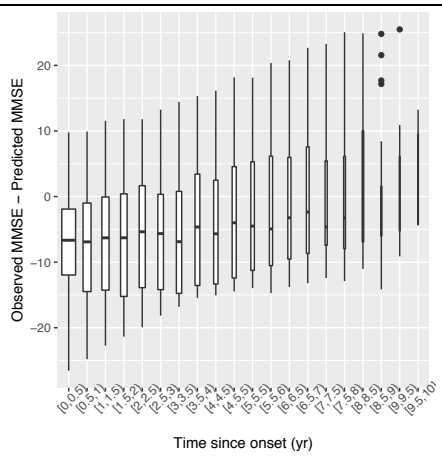
Kungsholmen



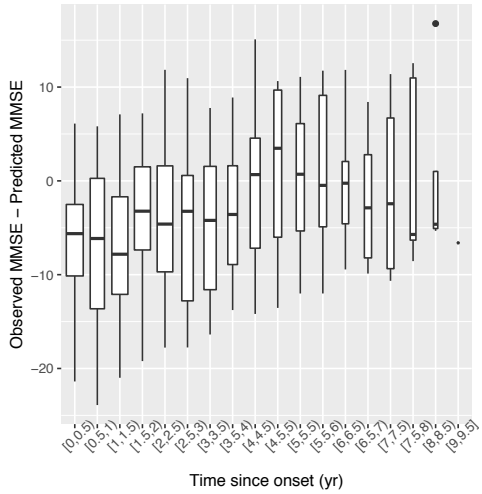
Gothenburg



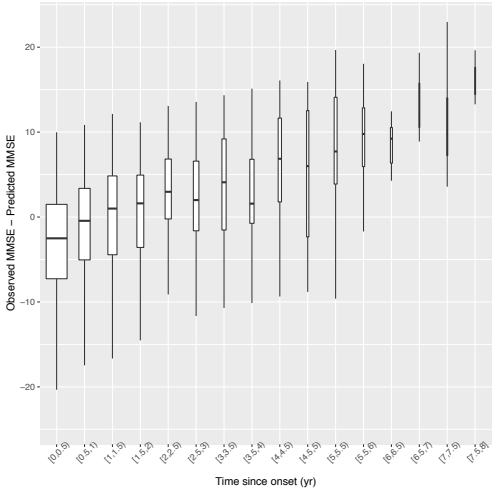
IPCI



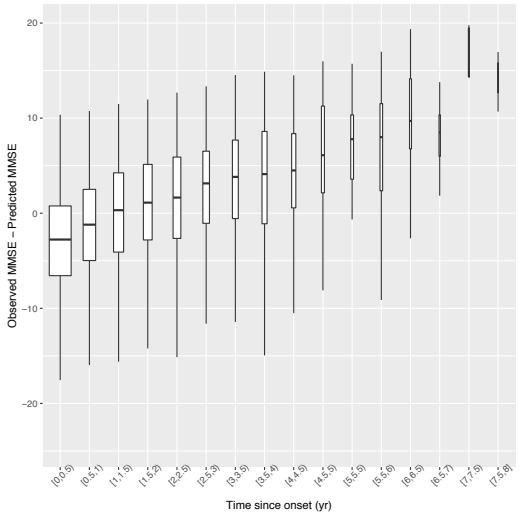
SIDIAP



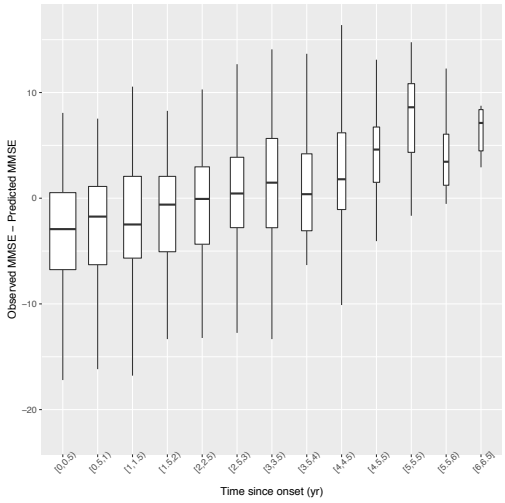
SIDIGI



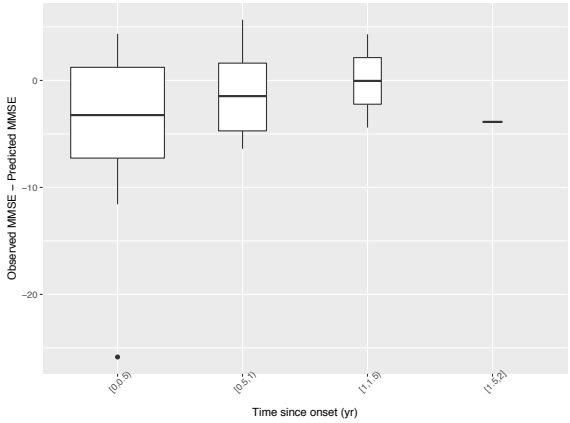
Copenhagen (0 drug)



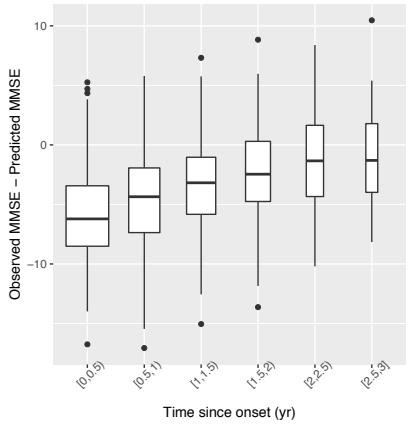
Copenhagen (1 drug)



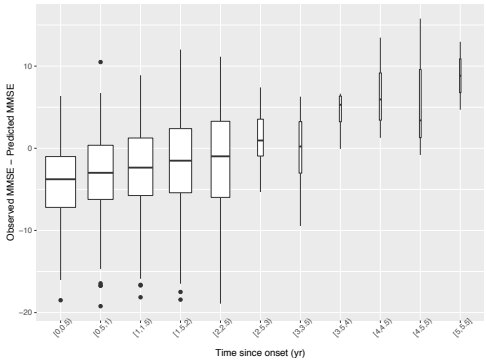
Copenhagen (>1 drug)



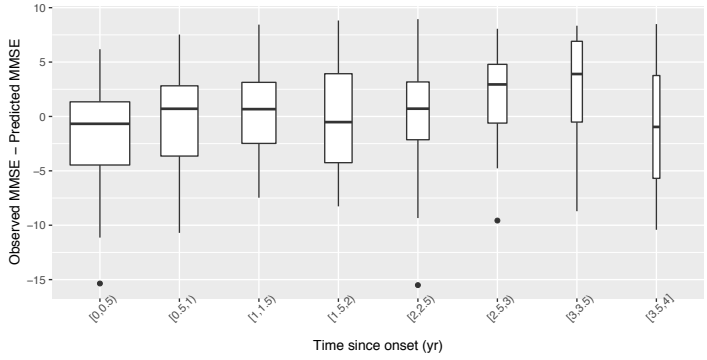
EDAR



Girona



ICTUS



MEMENTO