SimulAD: A dynamical model for personalized simulation and disease staging in Alzheimer's disease

Clément Abi Nader, Federica Ribaldi, Giovanni B Frisoni, Valentina Garibotto, Philippe Robert, Nicholas Ayache, Marco Lorenzi

SimulAD: A dynamical model for personalized simulation and disease staging in Alzheimer's disease

Clément Abi Nader *[1], Federica Ribaldi[2], Giovanni B. Frisoni[2], Valentina Garibotto[3,4], Philippe Robert[5], Nicholas Ayache[1] and Marco Lorenzi[1]

[1]Université Côte d'Azur, Inria Sophia Antipolis, Epione Research Project, France.
[2]Memory Clinic and LANVIE-Laboratory of Neuroimaging of Aging, Hospitals and University of Geneva, Geneva, Switzerland.
[3]Faculty of Medicine, Geneva University, Geneva, Switzerland.
[4]Nuclear Medicine, Geneva University, Geneva, Switzerland.
[5]Université Côte d'Azur, CoBTeK lab, MNC3 program, France.

Tuesday 16th November, 2021

Abstract

SimulAD is a computational disease progression model (DPM) originally developed on the ADNI database to simulate the evolution of clinical and imaging markers characteristic of AD, and to quantify the disease severity (DS) of a subject. In this work, we assessed the validity of this estimated DS, as well as the generalization of the DPM, by applying SimulAD on a new cohort from the Geneva Memory Center (GMC). The differences between the estimated DS of healthy, mild cognitive impairment and AD dementia groups were statistically significant (p-values < 0.05; $d \geq 0.8$). DS correlated with MMSE (ρ = -0.55), hippocampal atrophy (ρ = -0.62), glucose hypometabolism (ρ = -0.67), amyloid burden (ρ = 0.31) and tau deposition (ρ = 0.62) (p-values < 0.01). Based on the dynamics estimated on the ADNI cohort, we simulated a DPM for the subjects of the GMC cohort. The difference between the temporal evolution of similar biomarkers simulated on the ADNI and GMC cohorts remained below 10%. This study illustrates the robustness and good generalization of SimulAD, highlighting its potential for clinical and pharmaceutical studies.

***Keywords*—** Alzheimer's disease, Disease progression models, Clinical trials, Biomarkers

*Corresponding author at: Epione Research Project, INRIA Sophia-Antipolis, 2004, route des Lucioles, 06902 Sophia-Antipolis, France, clement.abi-nader@inria.fr.

1 Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder whose evolution has been hypothesized to follow a cascade of events (Jack, Knopman, et al. 2013). Deposition of the beta-amyloid protein in the brain cortex is believed to initiate this cascade, and to subsequently cause the aggregation of hyperphosphorylated tau protein in neurofibrillary tangles. This is followed by a process of neurodegeneration (i.e glucose hypometabolism and gray matter atrophy) ultimately leading to dementia. An inherent difficulty in diagnosing AD is that patients go through a long asymptomatic phase spanning approximately 10 to 20 years (Sperling et al. 2011) before showing clinical symptoms. To provide a biological assessment of the disease, AD has been recently defined as a pathology characterized by three main biomarkers categories, namely: amyloid, tau and neurodegeneration (Jack, Bennett, et al. 2018). These three biomarkers can be measured thanks to imaging techniques, such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), or in the case of amyloid and tau, also by lumbar puncture and blood collection. Monitoring these biomarkers is paramount in order to track the disease progression, and to potentially facilitate prevention or assessment of drug efficacy.

In the past years, the proliferation of studies collecting large amounts of biomarkers, combined with the growth of machine learning, fostered the development of computational models for automated AD diagnosis. For instance, many studies focused on the development of data-driven approaches for automatic assessment of clinical diagnosis (Davatzikos et al. 2009; Arbabshirani et al. 2017; Falahati et al. 2014). Based on the sole analysis of imaging-derived data, these methods showed that it is possible to automatically identify healthy controls, subjects with mild cognitive impairment, and patients suffering from AD dementia, some of them reporting results comparable to diagnosis rates obtained by expert physicians (Klöppel et al. 2008). However, most of these approaches have

been exclusively developed to solve a predictive task, and generally don't allow to understand the mechanisms relating the different biomarkers throughout AD evolution. As these mechanisms still remain partially unknown, different methods known as disease progression models were therefore introduced in order to estimate, in a data-driven fashion, the long-term progression of biomarkers (Jedynak et al. 2012). Due to the lack of an absolute time-line describing AD evolution, these models usually assume that the disease is characterized by monotonic changes, such that the modelled biomarkers steadily evolve from normal to pathological values. This assumption allows to reconstruct a time-line on which we can track the disease progression (Lorenzi et al. 2017). Moreover, these methods can be applied to a variety of data types, such as cortical and subcortical shapes (Marinescu et al. 2019) or volumetric images (Khanal, Lorenzi, et al. 2016; Khanal, Ayache, et al. 2017), thus offering a fine-grained spatial description of the changes affecting the brain. These models also allow to automatically assess the individual disease severity by comparing the clinical and imaging measurements of a given subject to the estimated disease progression. This latter capability of disease progression models is usually referred as disease staging. Ultimately, these approaches could potentially be used for identifying individuals at risk of cognitive decline, or for assessing drug efficacy in clinical trials.

Since these statistical models have been mostly developed on publicly available research datasets, such as the one provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI), their generalization to external cohorts from memory clinics still requires additional testing and validation (Mendelson et al. 2017). As clinical cohorts may fundamentally differ from the ADNI one, whether it be in terms of data acquisition or study population, automated diagnosis pipelines usually show a prominent decrease in performances. Regarding disease progression models, it is conceivable that biomarkers' trajectories estimated solely via the analysis of a single cohort may not be fully

representative of the natural disease course. This aspect would question the generalization of the progression model when tested on subjects from independent clinical cohorts. It is therefore essential to assess the generalization of this kind of models on independent datasets, with respect to their specific biases which can encompass a broad range of differences between cohorts such as data acquisition, missing data or data heterogeneity. These differences need to be addressed to finally deploy disease progression models in a practical clinical setting (Castro et al. 2020).

Recently, the event-based-model (EBM) (Fonteijn et al. 2012) and the Discriminative EBM (DEBM) (Venkatraghavan et al. 2019) underwent an extensive evaluation effort. These approaches model AD progression as a sequence of events representing the transition of a set of biomarkers from a normal to an abnormal state. Both EBM and DEBM have been applied on subjects from independent cohorts, providing accurate patients staging (Archetti et al. 2019). However, the kind of progression model estimated by these methods presents important limitations. First, both approaches are based on the simplistic assumption describing the pathological progression as a discrete sequence of biomarkers transitions from normal to abnormal states, which doesn't reflect the continuous nature of the changes affecting the brain during the disease. Second, they allow the analysis of summary measures only, such as regional brain uptake of grey matter density values, and thus don't enable the fine-grained quantification of the spatial patterns of changes associated with the disease. Third, while these two models inform us about the sequence of events characterizing AD, they don't provide insights about the dynamical interplay between biomarkers. Investigating such interactions would allow a deeper understanding of how the pathological processes at stake during the disease affect each other. Fourth, the EBM and the DEBM don't allow to simulate hypothetical scenarios of disease progression. Yet, such capability could be used to assess the effect of drug intervention on the disease evolution *in silico*, which could help planning and monitoring clinical trials.

To address these limitations, more refined approaches to disease progression modeling such as the Subtype and Stage Inference (SuStaIn) (Young et al. 2018) method have been proposed. This method reformulates the EBM and DEBM in order to model the disease progression over a continuous timeline, thus tackling the main limitation of the EBM and the DEBM which described the pathological progression as a sequence of discrete events. In addition, we note that SuStaIn allows to identify subtypes of disease progression, thus addressing the problem of AD heterogeneity which is one of the major challenges for the development of personalized disease progression models and their application in clinical practice for staging patients. Besides SuStaIn, many others disease progression models have been proposed to enable the fine-grained description of the pathological evolution in space and time (Marinescu et al. 2019; Koval et al. 2018; Bilgel et al. 2016). Within this context, SimulAD (Abi Nader et al. 2021) is a recent method allowing the analysis of clinical scores and multivariate imaging data extracted from MRI and PET scans to estimate a continuous spatio-temporal model of disease progression. Compared to the EBM, DEBM and SuStaIn, this approach offers a higher resolution for the imaging biomarkers, allowing to track the evolution of regional changes affecting the brain along a continuous temporal scale describing the disease course over 30 years. The trajectories estimated for clinical and imaging markers can subsequently be used as a reference to assess the individual disease severity, by locating subjects along the temporal scale describing the disease evolution. Moreover, SimulAD estimates the dynamical relationships between key biomarkers at stake during AD progression, namely: amyloid deposition, glucose hypometabolism, cerebral atrophy, cognitive and behavioural decline. Based on these relationships, the model can be applied to simulate the personalized evolution of any patient or group of patients only from the knowledge of their baseline clinical and imaging measurements. Finally, SimulAD enables to assess the impact of therapeutic intervention, such as anti-amyloid treatment, on cognitive

outcomes depending on the intervention time (Abi Nader et al. 2021).

While SimulAD was previously trained on the ADNI cohort, its generalization to independent datasets has not been evaluated. As it generally holds for statistical and machine learning models, one of the main difficulty for the generalization of SimulAD to independent cohorts lies in the challenge of accounting for potential missing data, heterogeneity due to different acquisition protocols or even data incompatibility between cohorts. Yet, evaluating the generalization of SimulAD to independent cohorts is essential to demonstrate the reliability of the dynamics allowing to personalize the models of disease progression, as well as the validity of the approach for providing accurate disease staging.

Within this context, we assess in this study the generalization capabilities of SimulAD. To this end, we test the robustness and reliability of this approach when applied to an independent dataset from a memory clinic, namely the Geneva Memory Center (GMC). This cohort includes patients with cognitive complaints, who underwent a baseline clinical and neuropsychological evaluation, MRI, amyloid-PET, (18)F-fluorodeoxyglucose-PET (FDG-) and (18)F-flortaucipir-PET (tau-) scans. Our evaluation procedure relies on three key aspects: (i) Development of a pre-processing pipeline allowing to apply SimulAD on the GMC cohort; (ii) Assessment of SimulAD validity for individual disease staging. (iii) Evaluation of the reliability of the progression of imaging and clinical markers estimated by SimulAD.

2 Material and methods

In this work subjects were divided in five clinical groups: cognitively healthy (NL stable), individuals diagnosed with mild cognitive impairment (MCI stable), patient suffering from Alzheimer's disease

dementia (AD dementia), subjects progressing from NL to MCI or AD dementia (NL converters), and finally subjects progressing from MCI to AD dementia (MCI converters).

2.1 Experimental cohorts

The ADNI and GMC cohorts respectively included 442 and 93 subjects. In Table 1, we provide socio-demographic information across clinical groups for these datasets. The clinical spectrum of the cohorts spans a broad range of cognitive severity, from healthy to moderate and severe dementia. Conversion to AD dementia was determined using the last available follow-up information. In both the ADNI and GMC cohorts, MCI converters are subjects who were diagnosed as MCI at baseline and subsequently progressed to AD. In the case of the ADNI database, diagnosis was established using the DX column from the ADNIMERGE file (https://adni.bitbucket.io/index.html), which reflects the standard ADNI clinical assessment based on Wechsler Memory Scale, Mini-Mental State Examination, and Clinical Dementia Rating. Concerning the subjects from the GMC cohort, diagnosis was established after that the patients underwent neuro-psychological tests under the supervision of a psychologist. MCI converters from the ADNI database had a mean follow-up of 9 years with a standard deviation of 1.5 years, while the ones from the GMC cohort were followed in average 2.4 years with a standard deviation of 1.7. All the participants were amyloid positive at baseline. In the case of the ADNI cohort, "amyloid positive" subjects are patients whose amyloid level in the CSF was below the nominal cutoff of 192 pg/ml (Gamberger et al. 2017) either at baseline or during any follow-up visit. Concerning the GMC cohort, "amyloid positivity" was evaluated using visual assessment performed by an expert nuclear medicine physician and following the tracer manufacturers guidelines. Concerning the ADNI cohort, multi-modal biomarkers consisting of neuropsychological tests and measures derived from MRI, FDG-PET, and Amyloid-PET scans

were collected at baseline and during follow-up visits. In the case of the GMC cohort, each participant underwent the Mini-Mental State Examination (MMSE), as well as an MRI, FDG-PET and amyloid-PET scan at baseline. In addition, a tau-PET scan was acquired for 50 subjects from this cohort at baseline. During subsequent visits, only the MMSE was assessed. Summary statistics about clinical and imaging-derived information across clinical groups are reported in Table 1.

2.2 Image preprocessing

We derived volumes of gray matter density in a standard anatomical space by relying on the longitudinal pipeline of Freesurfer (Reuter et al. 2012). Regional gray matter density was extracted from the Desikan-Killiany parcellation (Desikan et al. 2006). amyloid-PET, FDG-PET and tau-PET images were registered to their corresponding T1-MRI acquisition, and normalized to the cerebellum uptake. Regional amyloid load, glucose metabolism and tau burden were computed thanks to the PetSurfer software (Greve et al. 2014). For every imaging modality we discarded white-matter, ventricular, and cerebellar regions, thus obtaining 82 regions that were averaged across hemispheres.

2.3 Modeling framework

SimulAD is based on the hypothesis that AD evolution can be mathematically modelled by a set of key biomarkers following a dynamical system. These biomarkers are namely *clinical scores*, *gray matter atrophy*, *amyloid load* and *glucose metabolism*. This assumption has two consequences: the first one is that, at any given time, AD severity is uniquely associated with the values of these biomarkers. The second one is that past and futures states of the disease can be computed from the current ones thanks to mathematical relationships linking the biomarkers evolutions.

Table 1: Baseline characteristics of the ADNI and GMC cohorts. Average values, standard deviation in parenthesis. Acronyms: ADNI: Alzheimer's Disease Neuroimaging Initiative; GMC: Geneva Memory Center; NL: cognitively healthy; MCI: mild cognitive impairment; AD: Alzheimer's dementia; FDG: (18)F-fluorodeoxyglucose Positron Emission Tomography (PET) imaging; SUVR: Standardized Uptake Value Ratio; MMSE: Mini Mental State Examination; Tau: (18)F-flortaucipir PET imaging. Converters are NL patients progressing to MCI or AD dementia, or MCI individuals progressing to AD dementia. Hippocampal volume: extracted with Freesurfer. Amyloid burden: voxel-number weighted average of the amyloid uptake in the frontal, anterior/posterior cingulate, lateral parietal, and lateral temporal regions normalized to the cerebellum. Early amyloid: voxel-number weighted average of the uptake extracted from the early-phase (6 min) of amyloid-PET in the frontal, anterior/posterior cingulate, lateral parietal, and lateral temporal regions normalized to the cerebellum. Glucose metabolism: voxel-number weighted average of the FDG uptake in the angular, temporal, and posterior cingulate cortex normalized to the cerebellum. Tau burden: voxel-number weighted average of the tau uptake in the entorhinal, amygdala, parahippocampal, fusiform, inferior temporal, and middle temporal regions normalized to the cerebellum. Amyloid-corrected and FDG-corrected indicate the values obtained after performing the data adjustment presented in Section 2.6. Missing data for 36^a and 43^b subjects.

Group	NL stable		MCI stable		MCI converters		AD dementia	
Cohort	GMC	ADNI	GMC	ADNI	GMC	ADNI	GMC	ADNI
N	23	71	28	131	25	105	17	102
Female (%)	61	62	61	37	65	47	53	45
Age (years)	69.1 (7.5)	73.7(6.0)	74.3 (6.5)	72.0 (7.6)	73.8 (4.6)	72.6 (6.7)	70.4 (11.1)	73.6 (8.2)
Education (years)	17.3 (3.9)	16.3 (2.5)	14.0 (3.1)	16.3 (2.7)	12.5 (4.7)	16.2 (2.6)	11.6 (3.8)	15.6 (2.5)
MMSE	28.5 (1.0)	29.2 (1.1)	25.2 (3.1)	28.1 (1.8)	23.7 (4.3)	26.4 (2.7)	18.2 (6.5)	23.0 (2.1)
Hippocampus (mm^3)	4271 (435)	4047 (455)	3621 (534)	3907 (541)	3634 (448)	3517 (515)	3409 (436)	3401 (472)
Amyloid (SUVR)	0.67 (0.13)	0.74 (0.17)	0.96 (0.17)	0.79 (0.18)	0.93 (0.14)	0.92 (0.17)	0.89 (0.12)	0.92 (0.16)
Amyloid-corrected (SUVR)	0.70 (0.12)	/	0.93 (0.13)	/	0.91 (0.13)	/	0.89 (0.11)	/
Early amyloid (SUVR)	0.53 (0.03)	/	0.50 (0.03)	/	0.47 (0.02)	/	0.45 (0.03)	/
FDG^a (SUVR)	0.61 (0.05)	0.62 (0.04)	0.52 (0.04)	0.61 (0.06)	0.52 (0.04)	0.56 (0.06)	0.49 (0.06)	0.52 (0.05)
FDG-corrected SUVR	0.57 (0.03)	/	0.54 (0.04)	/	0.53 (0.04)	/	0.50 (0.05)	/
Tau^b (SUVR)	0.60 (0.06)	/	0.81 (0.20)	/	0.94 (0.27)	/	1.21 (0.34)	/

To estimate the complex relationships between high-dimensional imaging and clinical measures, the model first transforms baseline neuropsychological assessments and measures derived from MRI, amyloid-PET and FDG-PET data in a set of four corresponding z-scores. The transformation consists in a weighted average of the measurements derived from each type of data modality (i.e regional grey matter measurements in the case of *atrophy*). The obtained z-scores are respectively denoted z^{cli}, z^{atr}, z^{amy} z^{met}, and describe the overall pathological status of an individual. We hypothesize that these four z-scores are related by a set of relationships driving the disease progression, which are mathematically modelled by a system of Ordinary Differential Equations (ODEs). This system of ODEs provides us with an interaction rule that describes how the z-scores jointly evolve over time. The parameters controlling the system of ODEs are optimized such that the predicted evolution of the z-scores best matches the available follow-up clinical and imaging measurements of each individual. An overview of the model is provided in Supplementary Section 1. Further details about model optimization and mathematical transformations allowing to map the z-scores and raw measures are given in (Abi Nader et al. 2021).

Trajectory modelling. Thanks to this mathematical formulation, SimulAD can be used to simulate the progression of changes characterizing AD by considering the subjects diagnosed with AD dementia, and for whom we compute corresponding z-scores based on their baseline measures. Relying on the estimated set of relationships between z-scores we follow their evolution forward and backward in time, thus simulating the subjects' evolution from their original healthy condition to their current pathological state. We obtain z-scores trajectories summarizing the overall progression of AD, and from which we can estimate the long-term evolution of the corresponding clinical and imaging measurements.

Disease severity quantification. Relying on the reference trajectory estimated for the four z-scores summarizing AD evolution, we can subsequently perform individual disease staging. Based on the multi-modal imaging and clinical data of a given subject collected at any visit, we compute z-scores for each type of marker, and find the time-point τ jointly minimizing the distance between the individual z-scores and the reference trajectory. In the rest of the paper, we will refer to this time-point as the disease severity. The estimated disease severity τ locates a subject on the reference trajectory, thus quantifying its pathological state. It is also important to note that the disease severity can still be estimated even in the case of missing data, by only computing the z-scores of the available measures of the observed subject. We provide mathematical details on the disease severity estimation in Supplementary Section 1.

2.4 Estimated model

The parameters of the resulting model, presented in (Abi Nader et al. 2021), were estimated through the analysis of multi-modal longitudinal data from the ADNI cohort. The clinical scores consisted in the Clinical Dementia Rating Scale Sum of Boxes (CDRSB), Alzheimer's Disease Assessment Scale (ADAS11), Functional Assessment Questionaire (FAQ), Rey Auditory Verbal Learning Test (RAVLT) learning, RAVLT immediate, RAVLT forgetting and MMSE. Regional gray matter density, amyloid load and glucose metabolism were derived following the procedure detailed in Section 2.2. No tau-PET data was included in the model. Baseline socio-demographic information and summary statistics about clinical and imaging data for the subjects from the ADNI cohort are provided in Table 1. The disease progression previously estimated by SimulAD on the ADNI database is illustrated in Supplementary Figure 2 in which we show the evolution of the z-scores and their associated imaging and clinical measures.

2.5 Evaluation strategy

We considered the model of evolution estimated by SimulAD on the ADNI cohort as the reference progression for AD. We evaluated SimulAD by conducting a series of experiments on both ADNI and GMC cohorts which aimed at demonstrating respectively the known-groups validity, the concurrent validity and the reliability of the model.

Known-groups validity. We evaluated how the estimated disease severity discriminates subjects across clinical groups within each cohort. It is expected that values of disease severity should increase along with the severity of the clinical status. We further assessed the group-wise consistency of the disease severity distribution, by comparing its values for similar clinical groups between ADNI and GMC cohorts. Differences between groups were assessed using Student's t-test and Cohen's d effect size.

Concurrent validity. We assessed the extent to which the estimated disease severity correlates with validated clinical and imaging assessments. For each subject MMSE score was available and imaging-biomarkers assessment as follows: hippocampal atrophy evaluated with Freesurfer, glucose metabolism and amyloid burden computed by extracting standardized uptake value ratio in a composite mask of regions of interest (MetaROI approach (Jagust et al. 2009; Landau, Harvey, et al. 2010; Landau, Mintun, et al. 2012)). In the case of GMC, we also had 50 subjects who underwent a tau-PET scan. We compared their estimated disease severity with respect to their tau burden computed in a composite mask of relevant regions (MetaROI approach (Jack, Wiste, et al. 2017)). Correlation between the estimated disease severity and the different variables was assessed using Spearman rank correlation (ρ).

Reliability. The two previous experiments aimed to quantify the validity of SimulAD in terms of disease staging based on the reference progression previously simulated on the ADNI database. We verified the consistency and robustness of the dynamics estimated by SimulAD by simulating the evolution of clinical and imaging-derived markers based on the GMC data. To this end, we applied the procedure described in Section 2.3 on the AD dementia subjects from the GMC cohort, thus providing us with a new model of progression for clinical and imaging measurements, as well as specific z-scores trajectories personalized to the GMC cohort. We compared the disease progression models obtained on ADNI and GMC cohorts by computing the average error between their z-scores trajectories over time. In addition to comparing the disease progression models obtained on both cohorts, we proposed to demonstrate the reliability of SimulAD by evaluating its prediction abilities at the individual level. To do so, we simulated the evolution of the MMSE for the subjects from the GMC cohort and compared the estimated values with the MMSE assessed by the physicians during the follow-up clinical visits.

2.6 Data adjustment

In order to implement the aforementioned assessment strategy, a number of additional pre-processing steps had to be carried out.

Missing measures imputation. We recall that SimulAD relies on 7 neuropsychological tests (CDRSB, MMSE, ADAS11, FAQ, RAVLT learning, RAVLT immediate and RAVLT forgetting) to compute the score z^{cli}, and that the only common clinical test between the ADNI and GMC cohorts is the MMSE. However, relying only on the MMSE would bias the computation of z^{cli} for the subjects from the GMC cohort. To overcome this issue, we imputed the 6 missing clinical

scores for all the subjects from the GMC cohort. Imputation was carried out through k-neighbors regressions trained on the ADNI database to predict each clinical score based on the MMSE and the measures of regional grey matter volume. The trained models were subsequently applied on the GMC cohort to estimate the associated clinical scores. We performed a 10-fold cross validation on the ADNI database to evaluate the prediction performances of the models. We show in Table 2 the average and the 95% confidence interval of the relative error between the ground truth and predicted score. The average relative error remains below 10% for the CDRSB, ADAS11, RAVLT immediate and FAQ, while not exceeding 20% for the RAVLT learning and RAVLT forgetting. We provide additional information about the estimation error in Supplementary Section 3.1.

FDG data harmonization. We can observe in Table 1 that in the GMC cohort 36 subjects out of 93 are missing an FDG-PET scan, thus preventing the computation of their score z^{met}. For these subjects their disease severity can therefore be estimated only based on three z-scores (z^{cli}, z^{atr}, z^{cli}), leading to potential bias and mis-estimation compared to the rest of the cohort. To prevent this issue we computed the regional FDG uptake for every subjects of the GMC cohort based on their corresponding early-phase (6 min) of amyloid-PET scan (Daerr et al. 2017). This was done by fitting a linear regression between the early-amyloid and FDG uptake of all the subjects for each brain region. Figure 1 illustrates the linear fit between early-amyloid and FDG regional uptake for three brain regions. We show similar relationships for additional brain regions in Supplementary Section 3.2.

Tracer bias correction. The amyloid-PET scan of 17 subjects from the GMC cohort was acquired using the flutemetamol tracer, while the amyloid-PET scans of the remaining subjects from both ADNI and GMC cohorts were acquired using florbetapir. To compensate the effect of the tracer

on the regional uptake for these subjects, we converted the extracted amyloid burden on the centiloid scale and back to a florbetapir scale (Klunk et al. 2015; Battle et al. 2018; Navitsky et al. 2018). We observe in Figure 2 that the correction reduces the variability of the amyloid uptake values for the flutemetamol group, increasing the overlapping with the florbetapir one. Supplementary Section 3.3 provides histograms illustrating the same effect in other brain regions.

Table 2: Relative error between the ground truth and k-neighbors prediction of the different clinical scores in the ADNI cohort. Average values and 95% confidence interval. CDRSB: Clinic Dementia Rating Scale Sum of Boxes; ADAS11: Alzheimer's Disease Assessment Scale; FAQ: Functional Assessment Questionnaire; RAVLT: Rey Auditory Verbal Learning Test. ADNI: Alzheimer's Disease Neuroimaging Initiative. CI: Confidence interval.

Score	CDRSB	ADAS11	RAVLT immediate	RAVLT learning	RAVLT forgetting	FAQ
Relative error (%)	5.2	6.0	7.5	14.2	17.8	9.1
95% CI	[1.4 ; 10.3]	[3.2 ; 10.6]	[5.2 ; 10.2]	[10.7 ; 18.8]	[10.3 ; 25.8]	[1.5 ; 18.8]

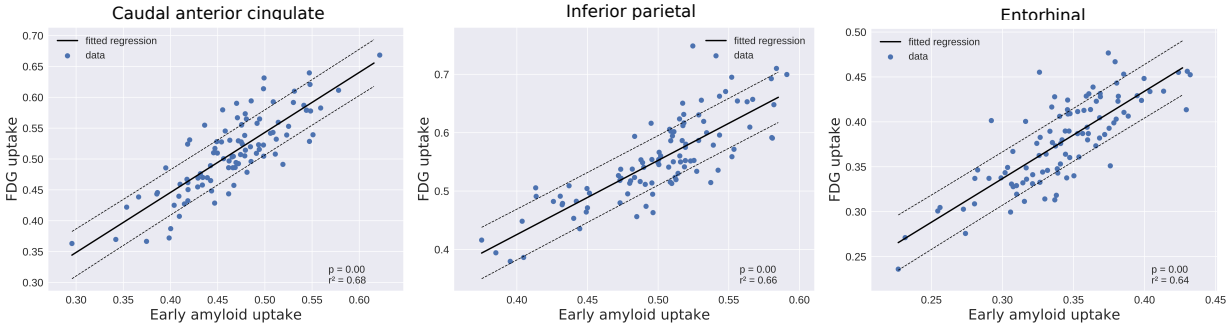


Figure 1: Scatter plot between the regional early-amyloid uptake and the corresponding FDG uptake for 57 patients of the GMC cohort. Solid black lines show the fitted linear model between regional FDG and early-amyloid. The dashed-lines represent 95% confidence interval. GMC: Geneva Memory Center FDG: (18)F-fluorodeoxyglucose.

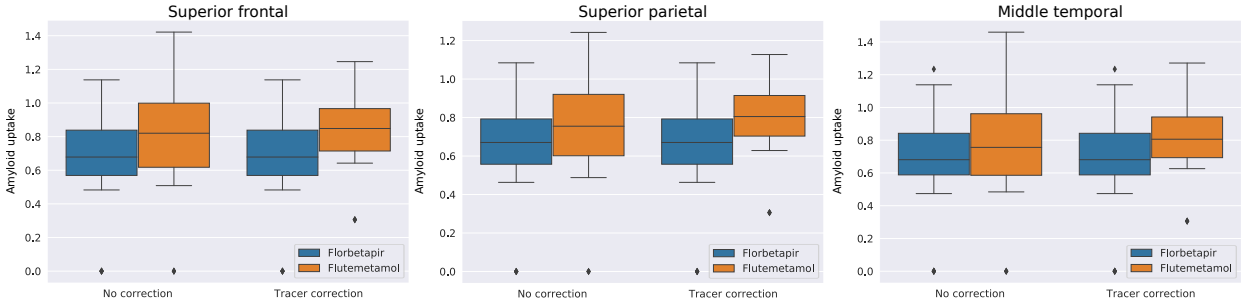


Figure 2: Distribution of the regional amyloid uptake depending on the tracer used during acquisition for the subjects from the GMC cohort. 76 amyloid-PET scans were acquired using florbetapir and 17 using flutemetamol. Tracer correction indicates that the regional amyloid uptake of subjects whose PET scan was acquired using flutemetamol was converted to a florbetapir scale. GMC: Geneva Memory Center; SUVR: Standardized Uptake Value Ratio; PET: Positron Emission Tomography.

3 Results

3.1 Known-groups validity

In this section, we considered the pathological progression previously estimated in (Abi Nader et al. 2021) by SimulAD as a reference trajectory, and computed the disease severity of the individuals from both cohorts based on the procedure of Section 2.3. We show in Figure 3 the group-wise distribution of the disease severity estimated by SimulAD for each subject in the ADNI and GMC datasets. We observe that for both cohorts the disease severity increases when going from healthy to pathological stages. The group-wise difference of disease severity across clinical groups is statistically significant for each comparison (Student's t-test $p < 0.05$) except in the case of MCI stable vs MCI converters for the GMC cohort (*cf.* Table 2a). We also notice rather large differences between clinical groups ($d > 0.7$, *cf.* Table 2a) for both cohorts except in the case of NL stable vs MCI stable for the ADNI cohort and MCI stable vs MCI converters for the GMC database. We also evaluated the consistency of the disease severity by comparing its distribution for similar clinical

groups across cohorts. We observe in Table 2b that the estimated disease severity of similar clinical groups is not significantly different ($p > 0.05$), and differences between cohorts are rather small ($d \leq 0.1$), except in the case of MCI stable. We recall that, apart from the MMSE, the clinical scores of the subjects from the GMC cohort were imputed based on the procedure detailed in Section 2.6. We show in Supplementary Figure 7 that the estimation of the individual disease severity is robust to this approximation.

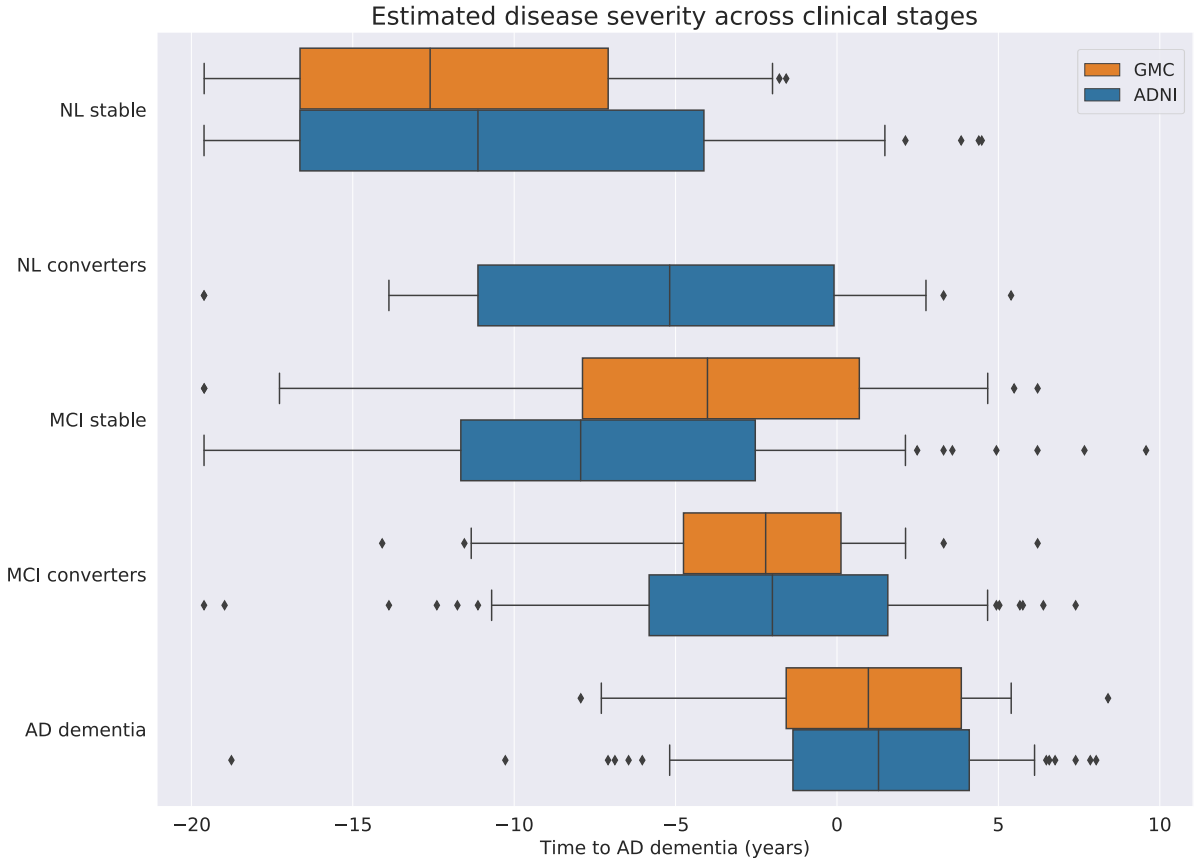


Figure 3: Distribution of the disease severity estimated by SimulAD across clinical stages for the ADNI and GMC cohorts relatively to the estimated model of disease progression (Supplementary Figure 2). ADNI: Alzheimer's Disease Neuroimaging Initiative; GMC: Geneva Memory Center; NL: cognitively healthy; MCI: mild cognitive impairment; AD: Alzheimer's dementia. Converters are NL patients progressing to MCI or AD dementia, or MCI individuals progressing to AD dementia.

Table 3: Comparison of the estimated disease severity distribution between clinical groups within each cohorts (a) and between similar clinical groups across cohorts (b); We report p-values of Student's *t*-test as well as the associated effect size (Cohen's *d*).

(a)

Within cohorts disease severity comparison

	NL stable vs NL converters		NL stable vs MCI stable		MCI stable vs MCI converters		MCI stable vs AD dementia		MCI converters vs AD dementia	
Cohort	ADNI	GMC	ADNI	GMC	ADNI	GMC	ADNI	GMC	ADNI	GMC
p-value	$7.5 \cdot 10^{-4}$	/	$1.8 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$	$1.2 \cdot 10^{-11}$	$4.4 \cdot 10^{-1}$	$1.1 \cdot 10^{-26}$	$3.8 \cdot 10^{-3}$	$3.1 \cdot 10^{-7}$	$1.6 \cdot 10^{-2}$
Cohen's d	0.75	/	0.35	1.1	0.91	0.21	1.5	0.80	0.74	0.81

(b)

Between cohorts disease severity comparison

	NL stable	MCI stable	MCI converters	AD dementia
p-value	$8.2 \cdot 10^{-1}$	$7.0 \cdot 10^{-3}$	$9.5 \cdot 10^{-1}$	$7.3 \cdot 10^{-1}$
Cohen's d	0.03	0.57	0.10	0.09

3.2 Concurrent validity

We show in Figure 4 the progression of standard clinical and imaging markers with respect to the disease severity estimated by SimulAD for subjects from the ADNI and GMC databases. For both cohorts, the disease severity significantly correlates with the MMSE score (ADNI: $\rho = -0.58$, $p < 0.01$; GMC: $\rho = -0.55$, $p < 0.01$). Regarding imaging-biomarkers, in both ADNI and GMC datasets the estimated disease severity correlates with hippocampal volume (ADNI: $\rho = -0.57$, $p < 0.01$; GMC: $\rho = -0.62$, $p < 0.01$), glucose metabolism (ADNI: $\rho = -0.80$, $p < 0.01$; GMC: $\rho = -0.67$, $p < 0.01$) and amyloid burden (ADNI: $\rho = 0.44$, $p < 0.01$; GMC: $\rho = 0.31$, $p < 0.01$). Since 50 individuals from the GMC cohort underwent a tau-PET scan, we also compare their estimated disease severity with their tau burden and show a significant correlation between them in Figure 4 ($\rho = 0.62$, $p < 0.01$). This latter correlation of the disease severity with a typical biomarker of AD that was not used for building the model supports the reliability of such a measure to summarize the overall severity of AD.

3.3 Model reliability

In the previous sections, we assessed the disease severity of the subjects based on the model of progression simulated on the ADNI cohort (Abi Nader et al. 2021). In order to demonstrate the reliability of the dynamics estimated by SimulAD, we simulated a new model of disease progression personalized to the GMC cohort. We show in Figure 5 the predicted evolution of imaging and clinical measurements based on this dataset. Similarly to what has been observed on the model previously trained on the ADNI database (Supplementary Figure 2), amyloid load increases and saturates early while following a uniform spatial pattern. Amyloid deposition is followed by a delayed process of neurodegeneration, more specifically a decrease of glucose metabolism and gray

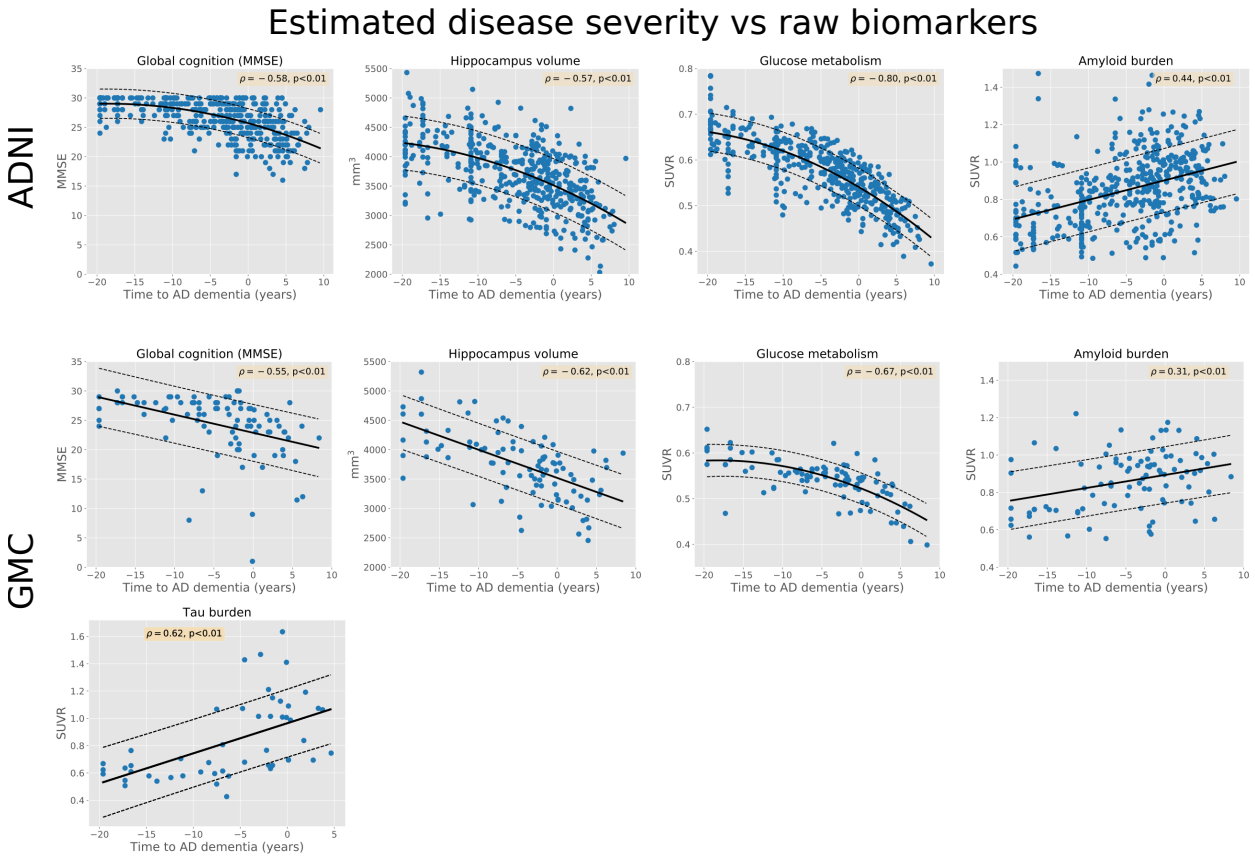


Figure 4: Relationships between the estimated disease severity and global cognition (MMSE), hippocampal volume, glucose metabolism, amyloid burden and tau burden (only for the GMC cohort). For each subject the estimated disease severity quantifies their position with respect to the model of disease progression (Supplementary Figure 2). The dashed-lines represent 95% confidence interval. ADNI: Alzheimer's Disease Neuroimaging Initiative; GMC: Geneva Memory Center; SUVR: Standardized Uptake Value Ratio.

matter atrophy, mostly affecting temporal and parietal regions. Finally, clinical scores such as the MMSE or the CDRSB show a non-linear evolution accelerating during the latest stages of the disease.

We also compared the models of progression obtained on the ADNI and GMC cohorts. Figure 6 shows the evolution of the different z-scores depending on the cohort used to estimate the disease progression. These z-scores indicate the overall evolution of clinical scores, cerebral atrophy, amyloid deposition and glucose metabolism during AD. Given that the z-scores are not related to a physical unit but rather quantify the abnormality of a particular process, they were re-scaled between 0 and 1 to illustrate the progression from healthy towards pathological stages. We observe that the four z-scores exhibit similar evolution patterns, whether they have been estimated on the ADNI or the GMC cohort. When averaged across time, the error between the z-scores of the two cohorts is of 6%, 7%, 8% and 9% for z^{cli}, z^{atr}, z^{met} and z^{amy} respectively. We provide in Supplementary Section 5 the evolution of the error between the evolutions of clinical scores and imaging regional measurements estimated based on the ADNI and the GMC cohort. When averaged over time, brain regions and clinical scores, the error is of 3%, 6%, 7% and 12% for MRI, FDG-PET, amyloid-PET derived regional measurements and clinical scores respectively. Finally, in spite of the fact that most of the clinical scores were imputed in the GMC cohort, Supplementary Section 4 shows that the resulting z-scores trajectories are robust to this estimation. Indeed, we observe that adding an additional error when imputing the clinical scores in the GMC cohort leads to rather small changes for the estimated z-scores trajectories compared to the results presented in Figure 6.

Finally, we applied SimulAD individually on all the patients from the GMC cohort to simulate their evolution over five years. We computed the error between their simulated MMSE and the one assessed by the physicians for the patients with available follow-up visits. The results are presented in Table 4, where we observe that the mean error ranges from 1.8 to 2.4 across clinical groups

highlighting the reliability of the model for estimating clinical outcomes during follow-up. For illustrative purposes, we also provide in Supplementary Figure 10 the simulated evolution of the MMSE, hippocampus volume, glucose metabolism and amyloid burden for four patients from the GMC cohort.

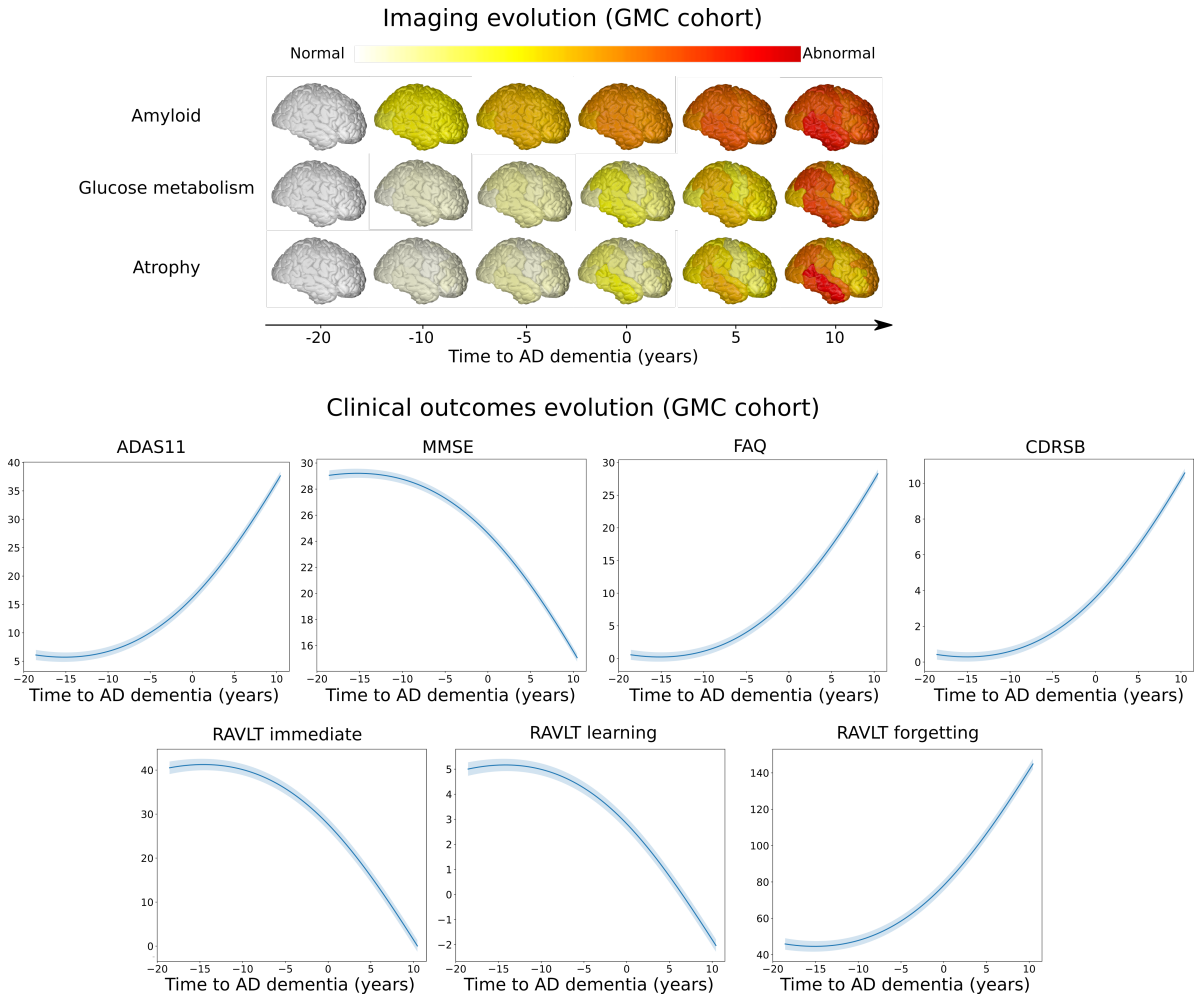


Figure 5: Simulated long-term evolution of cortical measurements for the different types of imaging markers and clinical scores based on the GMC cohort. Shadowed areas represent the standard deviation of the average trajectory. GMC: Geneva Memory Center.

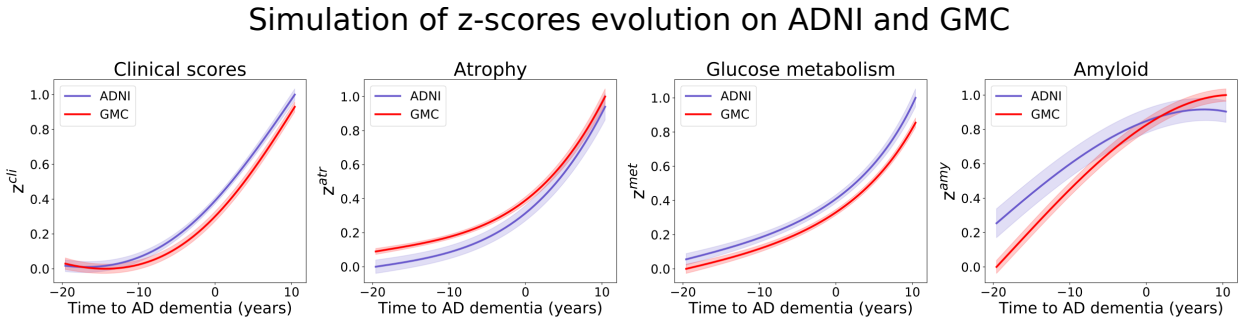


Figure 6: Estimated long-term dynamics depending on the cohort used for simulating the z-scores trajectories (time is relative to conversion to Alzheimer's dementia). The z-scores have been re-scaled between 0 and 1 to illustrate the progression of each process from normal to pathological stages. ADNI: Alzheimer's Disease Neuroimaging Initiative; GMC: Geneva Memory Center.

Table 4: For each clinical group: follow-up time in years (mean and standard deviation), total number of times the MMSE was assessed, and absolute error between the MMSE predicted by our method and the ground truth value from clinical evaluation (mean and standard deviation). NL: cognitively healthy; MCI: mild cognitive impairment; AD: Alzheimer's dementia. Converters are MCI individuals progressing to AD dementia.

Group	NL	MCI stable	MCI converters	AD dementia
Follow-up (years)	1.9 (1.2)	2.3 (1.5)	2.4 (1.7)	1.8 (1.4)
# MMSE assessments	48	63	59	31
Error	2.1 (1.5)	2.4 (1.9)	2.3 (2.0)	1.8 (1.6)

4 Discussion

In this study, we presented a thorough assessment of SimulAD on the independent GMC clinical cohort. The model was initially estimated based on the analysis of longitudinal imaging and clinical data from a subset of the ADNI database (Abi Nader et al. 2021). Due to a mismatch between cohorts in terms of missing measurements and data acquisition, we designed a pre-processing pipeline in order to enable the application of the model. We subsequently carried out a set of experiments to evaluate the validity of our approach and showed that the results obtained on the ADNI database were reproducible on the GMC cohort, outlining the robustness and generalization properties of SimulAD.

The present results underlined the feasibility of applying our model in a clinical context. We recall that throughout the experiments the disease severity was only assessed at baseline. Indeed, in the GMC cohort the subjects underwent the MMSE as well as an MRI, FDG-PET and amyloid-PET scan at baseline only. During the subsequent visits, only the MMSE was assessed in order to determine the clinical status of the subjects, however no imaging data was available. While the model is able to assess the disease severity even if many data modalities are missing, in this particular case the disease severity estimation would be entirely based on clinical information without any input from the imaging modalities. Given this context, our model wouldn't bring any additional insight about the pathological state of a subject compared to a battery of neuro-psychological tests. We showed that SimulAD was able to differentiate subjects across clinical groups for both cohorts (Figure 3, Table 3). Known-groups validity was established by assessing the similarity between cohorts of the estimated disease severity distribution for the NL stable, MCI converters and AD dementia groups (Table 2b). Moreover, the estimated disease severity correlated with clinical and

imaging assessments in both cohorts, and especially with tau burden on the GMC cohort. We recall that AD is defined by evidence of an abnormal load of both amyloid and tau (Jack, Bennett, et al. 2018). Therefore, given that the model was estimated based on a subset of the ADNI cohort without accounting for tau-PET data, the correlation between the disease severity and tau burden on an external dataset is an appreciable demonstration of the concurrent validity of SimulAD. We also observed that independently from the cohort used to simulate the disease progression, the long-term evolution of the z-scores were similar between cohorts (Figure 6), thus outlining the reliability of the proposed model of AD progression. Finally, we showed that when applied on the GMC cohort, SimulAD was able to reliably predict the evolution of the MMSE of the subjects, highlighting its potential as a tool for individual monitoring of clinical outcomes. Overall, these results indicate general robustness of SimulAD when applied to independent cohorts characterized by different imaging acquisition protocols.

Some results highlighted in the proposed analysis deserve further discussion. There is no statistically significant difference between the estimated severity of the MCI stable and converters in the GMC cohort, while there is a significant difference of the estimated disease severity between the MCI stable of the ADNI and GMC cohorts. Concerning the first remark, it is important to note that among the 28 MCI stable subjects of the GMC cohort, a single visit only was available for 16 of them. It is therefore likely that this sub-group contains potential converters who may bias the MCI stable group with non-representative measurements. Regarding the second remark, we note in Figure 3 that the median disease severity of the MCI stable group of the ADNI dataset is approximately of -8 years, while it is close to -4 years for the GMC cohort. This means that, according to our model, the pathological condition of the group of MCI stable from the GMC cohort is more severe than the one of the ADNI subset considered in the study. This is in agreement with the fact that

compared to the ADNI database, the MCI stable subjects from the GMC cohort exhibit a statistically significant lower MMSE score, lower volume of the hippocampus, lower glucose metabolism and higher amyloid burden (p-values<0.05, Student's *t*-test). Moreover, we note that the distribution of the disease severity is conserved between both cohorts for NL stable, MCI converters and AD dementia groups. This result may point to the generalization capabilities of the disease severity estimation, which seems to be solely influenced by the clinical status. It also shows the ability of the model to identify specific clinical sub-types.

To our knowledge, the EBM and the DEBM are the only data-driven models of disease progression which have been evaluated on several external cohorts (Archetti et al. 2019). The type of study here presented is therefore of relevant experimental value to enable the future application of disease progression models. We note that SimulAD presents certain advantages compared to standard event-based models, as it provides a finer-grained description of the disease for both spatial and temporal scales, in which regional changes affecting the brain are modelled on a continuous long-term time span. In addition, the method allows to personalize the evolution of clinical and imaging measurements for any patient or group of patients. Finally, the proposed mathematical formulation of AD progression as a system of ODEs allows to go beyond a simple description of the dynamics at stake during the disease evolution. In particular, we can simulate the impact of an anti-amyloid treatment on cognitive outcomes (Abi Nader et al. 2021).

As data-driven models are becoming more popular in healthcare thanks to their ability to leverage large scale clinical data, it is of utmost importance to facilitate their transfer from a research context to clinical practice. A growing number of regulatory institutions provided guidelines to help designing machine learning models that could be applied in clinical practice (Health 2019). The major problem

revolves around the generalization of the models beyond the dataset used to develop them, the main obstacle being their robustness to biases (Ghassemi et al. 2019). In this study, we developed a pipeline to mitigate the biases due to data heterogeneity and missing measurements. This work included data imputation for clinical data, and data standardization for PET imaging scans. Even though we carried out different tests to assess the reliability of our imputation strategy in Section 2.6, it would be interesting in a future work to better evaluate the effect of the data imputation on the model performances. This could be done for instance by performing it on the ADNI cohort and re-assess the disease severity to observe if it led to substantial changes. Unfortunately, this cannot be done at the moment due to the choice of the method we used for approximating the missing neuro-psychological scores (k-neighbors), and also because no early-phase amyloid-PET was available to predict the corresponding FDG uptake on the sub-sample of the ADNI database used to train SimulAD. A more radical approach could be to simply discard all the missing data modalities, and re-train a new model with only the available data. We explored this possibility in Supplementary Section 7, and showed how it impacts the interpretability of the model as well as the disease severity estimation. Overall, even if SimulAD proved robust to these approximations, one of the main challenge highlighted in this study points to the complexity of data integration across cohorts and studies. In our particular case, neuropsychological tests could be standardized on a common scale, thus allowing to replace a score by another if they assess similar functions, while PET data could be systematically converted to the centiloid scale.

Finally, this study motivates further extensions of SimulAD that would foster its adoption in clinical practice. For instance, a limitation of our model is the underlying hypothesis that there exists a unique progression of AD which is common across individuals, while the disease is in fact highly heterogeneous. SimulAD could be extended in the future to account for multiple

risk factors, such as the presence of APOE4 (Kim et al. 2009), thus leading to a higher level of personalization of the predictions. Another avenue of improvement would be to account for a larger panel of biomarkers, such as tau (Pontecorvo et al. 2019), in order to better comprehend the disease progression. Currently, SimulAD is still a research software and future efforts should focus on the development of a user-friendly platform that could be deployed in clinical routine.

5 Conclusion

We presented a preliminary validation of SimulAD on the clinical cohort of the GMC. The results highlighted the reliability of the dynamics simulated by SimulAD for the disease key biomarkers, and showed encouraging performances in terms of disease staging on both cohorts. SimulAD is a promising modeling tool that may enable in the future the identification of subjects for enrollment in clinical trials, or the monitoring of the efficacy of disease modifying drugs.

References

Abi Nader, Clément, Nicholas Ayache, Giovanni B Frisoni, Philippe Robert, and Marco Lorenzi (2021). "Simulating the outcome of amyloid treatments in Alzheimer's Disease from multi-modal imaging and clinical data". In: *Brain communications*.

Arbabshirani, Mohammad R., Sergey Plis, Jing Sui, and Vince D. Calhoun (2017). "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls". In: *NeuroImage* 145, pp. 137–165. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2016.02.079.

Archetti, D., S. Ingala, V. Venkatraghavan, V. Wottschel, A. L. Young, M. Bellio, E. E. Bron, S. Klein, F. Barkhof, D. C. Alexander, N. P. Oxtoby, G. B. Frisoni, and A. Redolfi (2019). "Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in Alzheimer's disease". In: *Neuroimage Clin* 24, p. 101954.

Battle, Mark R., Lovena Chedumbarum Pillay, Val J. Lowe, David Knopman, Bradley Kemp, Christopher C. Rowe, Vincent Doré, Victor L. Villemagne, and Christopher J. Buckley (2018). "Centiloid scaling for quantification of brain amyloid with [18F]flutemetamol using multiple processing methods". In: *EJNMMI Research* 8.1, p. 107. ISSN: 2191-219X. DOI: 10.1186/s13550-018-0456-7.

Bilgel, M., J. L. Prince, D. F. Wong, S. M. Resnick, and B. M. Jedynak (2016). "A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging". In: *Neuroimage* 134, pp. 658–670.

Castro, Daniel C., Ian Walker, and Ben Glocker (2020). "Causality matters in medical imaging". In: *Nature Communications* 11.1, p. 3673. ISSN: 2041-1723. DOI: 10.1038/s41467-020-17478-w.

Daerr, S., M. Brendel, C. Zach, E. Mille, D. Schilling, M. J. Zacherl, K. Bürger, A. Danek, O. Pogarell, A. Schildan, M. Patt, H. Barthel, O. Sabri, P. Bartenstein, and A. Rominger (2017). "Evaluation of early-phase [18F]-florbetaben PET acquisition in clinical routine cases". In: *NeuroImage Clinical* 14, pp. 77–86.

Davatzikos, Christos, Feng Xu, Yang An, Yong Fan, and Susan M. Resnick (2009). "Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index". In: *Brain* 132.8, pp. 2026–2035. DOI: 10.1093/brain/awp091.

Desikan, R. S., F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, M. S. Albert, and R. J. Killiany (2006). "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest". In: *NeuroImage* 31.3, pp. 968–980.

Falahati, Farshad, Eric Westman, and Andrew Simmons (2014). "Multivariate Data Analysis and Machine Learning in Alzheimer's Disease with a Focus on Structural Magnetic Resonance Imaging". In: *Journal of Alzheimer's Disease* 41. 3, pp. 685–708. ISSN: 1875-8908. DOI: 10.3233/JAD-131928.

Fonteijn, H. M., M. Modat, M. J. Clarkson, J. Barnes, M. Lehmann, N. Z. Hobbs, R. I. Scahill, S. J. Tabrizi, S. Ourselin, N. C. Fox, and D. C. Alexander (2012). "An event-based model for

disease progression and its application in familial Alzheimer's disease and Huntington's disease". In: *NeuroImage* 60.3, pp. 1880–1889.
Gamberger, D., N. Lavrač, S. Srivatsa, R. E. Tanzi, and P. M. Doraiswamy (2017). "Identification of clusters of rapid and slow decliners among subjects at risk for Alzheimer's disease". In: *Sci Rep* 7.1, p. 6763.
Ghassemi, Marzyeh, Tristan Naumann, Peter Schulam, Andrew L. Beam, Irene Y. Chen, and Rajesh Ranganath (2019). "Practical guidance on artificial intelligence for health-care data". In: *The Lancet Digital Health* 1.4, e157–e159. ISSN: 2589-7500. DOI: 10.1016/S2589-7500(19)30084-6.
Greve, D. N., C. Svarer, P. M. Fisher, L. Feng, A. E. Hansen, W. Baare, B. Rosen, B. Fischl, and G. M. Knudsen (2014). "Cortical surface-based analysis reduces bias and variance in kinetic modeling of brain PET data". In: *Neuroimage* 92, pp. 225–236.
Health, The Lancet Digital (2019). "Walking the tightrope of artificial intelligence guidelines in clinical practice". In: *The Lancet Digital Health* 1.3, e100. ISSN: 2589-7500. DOI: 10.1016/S2589-7500(19)30063-9.
Jack, C. R., D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish, E. Liu, J. L. Molinuevo, T. Montine, C. Phelps, K. P. Rankin, C. C. Rowe, P. Scheltens, E. Siemers, H. M. Snyder, R. Sperling, C. Elliott, E. Masliah, L. Ryan, and N. Silverberg (2018). "NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease". In: *Alzheimers Dement* 14.4, pp. 535–562.
Jack, C. R., D. S. Knopman, W. J. Jagust, R. C. Petersen, M. W. Weiner, P. S. Aisen, L. M. Shaw, P. Vemuri, H. J. Wiste, S. D. Weigand, T. G. Lesnick, V. S. Pankratz, M. C. Donohue, and J. Q. Trojanowski (2013). "Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers". In: *Lancet Neurol* 12.2, pp. 207–216.
Jack, C. R., H. J. Wiste, S. D. Weigand, T. M. Therneau, V. J. Lowe, D. S. Knopman, J. L. Gunter, M. L. Senjem, D. T. Jones, K. Kantarci, M. M. Machulda, M. M. Mielke, R. O. Roberts, P. Vemuri, D. A. Reyes, and R. C. Petersen (2017). "Defining imaging biomarker cut points for brain aging and Alzheimer's disease". In: *Alzheimers Dement* 13.3, pp. 205–216.
Jagust, W. J., S. M. Landau, L. M. Shaw, J. Q. Trojanowski, R. A. Koeppe, E. M. Reiman, N. L. Foster, R. C. Petersen, M. W. Weiner, J. C. Price, and C. A. Mathis (2009). "Relationships between biomarkers in aging and dementia". In: *Neurology* 73.15, pp. 1193–1199.
Jedynak, B. M., A. Lang, B. Liu, E. Katz, Y. Zhang, B. T. Wyman, D. Raunig, C. P. Jedynak, B. Caffo, and J. L. Prince (2012). "A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease Neuroimaging Initiative cohort". In: *NeuroImage* 63.3, pp. 1478–1486.
Khanal, Bishesh, Nicholas Ayache, and Xavier Pennec (2017). "Simulating Longitudinal Brain MRIs with known Volume Changes and Realistic Variations in Image Intensity". In: *Frontiers in Neuroscience* 11.Article 132, p. 18. DOI: 10.3389/fnins.2017.00132.
Khanal, Bishesh, Marco Lorenzi, Nicholas Ayache, and Xavier Pennec (2016). "A biophysical model of brain deformation to simulate and analyze longitudinal MRIs of patients with Alzheimer's disease". In: *NeuroImage* 134, pp. 35–52. DOI: 10.1016/j.neuroimage.2016.03.061.

Kim, J., J. M. Basak, and D. M. Holtzman (2009). "The role of apolipoprotein E in Alzheimer's disease". In: *Neuron* 63.3, pp. 287–303.

Klöppel, Stefan, Cynthia M. Stonnington, Carlton Chu, Bogdan Draganski, Rachael I. Scahill, Jonathan D. Rohrer, Nick C. Fox, Jr Jack Clifford R., John Ashburner, and Richard S. J. Frackowiak (2008). "Automatic classification of MR scans in Alzheimer's disease". In: *Brain* 131.3, pp. 681–689. ISSN: 0006-8950. DOI: 10.1093/brain/awm319.

Klunk, W. E., R. A. Koeppe, J. C. Price, T. L. Benzinger, M. D. Devous, W. J. Jagust, K. A. Johnson, C. A. Mathis, D. Minhas, M. J. Pontecorvo, C. C. Rowe, D. M. Skovronsky, and M. A. Mintun (2015). "The Centiloid Project: standardizing quantitative amyloid plaque estimation by PET". In: *Alzheimers Dement* 11.1, pp. 1–15.

Koval, Igor, Jean-Baptiste Schiratti, Alexandre Routier, Michael Bacci, Olivier Colliot, Stéphanie Allassonnière, and Stanley Durrleman (2018). "Spatiotemporal Propagation of the Cortical Atrophy: Population and Individual Patterns". In: *Frontiers in Neurology* 9, p. 235. ISSN: 1664-2295. DOI: 10.3389/fneur.2018.00235.

Landau, S. M., D. Harvey, C. M. Madison, E. M. Reiman, N. L. Foster, P. S. Aisen, R. C. Petersen, L. M. Shaw, J. Q. Trojanowski, C. R. Jack, M. W. Weiner, and W. J. Jagust (2010). "Comparing predictors of conversion and decline in mild cognitive impairment". In: *Neurology* 75.3, pp. 230–238.

Landau, S. M., M. A. Mintun, A. D. Joshi, R. A. Koeppe, R. C. Petersen, P. S. Aisen, M. W. Weiner, and W. J. Jagust (2012). "Amyloid deposition, hypometabolism, and longitudinal cognitive decline". In: *Ann Neurol* 72.4, pp. 578–586.

Lorenzi, Marco, Maurizio Filippone, Giovanni B. Frisoni, Daniel C. Alexander, and Sebastien Ourselin (2017). "Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease". In: *NeuroImage*. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.Neuroimage.2017.08.059.

Marinescu, R. V., A. Eshaghi, M. Lorenzi, A. L. Young, N. P. Oxtoby, S. Garbarino, S. J. Crutch, and D. C. Alexander (2019). "DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders". In: *NeuroImage* 192, pp. 166–177.

Mendelson, Alex F., Maria A. Zuluaga, Marco Lorenzi, Brian F. Hutton, and Sébastien Ourselin (2017). "Selection bias in the reported performances of AD classification pipelines". In: *NeuroImage: Clinical* 14, pp. 400–416. ISSN: 2213-1582. DOI: https://doi.org/10.1016/j.nicl.2016.12.018.

Navitsky, Michael, Abhinay D. Joshi, Ian Kennedy, William E. Klunk, Christopher C. Rowe, Dean F. Wong, Michael J. Pontecorvo, Mark A. Mintun, and Michael D. Devous (2018). "Standardization of amyloid quantitation with florbetapir standardized uptake value ratios to the Centiloid scale". In: *Alzheimer's & Dementia* 14.12, pp. 1565–1571. ISSN: 1552-5260. DOI: https://doi.org/10.1016/j.jalz.2018.06.1353.

Pontecorvo, M. J., M. D. Devous, I. Kennedy, M. Navitsky, M. Lu, N. Galante, S. Salloway, P. M. Doraiswamy, S. Southekal, A. K. Arora, A. McGeehan, N. C. Lim, H. Xiong, S. P. Truocchio, A. D. Joshi, S. Shcherbinin, B. Teske, A. S. Fleisher, and M. A. Mintun (2019). "A

multicentre longitudinal study of flortaucipir (18F) in normal ageing, mild cognitive impairment and Alzheimer's disease dementia". In: *Brain* 142.6, pp. 1723–1735.
Reuter, M., N. J. Schmansky, H. D. Rosas, and B. Fischl (2012). "Within-subject template estimation for unbiased longitudinal image analysis". In: *NeuroImage* 61.4, pp. 1402–1418.
Sperling, R. A., C. R. Jack, and P. S. Aisen (2011). "Testing the right target and right drug at the right stage". In: *Sci Transl Med* 3.111.
Venkatraghavan, Vikram, Esther E. Bron, Wiro J. Niessen, and Stefan Klein (2019). "Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling". In: *NeuroImage* 186, pp. 518–532. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2018.11.024.
Young, A. L., R. V. Marinescu, N. P. Oxtoby, M. Bocchetta, K. Yong, N. C. Firth, D. M. Cash, D. L. Thomas, K. M. Dick, J. Cardoso, J. van Swieten, B. Borroni, D. Galimberti, M. Masellis, M. C. Tartaglia, J. B. Rowe, C. Graff, F. Tagliavini, G. B. Frisoni, R. Laforce, E. Finger, A. de Mendonca, S. Sorbi, J. D. Warren, S. Crutch, N. C. Fox, S. Ourselin, J. M. Schott, J. D. Rohrer, and D. C. Alexander (2018). "Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference". In: *Nat Commun* 9.1, p. 4273.

Acknowledgements

Funding This work has been supported by the French government, through the UCA^{JEDI} and 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ref.n ANR-15-IDEX-01 and ANR-19-P3IA-0002), the grant AAP Santé 06 2017-260 DGA-DSH. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

Data availability Data for this study were collected at the Centre de la mémoire, Geneva University and University Hospitals, thanks to funds from: Association Suisse pour la Recherche sur l'Alzheimer, Genève; Fondation Segré, Genève; Ivan Pictet, Genève; Fondazione Agusta, Lugano; Fondation Chmielewski, Genève; Fondation Privée des HUG, Genève, Velux Stiftung; Swiss National Science Foundation (projects n.320030_182772, 320030_169876 and 320030_185028); Horizon 2020 (projects n. 667375); Human Brain Project; Innovative Medicines Initiatives (IMI contract n. 115736 and 115952).

Supplementary Material

1 Overview of SimulAD

We illustrate in Supplementary Figure 1 how SimulAD works. A weighted average of baseline measurements of amyloid deposition, glucose metabolism, atrophy and clinical scores allows to transform them in a corresponding z-score. A system of Ordinary Differential Equations (ODEs) links these four z-scores, thus providing us with the interaction rule allowing us to extrapolate the evolution of the z-scores in time, and to predict the corresponding imaging and clinical measures. In the next paragraphs, we provide details about the mathematical formulation of SimulAD.

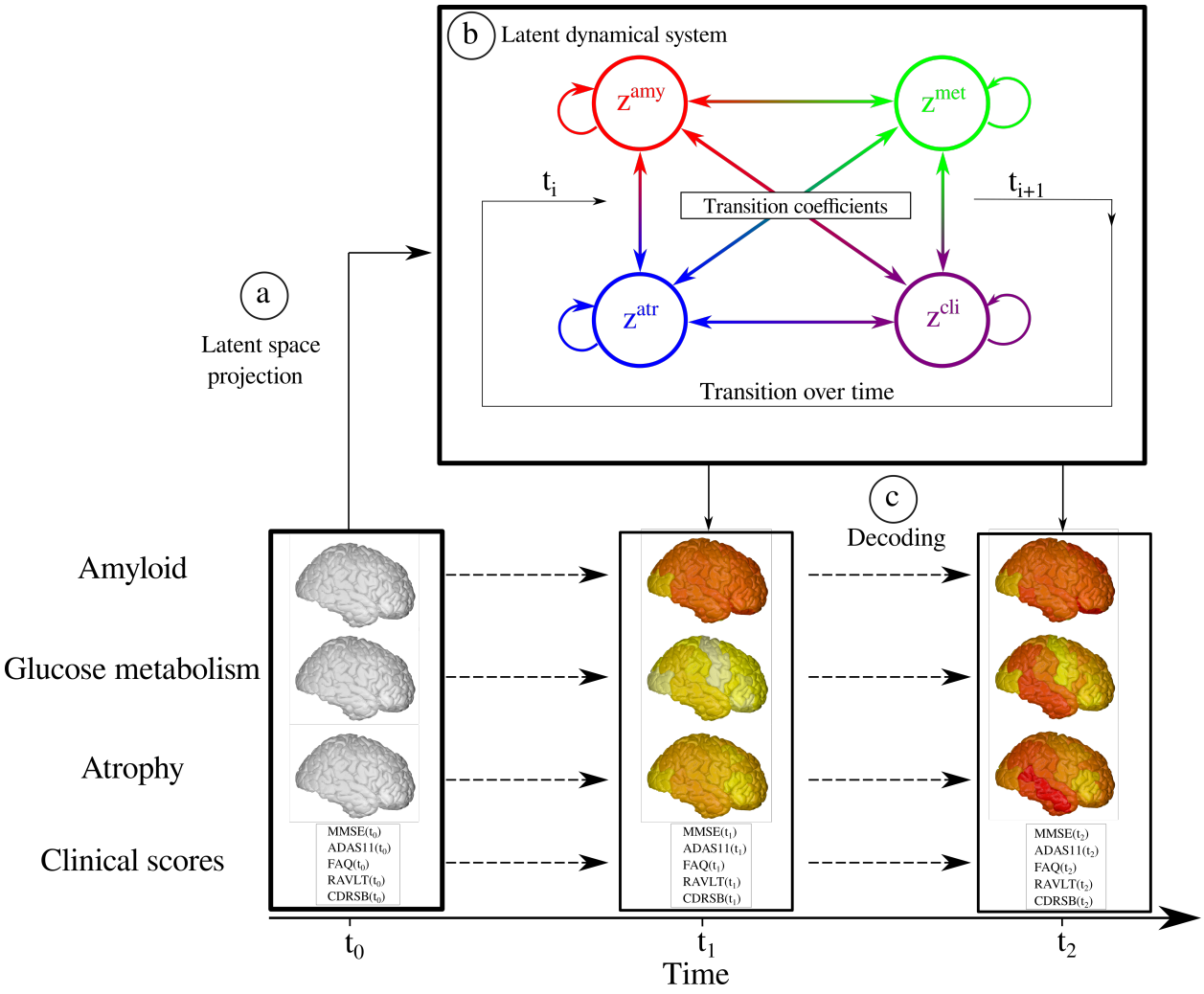
Measures from M data modalities at time t for a given subject i are stored in a matrix $\mathbf{X}_i(t) = [\mathbf{x}_i^1(t), \mathbf{x}_i^2(t), ..., \mathbf{x}_i^M(t)]^T$. It is assumed that these measures are independently generated by a common latent representation $\mathbf{z}_i(t)$ of the data such that:

$$\begin{aligned} p(\mathbf{X}_i(t)|\mathbf{z}_i(t), \boldsymbol{\sigma^2}, \boldsymbol{\psi}) &= \prod_m p(\mathbf{x}_i^m(t)|\mathbf{z}_i(t), \sigma_m^2, \psi_m) \\ &= \prod_m \mathcal{N}(\mu_m(\mathbf{z}_i(t), \psi_m), \sigma_m^2), \\ \mathbf{z_i}(t) &= \Lambda(\mathbf{z}_i(t_0), t), \\ \mathbf{z}_i(t_0) &\sim p(\mathbf{z}_i(t_0)), \end{aligned} \tag{1}$$

σ_m^2 is measurement noise and ψ_m are the parameters of the function μ_m that allows to compute the different measures of modality m for a subject based on its latent representation. Λ represents the interaction rule presented in the manuscript which allows to compute the latent variable at time t. It is defined as the solution of the following system of ODEs:

$$\begin{aligned} &\frac{d\mathbf{z}(t)}{dt} = \mathbf{W}\mathbf{z}(t) - \mathbf{V}\mathbf{z}^2(t) = g(\mathbf{z}(t), \theta_{ODE}) \text{ where,} \\ &\left(\mathbf{W}_{i,j}\right) = \begin{cases} k_i \text{ if i=j,} \\ \alpha_{i,j} \text{ otherwise;} \end{cases} \text{ and } \left(\mathbf{V}_{i,j}\right) = \begin{cases} k_i \text{ if i=j} \\ 0 \text{ otherwise,} \end{cases} \end{aligned} \tag{2}$$

where the parameters of the system of ODEs θ_{ODE} are the entries of the matrices $\mathbf{W}$ and $\mathbf{V}$. This system of ODEs enforces a sigmoidal evolution for each type of modality m, while accounting for the relationship between the different modalities. Based on this system of ODEs and given baseline measures $\mathbf{z}(0)$, the variable $\mathbf{z}(t)$ can be obtained by integration, $\mathbf{z}(t) = \mathbf{z}(0) + \int_0^t g(\mathbf{z}(x), \theta_{ODE})dx$. This mathematical operation is the interaction rule used to compute the long-term trajectories of the z-scores.



Supplementary Figure 1: Overview of SimulAD. a) Measures of amyloid deposition, glucose metabolism, atrophy and clinical scores are transformed in a corresponding z-score $z^{amy}, z^{met}, z^{atr}, z^{cli}$. b) The dynamical system describing the relationships between the z-scores allows to compute their transition across the evolution of the disease. c) Given the z-scores and the estimated dynamics, the follow-up measurements can be reconstructed to match the observed data.

Parameters' estimation is made by resorting to stochastic variational inference which is commonly used for this type of generative model. This gives us the following mathematical function that is maximized by gradient descent:

$$\mathcal{E} = \sum_m \mathbb{E}_{q(\mathbf{z}|\mathbf{X})} \left[\log p(\mathbf{x}^m|\mathbf{z}, \theta_{ODE}, \sigma_m^2, \psi_m) \right] - \mathcal{D}\left[q(z^m|\mathbf{x}^m) | p(z^m) \right], \tag{3}$$

where $\mathcal{D}$ refers to the Kullback-Leibler (KL) divergence, and q is an approximated posterior distribution that allows to sample values z^m for a given data modality m based on the raw measures $\mathbf{x}^m$.

Finally, we can estimate the individual disease severity of a given subject based on its raw measures $\mathbf{X}$ and the reference progression $\mathbf{z}(t)$. The individual disease severity τ is defined as the time-point jointly minimizing the distance between the raw measures of the subject and the reference progression:

$$\begin{aligned} \tau &= \arg\min_t ||f(\mathbf{X}, \boldsymbol{\phi}^1) - \mathbf{z}(t)||_1 \\ &= \sum_m |f(\mathbf{x}^m, \phi^1) - z^m(t)|. \end{aligned} \tag{4}$$

2 Model of disease progression

Once the model is optimized based on the ADNI data, we can simulate the evolution of four processes at stake during AD: cognitive and behavioural decline, gray matter atrophy, amyloid deposition and glucose metabolism. This is illustrated in Supplementary Figure 2 Panel I, in which we show the estimated evolution of the z-scores by relying on the measures from the AD dementia subjects of the ADNI cohort. The horizontal axis represents the course of the disease in years, and the reference point t=0 corresponds to the time of conversion to AD dementia. Given that the z-scores are not related to a physical unit but rather quantify the abnormality of a particular process, they were re-scaled between 0 and 1 to illustrate the progression from healthy towards pathological stages. Notably z^{amy}, which summarizes the amyloid deposition, shows a fast abnormal progression in the earliest stages before reaching a plateau. It is followed by z^{met} and z^{atr} respectively representing a global decrease of glucose metabolism and gray matter atrophy, and which evolve at a similar pace. Finally, we observe that the global clinical state quantified by z^{cli} shows a strong non-linear acceleration after conversion to AD dementia.

Based on the z-score trajectories of Supplementary Figure 2 Panel I, we can compute the evolution

of their associated clinical and regional imaging measurements shown in Supplementary Figure 2 Panel II. We note that the progression of the clinical and imaging markers is compatible with what has been observed for the z-scores. The predicted evolution of the remaining clinical scores is provided in Supplementary Figure 3.

3 Data correction

3.1 Clinical scores imputation

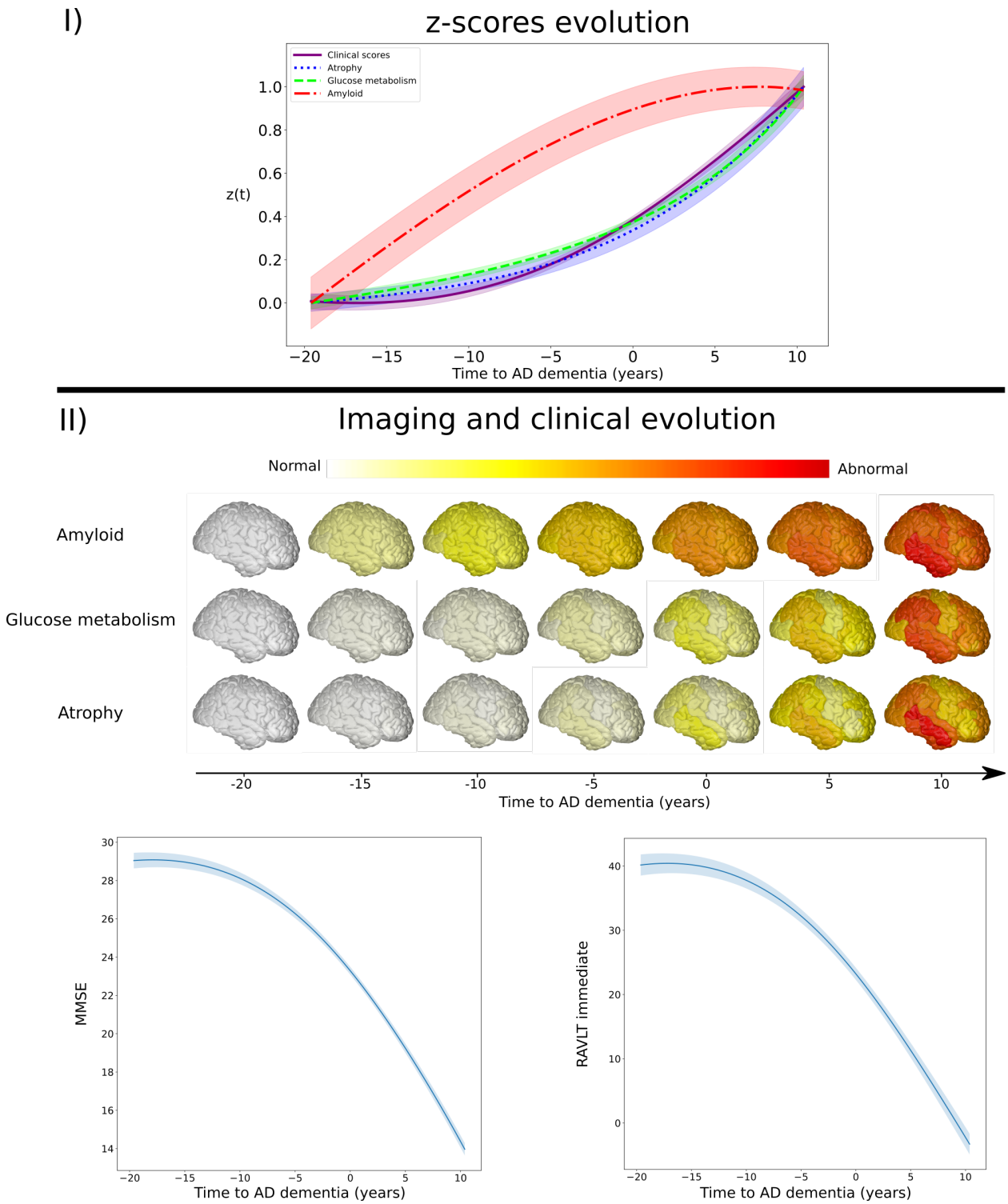
To complete what was presented in Section 2.6 concerning the imputation of the neuropsychological tests for the subjects of the GMC cohort, we show in Supplementary Table 1 the Krippendorff coefficient for each clinical score. This coefficient compares the inter-rater agreement, which would be here the agreement between the ground truth score and the estimated one. We observe that the value 0.8 is contained within the confidence interval of the CDRSB, ADAS11, RAVLT immediate and FAQ, which makes the associated k-neighbors models reliable predictors according to the thresholds defined by Krippendorff. The value 0.67 is contained within the confidence interval of the RAVLT learning and RAVLT forgetting, which makes the corresponding k-neighbors models acceptable for drawing prudent conclusions. We also provide the distribution of the absolute prediction error for all the scores across clinical groups in Supplementary Figure 4.

Supplementary Table 1: Krippendorff coefficient assessing the agreement between the ground truth and the k-neighbors prediction for all the clinical scores in the ADNI cohort. Average values and 95% confidence interval. CDRSB: Clinic Dementia Rating Scale Sum of Boxes; ADAS11: Alzheimer's Disease Assessment Scale; FAQ: Functional Assessment Questionnaire; RAVLT: Rey Auditory Verbal Learning Test. ADNI: Alzheimer's Disease Neuroimaging Initiative. CI: Confidence interval.

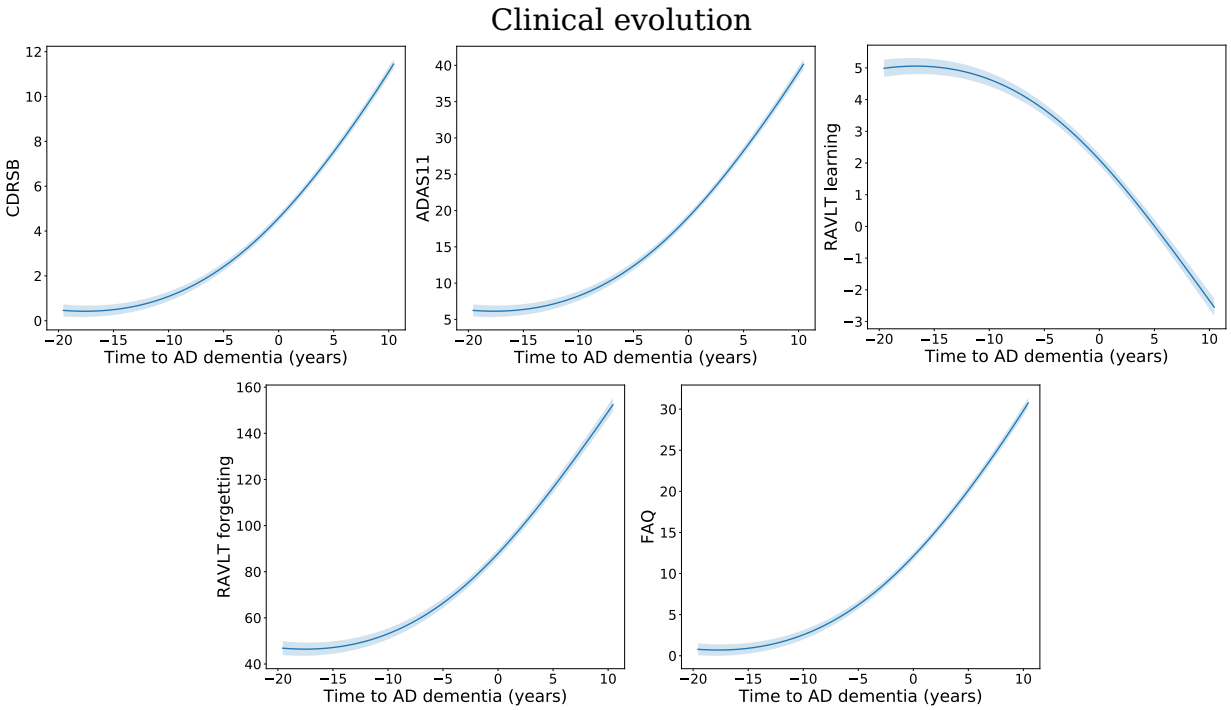
Score	CDRSB	ADAS11	RAVLT immediate	RAVLT learning	RAVLT forgetting	FAQ
Krippendorff	0.77	0.77	0.80	0.61	0.69	0.78
95% CI	[0.69 ; 0.83]	[0.70 ; 0.83]	[0.74 ; 0.85]	[0.54 ; 0.68]	[0.62 ; 0.75]	[0.72 ; 0.84]

3.2 FDG prediction

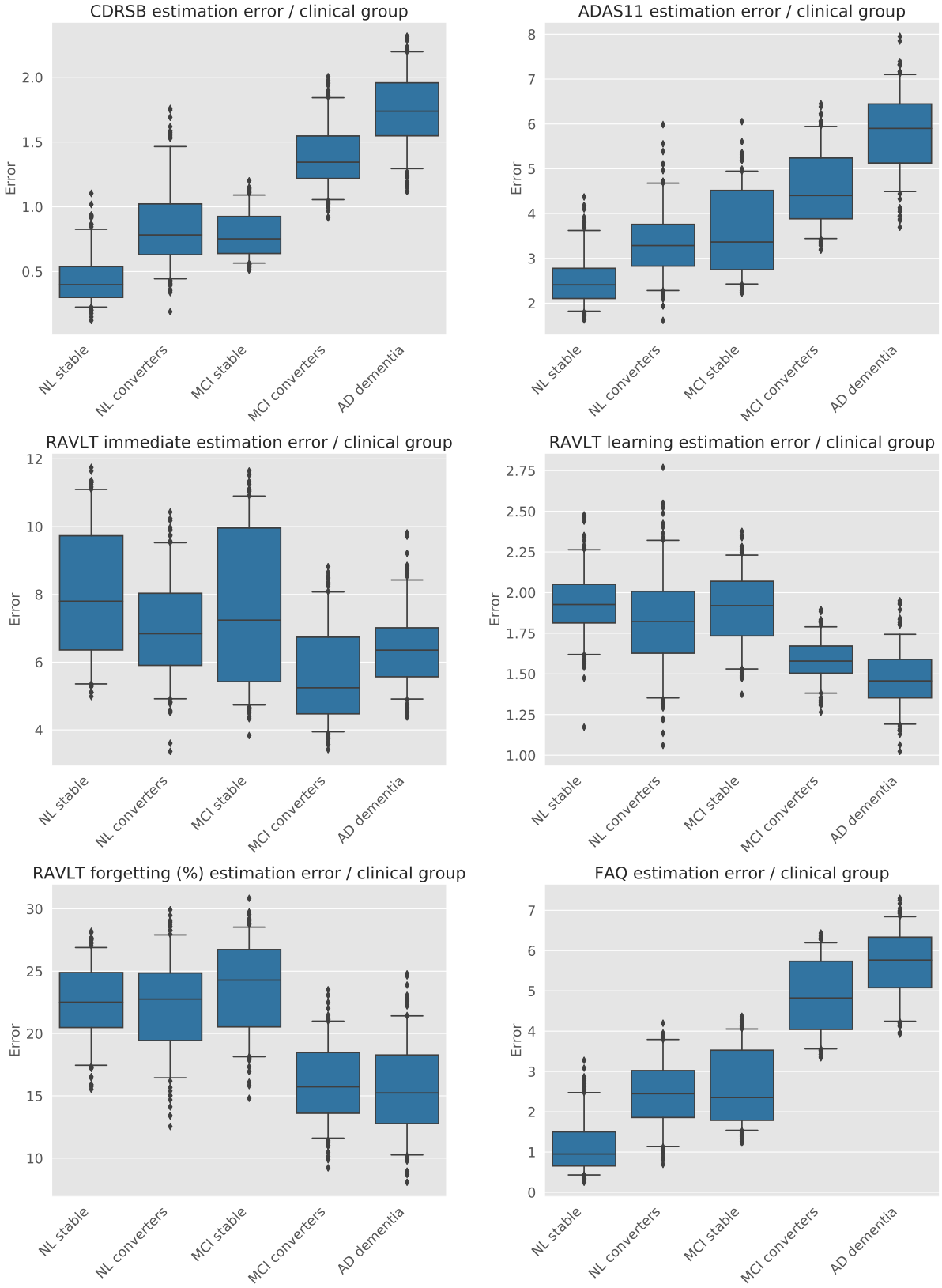
We show in Supplementary Figure 5 the fitted linear models between the FDG and early-AV45 uptake of 57 subjects from the GMC cohort for additional brain regions, as well as their raw values.



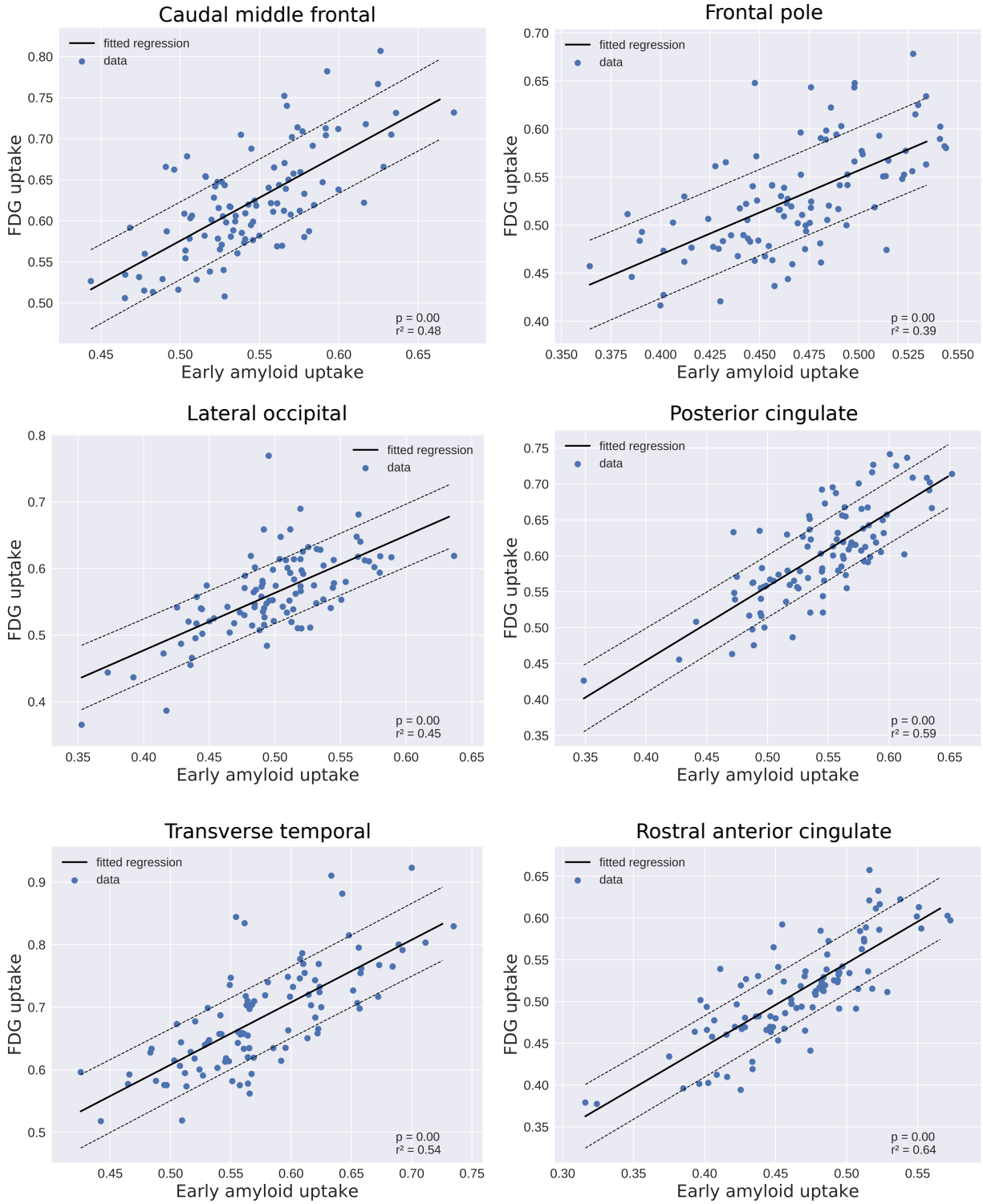
Supplementary Figure 2: Panel I: Estimated long-term dynamics of the z-scores (time is relative to conversion to Alzheimer's dementia) based on the ADNI cohort. The z-scores were re-scaled between 0 and 1 to illustrate the progression of each process from normal to pathological stages. Shadowed areas represent the standard deviation of the average trajectory. Panel II: Modelled long-term evolution of cortical measurements for the different types of imaging markers, and clinical scores based on the ADNI database. Shadowed areas represent the standard deviation of the average trajectory.



Supplementary Figure 3: Long-term modelled evolution of the missing clinical scores in Supplementary Figure 2. Shadowed areas represent the standard deviation of the average trajectory. CDRSB: Clinic Dementia Rating Sum of Boxes; ADAS11: Alzheimer's Disease Assessment Scale; FAQ: Functional Assessment Questionnaire; RAVLT: Rey Auditory Verbal Learning Test.



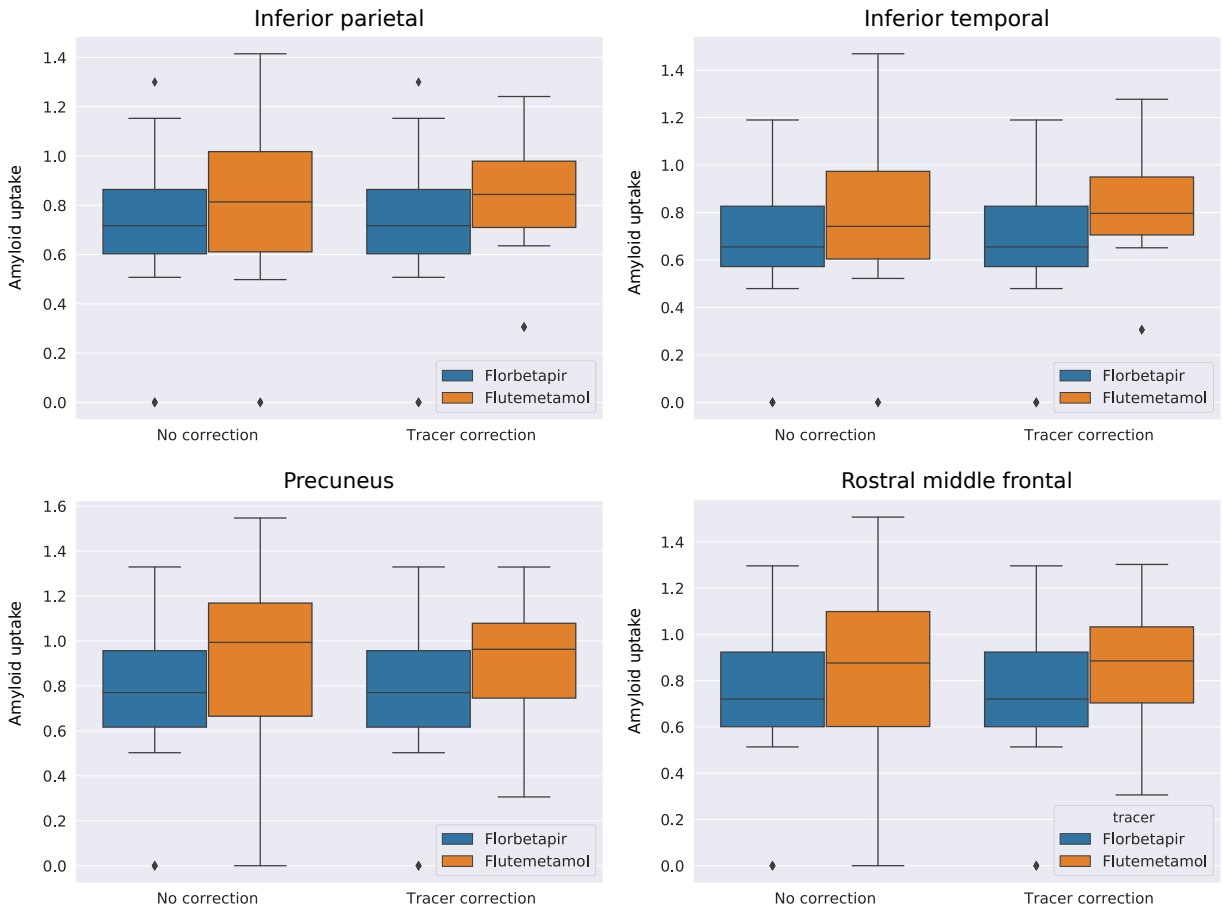
Supplementary Figure 4: Distribution of the absolute error between the ground truth and k-neighbors prediction for all the neuro-psychological tests across clinical groups. NL: cognitively healthy; MCI: mild cognitive impairment; AD: Alzheimer's dementia; CDRSB: Clinic Dementia Rating Sum of Boxes; ADAS11: Alzheimer's Disease Assessment Scale; MMSE: Mini-Mental State Examination;



Supplementary Figure 5: Scatter plot between the regional early-amyloid uptake and the corresponding FDG uptake for 57 patients of the GMC cohort. Solid black lines show the fitted linear model between regional FDG and early-amyloid. The dashed-lines represent 95% confidence interval. GMC: Geneva Memory Center FDG: (18)F-fluorodeoxyglucose Positron Emission Tomography (PET) imaging.

3.3 Tracer correction

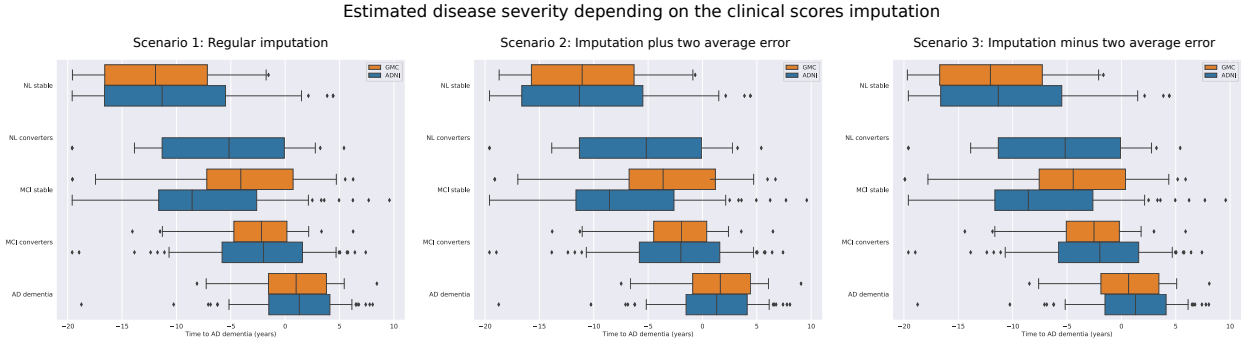
As mentioned in Section 2.6, we provide in Supplementary Figure 6 additional histograms showing the effect of the tracer correction in different brain regions. Consistently with what has been observed, we notice a reduction of the variability of the values for the flutemetamol group.



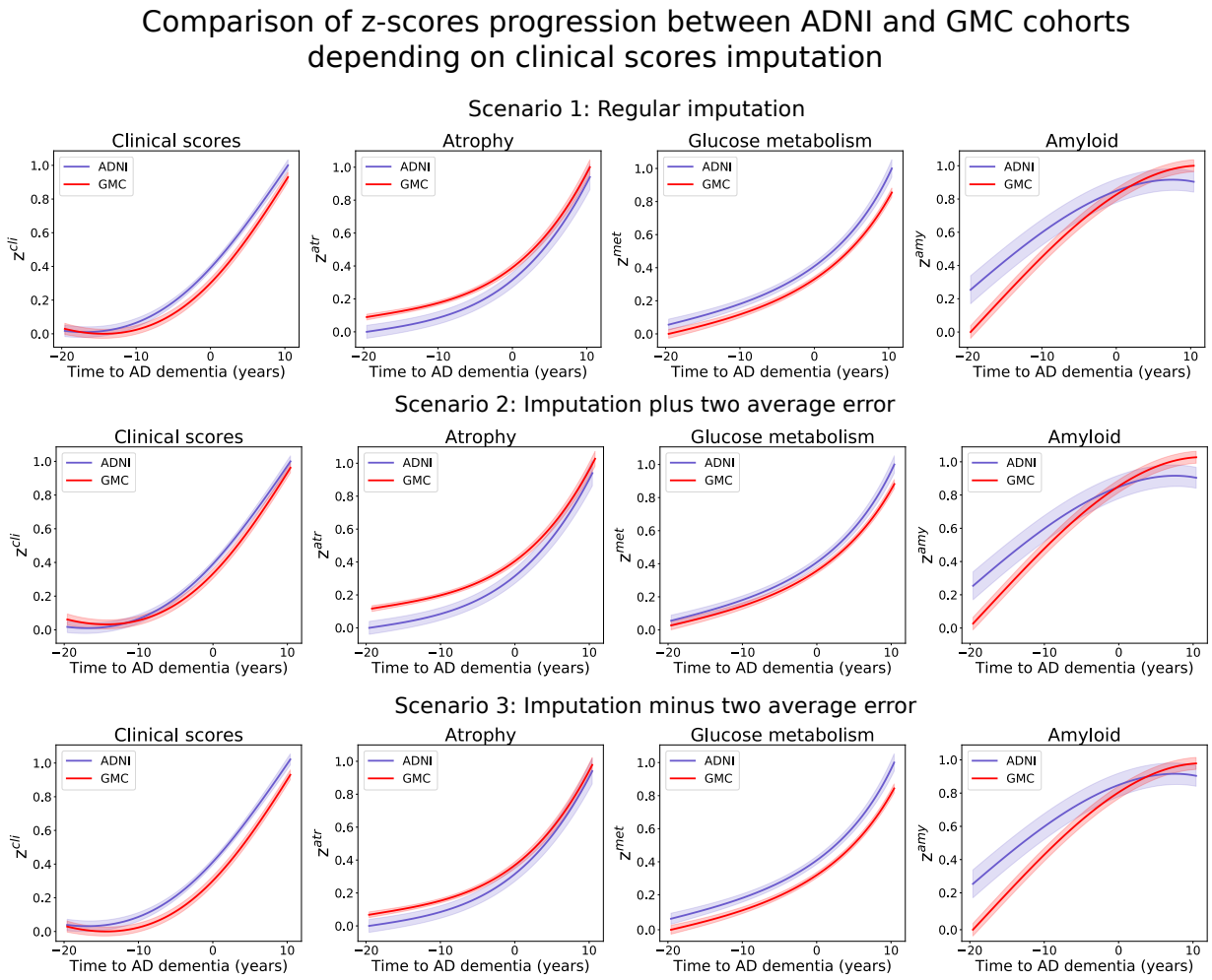
Supplementary Figure 6: Comparison of the distribution of the regional amyloid uptake before and after correction for the subjects from the GMC cohort, depending on the tracer used during acquisition. 76 Amyloid-PET scans were acquired using florbetapir and 17 using flutemetamol. GMC: Geneva Memory Center; SUVR: Standardized Uptake Value Ratio.

4 Robustness to the clinical scores prediction

In the case of the GMC cohort six clinical scores over seven were imputed based on the procedure detailed in Section 2.6. These approximated neuropsychological assessments are used to compute the z-scores z^{cli} of the individuals from the GMC cohort, thus affecting their estimated disease severity shown in Figure 3, but also the progression of the z-scores for the GMC cohort in Figure 6. We evaluate the impact of this approximation by computing the group-wise disease severity and the progression of the z-scores in three scenarios: (i) clinical scores are imputed following the procedure described in Section 2.6; (ii) clinical scores are imputed with an additional approximation by adding twice their respective average prediction error; (iii) clinical scores are imputed with an additional approximation by subtracting twice their average prediction error. We observe in Supplementary Figures 7 and 8 that introducing this additional approximation error when imputing the clinical scores leads to rather small changes compared with the results obtained with the regular imputation, for both the group-wise disease severity and the z-scores progression. This shows that SimulAD is robust to the imputation of the clinical scores for the GMC cohort, and that despite the induced margin of error the estimated disease severity and z-scores progression are reliable.



Supplementary Figure 7: Effect of the clinical scores imputation on the estimated disease severity for the subjects from the GMC cohort. Scenario 1: Clinical scores are imputed following the procedure described in Section 2.6. Scenario 2: Clinical scores are imputed with an additional approximation by adding twice their respective average prediction error. Scenario 3: Clinical scores are imputed with an additional approximation by subtracting twice their average prediction error. ADNI: Alzheimer's Disease Neuroimaging Initiative; GMC: Geneva Memory Center; NL: cognitively healthy; MCI: mild cognitive impairment; AD: Alzheimer's dementia. Converters are cognitively unimpaired and MCI subjects whose clinical diagnosis change during follow-up.



Supplementary Figure 8: Effect of the clinical scores imputation on the long-term dynamics for the subjects from the GMC cohort. Shadowed areas represent the standard deviation of the average trajectory. Scenario 1: Clinical scores are imputed following the procedure described in Supplementary Section 3.1. Scenario 2: Clinical scores are imputed with an additional approximation by adding twice their respective average prediction error. Scenario 3: Clinical scores are imputed with an additional approximation by subtracting twice their average prediction error. ADNI: Alzheimer's Disease Neuroimaging Initiative; GMC: Geneva Memory Center.

5 Comparison of the simulated evolution of clinical and imaging measures based on ADNI and GMC cohorts

The similarity between the two models of disease progression based on the ADNI and GMC cohorts (*cf.* Section 3.3) is confirmed in Supplementary Figure 9. In this figure, we show the temporal evolution of the error between the clinical and regional imaging measurements simulated based on the ADNI cohort and the ones based on the GMC cohort. We observe that for imaging markers the regional error remains below 10% for most regions across the different modalities, and reaches a maximum of approximately 20% in the case of amyloid deposition at the earliest stages of the disease. Concerning clinical scores, the error remains below 10% and 20% for MMSE and RAVLT immediate respectively. When averaged across time, brain regions and clinical scores, the error is of 3%, 6%, 7% and 12% for MRI, FDG-PET, amyloid-PET data and clinical scores respectively.

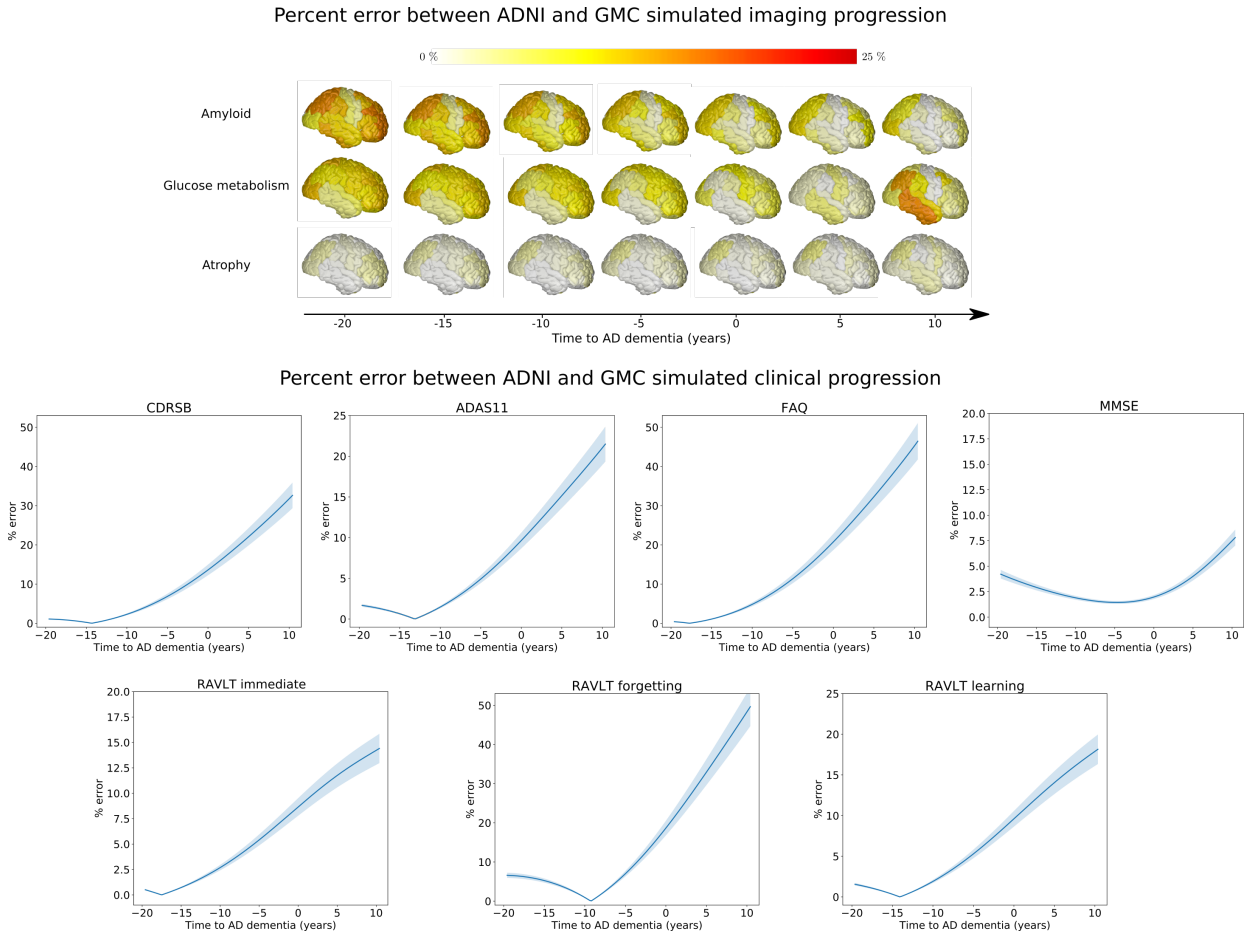
6 Simulating individual trajectories

In Supplementary Figure 10, we show the predicted evolution of the MMSE, hippocampus volume, glucose metabolism and amyloid burden for four subjects from the GMC cohort: one normal, one stable MCI, one MCI converter and one with AD dementia. We recall that the subjects from the GMC cohort underwent clinical assessments, MRI and PET scans at baseline only. During subsequent visits, the only available information were MMSE assessments. We observe that for these four patients, the model was able to accurately predict the MMSE. Concerning hippocampus volume, glucose metabolism and amyloid burden, we show the model prediction for the five next years but cannot compare it to the real values as this data was not available.

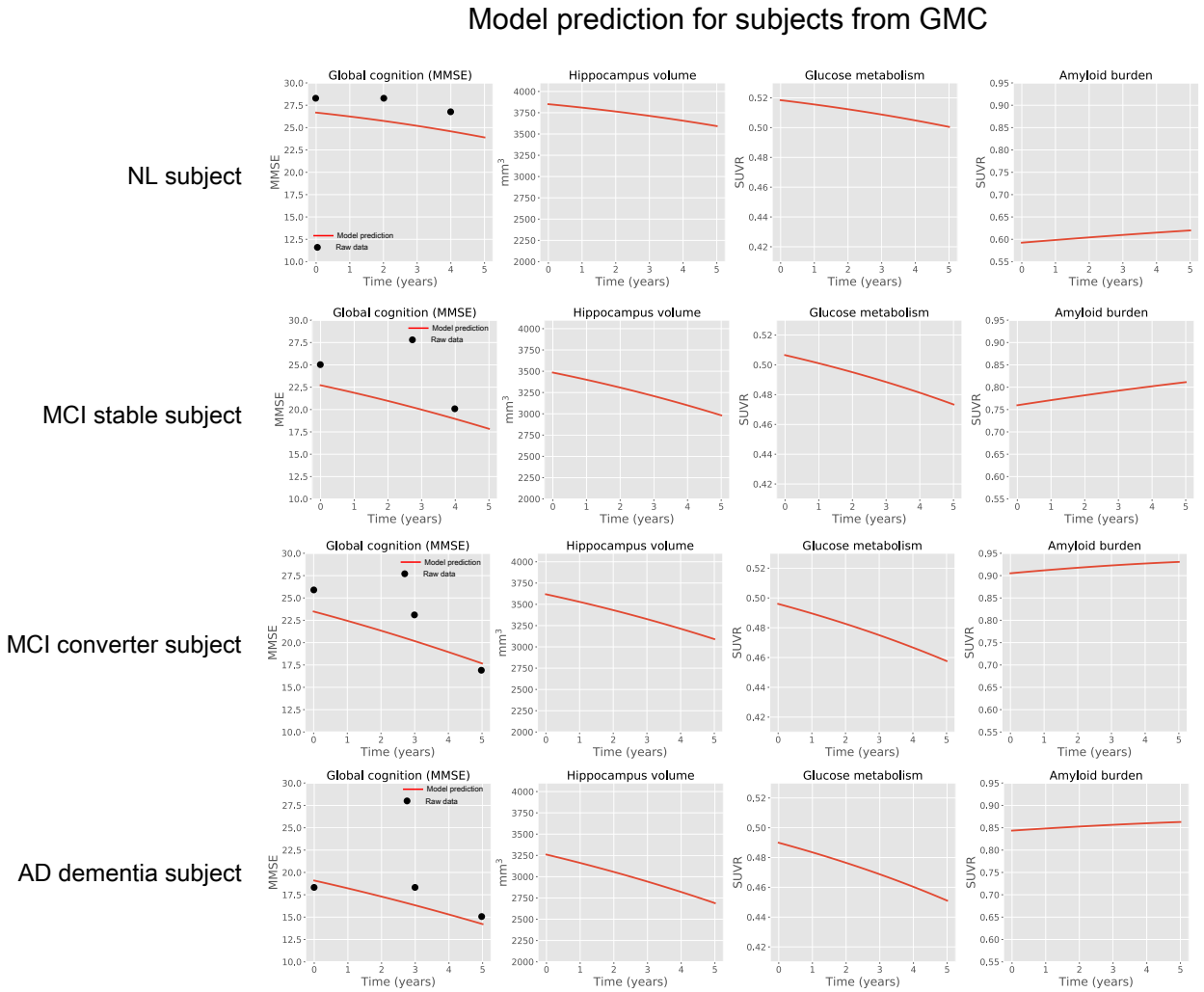
7 Discarding FDG data and missing clinical data

In this appendix, we propose to train a model on the ADNI cohort where the FDG data and the 6 missing clinical scores are removed. We show in Figure 11 below the obtained trajectories for the three remaining z-scores. We observe that the model now spans approximately 20 years. Amyloid deposits and saturates first, followed by cerebral atrophy and cognitive decline. This sequence of events is in agreement with the hypothesis of the pathological cascade. However, compared to the original model we miss the information about glucose metabolism which is known to play an important role in the pathogenesis of AD. Moreover, the clinical state of patients is now only determined by their MMSE, while in the original model the clinical z-score was computed based on a total of seven neuro-psychological assessments, thus allowing to give a broader view of the clinical status of a subject.

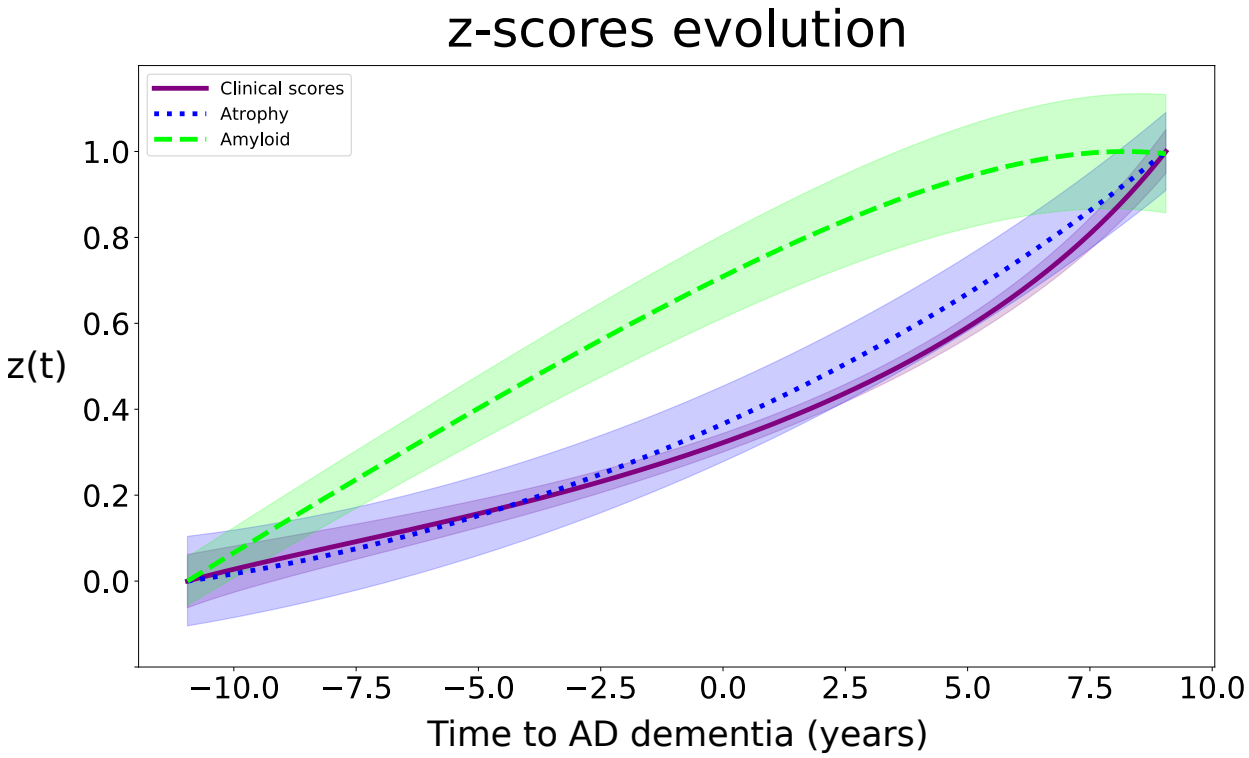
Based on this new model, we estimated the disease severity of the subjects from the ADNI and GMC cohorts, as shown in Supplementary Figure 12. Similarly to Figure 3, the disease severity increases



Supplementary Figure 9: Modelled long-term evolution of the error between the progressions simulated based on the ADNI and GMC cohorts in terms of cortical measurements and clinical scores. Shadowed areas represent the standard deviation of the average trajectory. ADNI: Alzheimer's Disease Neuroimaging Initiative; GMC: Geneva Memory Center; CDRSB: Clinic Dementia Rating Sum of Boxes; ADAS11: Alzheimer's Disease Assessment Scale; FAQ: Functional Assessment Questionnaire; RAVLT: Rey Auditory Verbal Learning Test; MMSE: Mini-Mental State Examination.



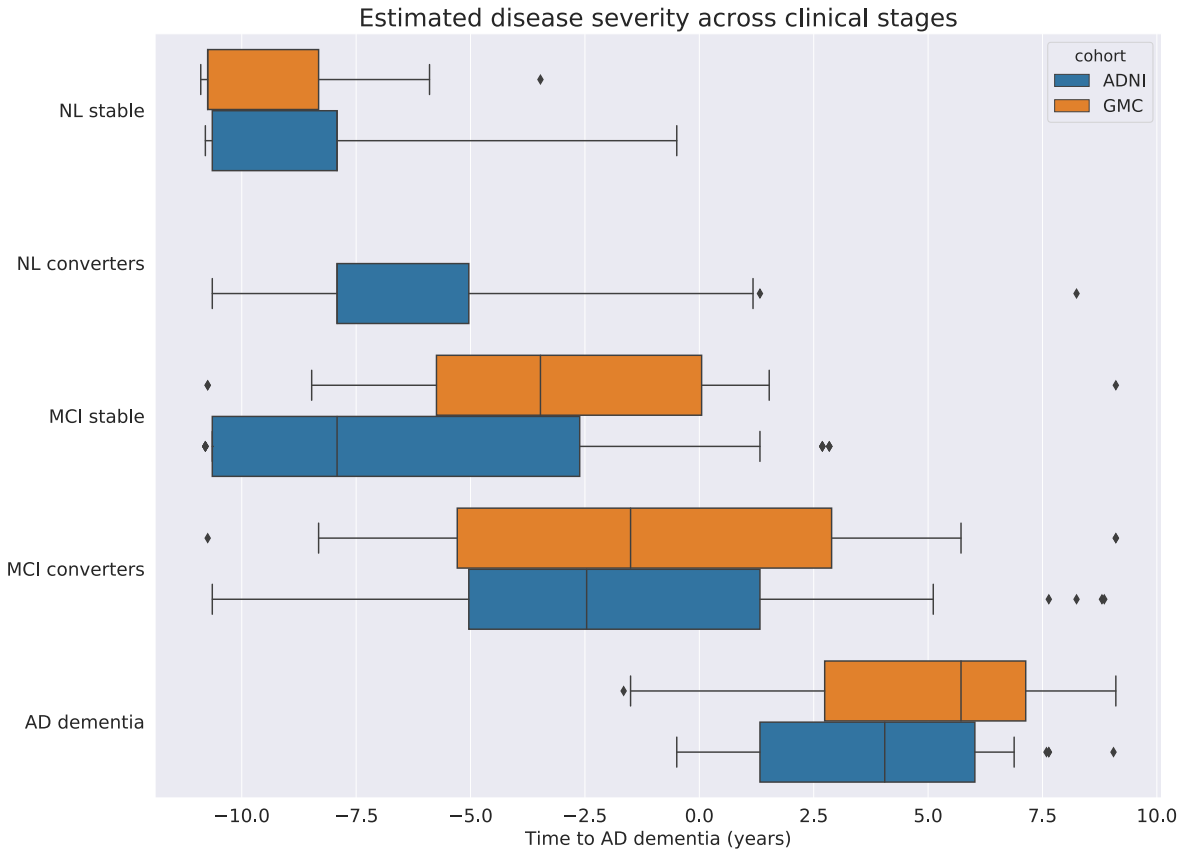
Supplementary Figure 10: Prediction of MMSE, hippocampus volume, glucose metabolism and amyloid burden for four subjects from the GMC cohort. Red lines represent the model prediction and black dots raw values. GMC: Geneva Memory Center; NL: cognitively healthy; MCI: mild cognitive impairment; AD: Alzheimer's dementia. Converters are NL patients progressing to MCI or AD dementia, or MCI individuals progressing to AD dementia.



Supplementary Figure 11: Estimated long-term dynamics of the z-scores (time is relative to conversion to Alzheimer's dementia) based on the ADNI cohort. The z-scores were re-scaled between 0 and 1 to illustrate the progression of each process from normal to pathological stages. Shadowed areas represent the standard deviation of the average trajectory.

when going from healthy towards pathological stages. The group-wise difference of disease severity between clinical groups is statistically significant except in the case of MCI stable and converters for the GMC cohort (*cf.* Supplementary Table 2). We also observe rather large differences between clinical groups ($d > 0.7$ *cf.* Supplementary Table 2) except for NL stable vs MCI stable for the ADNI cohort and MCI stable vs MCI converters for the GMC database, as shown in the manuscript where all the data modalities were included. Finally, if we look at the distribution of the disease severity for similar groups between cohorts, we notice that differences between cohorts are larger compared to the original experiment. Indeed, we see that the Cohen's d is approximately of 0.3, while before it was below 0.1 (Table 3) for NL stable, MCI converters and AD dementia, while in the case of the MCI stable it increased from 0.57 to 0.74.

Overall, this shows that even if we remove FDG data and the missing clinical scores, SimulAD is still able to learn a disease progression model that can accurately stage patients and differentiate between clinical groups. However, we obtain a less informative model which doesn't take into account the diversity of pathological processes affecting the brain during AD. We also noticed that the disease severity for similar groups across cohorts was not as consistent as in the original case. This might be due to the fact that despite the approximation due to the imputation of the FDG uptake and missing clinical scores, the inclusion of this data still allows a more global assessment of subjects. Indeed, the heterogeneity of the data on which our model relies is essential to fully appreciate the pathological state of a subject. Therefore, we would recommend to take advantage of the ease to include heterogeneous data in our model, and use SimulAD with as many data modalities as possible, even if this might require some pre-processing steps to impute missing data. However, in the case where imputation would be too cumbersome or impossible, we showed in this experiment that SimulAD would still provide a reliable estimation of the disease severity of patients when some data modalities are discarded.



Supplementary Figure 12: Distribution of the disease severity estimated by SimulAD across clinical stages for the ADNI and GMC cohorts relatively. The disease severity is estimated relatively to a progression model in which FDG data and the 6 missing clinical scores were discarded and no imputation step needs to be carried out. ADNI: Alzheimer's Disease Neuroimaging Initiative; GMC: Geneva Memory Center; NL: cognitively healthy; MCI: mild cognitive impairment; AD: Alzheimer's dementia. Converters are NL patients progressing to MCI or AD dementia, or MCI individuals progressing to AD dementia.

Supplementary Table 2: Comparison of the estimated disease severity distribution between clinical groups within each cohorts (a) and between similar clinical groups across cohorts (b); We report p-values of Student's *t*-test as well as the associated effect size (Cohen's *d*).

(a)

Within cohorts disease severity comparison

	NL stable vs NL converters		NL stable vs MCI stable		MCI stable vs MCI converters		MCI stable vs AD dementia		MCI converters vs AD dementia	
Cohort	ADNI	GMC	ADNI	GMC	ADNI	GMC	ADNI	GMC	ADNI	GMC
p-value	$1,1 \cdot 10^{-2}$	/	$2.5 \cdot 10^{-6}$	$1.1 \cdot 10^{-7}$	$5.4 \cdot 10^{-6}$	$1.9 \cdot 10^{-1}$	0.0	$3.3 \cdot 10^{-7}$	0.0	$3.3 \cdot 10^{-4}$
Cohen's d	0.62	/	0.65	1.73	0.75	0.38	2.60	1.81	1.52	1.21

(b)

Between cohorts disease severity comparison

	NL stable	MCI stable	MCI converters	AD dementia
p-value	$1.1 \cdot 10^{-1}$	$4.4 \cdot 10^{-3}$	$2.9 \cdot 10^{-1}$	$3.5 \cdot 10^{-1}$
Cohen's d	0.32	0.74	0.30	0.33